中国科学院软件研究所学术年会暨重点实验室科技活动周 2025 第十届
学术论文

通用特征引导的零样本类别级物体姿态估计

作者：曲文天，蒙宸宇，李衡，程坚，马翠霞，王宏安，周晓，邓小明，谭平

Universal Features Guided Zero-Shot Category-Level Object Pose Estimation
Proceedings of the AAAI Conference on Artificial Intelligence, 2025

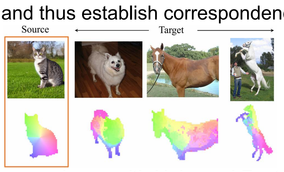
联系方式：邓小明，xiaoming@iscas.ac.cn

Motivation

The diversity of object categories limits the generalization ability of pose estimation models.

Various Object Categories[1,2]

The universal features extracted by the foundation models can learn the semantic similarities between objects and thus establish correspondences.



Correspondence with Universal Features[3]

Introduction

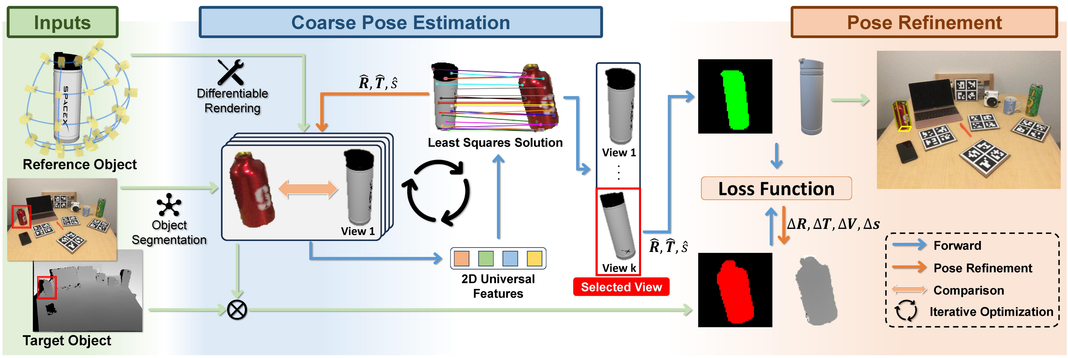
We propose a zero-shot method to achieve category-level 6-DOF object pose estimation, which exploits both 2D and 3D universal features of input RGB-D image to establish semantic similarity-based correspondences and can be extended to unseen categories without additional model fine-tuning:

Our main contributions:

- A 2D/3D universal features guided zero-shot category-level object pose estimation with coarse-to-fine optimization.
- In order to handle pose ambiguity due to intra-category shape difference, we employ 3D universal features to refine the 6-DOF object and the shape of reference model by dense pixel-level registration.

Project Webpage:
https://iscas3dv.github.io/universal6dpose/

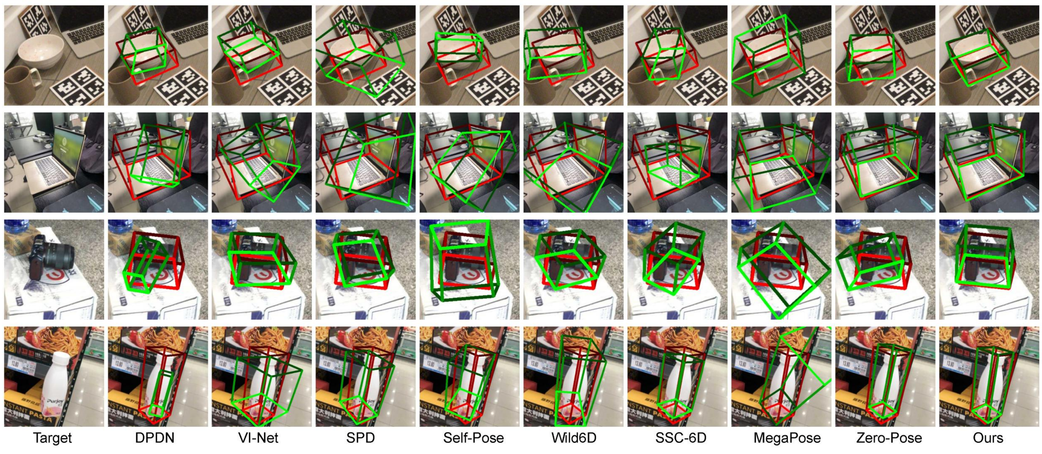
Method



Our framework includes a coarse pose estimation module and a pose refinement module. In the first module, we establish the correspondences between image pairs based on the 2D universal features and calculate the coarse pose using least squares in an iterative manner. In the second module, we use pixel-level optimization combined with 3D universal features to refine the pose and shape of reference model to obtain the fine pose.

Experiment

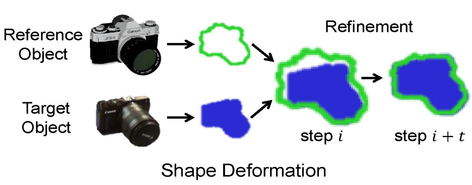
Qualitative results on REAL275 and Wild6D



Effect of Iterative Estimation

When there are large pose differences between objects, the 2D universal features similarity degrade. After iterative optimization, as the objects are gradually aligned, the correspondence between the objects become smoother, which support to calculate an accurate pose.

Effect of Shape Deformation



After the shape optimization, the reference object shape will become closer to the target object shape, resulting in more accurate pose.

Reference:

[1] Wang H, Sridhar S, Huang J, et al. Normalized object coordinate space for category-level 6d object pose and size estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2642-2651.
[2] Jung H J, Wu S C, Ruhkamp P, et al. Housecat6d-a large-scale multi-modal category level 6d object perception dataset with household objects in realistic scenarios[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 22498-22508.
[3] Zhang J, Herrmann C, Hur J, et al. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence[J]. Advances in Neural Information Processing Systems, 2023, 36: 45533-45547.
[4] Fu Y, Wang X. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset[J]. Advances in Neural Information Processing Systems, 2022, 35: 27469-27483.
[5] Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 9650-9660.
[6] Oquab M, Darcet T, Moutakanni T, et al. Dinov2: Learning robust visual features without supervision[J]. arXiv preprint arXiv:2304.07193, 2023.
[7] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695.
[8] Goodwin W, Vaze S, Havoutis I, et al. Zero-shot category-level object pose estimation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 516-532.