ZS6D-SD-DINO: Zero-shot 6D Object Pose Estimation with Stable Diffusion and DINOv2

Kilian Kliegel

***Abstract*—The increasing use of vision-based robotic systems in dynamic environments requires flexible solutions for object pose estimation. A key challenge is the use of methods that are accurate without relying on object-specific pre-training or task-specific fine-tuning. One potential solution are the pre-trained models not designed for a specific object category, allowing a generalized feature extraction, which can be used for the object pose estimation of unseen objects. This concept can be further enhanced by combining two models and fusing their extracted features. This paper introduces ZS6D-SD-DINO, a novel approach of concatenating features from Stable Diffuison (SD) and DINOv2 to accomplish zero-shot 6D object pose estimation. Not as expected, experiments conducted on the BOP YCBV dataset show that the combined features underperform compared to existing methods and are not suitable for 6D object pose estimation.**

I. Introduction

Vision-based applications play a crucial role in the future development of robotic systems. For a wide range of tasks, such as grasping and manipulation, accurate object pose estimation is a key element. As this area needs to adapt to diverse and changing environments, previous approaches with object-specific pre-training and task-specific fine-tuning are no longer sufficient. More flexible and generalized methods are needed to perform zero-shot 6D object pose estimation.

To address this challenge, one promising approach is the use of self-supervised pre-trained Vision Transformers (ViT). A concept of this solution is the ZS6D framework [1]. Descriptive features of an unseen object are extracted by a single forward pass through the DINOv1-ViT-S/8 model and then used for the pose estimation [1].

As the final result is strongly influenced by the quality of the extracted features, this paper explores the potential advantages of combining two pre-trained models [2]: Stable Diffusion (SD) [3] and DINOv2 [4]. This concatenation is shown in Figure 1, allowing DINOv2 features to be complemented with high quality spatial features from SD [2].

To use this process for the zero-shot 6D object pose estimation, this paper introduces ZS6D-SD-DINO, a novel approach to utilize SD and DINOv2 [2] within the ZS6D framework [1]. The proposed method aims to enhance the descriptive quality of extracted features, which can lead to more accurate semantic correspondences and an improved 6D object pose estimation.

This paper has the following structure: Section II reviews state of the art methods. Section III introduces the proposed solution. The implementation and experimental results are discussed in Section IV. Finally, Section V summarizes the main findings.

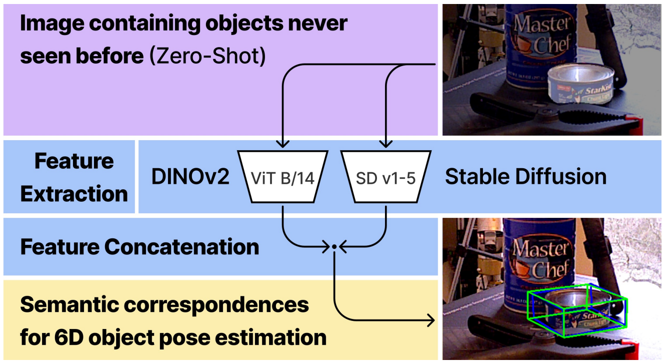
Image containing objects never seen before (Zero-Shot)
Feature Extraction
DINOv2
ViT B/14
SD v1-5
Stable Diffusion
Feature Concatenation
Semantic correspondences for 6D object pose estimation

Fig. 1: **Zero-shot 6D object pose estimation with SD and DINOv2** Concatenated features are used to establish semantic correspondences for 6D object pose estimation.

II. Related Work

Convolutional neural networks (CNNs) used to be the state of the art solution for 3D pose estimation of novel objects, but the recent work of [5] shows that ViTs can be more accurate. In addition to that, the self-supervised training of ViTs does not require object-specific fine-tuning, making them more convenient to use in real world scenarios [6]. Further research has shown that the capabilities of ViTs can also be used for zero-shot 6D object pose estimation [1].

In addition to the recent developments of ViTs, diffusion models have also been used for descriptive tasks. Despite the intended application for generative image creation [3], pre-trained diffusion models showed promising results for image segmentation [7], [8], classification [9] and semantic correspondences [10], [2]. The research presented in [11] demonstrates how pre-trained diffusion models can be leveraged for a reverse diffusion process and are able to establish noise robust correspondences to do 6D objec pose estimation.

