Bridging the Last Mile in Sim-to-Real Robot Perception via Bayesian Active Learning

[1]German Aerospace Center (DLR) [2]Technical University of Munich (TUM)

Jianxiang Feng[1,2], Jongseok Lee[1], Maximilian Durner[1,2], Rudoplh Triebel[1,2]

Deutsches Zentrum für Luft- und Raumfahrt
German Aerospace Center

MUNICH SCHOOL FOR DATA SCIENCE
HELMHOLTZ | TUM | LMU

Motivation

Simulation tools such as BlenderProc [2] can provide **a large amount** of photo-realistic synthetic data with **annotations** required by robotic vision tasks such as object detection. However, when relying only on simulation data, it's hard to resolve the problem of the **simulation-to-reality (Sim-to-Real) gap** (Fig.1).

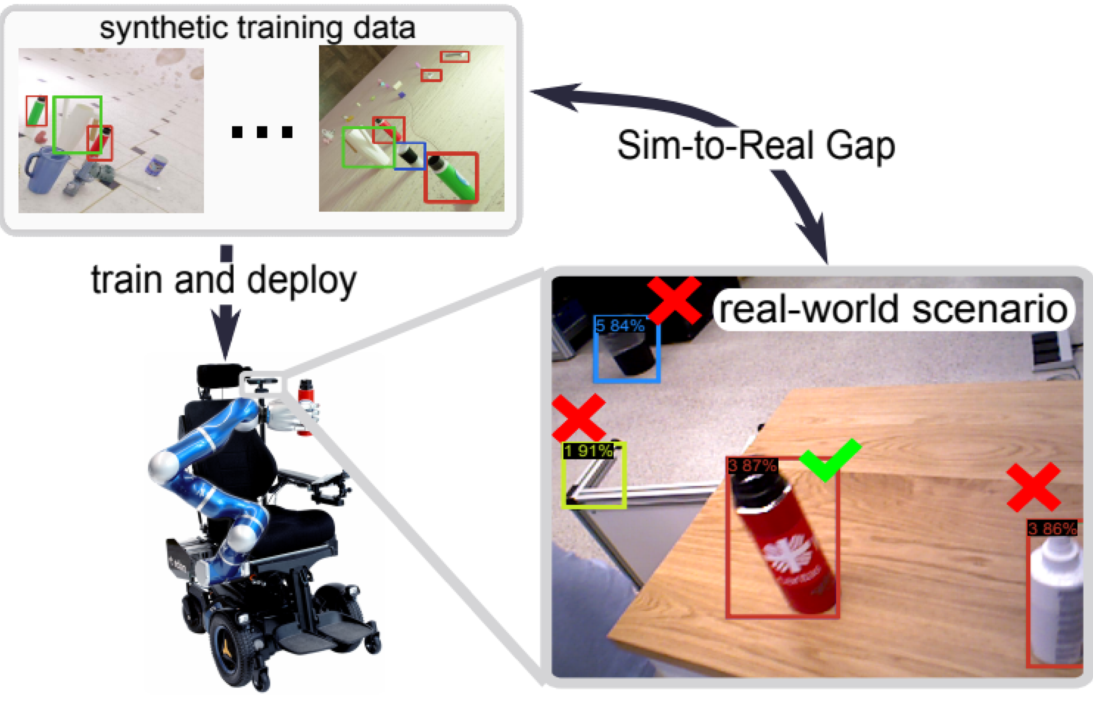

Fig.1: Illustration of Sim-to-Real gap.

Idea

- **Given:** An object detector trained with photo-realistic synthetic annotated data (Fig. 2).

