FoundPose

Unseen Object Pose Estimation with Foundation Features

Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, Tomas Hodan

Technical University of Munich

Example FoundPose results on datasets HB, LM-O, IC-BIN, TUD-L, ITODD and T-LESS, showing that our method can handle a broad range or objects, including textured, texture-less and symmetric ones. Each example shows the query image crop with the CNOS mask in white (top left), retrieved templates (middle row), matched patch descriptors of the crop and the template that led to the top-quality pose estimate (bottom row), and the contour of the ground-truth pose in red, the coarse pose estimate in blue, and the refined pose estimate in green (top right).

CAD-based object pose estimation

1. **How to quickly onboard a new object just from its CAD model (without any training)?**
2. **How to bridge CAD-to-image domain gap?**

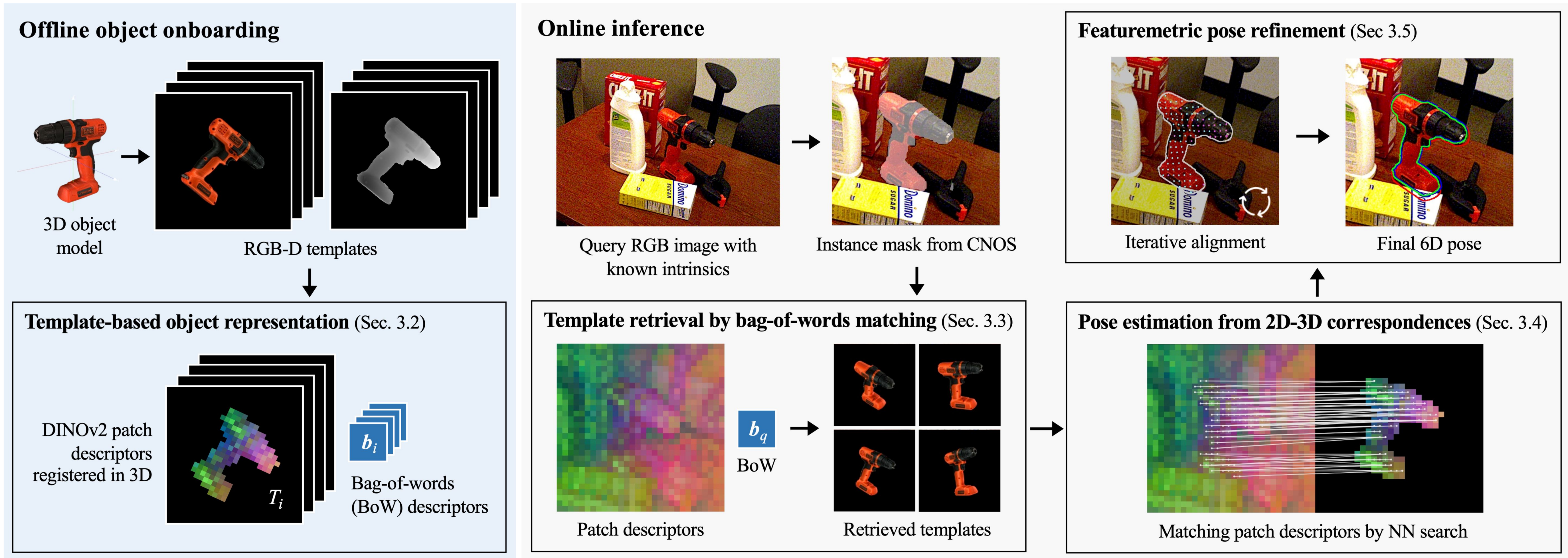
Existing methods (MegaPose, GenFlow, etc.) pre-train on large-scale, task-specific datasets

Instead, *FoundPose relies on the all-purpose DINOv2 features integrated into classical techniques* (BoW, PnP from 2D-3D corresp.), and achieves SOTA without any training

