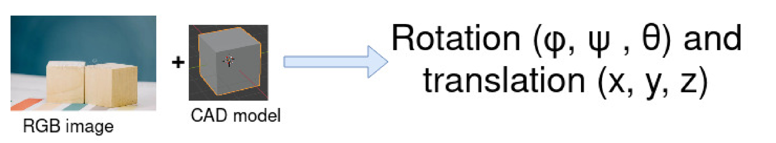
Towards Implicit Representation of a Pose: a New Improved Pipeline of Augmented Autoencoder

Elena Govi, Davide Sapienza, Giorgia Franchini, Carmelo Scribano, Marko Bertogna

University of Modena and Reggio Emilia - University of Parma

6D Pose Estimation

- The '6d pose estimation' problem consists of both **detection** and **localization** (also orientation) of an object in a scene, only by using cameras.
- It is a crucial task for several computer vision applications, such as **autonomous object picking** in the Industry 4.0 era, just to mention one.



Industrial Scenario Challenges

- Cluttered scenario
- Reflecting Textures
- Symmetries and self-occlusions
- Thin and elongated shape

Proposed Pipeline

- **Segmentation** with YOLOv7 [3]
- **6D estimation** with a less augmented version of Augmented Autoencoder(AAE) [2]

Segmentation

- **Dataset Creation** with a *geometric background*.
- YOLOv7-seg trained for 500 epochs, 17500 simulated images during training.

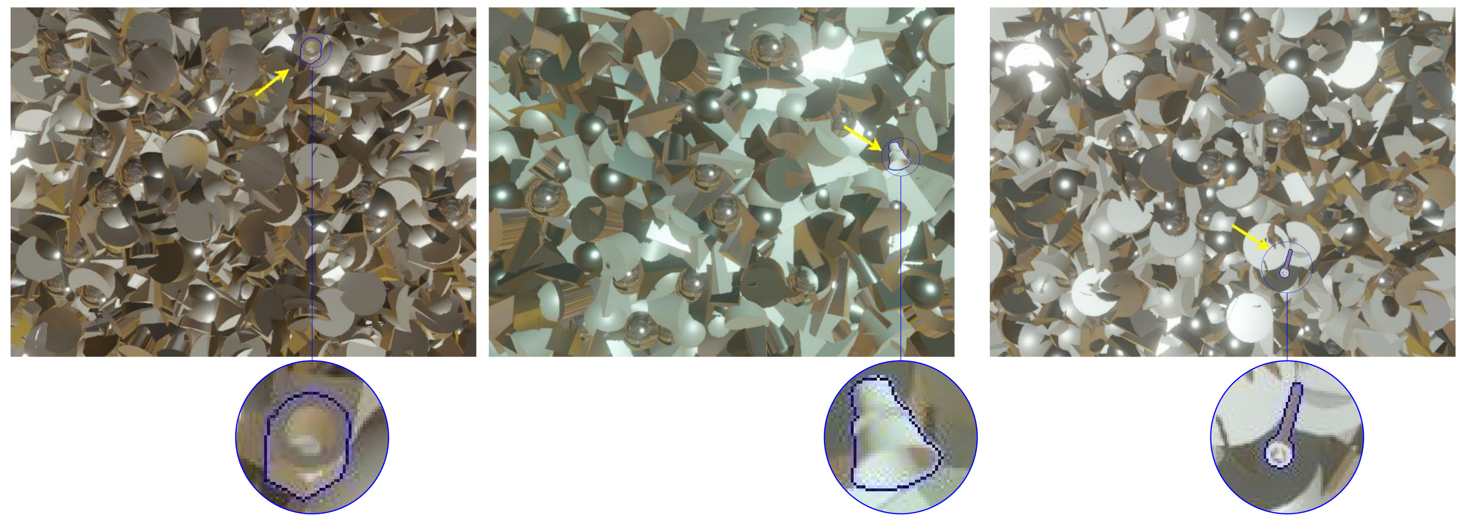


Figure 1. Examples of simulated scenes with geometric background.

Precision ,Recall and mean Average Precision (mAP) with a threshold of 0.5, which respectively achieve the following scores: (P) $\mathbf{0.997}$, (R) $\mathbf{0.997}$, (mAP) $\mathbf{0.994}$.

