CVPR SEATTLE, WA JUNE 17-21, 2024

Website, code, data

Open-vocabulary object 6D pose estimation

Jaime Corsetti[1,2] Davide Boscaini[1] Changjae Oh[3] Andrea Cavallaro[4,5] Fabio Poiesi[1]

[1]Fondazione Bruno Kessler [2]University of Trento [3]Queen Mary University of London [4]IDIAP Research Institute [5]EPFL

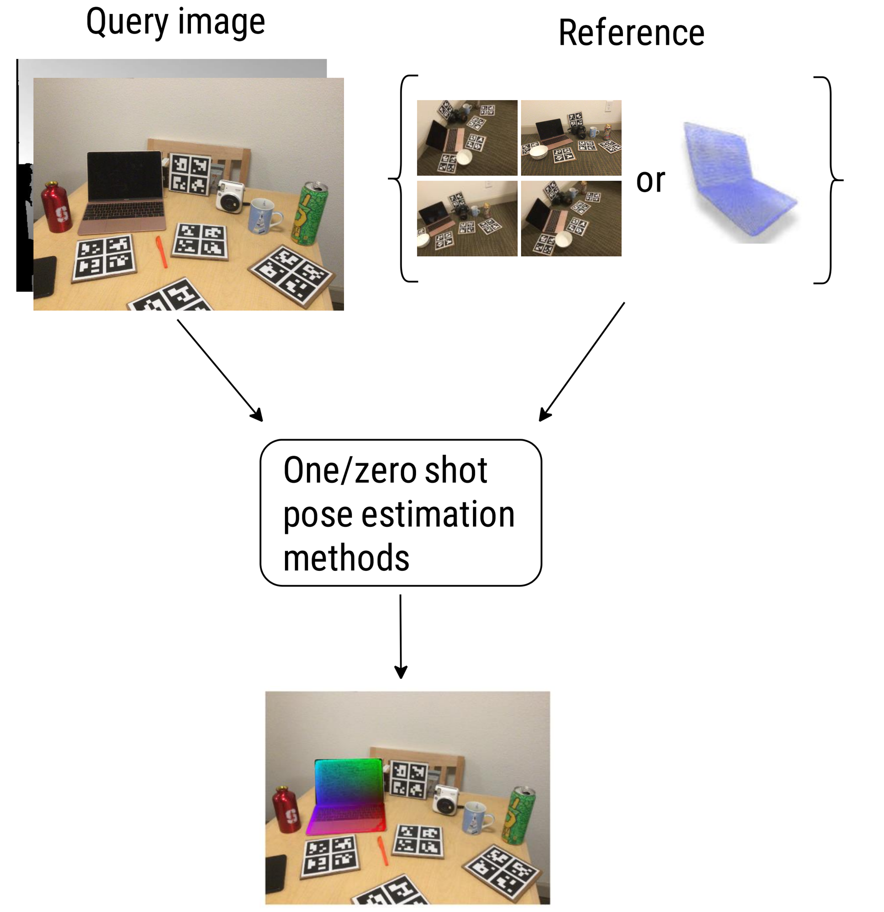
jcorsetti@fbk.eu github.com/jcorsetti/oryon

Motivation

Why: Relax the assumptions of unseen-object 6D pose estimation methods, by removing the need for 3D models or video sequences showing the unseen object