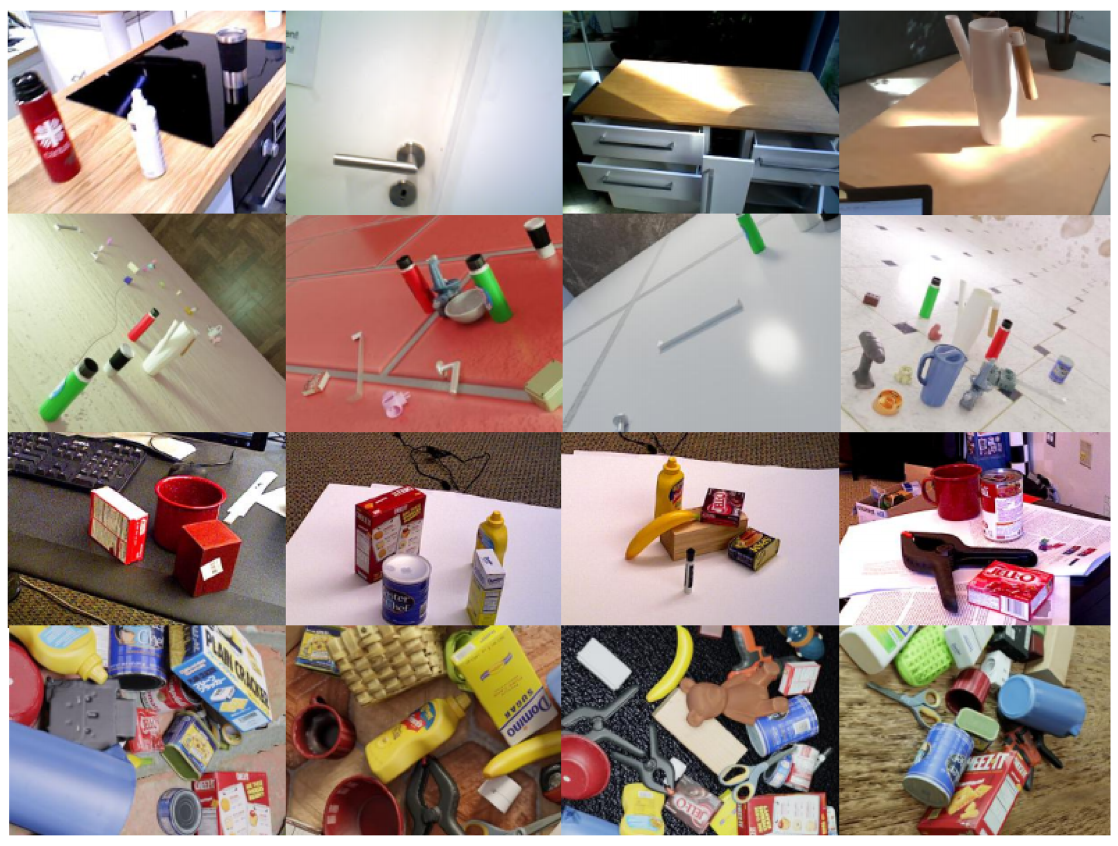

Fig.2: Examples of real and synthetic data sets.

- **Goal:** To bridge the Sim-to-Real gap with as few real annotated data as possible.
- **Method:** Active learning with a Bayesian object detector (Fig. 3).

