29th RoboCup International Symposium, Incheon, Korea, 2026.
Best Paper Award.

Efficient Image Annotation via Semi-Supervised Object Segmentation with Label Propagation

Vitalii Tutevych*, Raphael Memmesheimer*, Luca Eichler, Dmytro Pavlichenko, Fynn Schilke, Rodja Krudewig, and Sven Behnke

Autonomous Intelligent Systems Group, University of Bonn, Germany
memmesheimer@ais.uni-bonn.de

Abstract. Reliable object perception is necessary for general-purpose service robots. Open-vocabulary detectors struggle to generalize beyond a few classes and fully supervised training of object detectors requires time-intensive annotations. We present a semi-supervised label propagation approach for household object segmentation. A segment proposer generates class-agnostic masks, and an ensemble of Hopfield networks assigns labels by learning representative embeddings in complementary foundation model embedding spaces (CLIP, ViT, Theia). Our approach scales to 50 object classes with limited annotation overhead and can automatically label 60% of the data in a RoboCup@Home setting, where preparation time is severely constrained. Dataset and code are publicly available at https://github.com/ais-bonn/label_propagation.

1 Introduction

Robots interacting with objects in an environment require them to perceive these objects. The creation of accurately labeled object segmentation datasets is a time-intensive and exhaustive effort. This is especially a problem in time-constrained settings with many class instances, e.g., in robot competitions such as RoboCup@Home [9]. RoboCup@Home is a robot competition where robots perform household tasks in realistic apartments autonomously. Upon arrival, teams are provided with a list of ca. 50 object classes that need to be recognized and manipulated, with limited preparation time (commonly one or two days) to record training data and train a model for the robot to perform the tasks. Recent advances in contrastive language-image pre-training [16] laid the foundation for open-vocabulary object detection [7] and segmentation approaches [21], and foundation models for segmentation have been highly influential [5]. In previous competition deployments [12, 10], however, such open-vocabulary approaches performed well for small numbers of classes but did not generalize to the larger object vocabularies required in this setting.

In this paper, we present a two-stage process to close this gap. First, a segment proposal model is trained on a large object segmentation dataset gathered over multiple competition attendances and used to estimate class-agnostic object

* These authors contributed equally to this work.

segments. Second, a set of reference images per object class is used to learn a metastable representation of each class, which is compared to the input embedding in Foundation Model representation space for label association. This approach substantially reduces manual labeling effort by automatically suggesting 60% of the labels. Finally, a fully supervised model is trained that performs well on many object classes, including visually similar ones. The focused propagation reduces the influence of outliers and improves the accuracy of label assignments.

We summarize the contributions of this paper as follows:

- We propose a semi-automatic approach that guides the labeling process by estimating general object segments and propagating reference labels for classes among the whole dataset.
- We publicly provide the code and the dataset that we utilized for training the general object segmentation model and for evaluating the semi-automatic label propagation approach.

2 Related Work

Currently, segmentation models following a supervised training scheme [1, 4, 6] still define the standard for robot competitions [12, 10, 11]. Open-vocabulary approaches [21, 7] applied in these competitions performed well for small class sets, but their accuracy degraded as the vocabulary approached the full competition object set [12, 10].

Semantic Segmentation. Traditional semantic segmentation approaches rely on manually annotated segments and are trained in a supervised manner. A common representative is YOLO [4], which is widely used in robot competitions due to its ease of use for both training and inference. DETR [1] reformulated object detection as a transformer-based set prediction problem, while MaskDINO [6] extended this approach to pixel-wise semantic segmentation. Both models improved the state-of-the-art detection and segmentation results. Recently, open-vocabulary approaches like Detic [21] and Grounding-DINO [7] enabled promptable semantic segmentation. Gouda et al. [3] integrate a zero-shot segmentation model (like SAM) with a Centroid Triplet Loss (CTL) trained backbone to match query images of objects. This approach is related to ours, which instead uses a Segment Proposer to generate class-agnostic masks and associates labels from reference sets with an ensemble of Hopfield networks. Pätzold, Nogga, and Behnke [15] guide open-vocabulary detectors with input prompts generated by visual language models.