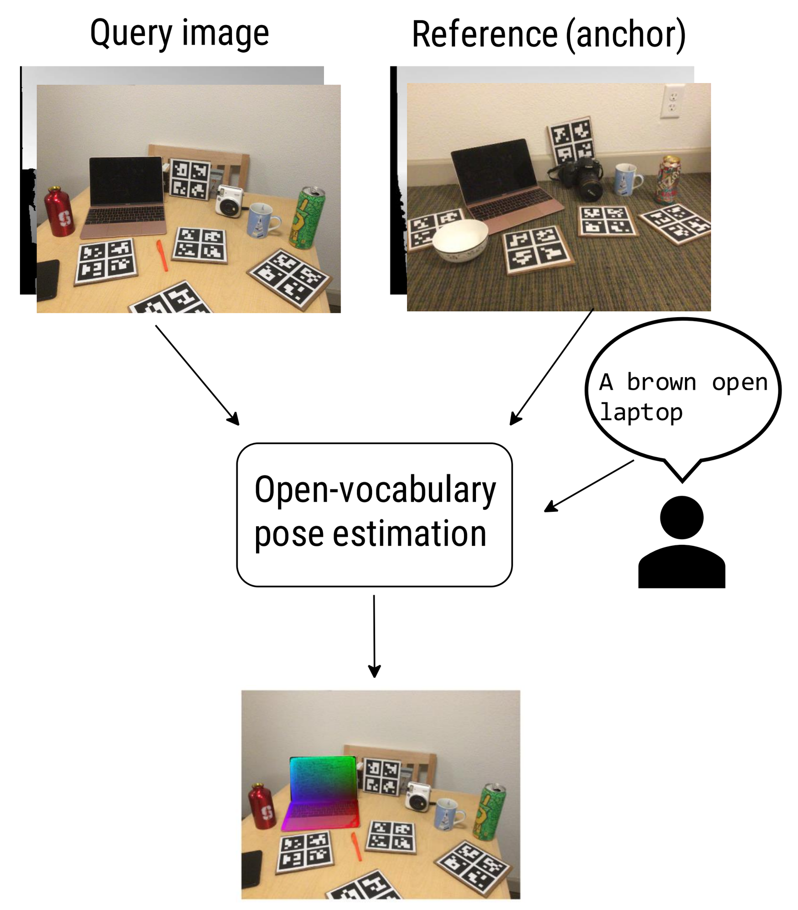
How: Leverage on pretrained Vision-Language Models

State-of-the-art assumptions[1,2]



- ☹ 3D model or multiple views required for the unseen object
- ☹ Complex preprocessing procedures for evaluation
- ☹ Detector/segmentor trained to localize the unseen object

Our assumptions



- ☺ Single RGBD reference view of the unseen object
- ☺ Requires a textual description of the unseen object
- ☺ Joint localization and pose estimation

Our contributions

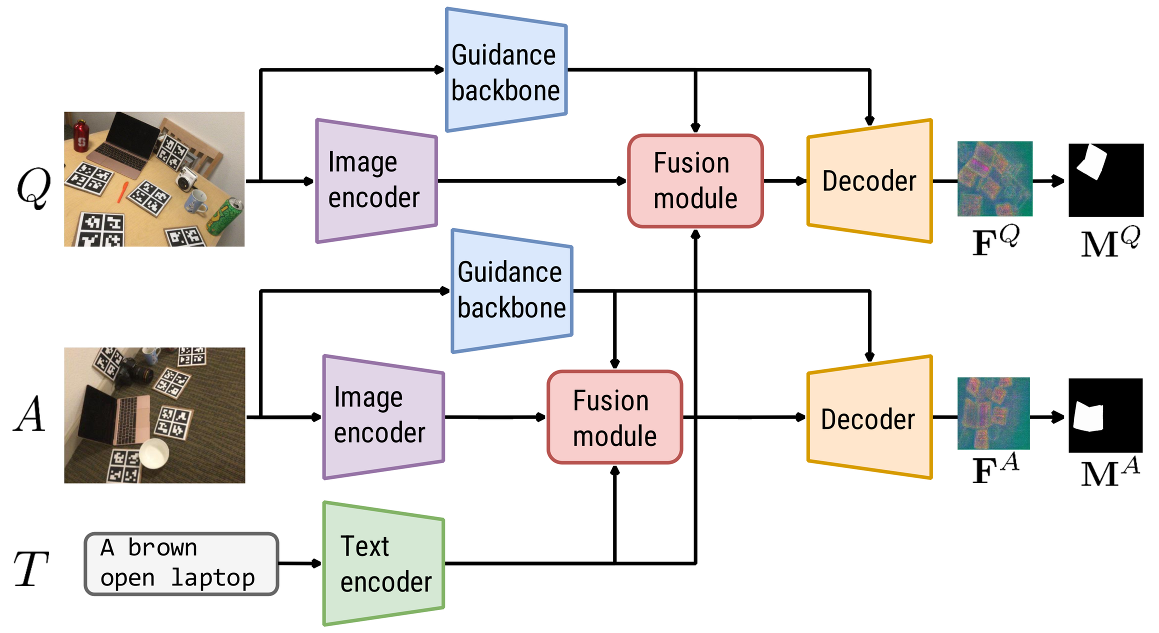
- Novel open-vocabulary **formulation** for unseen-object 6D pose estimation
- Novel **architecture** based on Visual-Language-Models to tackle this task
- New **benchmark** of 4K images and 34 objects with textual annotations