References

[1] D. Sapienza, E. Govi, S. Aldhaheri, G. Franchini, M. Bertognaz, E. Roura, È. Pairet, M. Verucchi, and P. Ardón. Model-based underwater 6d pose estimation from rgb. *arXiv preprint arXiv:2302.06821*, 2023.

[2] M. Sundermeyer, Z.-C. Marton, M. Durner, and R. Triebel. Augmented autoencoders: Implicit 3d orientation learning for 6d object detection. *International Journal of Computer Vision*, 128:714–729, 2020.

[3] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022.

Augmented Autoncoder

- Encoder(ϕ)-Decoder(ψ) structure
- Random Data Augmentation Function f_{aug}

The corrected equation is as follows: $x = (\psi \circ \phi \circ f_{aug})(x) = (\psi \circ \phi)(x')$ where x is the original input and x' is the augmented image.

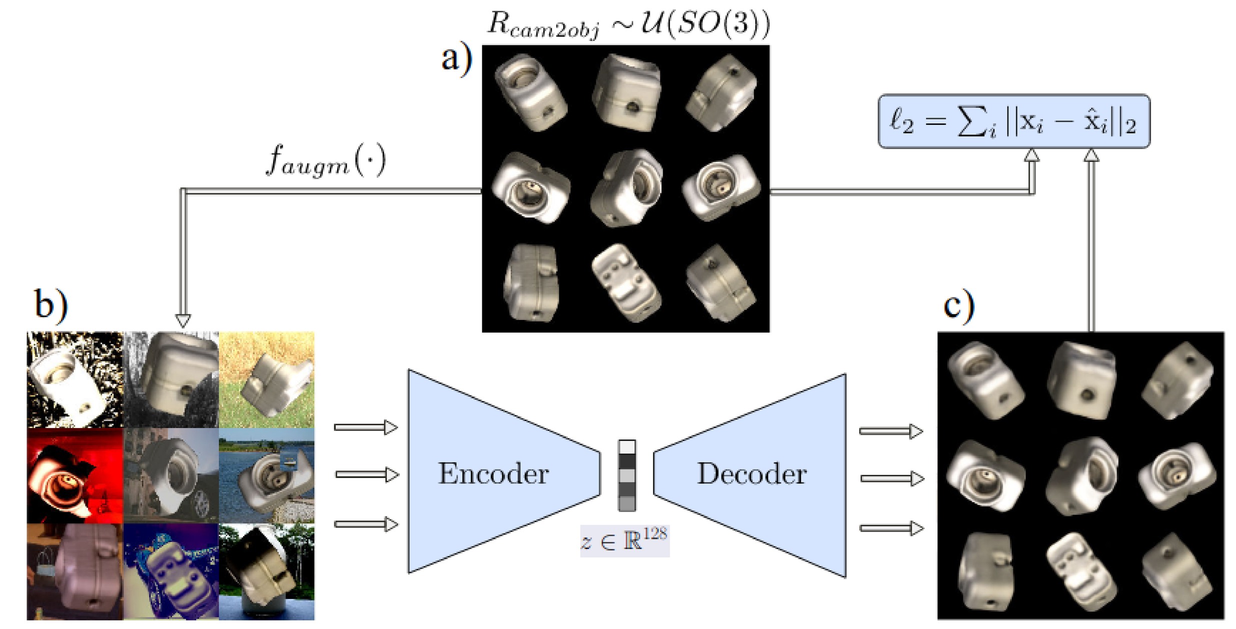


Figure 2. Training strategy: a) reconstruction target batch x of uniformly sampled SO(3) object views; b) geometric and color augmented input; c) reconstruction $\hat{x}$ after 40000 iterations

- After **training** (figure 5), a *codebook* is created by generating a **latent representation** $z \in \mathbb{R}^{128}$ of each possible object view, and its correspondent **P** matrix $\mathcal{R}_{cam2obj}$.
- At **test time**, first the object is segmented and masked. Secondly, the encoder gives its latent space features. Then, cosine similarity is computed between the input latent representation code $z_{test} \in \mathbb{R}^{128}$ and all codes $z_i \in \mathbb{R}^{128}$ from the codebook:

$$cos_i = \frac{z_i z_{test}}{\|z_i\| \|z_{test}\|}$$

- The highest similarity is chosen and the corresponding rotation matrix from the codebook is returned as 3D object orientation.