Datasets. Massouh, Brigato, and Iocchi [8] proposed a large-scale object recognition dataset of household objects. The dataset, collected from web image searches, consists of 196,000 images in the training set and spans 180 classes. It provides class annotations per image only and is therefore not suitable for object detection and segmentation approaches. Meneghetti et al. [13] provide

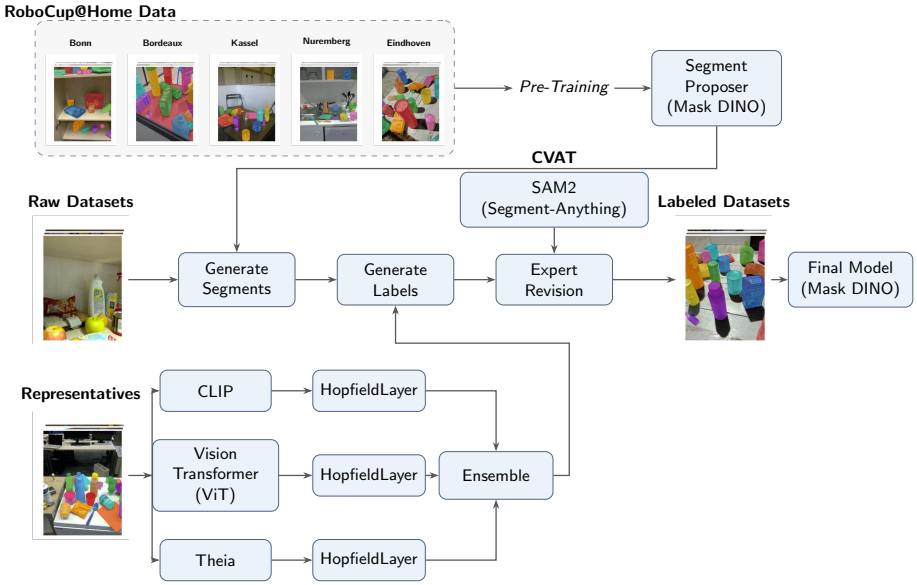


Fig. 1: Labeling and training pipeline. To train the final object detector, we first utilize a Segment Proposer model pre-trained on the data from previous competitions. After initial pre-labeling, a human expert verifies and corrects the generated labels for the representative dataset. The labeler is then used with a general object detector to generate improved annotations for the final detector.

a dataset of annotated images targeting RoboCup@Home objects, consisting of two sets with a total of 554 annotated images and 2765 labeled bounding boxes. Novkovic et al. [14] proposed a dataset containing scenes with single objects and box scenes with cluttered household object arrangements, collected as multimodal data from four different sensors. Tyree et al. [20] present the HOPE dataset, which consists of 2088 images from 10 different scenes labeled on an object segment and pose level. In contrast to these datasets, ours provides pixel-level segmentation annotations of local household objects captured in multiple countries, enabling robotic downstream tasks such as mobile manipulation.

3 Approach

We first train a Segment Proposer to generate class-agnostic masks, and then train an ensemble of Hopfield networks to assign labels by learning representative embeddings in complementary Foundation Model embedding spaces (CLIP, ViT, Theia). The approach is summarized in Figure 1. In the following, we describe the data collection process, the Segment Proposer, the Hopfield Labeler, and the final detector training.

Fig. 2: Data-recording setup: We utilized an Orbbec Gemini 2 connected to a Valve Steam Deck supported by a guided data-recording procedure.

3.1 Data Collection

We collect two types of data at each competition venue using the setup shown in Figure 2. For the *representative dataset*, each of the around 50 objects is recorded individually in isolation, but in front of varying backgrounds, with 4–8 images per object captured from four cardinal directions at two distances, yielding 1600–2000 instances per venue. The *training/validation datasets* consist of 144 training and 16 validation images per set, with 2–3 sets per venue, featuring cluttered scenes with object overlaps, occlusions, and challenging lighting. Throughout this paper, datasets are named after the venue at which they were recorded: the RoboCup world championships in Bordeaux (2023), Eindhoven (2024), and Salvador (2025), the RoboCup German Open competitions in Kassel (2023), Nuernberg (2025), and Cologne (2026), and recordings from our lab in Bonn.