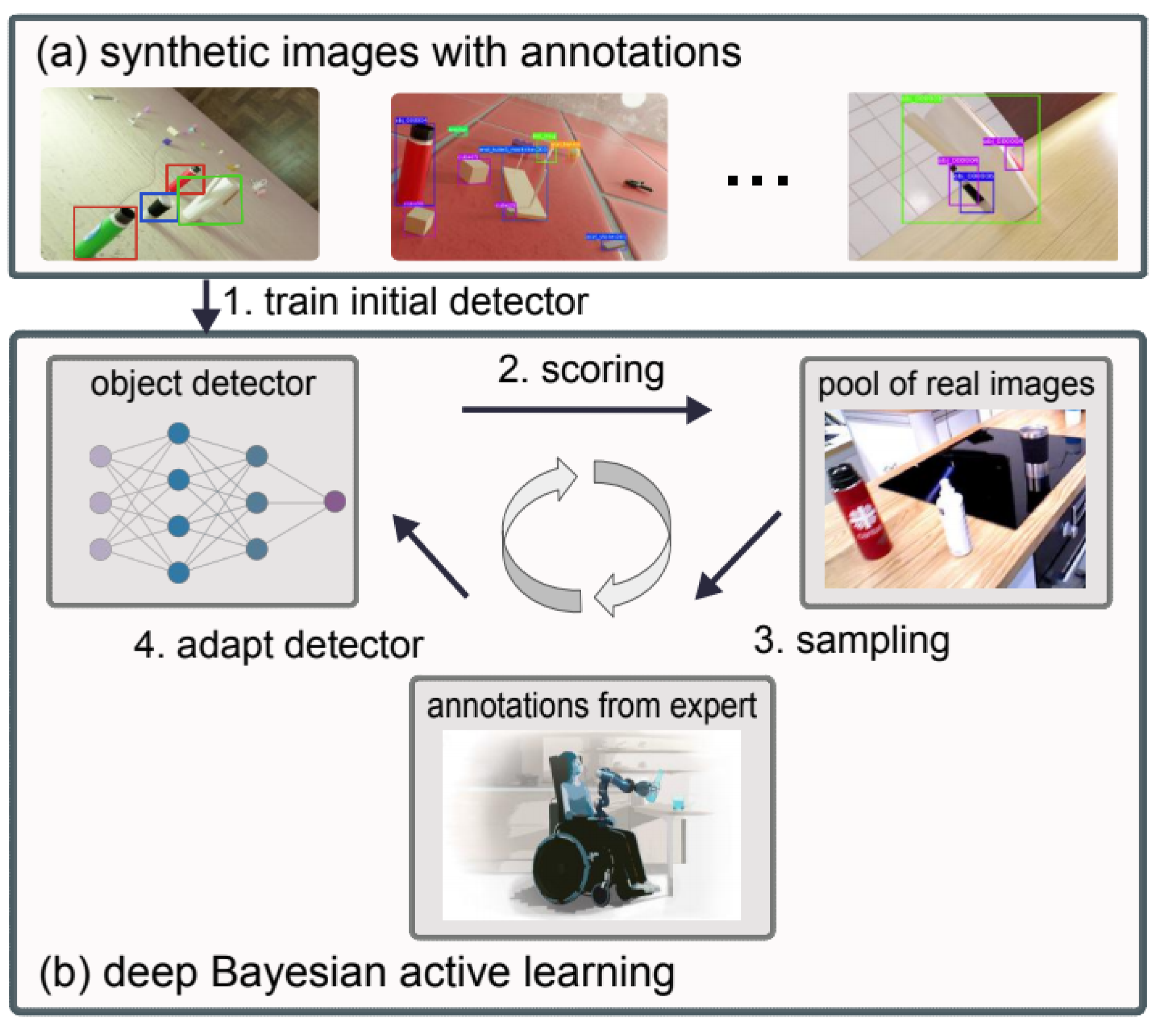

Fig.3: Pipeline overview.

Approach

Bayesian Neural Network (BNN) object detector:

- Monte Carlo Dropout for BNN posterior predictive inference at anchor-level:

$$p(\mathbf{y}^* \mid \mathbf{x}^*, \mathscr{D}_{train}) = \int p(\mathbf{y}^* \mid \mathbf{x}^*, \theta) p(\theta \mid \mathscr{D}_{train}) d\theta. \quad (1)$$