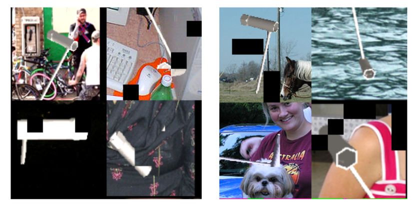


Figure 3. Different Data Augmentation with VOC dataset .

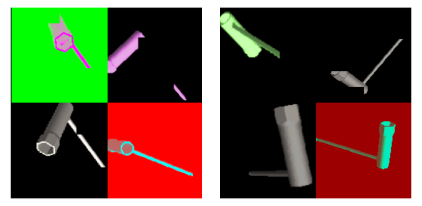


Figure 4. Less Augmented Data (without VOC dataset).

Two different models have been trained, where different data augmentation:

- **Original AAE** (with VOC dataset images as background), as in figure 3;
- **Less AAE** (without VOC dataset images as background), as in figure 4.

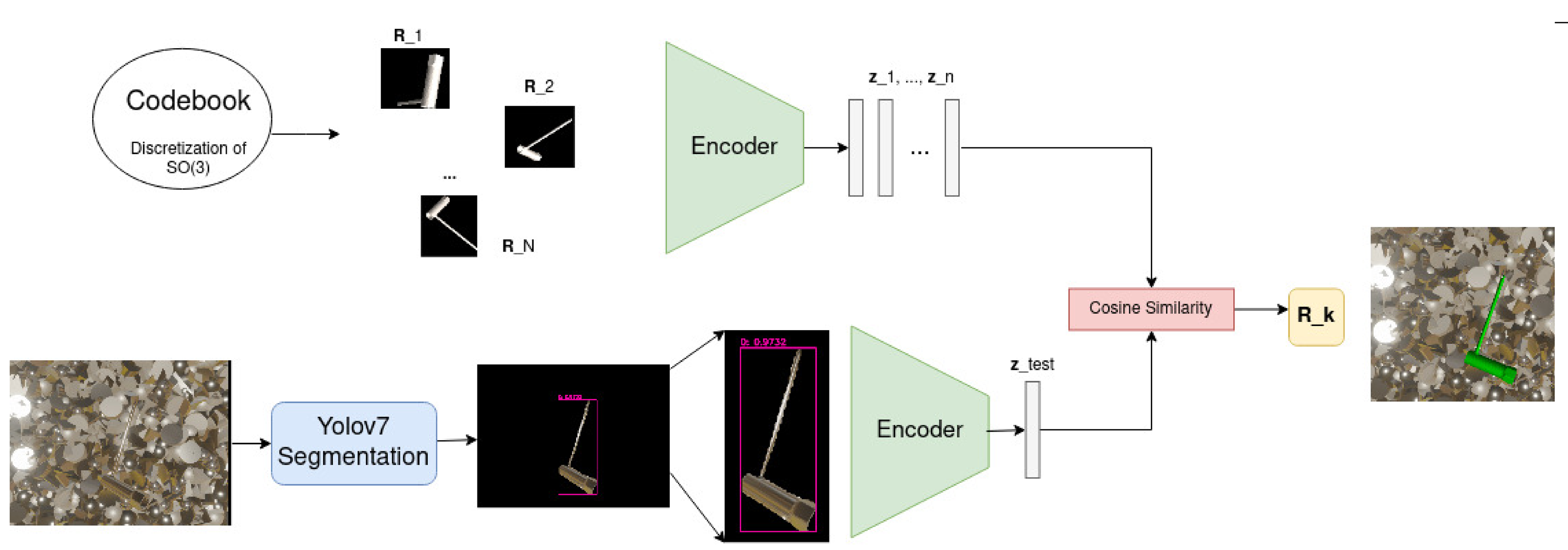


Figure 5. Pipeline with segmentation

Evaluation

- Given the model $\mathcal{M}$, the estimated pose $\hat{\mathbf{P}} = [\hat{\mathbf{R}}, \hat{\mathbf{t}}; \mathbf{0}, 1]$ and the ground-truth $\overline{\mathbf{P}}$, the *Average Distance to the Correspondent Model Point* is $e_{ADD} = avg_{x\in\mathcal{M}} \|\hat{\mathbf{P}}\mathbf{x} - \overline{\mathbf{P}}\mathbf{x}\|$
- If the model $\mathcal{M}$ has indistinguishable views, the error is calculated as the Average Distance to the *Closest Model Point*: $e_{ADI} = avg_{\mathbf{x_1}\in\mathcal{M}} \min_{\mathbf{x_2}\in\mathcal{M}} \|\hat{\mathbf{P}}\mathbf{x_1} - \overline{\mathbf{P}}\mathbf{x_2}\|$
- **Criterion of Correctness** The estimated pose is considered correct if $e < \theta_{AD} = k_m d$ where k_m constant generally equal to 0.1, d = object diameter