Benchmark examples

Our method: Oryon

Formulation

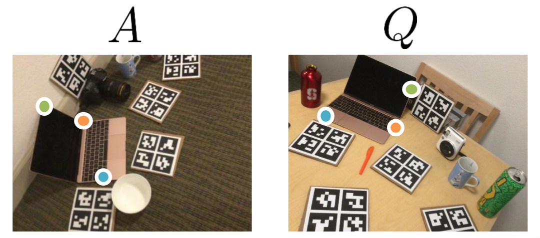
- Given an object O present in A and Q, Oryon estimates $T_{A \rightarrow Q}$: the 6D pose of O in Q with respect to A
- O is described by a textual prompt T given by the user
- At train time $O \in O_{train}$, at test time $O \in O_{test}$, with $O_{train} \cap O_{test} = \varnothing$
- A and Q show different scenes, and are represented as RGBD images



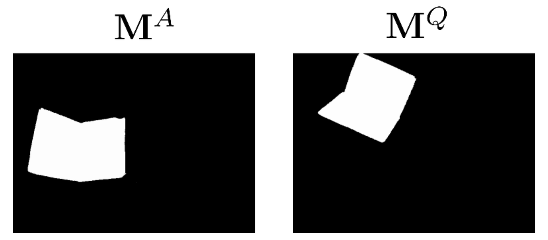
Method highlights

- Fusion and upsampling of textual and visual features from CLIP
- Joint prediction of feature map and segmentation mask for each scene
- Match-based registration algorithm to obtain the final object 6D pose

Ground-truth matches



Mask optimization

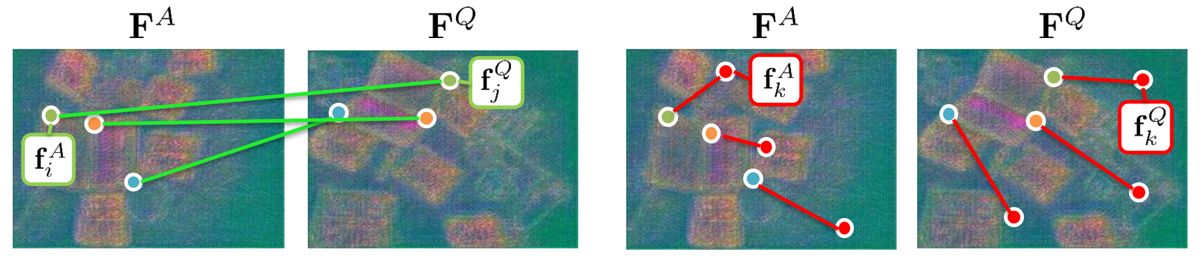


Dice mask loss ℓ_M

Feature optimization

$$\ell_P = \sum_{(i,j)\in\mathcal{P}} \frac{1}{|\mathcal{P}|} \left(\mathrm{dist}(\mathbf{f}_i^A, \mathbf{f}_j^Q) - \mu_P\right)_+$$