- Bayesian Inference to replace non-maximum-suppression (NMS) [1] :

$$p_{[\hat{\mathbf{c}}_1,\dots,\hat{\mathbf{c}}_M]}(\mathbf{c}|\mathbf{x}^*, \mathscr{D}_{train}) \propto p_{\hat{\mathbf{c}}_1}(\mathbf{c}|\mathbf{x}^*, \mathscr{D}_{train}) \prod_{i=2}^{m} p(\hat{\mathbf{c}}_i|\mathbf{c}, \mathbf{x}^*, \mathscr{D}_{train}), \quad (2)$$

$$p_{[\hat{\mathbf{b}}_1,\dots,\hat{\mathbf{b}}_M]}(\mathbf{b}|\mathbf{x}^*, \mathscr{D}_{train}) \propto p_{\hat{\mathbf{b}}_1}(\mathbf{b}|\mathbf{x}^*, \mathscr{D}_{train}) \prod_{i=2}^{m} p(\hat{\mathbf{b}}_i|\mathbf{b}, \mathbf{x}^*, \mathscr{D}_{train}). \quad (3)$$

- **Scoring:** to compute the informativeness (entropy of the predictive distribution) of j-th detected instance on k-th image and aggregate them into one score representing the informativeness of the k-th image:

- Score of category classification:

$$\begin{aligned}\mathscr{U}_{j,cls} &= \sum_{i=1}^{|\mathscr{C}|} \mathscr{H}(p(c_i|\mathbf{x}^*, \mathscr{D}_{train})), \\ &= \sum_{i=1}^{|\mathscr{C}|} [-p(c_i|\mathbf{x}^*, \mathscr{D}_{train}) \log p(c_i|\mathbf{x}^*, \mathscr{D}_{train}) \\ &\quad - (1 - p(c_i|\mathbf{x}^*, \mathscr{D}_{train})) \log (1 - p(c_i|\mathbf{x}^*, \mathscr{D}_{train}))].\end{aligned} \quad (4)$$

- Score of bounding box regression:

$$\begin{aligned}\mathscr{U}_{j,reg} &= \mathscr{H}(p(\mathbf{b}|\mathbf{x}^*, \mathscr{D}_{train})) \\ &= \frac{k}{2} + \frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln(|\mathbf{C}_b|),\end{aligned} \quad (5)$$

- Acquisition function comprises of a combination function and an aggregation function:

$$\mathscr{A}(\mathbf{x}_k) = agg_{j \in N_k}(comb(\mathscr{U}_{j,cls}, \mathscr{U}_{j,reg})), \quad (6)$$

Options for combination function:
1. maximum; 2.weighted sum;
Options for aggregation function:
1. average; 2. summation.

- **Sampling:** To select a subset of data from the pool set to query from human.
Problem: when employing naive ranking of scores from the scoring step and selecting the highest N ones, the problem of fore-ground class imbalance can cause under-performance.
Solutions: employ two following sampling strategies.

- Core-set [3]: to select points that can best represent the pool set based on a distance function between the data points in the pool set and previously selected set:

$$\triangle(\mathbf{x}_i, \mathbf{x}_j)_{\mathbf{x}_i \in \mathscr{D}_{pool}, \mathbf{x}_j \in \mathbf{s}_0} = || \sum_{k=1}^{N_i} p(\mathbf{c}_k|\mathbf{x}_i) - \sum_{k=1}^{N_j} p(\mathbf{c}_k|\mathbf{x}_j) ||_2 + w\mathscr{A}(\mathbf{x}_i) \quad (7)$$

- Ranking after sub-sampling (Fig.4): by assuming certain degree of redundancy in the data set, we propose to do ranking after uniform sub-sampling which can generate more balanced data set:

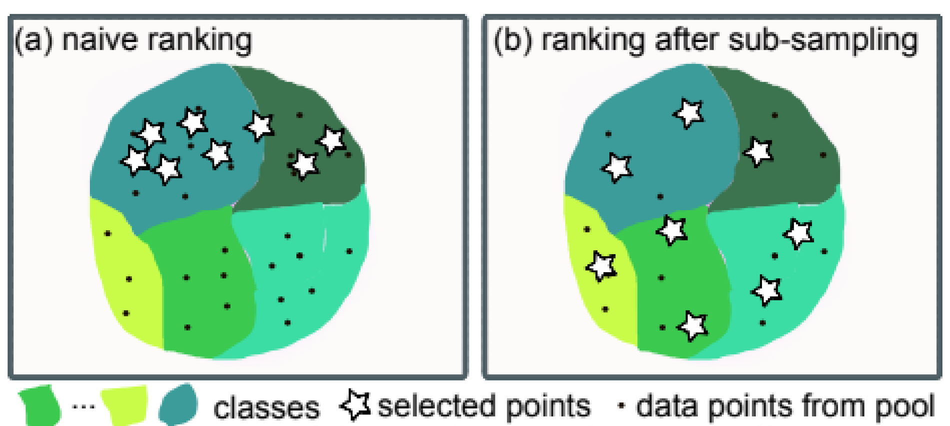

Fig.4: Ranking after sub-sampling.

Experiment

- Data sets (Fig. 2): 1. self-collected daily object data set (5 categories); 2. sub-sampled YCBV data set (21 categories) [5];
- Implementation details: RetinaNet [4] for object detection; Domain randomization for synthetic data generation.
- Active learning hyper-parameters: #iteration: 10; #acquisition: 20 for daily object and 50 for sub-sampled YCBV data set .

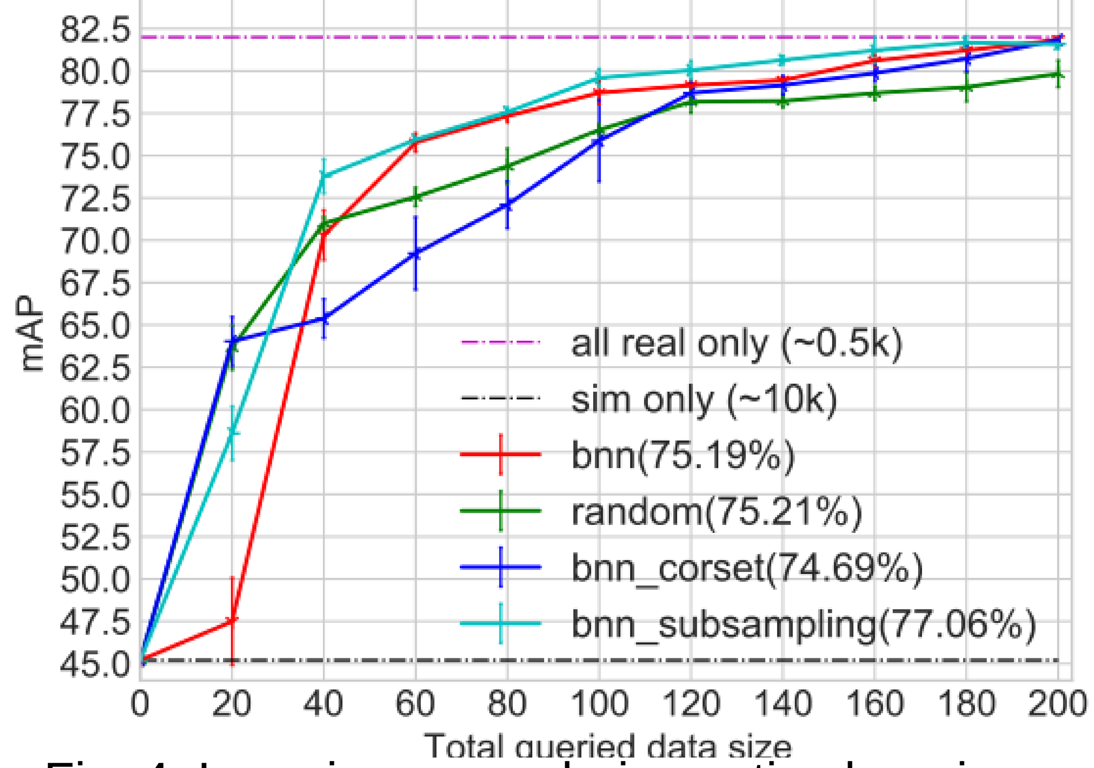

Fig. 4: Learning curve during active learning on daily object data set for 3 random runs.