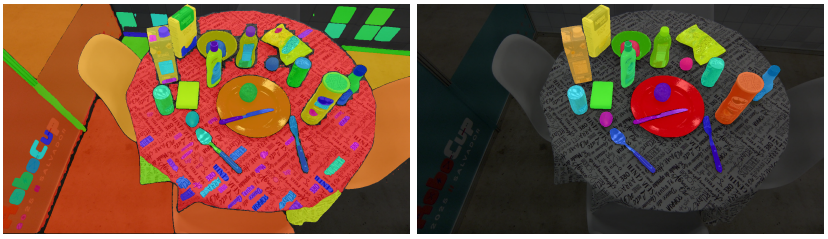
3.2 Segment Proposer

We introduce a Segment Proposer model that automates object segmentation in images, eliminating the need for manual mask annotation by human annotators. In contrast to the guided Segment Anything approach [5], which requires the manual selection of positive and negative points, our model is trained specifically on household objects and operates without manual interaction. The model is trained on annotated data from the Bordeaux, Eindhoven, and Kassel datasets and is used to generate segment proposals for the Nuernberg and Salvador sets. Figure 3 shows a qualitative comparison of our approach to Segment Anything. Coupling Segment Anything with additional filtering or detection models would be possible, but adds components and computational overhead.

3.3 Hopfield Labeler

The Labeler (Figure 4) assigns a class label to each proposed mask by comparing its Foundation Model embedding against a set of learned *representative* embeddings — one per class. We decouple the Labeler from the Segment Proposer: labels are inferred from small representative crops; the Labeler is never fine-tuned on full detection datasets.

Foundation Model embeddings. Each annotated patch is encoded by a frozen Foundation Model. We use OpenAI CLIP, Vision Transformer (ViT),



(a) Segment Anything (b) Segment Proposer (Ours)

Fig. 3: Qualitative example of the Segment Anything model in comparison to our proposed Segment Proposer for household objects.

and Theia [19]. These models were selected to cover complementary embedding spaces: CLIP provides language-aligned visual features, ViT provides purely visual features from supervised classification training, and Theia is distilled from multiple vision foundation models (CLIP, ViT, SAM, DINOv2, and Depth Anything) for robot learning. For transformer-based models the `[CLS]` token serves as the patch representation; Theia provides a direct bottleneck vector.

Hopfield Memory. We store one learned representative embedding per class in a Hopfield Memory [17]. At inference the attention score between a query embedding R and the memory matrix Y is:

$$\text{scores} = \text{softmax}\big(\beta\, R W_Q W_K^\top Y^\top\big)\,, \tag{1}$$

where R denotes the Foundation Model embedding of the query patch, Y the matrix of stored class representatives, W_Q and W_K learned query and key projection matrices, and β the inverse temperature of the softmax. Training minimizes the MSE between predicted scores and one-hot class targets:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\big(\text{scores}(x_i) - \mathbb{1}_{y_i}\big)^2, \tag{2}$$

where N is the number of training samples, x_i the embedding of the i-th sample, y_i its class label, and $\mathbb{1}_{y_i}$ the one-hot encoding of y_i. This is equivalent to simultaneously pulling each class representative toward its positive examples while pushing all other representatives away — a supervised analogue of contrastive learning.

Bank subdivision and regularization. To improve robustness we split the Hopfield Memory into m banks, each operating in a lower-dimensional projection space similar to multi-head attention; we use $m = 4$ banks in all experiments. Final scores are the mean across banks. Two regularizers enforce complementarity: an *intra-bank* term clusters representatives within each bank, and an *inter-bank* term pushes the same-class representatives of different banks apart. Empirically, this subdivision together with the regularizers consistently improves recognition accuracy.