Contributions

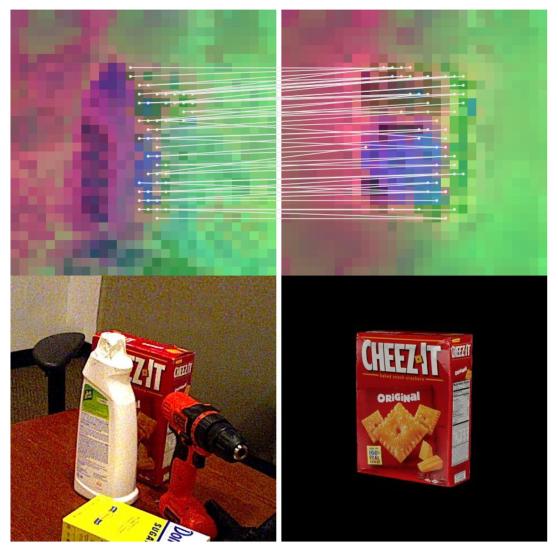
1. **Simple training-free method** for CAD-based object pose estimation, achieving SOTA
2. **Efficient template retrieval approach** which requires 100X fewer templates than competitors and is robust to partial occlusion
3. **Lightweight object representation** which is fast to build and has a 25X lower memory footprint than competitors, enabling scaling to large numbers of objects
4. **Demonstrated importance of intermediate DINOv2 features** for handling symmetric and texture-less objects, also outperforming descriptors from other foundation models

Method overview



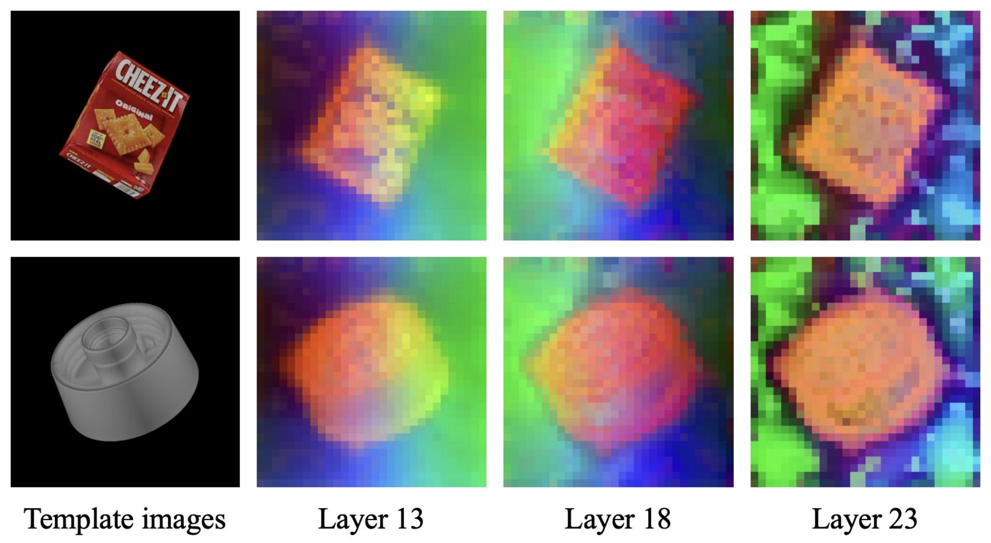
1. Render ~400 RGB-D object templates
2. Extract DINOv2 patch descriptors
3. Register descriptors in 3D using the depth channel (for 2D-3D correspondences)
4. Calculate bag-of-words (BoW) descriptors

1. Detect the target objects using CNOS (based on FastSAM)
2. Perspectively crop detected regions and extract DINOv2 patch descriptors
3. Retrieve similarly-looking templates (BoW vectors are compared by cosine similarity)
4. Establish 2D-3D correspondences by nearest-neighbor matching of DINOv2 descriptors
5. Estimate the object pose by the PnP-RANSAC algorithm



Bridging CAD-to-real gap

Reliable correspondences between a *real query image* (left) and a *synthetic template* (right) can be established by a simple nearest-neighbor matching of DINOv2 features



Handling symmetric and texture-less objects

Features from an intermediate DINOv2 layer yield consistent correspondences even when the semantic information is ambiguous due to symmetries or a lack of texture

