RÉPUBLIQUE FRANÇAISE

Liberté
Égalité
Fraternité

Thomas Rey

ONERA

THE FRENCH AEROSPACE LAB

www.onera.fr

RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

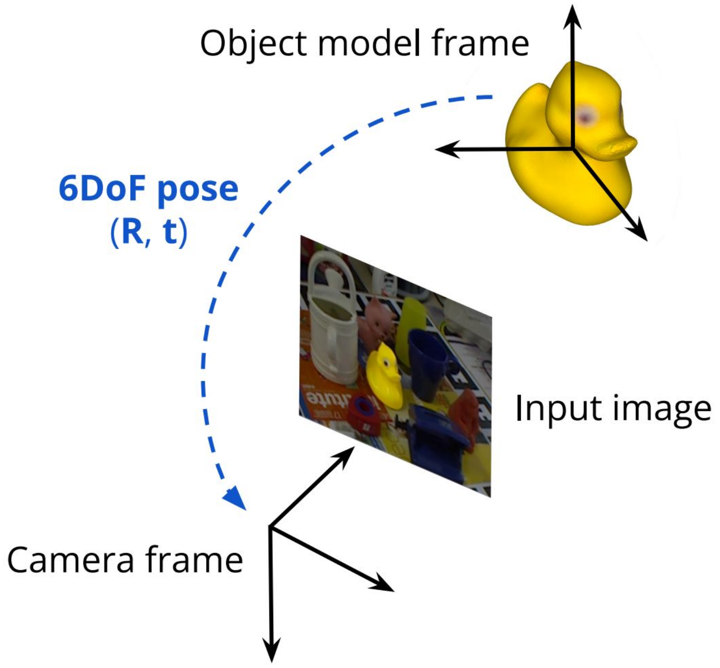
ONERA
THE FRENCH AEROSPACE LAB

Object pose estimation

Summary

Introduction

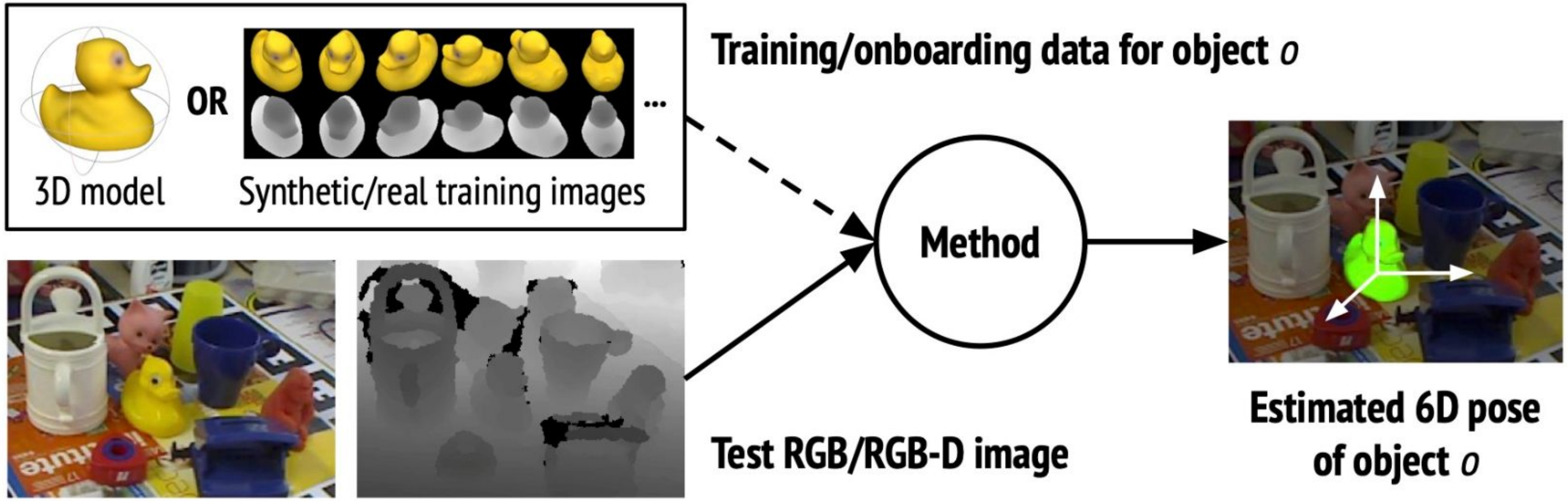
What is intended in the pose estimation task?



- Determine the **3D translation vector**
- Determine the **3D rotation matrix**
- Pose **relative to** the camera frame
- Possible inputs:
  1. RGB image
  2. RGB-D image
  3. Point cloud
  4. CAD model of the object

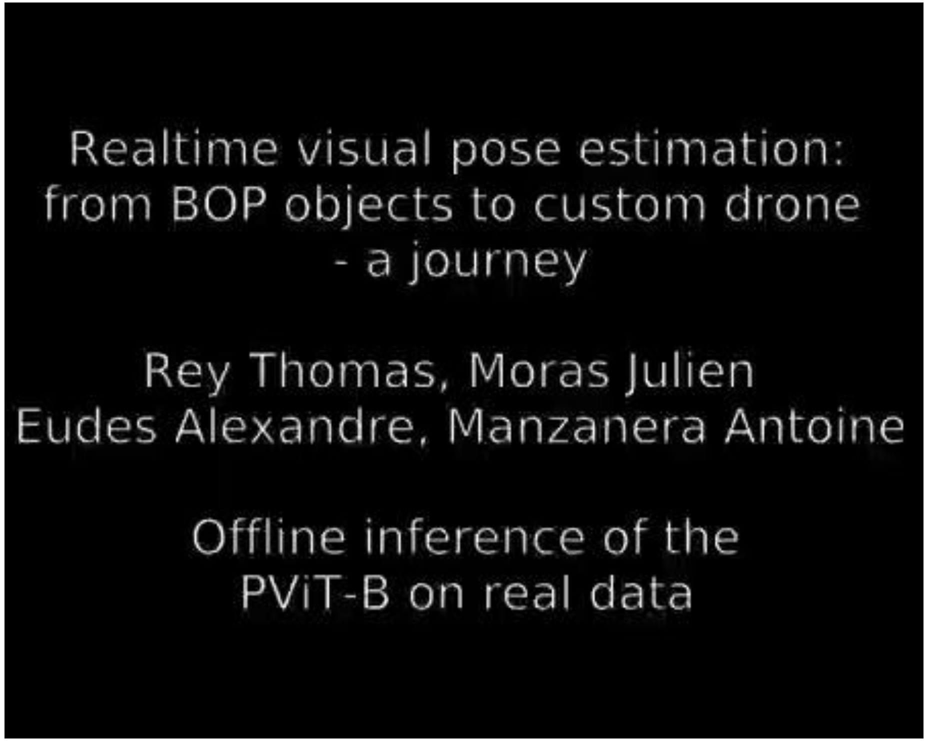
Hodan, Tomas, et al. "BOP Challenge 2023 on Detection Segmentation and Pose Estimation of Seen and Unseen Rigid Objects." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

How does it work?


3D model
OR
Synthetic/real training images
...
Training/onboarding data for object o
Method
Test RGB/RGB-D image
Estimated 6D pose of object o

Why is it important to estimate the pose?

Estimating the pose of any drone seen by another drone equipped with a camera.



Realtime visual pose estimation: from BOP objects to custom drone - a journey

Rey Thomas, Moras Julien
Eudes Alexandre, Manzanera Antoine

Offline inference of the PViT-B on real data

Why is it important to estimate the pose?

Robotics: Precise information needed for the prehension task

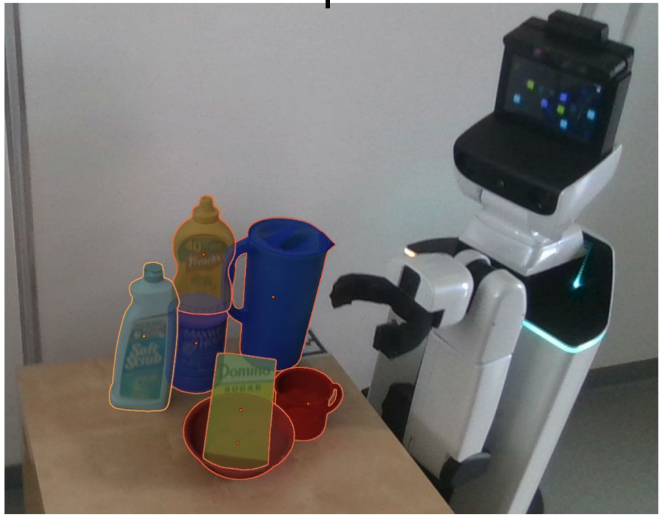


Fig. 1. **Object poses** YCB-V [6] objects overlaid with their annotated poses.