Ensemble. One Hopfield head is trained independently per Foundation Model embedding space. At inference the predictions of all three heads are com-

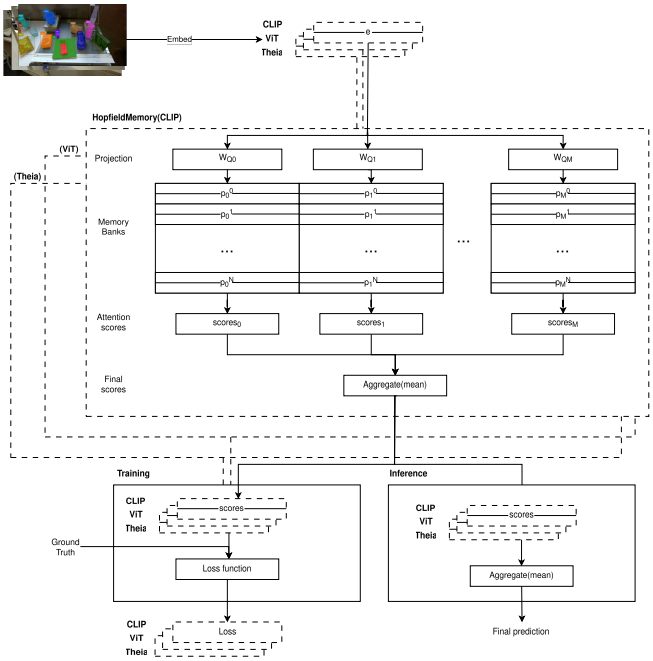


Fig. 4: Labeler architecture. One Hopfield head is trained per foundation model; outputs are combined by mean aggregation at inference time.

bined by mean aggregation; the Ensemble is not trained or fine-tuned after the individual heads are trained. Segments whose ensemble confidence falls below a threshold — including false-positive masks of the Segment Proposer, such as background items — are not force-assigned the closest label, but rejected and left to the human annotator.

3.4 Final Detector Training

The Segment Proposer and Hopfield Ensemble generate annotations for the training/validation sets. A human expert reviews and corrects these annotations in CVAT [2] before training the final detector (MaskDINO via Detectron2) used on the robot. We study the impact of skipping this review step in **Q4**.

4 Experiments

We address four questions: **Q1** Does learning representative embeddings via Hopfield Memory outperform fixed cosine-similarity retrieval? **Q2** Is an ensemble of independently trained heads more accurate than any single head? **Q3** Does the full pipeline reduce annotation effort in practice? **Q4** Can pipeline-generated annotations replace expert annotations for detector training?

(a) Nuernberg (b) Eindhoven (c) Salvador

Fig. 5: Labeled example images from different competition venues.

4.1 Dataset

Following the collection procedure of Section 3.1, we assembled a dataset of 118102 segmentation annotations across 216 object categories in 12506 images. Data were collected at seven venues across four countries — France, Germany, the Netherlands, and Brazil — reflecting diverse local household products. Example annotations are shown in Figure 5.

4.2 Training and Deployment Details

Training uses a PC with a 24 GB NVIDIA GeForce RTX 4090. The Labeler requires ∼10 GB VRAM and trains in 1.5 h; the Final Detector (MaskDINO) uses the full card and trains in 30 min with 2400 iterations, batch size 4. Full machine annotation of one dataset with ca. 160 images takes 2–3 minutes. Images are resized to 224×224 and augmented with random crops, flips, affine transforms, colour jitter, grayscale, and Gaussian blur during Labeler training. Key hyperparameters: learning rate 0.001, 20 epochs, batch size 16, $\lambda_{\text{inter}} = 0.01$, $\lambda_{\text{intra}} = 0.1$.

4.3 Evaluation and Analysis

Evaluation Setup. The trained models are evaluated using the metrics of Mask Precision, Mask Recall, and Mask Mean Average Precision (mAP). We use FiftyOne for model evaluation on the gathered data and report the micro-averaged metrics provided by the library. Since the Labeler assigns exactly one label per segment, each misclassification counts as both a false positive and a false negative, so the micro-averaged Precision, Recall, and F1-score coincide (Tables 1 and 2).

Q1 For the first experiment we compare the respective Hopfield heads with a cosine similarity baseline that stores 4–5 fixed prototype embeddings per class, extracted from the representative dataset or sampled from the training set if none is available, and assigns the nearest class at inference time. In this and the next experiment we use ground truth segmentations; the performance gain averaged over all city datasets is presented in Table 1.

Hopfield heads consistently outperform cosine similarity across all six city datasets and all three architectures, supporting **Q1**. CLIP benefits most, with average mAP rising from **0.521** to **0.699**. Theia reaches the highest individual

Table 1: Hopfield Heads with average performance gain over Cosine.