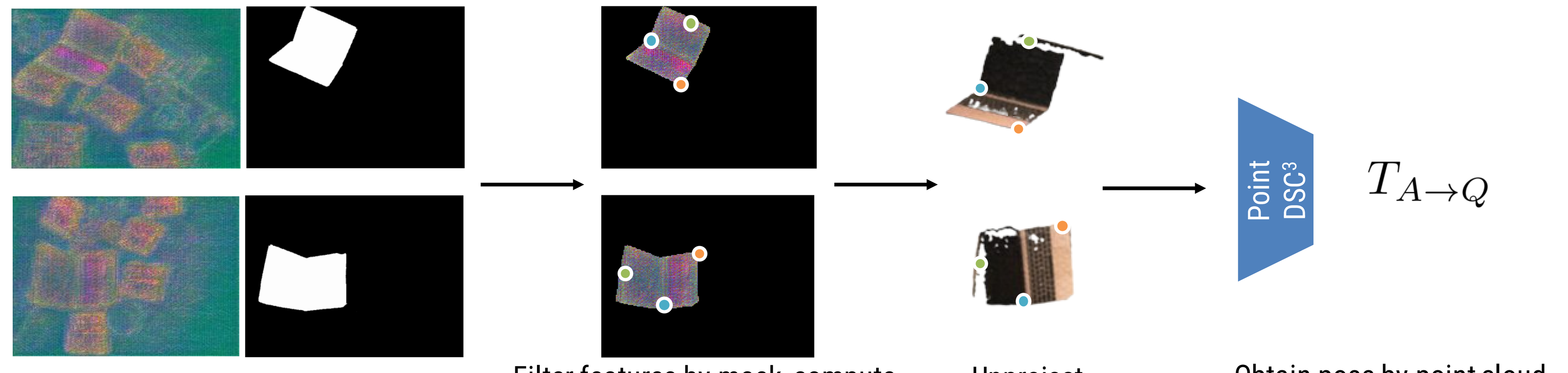
$$\ell_N = \sum_{(i,j)\in\mathcal{P}} \frac{1}{2|\mathcal{P}_i|} \left(\mu_N - \min_{k\in\mathcal{N}_i} \mathrm{dist}(\mathbf{f}_i^A, \mathbf{f}_k^A)\right)_+ + \frac{1}{2|\mathcal{P}_j|} \left(\mu_N - \min_{k\in\mathcal{N}_j} \mathrm{dist}(\mathbf{f}_j^Q, \mathbf{f}_k^Q)\right)_+$$



Positive feature loss ℓ_P

Hardest-negative loss ℓ_N

Inference procedure



Infer features and masks

Filter features by mask, compute 2D matches by nearest neighbour

Unproject matches to 3D

Obtain pose by point cloud registration algorithm

Results

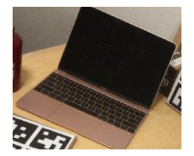
Quantitative results

Method	Mask	REAL275[4] AR	VSD	MSSD	MSPD	ADD	mIoU	Toyota-Light[5] AR	VSD	MSSD	MSPD	ADD	mIoU
SIFT[7]	Oracle	34.1	16.5	37.9	48.0	16.4	100.0	30.3	7.3	39.6	44.1	14.1	100.0
	OVSeg[6]	18.3	8.6	19.9	26.5	7.4	56.4	25.8	6.4	34.2	36.9	11.8	75.5
	Ours	24.4	12.2	27.3	33.8	12.8	66.5	27.2	5.7	35.4	40.6	9.9	68.1
ObjectMatch[8]	Oracle	26.0	15.5	31.7	30.8	13.4	100.0	9.8	2.4	13.0	14.0	5.4	100.0
	OVSeg[6]	14.9	9.1	18.8	16.8	7.8	56.4	9.2	2.6	12.1	13.0	5.3	75.5
	Ours	22.4	14.1	27.9	25.2	13.2	66.5	8.3	2.2	10.5	12.1	3.8	68.1
Oryon	Oracle	46.5	32.1	50.9	56.7	34.9	100.0	34.1	13.9	42.9	45.5	22.9	100.0
	OVSeg[6]	26.4	18.3	29.4	31.5	17.2	56.4	29.2	11.9	36.8	38.9	18.9	**75.5**
	Ours	**32.2**	**23.6**	**36.6**	**36.4**	**24.3**	**66.5**	**30.3**	**12.1**	**37.5**	**41.4**	**20.9**	68.1
Δ score		**+7.8**	**+9.5**	**+8.7**	**+2.6**	**+11.1**	**+10.1**	**+3.1**	**+6.4**	**+2.1**	**+0.8**	**+11.0**	**-7.4**

Prompt influence on REAL275

Prompt	Mask	AR	ADD	mIoU
No name	Oracle	38.6	21.5	100.0
	Ours	3.0	0.4	3.0
Misleading	Oracle	39.4	24.9	100.0
	Ours	25.4	14.5	56.4
Generic	Oracle	39.0	26.1	100.0
	Ours	30.0	19.0	63.0
Standard	Oracle	46.5	34.9	100.0
	Ours	**32.2**	**24.3**	**66.5**