Augmented reality: to superpose virtual objects onto the real world

We are going to focus on RGB methods, mostly used in robotics.

Bauer, Dominik, et al. "Challenges for monocular 6d object pose estimation in robotics." *IEEE Transactions on Robotics* (2024).
https://www.smartrental.fr/fr/blog/tablettes/la-realite-augmentee-un-monde-de-possibilites-a-portee-de-la-main

How to compare the different methods

Usual metrics:

- **ADD:** Average distance between 3D points from the CAD model transformed by the GT and estimated pose, usually used with a threshold of 10% of the object's diameter.
- **Visible Surface Discrepancy (VSD):** Compares the superposition of visible surfaces using the GT and estimated poses (more robust to occlusion).
- **X° or X cm:** Precision metrics using X (° or cm) as a threshold.

$$\frac{1}{m}\sum_{\mathbf{x}\in\mathcal{M}} \|(\mathrm{Rot}(\tilde{\mathbf{r}},\mathbf{x})+\tilde{\mathbf{t}})-(\mathrm{Rot}(\mathbf{r},\mathbf{x})+\mathbf{t})\|_2$$

$$e_{\mathrm{VSD}}(\hat{S},\bar{S},S_I,\hat{V},\bar{V},\tau) = \underset{p\in\hat{V}\cup\bar{V}}{\mathrm{avg}} \begin{cases} 0 & \text{if } p\in\hat{V}\cap\bar{V} \,\wedge\, |\hat{S}(p)-\bar{S}(p)|<\tau \\ 1 & \text{otherwise.} \end{cases} \quad (1)$$

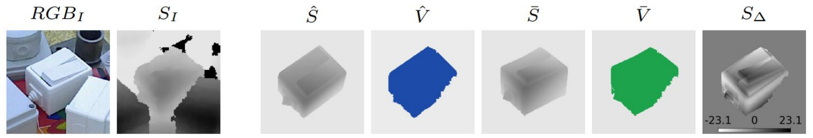


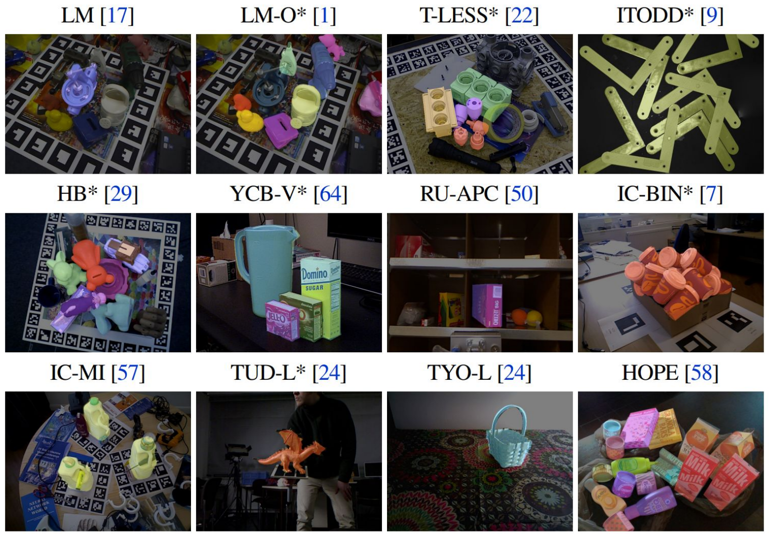
Fig. 2. Quantities used in the calculation of e_{VSD}. Left: Color channels RGB_I (only for illustration) and distance map S_I of a test image I. Right: Distance maps $\hat{S}$ and $\bar{S}$ are obtained by rendering the object model $\mathcal{M}$ at the estimated pose $\hat{\mathbf{P}}$ and the ground-truth pose $\bar{\mathbf{P}}$ respectively. $\hat{V}$ and $\bar{V}$ are masks of the model surface that is visible in I, obtained by comparing $\hat{S}$ and $\bar{S}$ with S_I. Distance differences $S_\Delta(p) = \hat{S}(p) - \bar{S}(p)$, $\forall p \in \hat{V}\cap\bar{V}$, are used for the pixel-wise evaluation of the surface alignment.

Hodan, Tomas, et al. "BOP Challenge 2023 on Detection Segmentation and Pose Estimation of Seen and Unseen Rigid Objects." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

BOP Challenge

The goal of BOP is to capture the state of the art in 6DoF object pose estimation and related tasks such as 2D object detection and segmentation. An accurate, fast, robust, scalable and easy-to-train method that solves this task will have a big impact in application fields such as robotics or augmented reality.

REAL DATASETS:



SYNTHETIC DATASETS:

Possible challenges : Multi-instance / objects / occlusion / category / symmetries / challenging materials

https://bop.felk.cvut.cz/home/

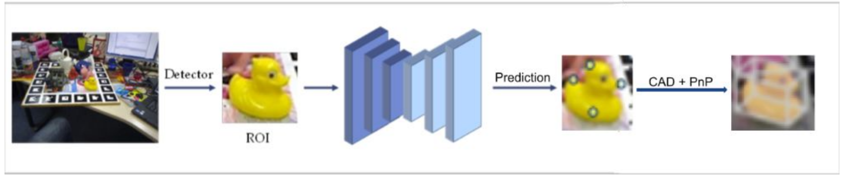
RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

ONERA
THE FRENCH AEROSPACE LAB

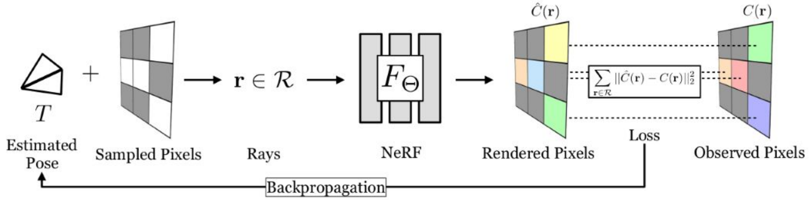
Instance pose estimation

4 paradigms

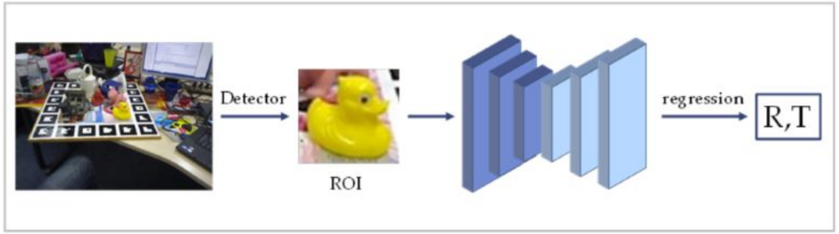
Correspondence methods:



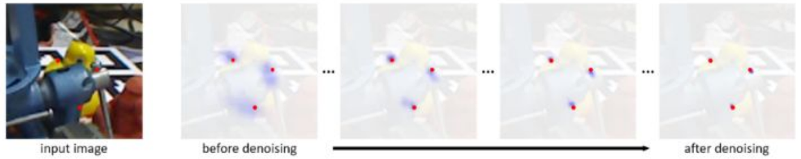
Render and compare methods:



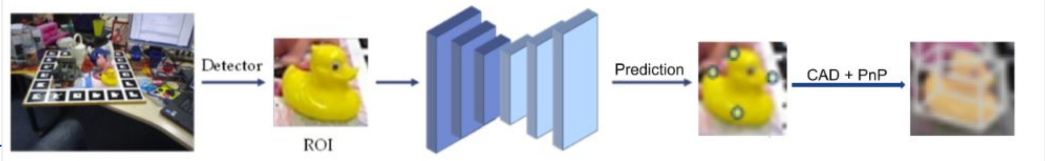
Regression methods:



Diffusion methods:

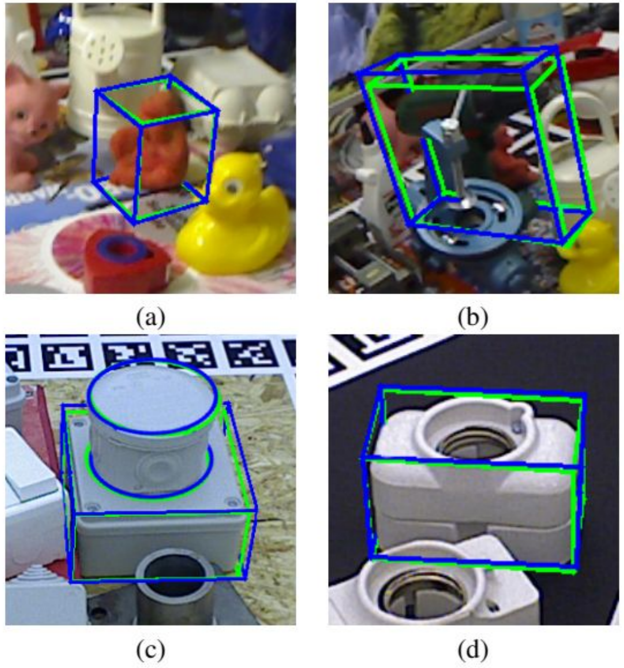


Correspondence methods



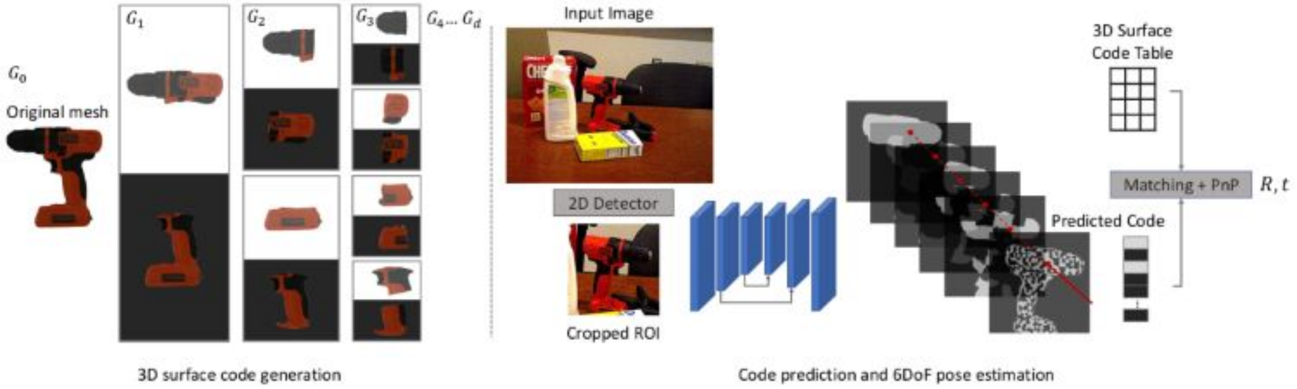
BB8:

- **“Old” method (2017)**
- Estimates the pose with **2D-3D correspondences**
- Uses the 8 corners of the 3D bounding box
- **Known** in 3D space, their projections are **predicted** on the image (2D space)
- Uses the PnP algorithm to find the 6D pose



Rad, Mahdi, and Vincent Lepetit. "Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth." *Proceedings of the IEEE international conference on computer vision*. 2017

Correspondence methods: ZebraPose



- Dense discrete descriptor to represent the surface of the object
- Hierarchical encoding and classification (binary code for each pixel)

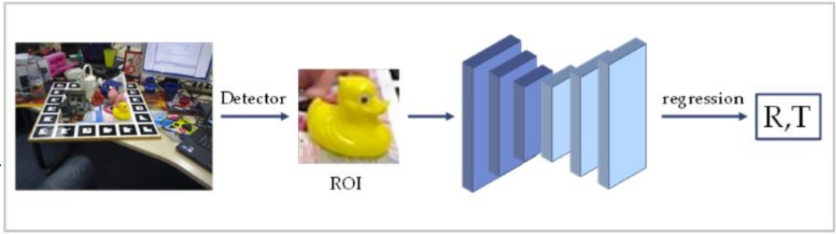
- Coarse-to-fine training (from coarse classification [low-level bits] to high-level bits)

- Hierarchical encoding is a unique descriptor enabling direct matching between pixels and the object surface
- Dense descriptors -> Robust to occlusion

- Uses the PnP algorithm -> Slow (~4 Hz)

Su, Yongzhi, et al. "Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Regression methods



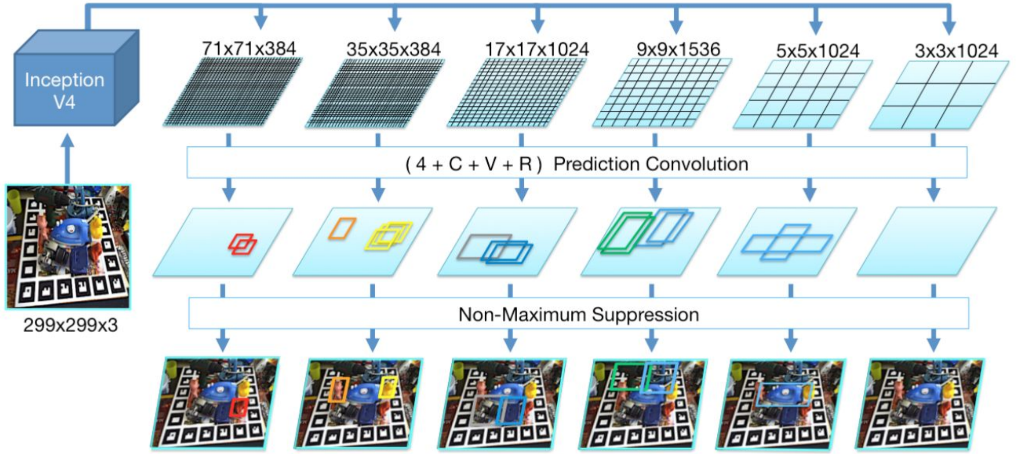


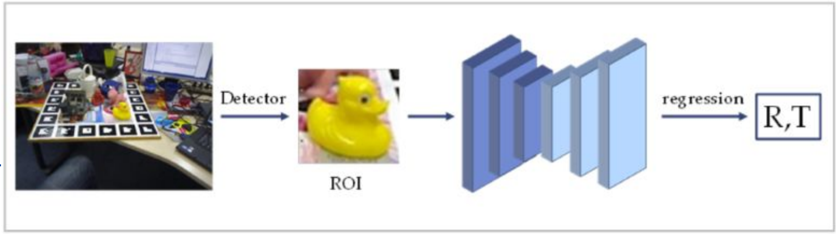
Figure 1: Schematic overview of the SSD-style network prediction. We feed our network with a 299 × 299 RGB image and produce six feature maps at different scales from the input image using branches from InceptionV4. Each map is then convolved with trained prediction kernels of shape (4 + C + V + R) to determine object class, 2D bounding box as well as scores for possible viewpoints and in-plane rotations that are parsed to build 6D pose hypotheses. Thereby, C denotes the number of object classes, V the number of viewpoints and R the number of in-plane rotation classes. The other 4 values are utilized to refine the corners of the discrete bounding boxes to tightly fit the detected object.

SSD-6D:

- **“Old” method (2017)**
- **Parameterization** of the 6D pose: viewpoint (V possibilities), in-plane rotation (R possibilities), and the 3D position
- **Fast** (~10Hz)
- **Discretization** of the rotation from SO(3) -> $\mathbb{R}^n$

Kehl, Wadim, et al. "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again." *Proceedings of the IEEE international conference on computer vision*. 2017

Regression methods



Pose parametrization:

Translation (TP):

- 2D position of the projected object's centroid
- Distance between the camera and the object

Scale-invariant TP:

$$\begin{cases} \gamma_x = (c_x - o_x)/s \\ \gamma_y = (c_y - o_y)/s \\ \gamma_z = t_z/r \end{cases}$$

r = Img_size / RoI_size

Size-invariant TP:

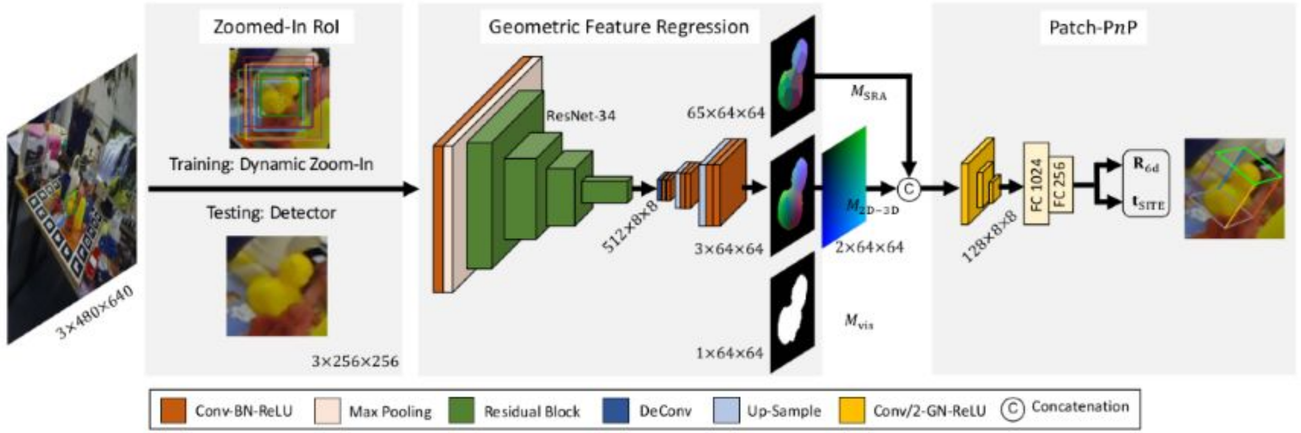
$$r = f \cdot s_{obj}/s$$

Rotation:

- The rotation representation needs to have no singularity (Euler angles / quaternions) -> requires more than 5 parameters
- The R6D representation is mostly used (first two columns of the rotation matrix) -> Gram-Schmidt is used to find the last one

Wang, Gu, et al. "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021
Stapf, Sebastian, Tobias Bauernfeind, and Marco Riboldi. "PViT-6D: Overclocking Vision Transformers for 6D Pose Estimation with Confidence-Level Prediction and Pose Tokens." *arXiv preprint arXiv:2311.17504* (2023)

Regression methods: GDR-Net

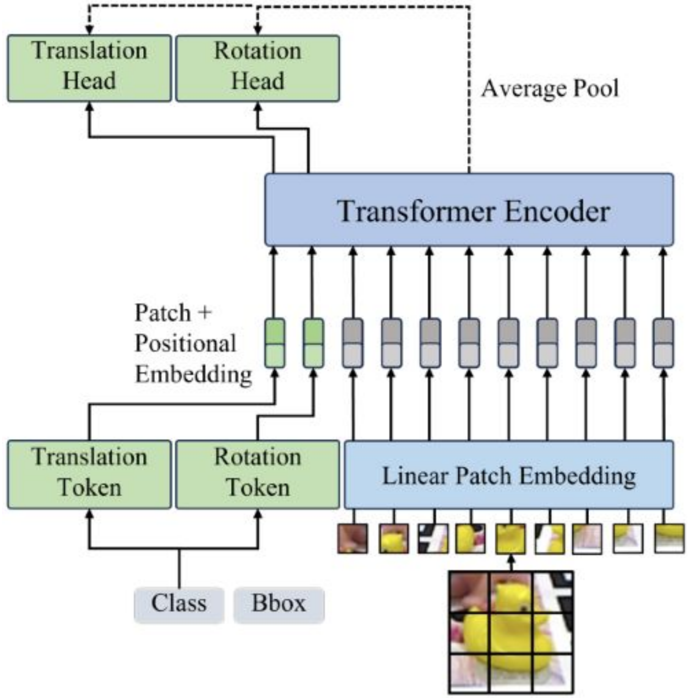


- **Mixture** of a correspondence-based and a regression-based method

- Predicts correspondence-like features (2D-3D / Visible surface / visible mask)
- Patch-PnP: Deep Learning method to replace the time-consuming PnP
- Loss function can be calculated in the pose space (not limited to the descriptors)
- Data augmentation is used to improve robustness to poor 2D detections

Wang, Gu, et al. "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021

Regression methods: PViT-6D

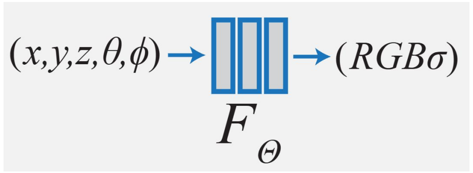


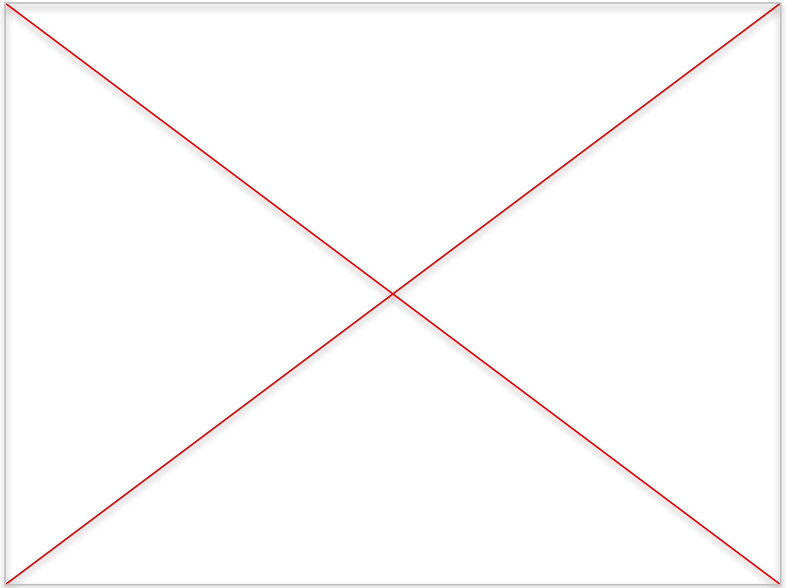
- One of the first methods using Transformers
- Tokens are used to collect specific features for each prediction (translation and rotation)
- One MLP is used for each prediction
- Fast (50Hz)
- Calculates a confidence score for the pose estimation

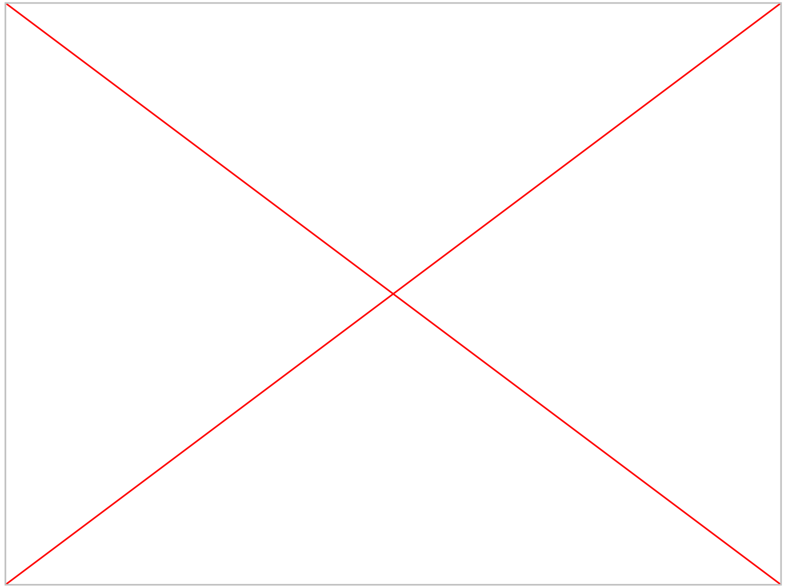
Stapf, Sebastian, Tobias Bauernfeind, and Marco Riboldi. "PViT-6D: Overclocking Vision Transformers for 6D Pose Estimation with Confidence-Level Prediction and Pose Tokens." *arXiv preprint arXiv:2311.17504* (2023)

Render and compare methods

What is a NeRF ? ➡ A method for synthesizing novel views of complex scenes by optimizing an underlying continuous volumetric scene function using a sparse set of input views

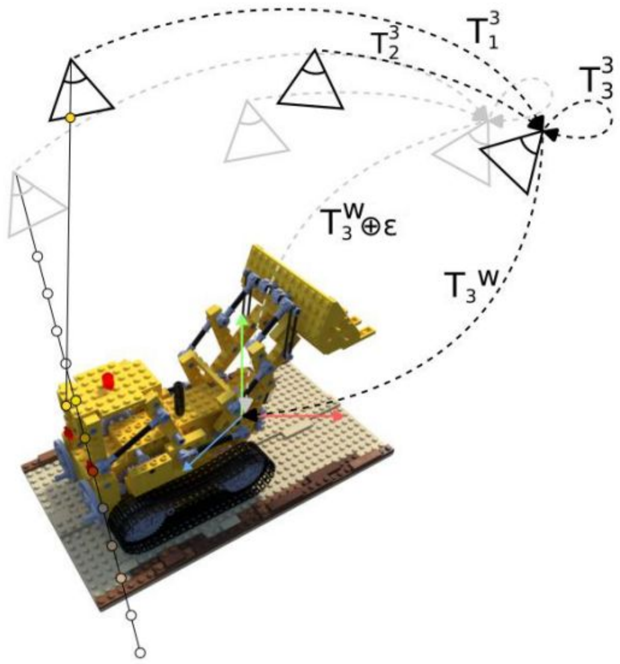






Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *Communications of the ACM* 65.1 (2021): 99-106