Based on these recent developments, this paper explores a new correspondence-based hybrid approach that combines ViTs and diffusion models for 6D pose estimation.

III. ZS6D-SD-DINO

This section presents the proposed solution to enhance ZS6D [1] for an even more robust and accurate object pose estimation. As shown in Figure 2, the existing framework is extended by implementing concatenated features of SD and DINOv2 [2].

In the first stage, as a standard tool for segmentation, CNOS [12], [15], [1] is used to isolate the relevant object from the input image. The features of this object are extracted

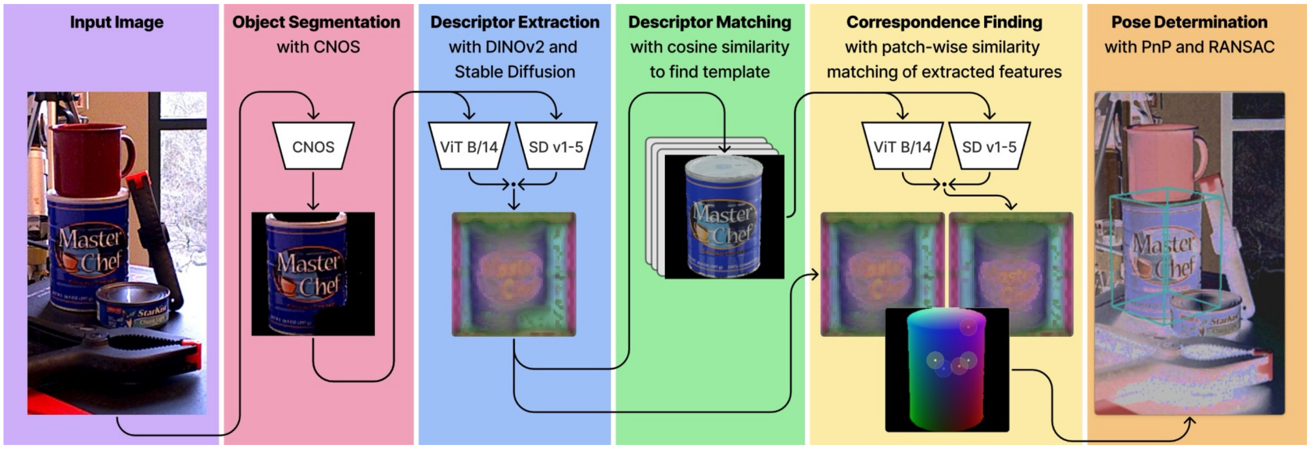

Fig. 2: **Pipeline of ZS6D [1] with SD-DINO [2]** CNOS [12] segments the target object from the input image. Descriptors are extracted using DINOv2 and SD, then concatenated [2]. This descriptor is matched against a template database using cosine similarity [1]. From the matched template, descriptors are extracted using DINOv2 and SD, then concatenated [2]. Patchwise similarity matching of the extracted features establishes correspondences between the segmented input and the matched template [2]. Finally, PnP [13] and RANSAC [14], [1] are applied for the 6D object pose determination [1].

separately by a single forward pass through SD and DINOv2 [2]. To ensure compatibility, the descriptors are normalized and fused together by concatenation [2], which creates the object descriptor D_{obj}. As mentioned by [2], this approach should keep the strengths of both models, sparse but accurate features from DINOv2 and high-quality spatial features from SD. Following the ZS6D [1] framework, from a set of pre-processed template descriptors D_t, the one with the smallest cosine similarity deviation to D_{obj} is selected. To establish semantic correspondences in the next step, the approach of patch-wise similarity matching from SD-DINO [2] was used. Each feature patch of the segmented object is compared to all feature patches of the template object. The one with the smallest Euclidean distance is selected as the best match. The found correspondences are then used for further processing. Together with the known pose of the template object, the Perspective-n-Point (PnP) [13] algorithm can be applied to retrieve the final 6D pose estimation of the segmented object, RANSAC iterations [14] help to eliminate outliers [1].

IV. Experiments

In this section, the implementation and results of the experiments are discussed. For evaluating the 6D object pose estimation of ZS6D-SD-DINO, the YCBV dataset of BOP [16] is utilized. This dataset contains 21 distinct objects captured under different illumination conditions.