Example prompts



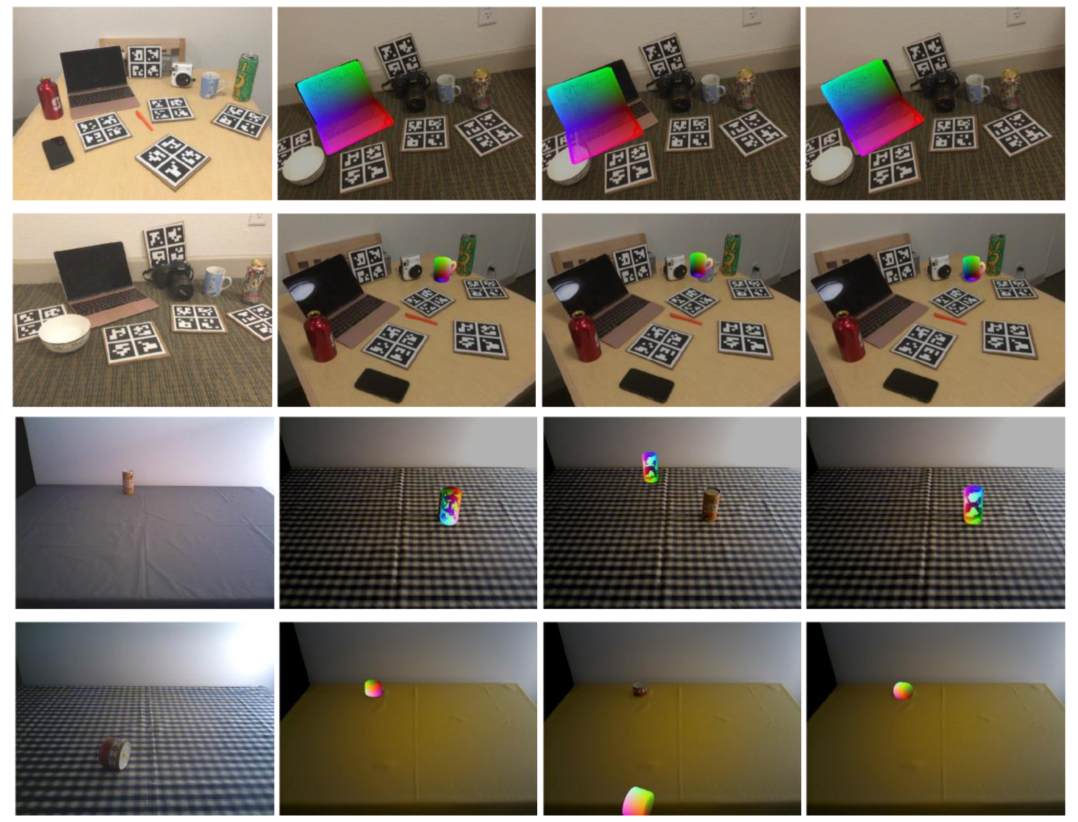
No name: A brown open object

Misleading: A white closed laptop

Generic: A laptop

Standard: A brown open laptop

Qualitative results



Limitations and future works

- Oryon requires camera intrinsics and depth information: a depth estimator could be used instead
- Object textual description could be difficult to provide for some objects (e.g., industrial components)
- Prompt detail is limited by training data: an ad-hoc dataset could be generated by using LLMs

1. Shugurov, Ivan, et al. "Osop: A multi-stage one shot object pose estimation framework." CVPR 2022.
2. He, Xingyi, et al. "Onepose++: Keypoint-free one-shot object pose estimation without CAD models." NeurIPS 2022.
3. Bai, Xuyang, et al. "Pointdsc: Robust point cloud registration using deep spatial consistency." CVPR 2021.
4. Wang, He, et al. "Normalized object coordinate space for category-level 6d object pose and size estimation." CVPR 2019.
5. Hodan, Tomas, et al. "Bop: Benchmark for 6d object pose estimation." ECCV 2018.
6. Liang, Feng, et al. "Open-vocabulary semantic segmentation with mask-adapted clip." CVPR 2023.
7. Lowe, David G. "Object recognition from local scale-invariant features." ICCV 1999.
8. Gümeli, Can, et al. "ObjectMatch: Robust Registration using Canonical Object Correspondences." CVPR 2023.

This work was supported by the European Union's Horizon Europe research and innovation programme under grant agreement No 101058589 (AI-PRISM), and it made use of time on the Tier 2 HPC facility JADE2, funded by EPSRC (EP/T022205/1).