Render and compare methods



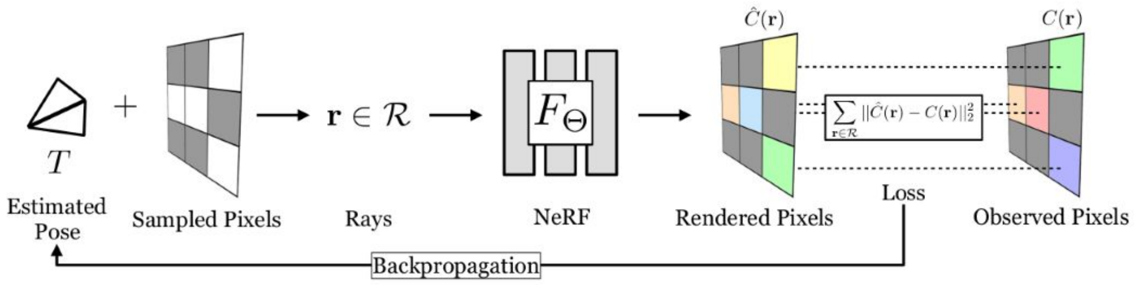
- **NeRF** trained to represent the object
- **Iterative** methods (compare the GT image with the image rendered from the estimated pose)
- Rendering is done by a **NeRF** (needs a **CAD** model for training)

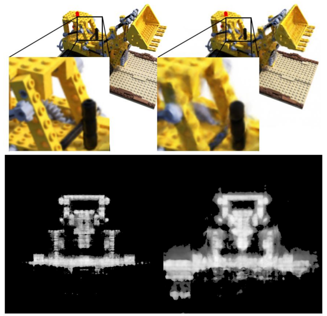
- Very precise (if the NeRF is well trained)
- Robust to clutter and occlusion

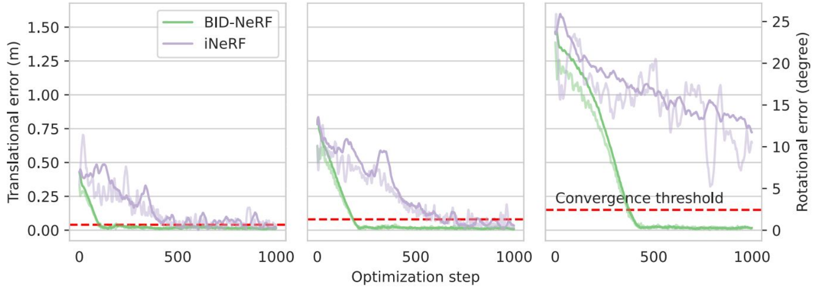
- Slow method
- Needs a good initial guess (prone to local minima)

Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *Communications of the ACM* 65.1 (2021): 99-106

Render and compare methods: iNeRF / BID-NeRF







- iNeRF **first** article to propose the approach
- BID-NeRF optimizes the technique to make it much **faster** (using a **coarse** model)
- Less prone to local minima
- New methods use **Gaussian splatting**

Yen-Chen, Lin, et al. "inerf: Inverting neural radiance fields for pose estimation." *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021

Render and compare methods: Gaussian Splatting

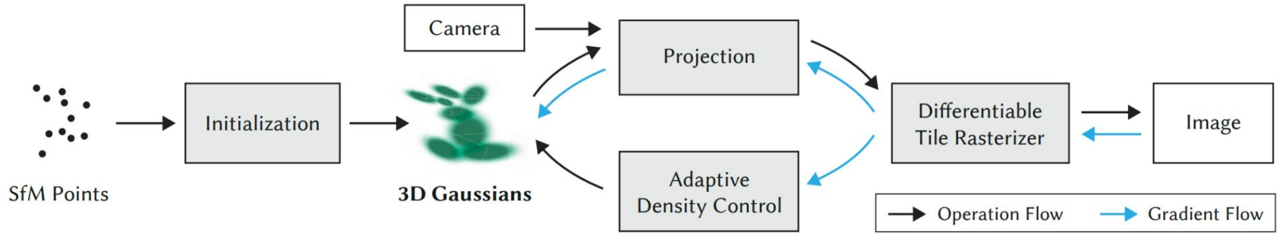
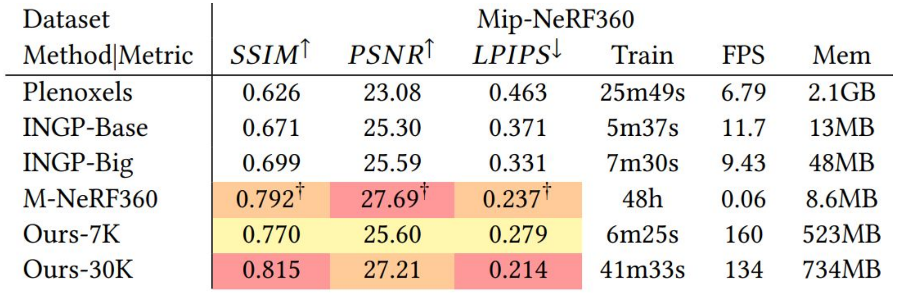
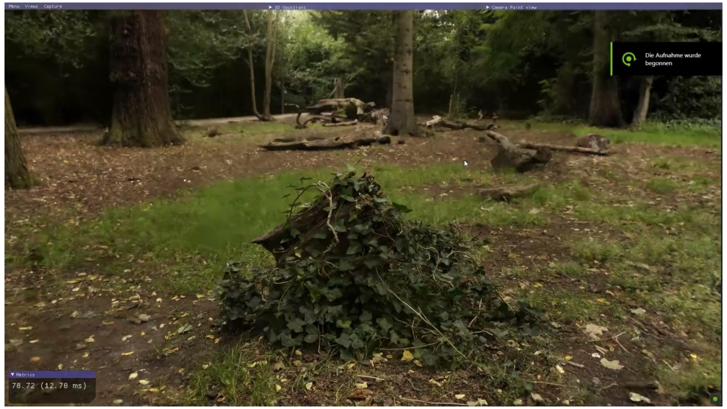


Fig. 2. Optimization starts with the sparse SfM point cloud and creates a set of 3D Gaussians. We then optimize and adaptively control the density of this set of Gaussians. During optimization we use our fast tile-based renderer, allowing competitive training times compared to SOTA fast radiance field methods. Once trained, our renderer allows real-time navigation for a wide variety of scenes.



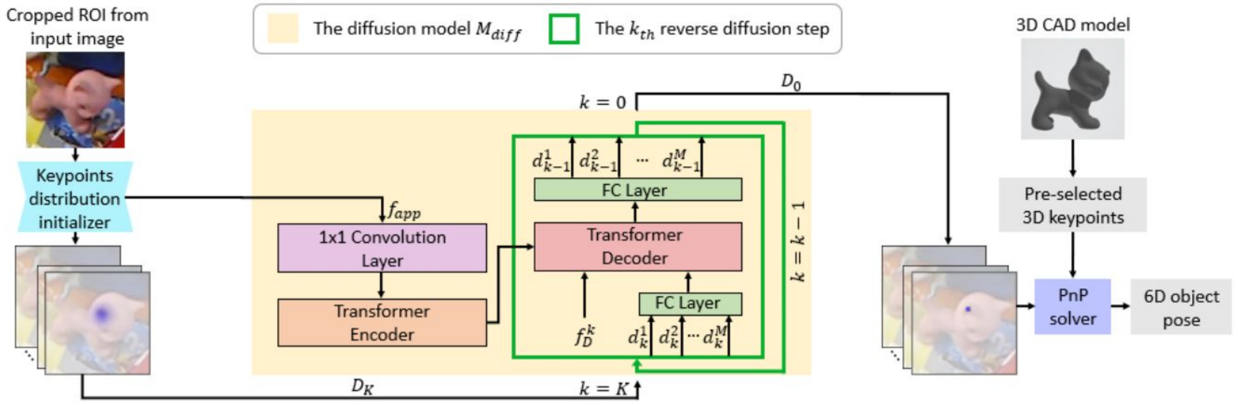
Dataset	Mip-NeRF360					
Method|Metric	*SSIM*↑	*PSNR*↑	*LPIPS*↓	Train	FPS	Mem
Plenoxels	0.626	23.08	0.463	25m49s	6.79	2.1GB
INGP-Base	0.671	25.30	0.371	5m37s	11.7	13MB
INGP-Big	0.699	25.59	0.331	7m30s	9.43	48MB
M-NeRF360	0.792†	27.69†	0.237†	48h	0.06	8.6MB
Ours-7K	0.770	25.60	0.279	6m25s	160	523MB
Ours-30K	0.815	27.21	0.214	41m33s	134	734MB