Qualitative results

#	Method	Pose refinement	No train.	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V	Average	Time
	Coarse pose estimation:											
1	FoundPose	–	✓	**39.6**	**33.8**	**46.7**	**23.9**	**20.4**	**50.8**	**45.2**	**37.2**	1.7
2	GigaPose [59]	–	✗	29.9	27.3	30.2	23.1	18.8	34.8	29.0	27.6	**0.9**
3	GenFlow [54]	–	✗	25.0	21.5	30.0	16.8	15.4	28.3	27.7	23.5	3.8
4	MegaPose [41]	–	✗	22.9	17.7	25.8	15.2	10.8	25.1	28.1	20.8	15.5
5	OSOP [75]	–	✗	31.2	–	–	–	–	49.2	33.2	–	–
6	ZS6D [3]	–	✓	29.8	21.0	–	–	–	–	32.4	–	–
	With pose refinement (a single hypothesis):											
7	FoundPose	Featuremetric	✓	39.5	39.6	56.7	28.3	26.2	58.5	49.7	42.6	2.6
8	FoundPose	MegaPose	✗	55.4	51.0	63.3	43.0	34.6	69.5	**66.1**	54.7	4.4
9	FoundPose	Feat. + MegaPose	✗	**55.7**	51.0	63.3	43.3	35.7	**69.7**	**66.1**	**55.0**	6.4
10	Gigapose [59]	MegaPose	✗	55.6	**54.6**	57.8	**44.3**	**37.8**	69.3	63.4	54.7	**2.4**
11	MegaPose [41]	MegaPose	✗	49.9	47.7	**65.3**	36.7	31.5	65.4	60.1	50.9	31.7
	With pose refinement (5 hypotheses):											
12	FoundPose	Featuremetric	✓	42.0	43.6	60.2	30.5	27.3	53.7	51.3	44.1	7.4
13	FoundPose	MegaPose	✗	58.6	54.9	65.7	44.4	36.1	70.3	67.3	56.8	11.2
14	FoundPose	Feat. + MegaPose	✗	**61.0**	**57.0**	**69.4**	**47.9**	**40.7**	72.3	**69.0**	**59.6**	20.5
15	GigaPose [59]	MegaPose	✗	59.9	**57.0**	64.5	46.7	39.7	72.2	66.3	57.9	**7.3**
16	GenFlow [54]	GenFlow	✗	56.3	52.3	68.4	45.3	39.5	**73.9**	63.3	57.1	20.9
17	MegaPose [41]	MegaPose	✗	56.0	50.7	68.4	41.4	33.8	70.4	62.1	54.7	47.4

- **SOTA on coarse pose estimation** – FoundPose outperforms GigaPose by +10, GenFlow by +14 and MegaPose by +16 AR
- Significantly faster than MegaPose-Coarse (render & compare)
- +10 AR w.r.t. to the only other training-free method (ZS6D)

#	Method	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V	Average	Time
	Backbones for extracting patch descriptors:									
1	DINOv2 ViT-L – layer 18	39.6	33.8	46.7	23.9	20.4	50.8	45.2	37.2	1.7
2	DINOv2 ViT-L – layer 23	23.2	22.8	31.2	10.3	9.7	33.0	34.0	23.5	1.5
3	DINOv2 ViT-S – layer 9	34.0	31.6	42.7	21.7	16.8	46.8	44.7	34.0	1.3
4	DINOv2 ViT-S – layer 11	22.8	24.2	29.8	11.9	10.5	30.4	36.4	23.7	1.3
5	SAM ViT-L [40] – layer 23	2.2	12.8	9.2	7.5	6.0	10.6	26.9	10.7	3.4
6	DenseSIFT – step size 7px	3.2	2.6	6.5	10.5	2.9	5.6	22.2	7.6	1.4
7	S2DNet [23]	0.8	1.2	0.8	1.4	1.2	1.2	1.3	1.1	1.8
	Template retrieval by matching cls token from layer 18 of DINOv2 ViT-L:									
8	Retrieval by cls token	19.9	17.8	24.6	10.3	13.6	17.7	23.6	18.2	1.6
9	Retrieval by cls token with black bg.	25.5	26.2	30.3	16.7	13.6	29.3	34.4	25.1	1.6
	Other ablations:									
10	Pose given by the top matched template	20.3	18.5	23.0	12.8	12.4	19.6	17.6	17.7	1.0
11	Ground-truth instead of CNOS masks	45.6	53.1	57.1	30.6	–	–	50.9	–	–

- Intermediate DINOv2 features noticeably outperform features from the last DINOv2 layer, SAM, CLIP, LoFTR, S2DNet, DenseSIFT
- BoW retrieval considerably outperforms CLS-based retrieval and reaches accuracy of MegaPose-Coarse while being 15X faster