Model	mAP	Accuracy	Precision	Recall	F1-Score
CLIP	0.699 (+0.178)	0.570 (+0.136)	0.720 (+0.121)	0.720 (+0.121)	0.720 (+0.121)
Theia	0.704 (+0.056)	0.565 (+0.037)	0.716 (+0.029)	0.716 (+0.029)	0.716 (+0.029)
ViT	0.701 (+0.058)	0.559 (+0.054)	0.710 (+0.043)	0.710 (+0.043)	0.710 (+0.043)

Table 2: Ensemble improvement rate.

Model	mAP	Accuracy	Precision	Recall	F1-score
CLIP	0.694	0.565	0.715	0.715	0.715
Theia	0.699	0.552	0.705	0.705	0.705
ViT	0.704	0.555	0.709	0.709	0.709
Ensemble	**0.821**	**0.703**	**0.824**	**0.824**	**0.824**

average (**0.704 mAP**), yet the gap between models hints at complementary representations — motivation for the ensemble in Q2.

Q2 For the second experiment we test each of the trained Hopfield Heads against their ensemble. Here, the distilled Theia Foundation Model is of particular interest.

The results in Table 2 clearly demonstrate the benefit of the Ensemble over the individual models, supporting our claim for **Q2**: a single distilled representation is not necessarily optimal for object recognition, and the chosen Foundation Models possess complementary representations that make them a good fit for an Ensemble.

Q3 The third experiment is two-fold: we first evaluate the full pipeline (Segment Proposer + Ensemble Head) against the ground truth data, and then evaluate its efficiency in aiding the annotation process.

To set up the experiment, we divide the object classes into three categories determined by their shape: Simple — spherical and boxy objects (apple, cereal box), Medium — non-simple volumetric shapes (pack of chips, bottles, cups), and Complex — stretched and precise shapes that take more variation to capture (cutlery). Five human experts annotated a synthetic dataset, and we measured the time necessary to correctly label objects of each difficulty class (Table 3).

We then evaluate our approach using the established shape categories of recognized objects. The results are available in Table 4.

The results confirm the ensemble benefit: mAP improves across all city datasets, with gains of up to 0.15 over individual heads (Eindhoven, Cologne). Even in Bordeaux, where complex objects are more varied, the ensemble maintains the lead. The final comparison averaged over city datasets is in Table 5.

The results also show the impact of Segment Proposer quality, with mAP dropping by around 0.2 compared to running the Ensemble on ground truth segmentation.

Finally, we combine the annotation time results (Table 3) with the pipeline accuracy (Table 5) to estimate the time our approach saves human annotators on site (Table 6).

Table 3: Average labeling time per complexity class.

Complexity	Per Dataset	Per Object
Simple (S)	09:05.98	00:02.27
Medium (M)	09:46.22	00:02.44
Complex (C)	11:17.18	00:02.82

Table 4: mAP by dataset, model, and object complexity.

	CLIP				Theia				ViT				Ensemble			
Dataset	S	M	C	**All**	S	M	C	**All**	S	M	C	**All**	S	M	C	**All**
Bonn	.65	.71	.21	**.624**	.67	.72	.19	**.637**	.65	.73	.23	**.633**	.70	.76	.24	**.671**
Bordeaux	.45	.44	.03	**.349**	.43	.44	.03	**.338**	.43	.43	.05	**.337**	.53	.54	.05	**.417**
Eindhoven	.55	.48	.39	**.509**	.56	.43	.43	**.505**	.56	.43	.41	**.499**	.67	.64	.51	**.638**
Kassel	.62	.64	.00	**.607**	.61	.66	.00	**.613**	.63	.66	.00	**.622**	.71	.73	.00	**.695**
Cologne	.44	.51	.19	**.443**	.51	.46	.22	**.451**	.54	.54	.27	**.507**	.64	.64	.29	**.601**
Nuernberg	.57	.64	.23	**.558**	.60	.60	.20	**.557**	.55	.62	.22	**.538**	.70	.73	.33	**.667**
Salvador	.63	.59	.33	**.563**	.66	.58	.31	**.569**	.62	.57	.34	**.554**	.69	.62	.36	**.604**

The results empirically show the benefit of our approach and confirm that it is capable of taking over roughly 60% of the data without correction, supporting our claim for **Q3**. The Bordeaux dataset is an exception, as it contains considerably more variation in the objects of the Complex category. The time saved for Complex objects also varies strongly between venues: in Kassel none of the 144 Complex instances were retrieved, and in Cologne only 73 of 253. Complex objects, predominantly cutlery, are thin and elongated; the Segment Proposer therefore often produces fragmented or incomplete masks whose embeddings receive low ensemble confidence and are rejected rather than propagated. Reflective metal surfaces, whose appearance changes strongly with the venue lighting, further degrade the proposals and the embedding similarity to the representative images.