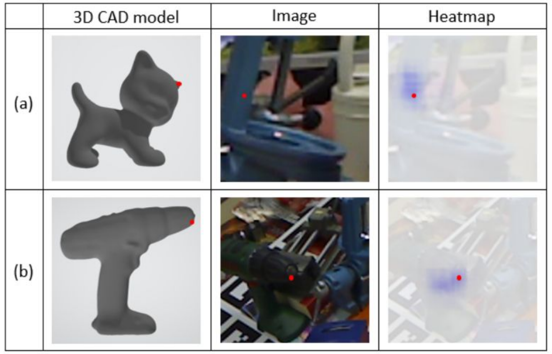


Kerbl, Bernhard, et al. "3D Gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.* 42.4 (2023): 139-1.

Diffusion methods

- Similar to the correspondence methods
- **Denoising** of the 2D position of keypoints, conditioned on the image


Cropped ROI from input image
The diffusion model M_{diff}
The k_{th} reverse diffusion step
3D CAD model
Keypoints distribution initializer
f_{app}
1x1 Convolution Layer
Transformer Encoder
FC Layer
Transformer Decoder
FC Layer
d_{k-1}^1 d_{k-1}^2 ... d_{k-1}^M
f_D^k d_k^1 d_k^2 ... d_k^M
$k = 0$
D_0
$k = k - 1$
D_K
$k = K$
Pre-selected 3D keypoints
PnP solver
6D object pose



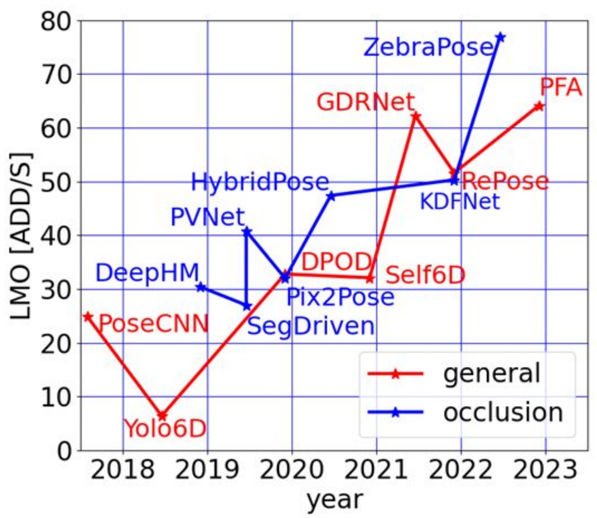
- Good for cluttered scenes
- Good performance

- Complex modeling of the noise (training phase)
- Quite slow as it is an iterative process

Xu, Li, et al. "6d-diff: A keypoint diffusion framework for 6d object pose estimation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024

Performance of differents methods

Fast evolution



Highly competitive research domain:

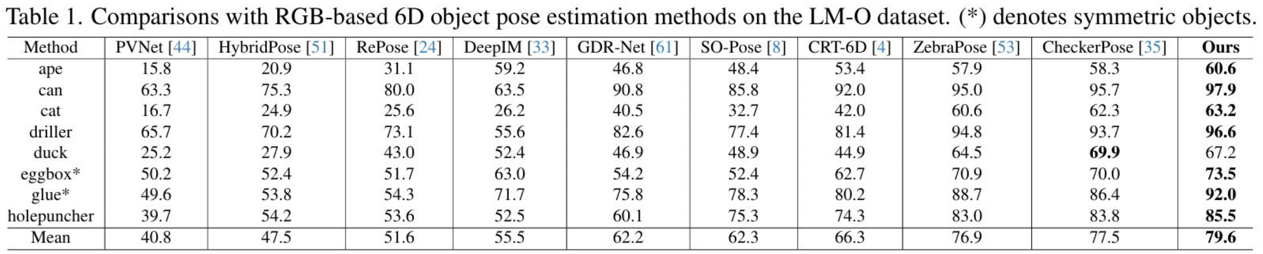
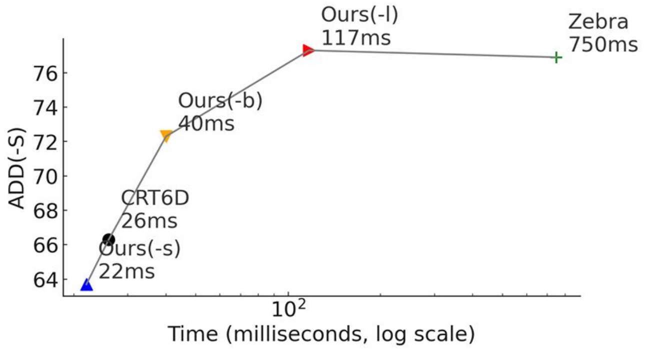


Table 1. Comparisons with RGB-based 6D object pose estimation methods on the LM-O dataset. (*) denotes symmetric objects.

Method	PVNet [44]	HybridPose [51]	RePose [24]	DeepIM [33]	GDR-Net [61]	SO-Pose [8]	CRT-6D [4]	ZebraPose [53]	CheckerPose [35]	Ours
ape	15.8	20.9	31.1	59.2	46.8	48.4	53.4	57.9	58.3	**60.6**
can	63.3	75.3	80.0	63.5	90.8	85.8	92.0	95.0	95.7	**97.9**
cat	16.7	24.9	25.6	26.2	40.5	32.7	42.0	60.6	62.3	**63.2**
driller	65.7	70.2	73.1	55.6	82.6	77.4	81.4	94.8	93.7	**96.6**
duck	25.2	27.9	43.0	52.4	46.9	48.9	44.9	64.5	**69.9**	67.2
eggbox*	50.2	52.4	51.7	63.0	54.2	52.4	62.7	70.9	70.0	**73.5**
glue*	49.6	53.8	54.3	71.7	75.8	78.3	80.2	88.7	86.4	**92.0**
holepuncher	39.7	54.2	53.6	52.5	60.1	75.3	74.3	83.0	83.8	**85.5**
Mean	40.8	47.5	51.6	55.5	62.2	62.3	66.3	76.9	77.5	**79.6**

Inference time is fairly important:



RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

ONERA
THE FRENCH AEROSPACE LAB

How to evaluate the confidence in the estimated pose ?

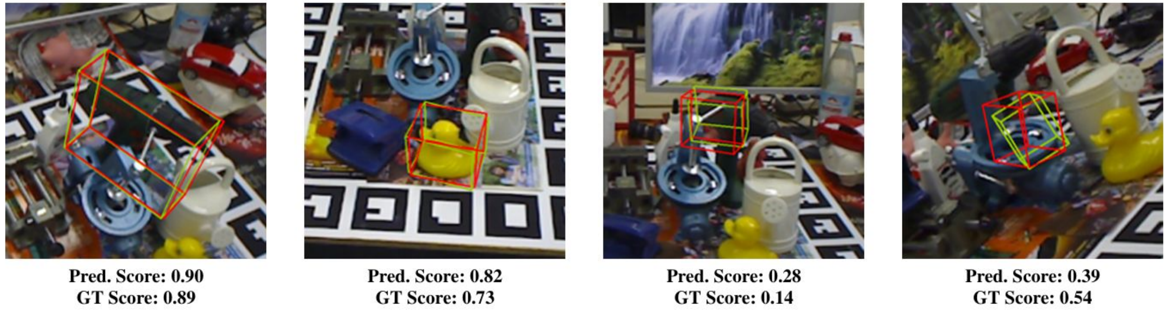
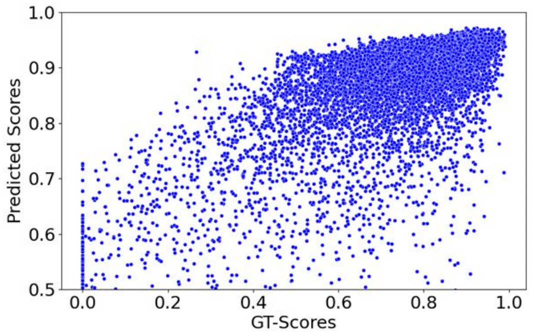


Figure 3. Visualization of predicted confidence score and ground truth 3D-IoU score. The green 3D bounding box depicts the box transformed with the ground truth pose, while the red box is transformed with the predicted pose. Below the imges the individual predicted scores and ground truth scores are shown.