Comparison with the original pipeline

We compared in Table 1 all possible combination of three detectors (YOLOv4 as proposed in [1], YOLOv7 as 2D detector and as segmentator) and two types of augmented autoncoder (AAE and LessAAE).

First Phase	Detector	6D Estimator	ADD(-S) recall
BBs	Yolov4	AAE	3.406%
BBs	Yolov7	AAE	3.39%
Segm	Yolov7-seg	AAE	7.218%
BBs	Yolov4	LessAAE	0.801%
BBs	Yolov7	LessAAE	2.12%
Segm	Yolov7-seg	LessAAE	$\mathbf{30.573\%}$

Table 1. ADD(-S) recall results on different pipelines on the spark plug key with geometric background.

Quantitative and qualitative results

	Yolov7-Seg + LessAAE	Yolov7-Seg + AAE
Spark Plug Key	**43.31%**	6.68%
Screw	12.79%	**15.40%**
Nozzle	**5.202%**	2.890%
Nut	**43.046**%	41.281%

Table 2. Final recall results on the four objects for the multi-objects 6D pose estimation.

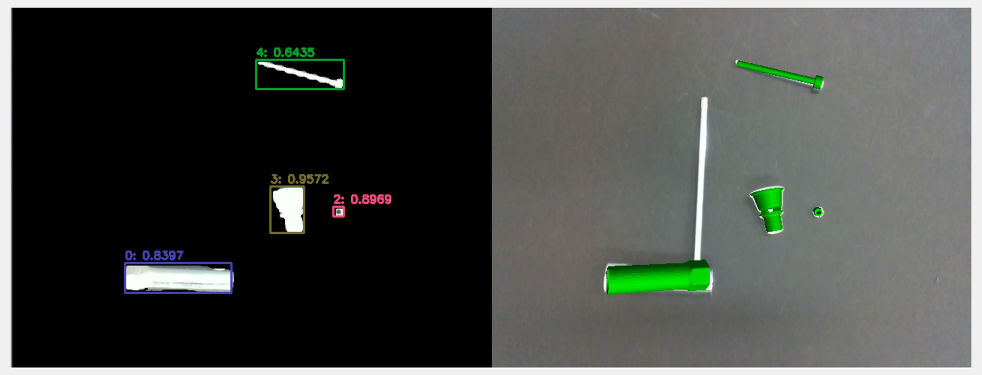


Figure 6. Results on a real image.

Conclusions an Applications

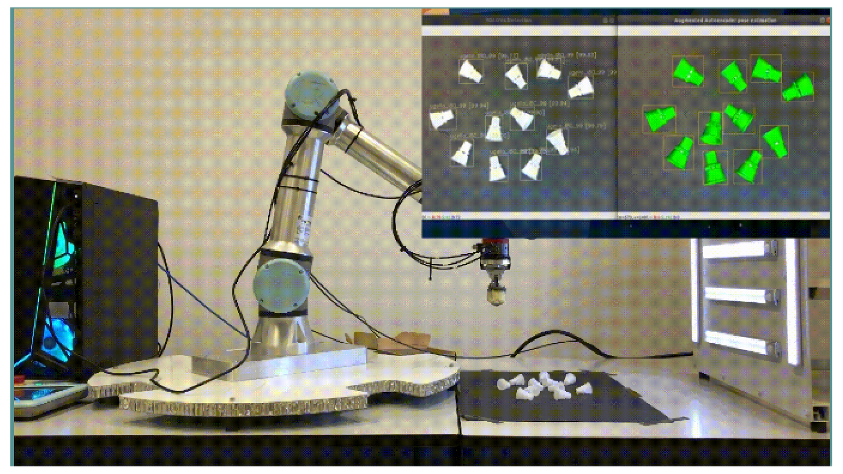


Figure 7. Universal Robot: e-Series UR5e

1. Object Localization and Recognition
2. 6D pose estimation
3. Pick and Place Task