Following the the feature extraction process of SD-DINO [2], the base model DINOv2-ViT-B/14 with a patch size of 14 is used. As mentioned in [2], the most distinctive features can be obtained from the ”token” facet of layer 11. In addition to that, the SD model v1-5 with a timestep of $t = 100$ is used as well [2]. To be able to fuse the extracted features of both models [2], the input image size is set to 840 pixels for DINOv2 and 960 pixels for SD. After the feature concatenation the descriptors have a shape of $[1, 1, 3600, 1152]$. These fused descriptors are then used for template matching and correspondence finding within the ZS6D-SD-DINO pipeline.

To extract semantic correspondences between features, different approaches were tested. One method focuses on grouping similar features through k-means clustering and then using the Hungarian algorithm [17] to match the cluster centroids. But this method was dismissed because it did not perform well with the YCBV dataset [16].

Instead, the approach described in Section III was used. Not as expected, this method achieved an average recall of only 0.08876 on the YCBV dataset [16] which is quite low compared to existing methods like ZS6D with DINOv1 [1]. Despite a few accurate pose estimations, many result images showed pose estimation errors, especially deviations for objects with rotationally symmetric shapes as shown in Figure 3. This indicates that the concatenated features might give more importance to object shapes because of the impact of SD features.

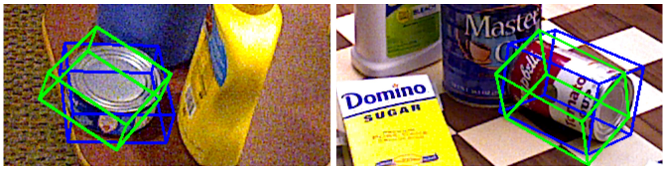

Fig. 3: **Pose estimation errors** Inaccurate 6D pose estimation for objects with rotationally symmetric shapes.

V. Conclusion

This work looked into zero-shot 6D object pose estimation [1] using concatenated features of SD and DINOv2 [2]. The evaluation result on the YCBV dataset [16] indicates that this approach is not suitable for the ZS6D [1] framework, as it significantly underperformed compared to existing methods. Nevertheless, since there is potential in semantic correspondences of concatenated features, future research can explore alternative methods that may be better suited for zero-shot 6D pose estimation.

References

[1] P. Ausserlechner, D. Haberger, S. Thalhammer, J.-B. Weibel, and M. Vincze, "Zs6d: Zero-shot 6d object pose estimation using vision transformers," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 463–469.

[2] J. Zhang, C. Herrmann, J. Hur, L. P. Cabrera, V. Jampani, D. Sun, and M.-H. Yang, "A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence," *arXiv preprint arxiv:2305.15347*, 2023.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[4] 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.

[5] S. Thalhammer, J.-B. Weibel, M. Vincze, and J. Garcia-Rodriguez, "Self-supervised vision transformers for 3d pose estimation of novel objects," *Image and Vision Computing*, vol. 139, p. 104816, 2023.

[6] 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.

[7] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2955–2966.

[8] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 5729–5739.

[9] A. C. Li, M. Prabhudesai, S. Duggal, E. Brown, and D. Pathak, "Your diffusion model is secretly a zero-shot classifier," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2206–2217.

[10] J. Zhang, C. Herrmann, J. Hur, E. Chen, V. Jampani, D. Sun, and M.-H. Yang, "Telling left from right: Identifying geometry-aware semantic correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3076–3085.

[11] L. Xu, H. Qu, Y. Cai, and J. Liu, "6d-diff: A keypoint diffusion framework for 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9676–9686.

[12] V. N. Nguyen, T. Groueix, G. Ponimatkin, V. Lepetit, and T. Hodan, "Cnos: A strong baseline for cad-based novel object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2134–2140.

[13] 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.

[14] 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.

[15] T. Hodan, M. Sundermeyer, Y. Labbe, V. N. Nguyen, G. Wang, E. Brachmann, B. Drost, V. Lepetit, C. Rother, and J. Matas, "Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5610–5619.

[16] 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.

[17] M. I. Malinen and P. Fränti, "Balanced k-means for clustering," in *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop*. Springer, 2014, pp. 32–41.