First authors to introduce the notion of confidence of the estimated pose

- A classification model is trained to evaluate the confidence score (~3D IoU)
- Finding the relation between easy and hard estimation
- Good interpretability
- Easy to implement in all pose estimators

Problems of the instance pose estimation techniques

- **Occlusion** / cluttered scenes
- Object **symmetries**
- The **rotation estimation** is complex
- New techniques use an **object detector**, needs to have a good **precision and recall** to be used in a real-world application
- Suffers from poor RoI detection, even with data augmentation during training
- Needs **precise CAD** models, which are costly and time-consuming
- **Deformable objects** are great challenge
- Estimating the pose of known object is **not transferable** to any real-world applications

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

ONERA
THE FRENCH AEROSPACE LAB

Category pose estimation

Category pose estimation

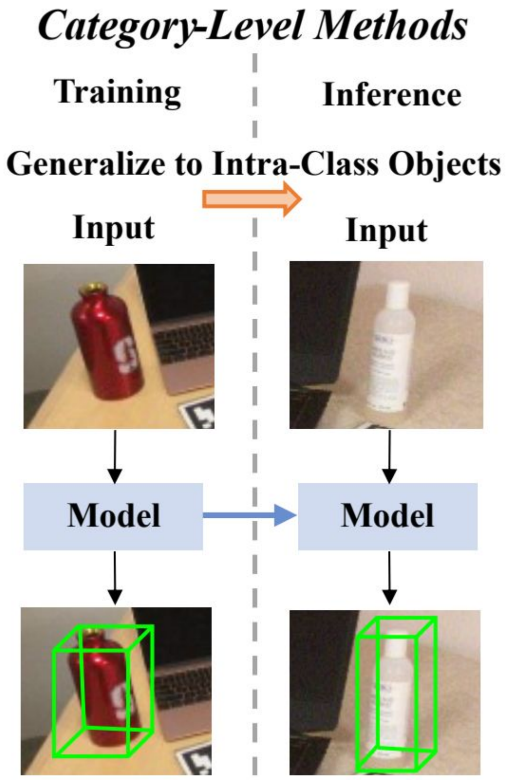
- Predicts the 6D pose of **unseen object instances** belonging to a **predefined category**

- Aims to generalize across **variations within a category**

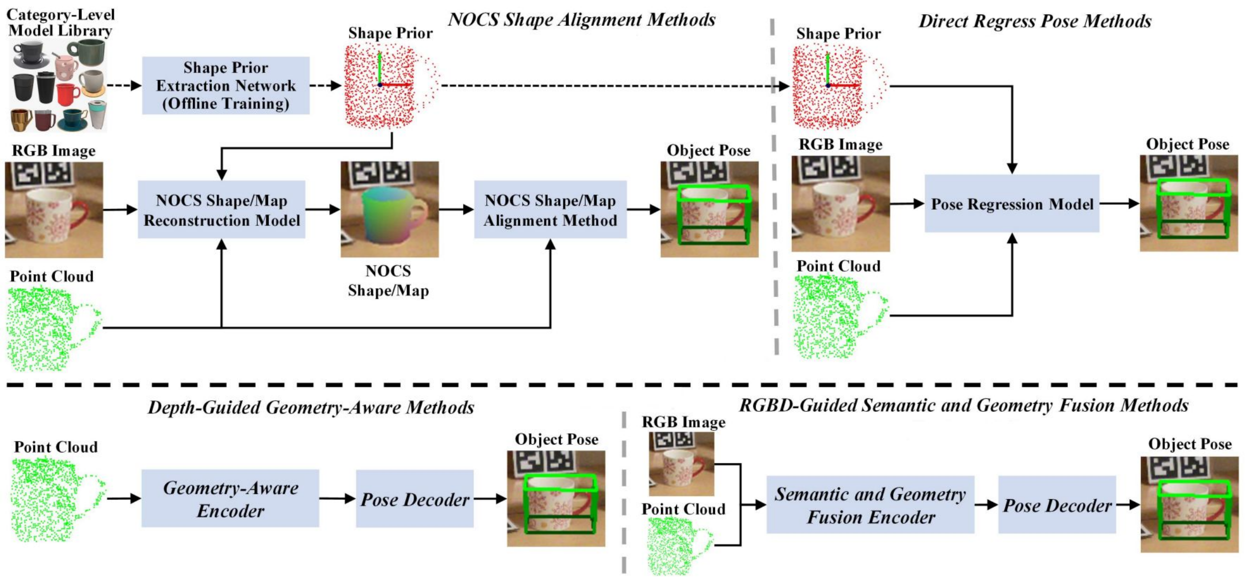
Challenges:

- Handling variations in shape, size, texture can be difficult

- Limited **training data** (CAD models)



Category pose estimation: Main approaches

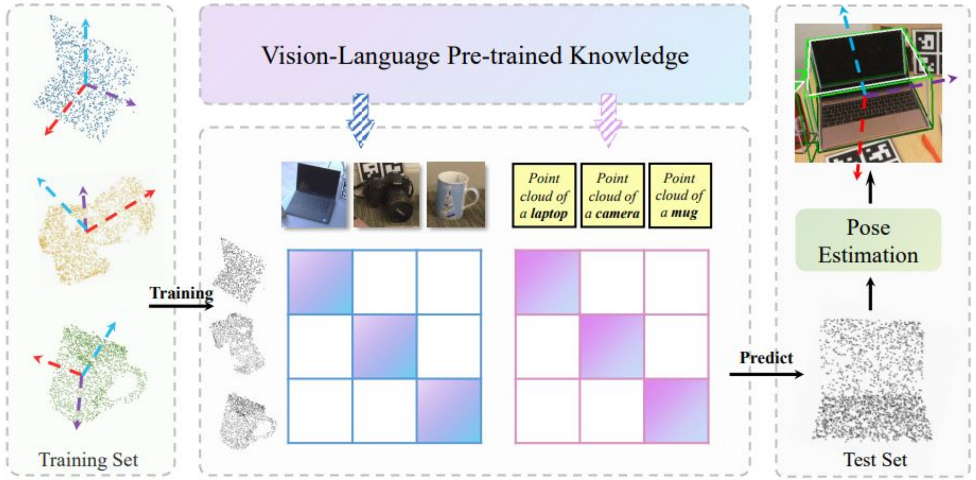

Category-Level Model Library
Shape Prior Extraction Network (Offline Training)
Shape Prior
NOCS Shape Alignment Methods
RGB Image
NOCS Shape/Map Reconstruction Model
NOCS Shape/Map
NOCS Shape/Map Alignment Method
Object Pose
Point Cloud
Direct Regress Pose Methods
Shape Prior
RGB Image
Pose Regression Model
Object Pose
Point Cloud
Depth-Guided Geometry-Aware Methods
Point Cloud
Geometry-Aware Encoder
Pose Decoder
Object Pose
RGBD-Guided Semantic and Geometry Fusion Methods
RGB Image
Point Cloud
Semantic and Geometry Fusion Encoder
Pose Decoder
Object Pose

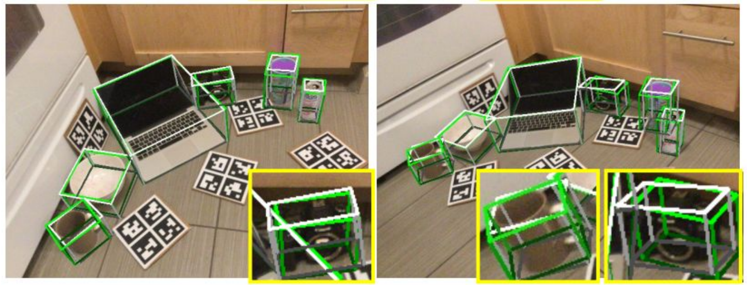
Key ideas:

- Understand the **characteristics** of each category (shape/symmetries)
- Mostly uses **point clouds**, RGB images can also be used
- Multi-modal learning (text/image)

Liu, Jian, Wei Sun, Hui Yang, Zhiwen Zeng, Chongpei Liu, Jin Zheng, Xingyu Liu, Hossein Rahmani, Nicu Sebe, and Ajmal Mian. "Deep Learning-Based Object Pose Estimation: A Comprehensive Survey." arXiv, May 31, 2024. https://doi.org/10.48550/arXiv.2405.07801.