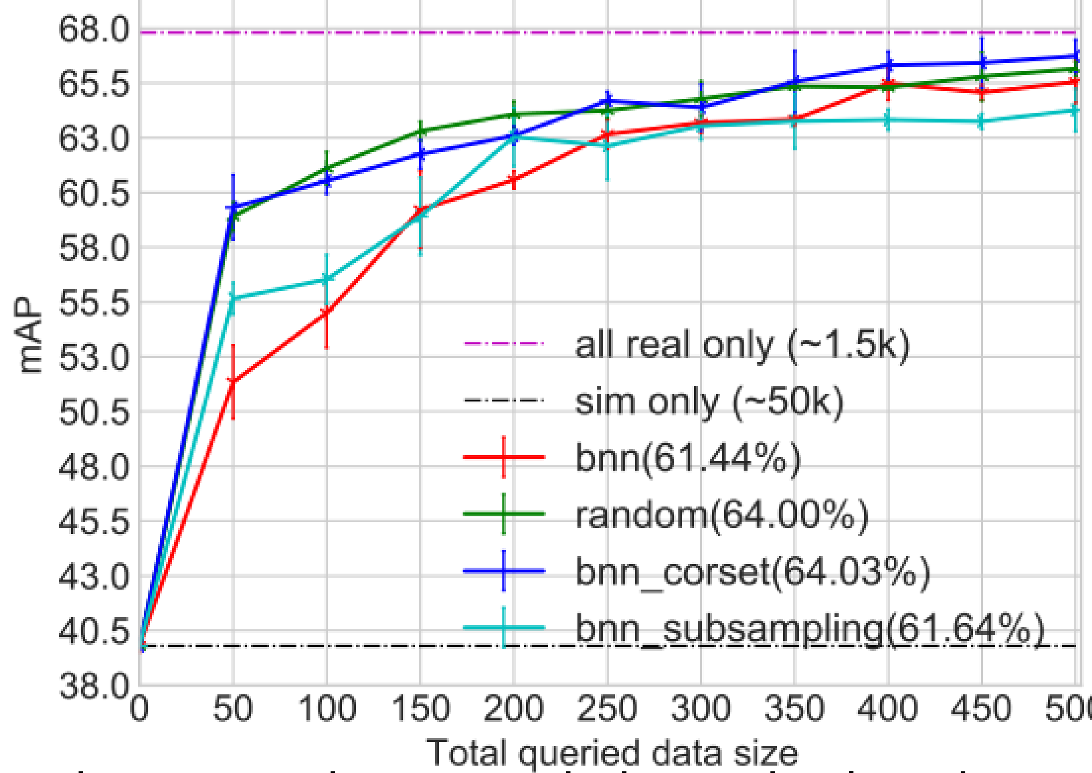

Fig. 5: Learning curve during active learning on YCBV data set for 3 random runs.

Conclusion

- We present a Sim-to-Real pipeline that can efficiently use real annotated data to bridge the gap based on deep Bayesian active learning.

- Empirically we show that the real annotated images can efficiently reduce the reality gap in practice by saving up to 60% data.

- Our experiments indicate that the foreground class imbalance can be one of the factors which can determine the success of our pipeline in practice.

References

[1] Harakeh, Ali, Michael Smart, and Steven L. Waslander. "Bayesod: A bayesian approach for uncertainty estimation in deep object detectors." 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020.
[2] Denninger, M., Sundermeyer, M., Winkelbauer, D., Zidan, Y., Olefir, D., Elbadrawy, M., ... & Katam, H. (2019). Blenderproc. arXiv preprint arXiv:1911.01911.
[3] Sener, Ozan, and Silvio Savarese. "Active learning for convolutional neural networks: A core-set approach." arXiv preprint arXiv:1708.00489 (2017).
[4] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.
[5] Xiang, Yu, et al. "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes." arXiv preprint arXiv:1711.00199 (2017).