Category pose estimation: CLIPose





CLIP is a pretrained vision-language model (classification task)

- Use 3 modalities text, point clouds, RGB (robust category-specific information)
- Contrastive learning to align the three modalities in a latent space (maximize the similarity within intra-category and minimize for different objects)
- Information about rotation and translation in the prompt (during the training phase to increase similarity between all the modalities)

Lin, Xiao, et al. "Clipose: Category-level object pose estimation with pre-trained vision-language knowledge." *IEEE Transactions on Circuits and Systems for Video Technology* (2024)

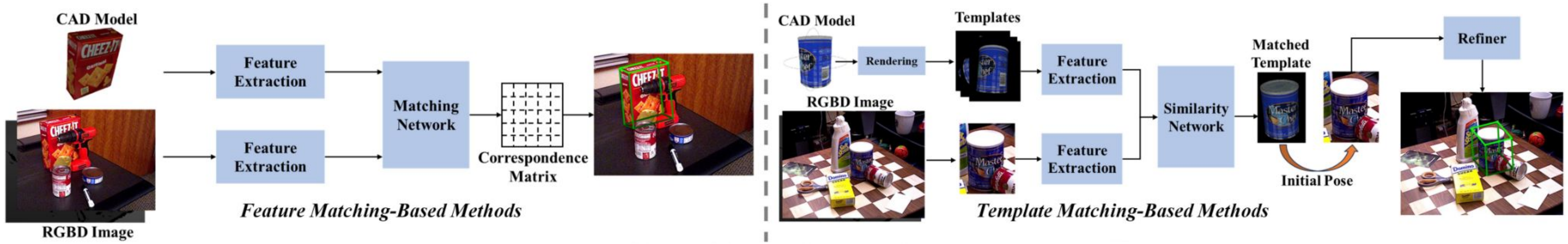
RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

ONERA
THE FRENCH AEROSPACE LAB

Unseen objects pose estimation

Unseen objects pose estimation

- Critical aspect of achieving robust pose estimation in **real-world** scenarios
- **Zero-shot** learning, poorer performance but useful in any situation
- Needs **more information** than just the image (CAD model + RGBD)



RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

ONERA
THE FRENCH AEROSPACE LAB

Opening

Opening : Symmetries

One of the major problem for pose estimators is symmetrical objects

Multiple sources of symmetries:

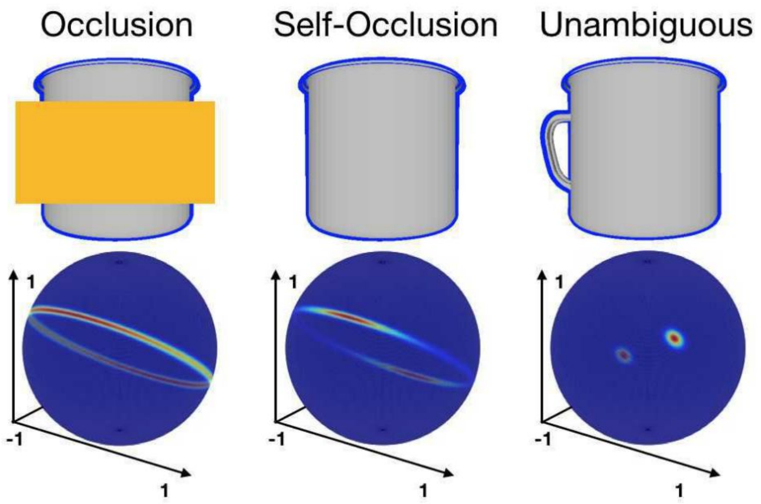
- texture (monochromatic)
- geometric (shape)
- occlusion

Type of symmetries:

- Continuous (Rotation around an axis)
- Discrete (Reflection around an axe or a plane)

- Lift the ambiguities

- Complexify the problem (architecture)
- Needs newer labels



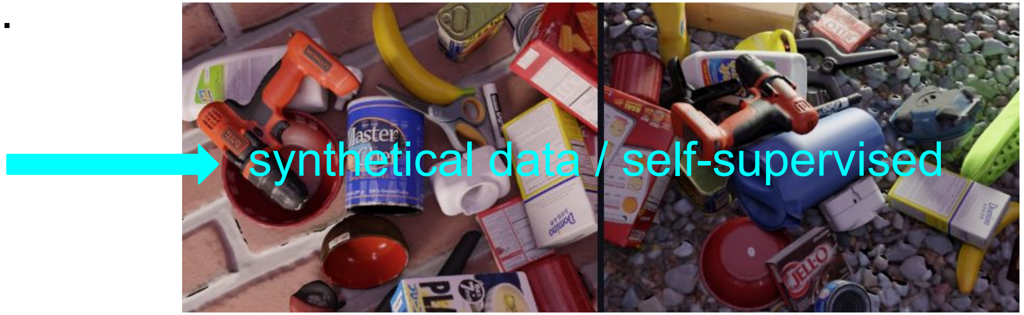
Manhardt, Fabian, et al. "Explaining the ambiguity of object detection and 6d pose from visual data." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

Opening : Lack of data

One of the major problem for pose estimators is the lack of data

Data is the most important aspect of AI.
Labeled data is:

- Costly
- Difficult to obtain
- Time consuming
- Not scalable to all objects



Synthetic data

- Use of 3D computer graphics software (Blender / Unreal)
- Domain gap (synthetic to real)

Self-supervised:

- Uses unlabeled data
- Uses an auxiliary task to train the model
- Mostly uses differentiable rendering

(image similarity / latent space similarity)

- Teacher / student paradigm

Denninger, Maximilian, et al. "Blenderproc2: A procedural pipeline for photorealistic rendering." *Journal of Open Source Software* 8.82 (2023)
Wang, Gu, et al. "Occlusion-aware self-supervised monocular 6D object pose estimation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.3 (2021)

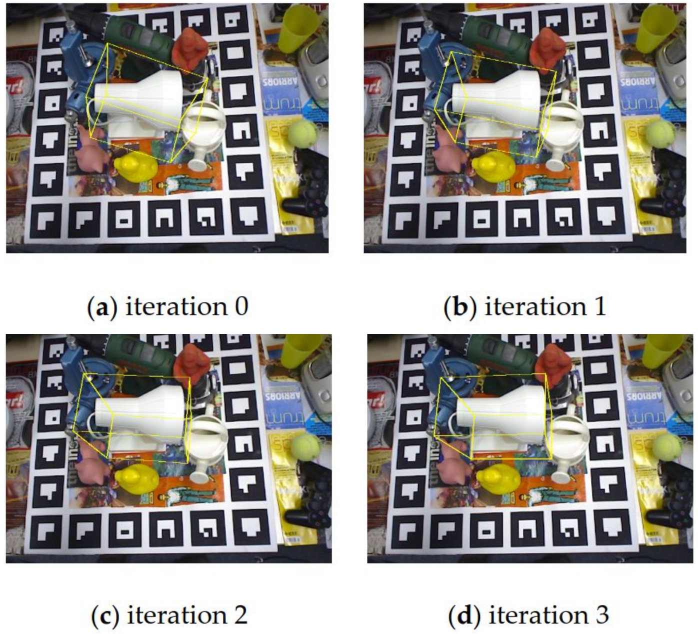
Opening : Refinement methods

Refinement methods aims at producing a better estimate by iteratively comparing the CAD model projected by the estimated pose and its real pose

- Alignment method (Iterative closest point)
- NeRF-based method

- Improves the precision
- May corrected errors

- Impacted by occlusions
- Very slow (greedy algorithm)
- Object specific


(a) iteration 0 (b) iteration 1 (c) iteration 2 (d) iteration 3

Huang, S.-K.; Hsu, C.-C.; Wang, W.-Y.; Lin, C.-H. Iterative Pose Refinement for Object Pose Estimation Based on RGBD Data. *Sensors* **2020**, *20*, 4114

Thank you, let’s dive in the practical session!