Q4 For the last experiment, we evaluate the quality of trained detectors. We compare MaskDINO models trained purely on the generated annotations to the ones that were post-corrected by human experts. We use both training and validation splits for the final evaluation and present our findings in Table 7.

Detectors trained purely on pipeline-generated annotations lose on average 0.15 mAP compared to their expert-corrected counterparts (0.634 vs. 0.788), answering **Q4**: pipeline annotations alone yield functional detectors, but the gap is not negligible. Since reduced precision and recall translate into misidentified or missed objects, and thus failed grasps in manipulation tasks, we recommend retaining the human review step of Section 3.4 for final deployment and regard fully automatic training as a fallback when preparation time is critically short.

4.4 Competition Deployment

The presented approach has been deployed for the NimbRo@Home team at the RoboCup@Home world championship 2025 in Salvador, Brazil and the RoboCup@Home

Table 5: Approach performance with Segment Proposer across datasets.

Model	Simple	Medium	Complex	All	Ground Truth
CLIP	0.560	0.572	0.199	0.522	0.713
Theia	0.579	0.557	0.198	0.524	0.716
ViT	0.567	0.568	0.216	0.527	0.715
Ensemble	**0.663**	**0.664**	**0.255**	**0.613**	**0.825**

Table 6: Objects retrieved and labeling time saved by the model.

	Objects Retrieved (of GT)			Time Saved (of GT)				%
Dataset	Simple	Medium	Complex	Simple	Medium	Complex	Total	Saved
Bonn	2850 (4065)	2425 (3203)	179 (732)	1:47:48 (2:33:47)	1:38:36 (2:10:15)	0:08:23 (0:34:24)	3:34:48 (5:18:27)	67.5%
Bordeaux	1315 (2471)	916 (1703)	86 (1714)	0:49:44 (1:33:29)	0:37:15 (1:09:15)	0:04:01 (1:20:33)	1:31:01 (4:03:17)	37.4%
Eindhoven	2305 (3445)	1095 (1721)	173 (342)	1:27:11 (2:10:20)	0:44:30 (1:09:59)	0:08:08 (0:16:04)	2:19:51 (3:36:23)	64.6%
Kassel	1581 (2230)	1356 (1858)	0 (144)	0:59:49 (1:24:22)	0:55:09 (1:15:33)	0:00:00 (0:06:46)	1:54:58 (2:46:41)	69.0%
Cologne	1509 (2340)	1335 (2082)	73 (253)	0:57:06 (1:28:31)	0:54:16 (1:24:40)	0:03:26 (0:11:53)	1:54:48 (3:05:05)	62.0%
Nuernberg	1484 (2120)	1538 (2115)	117 (353)	0:56:08 (1:20:12)	1:02:31 (1:26:00)	0:05:30 (0:16:35)	2:04:10 (3:02:48)	67.9%
Salvador	1992 (2900)	1842 (2962)	335 (929)	1:15:22 (1:49:43)	1:14:55 (2:00:27)	0:15:45 (0:43:39)	2:46:03 (4:33:50)	60.6%

German Open 2026 in Cologne, Germany. Impressions of the participations are shown in Figure 6. Task-specific segmentation models were trained using the pipeline, enabling on-site annotation and rapid model updates. The approach was successfully deployed for manipulation-centric tasks such as Storing Groceries, General Purpose Service Robot, and Restaurant, where the robot must interact with many objects.

5 Conclusion

In this paper, we presented a two-step label propagation approach that estimates general object segmentations for household objects: a Segment Proposer is trained to establish a shape bias, and an ensemble of Hopfield networks, trained only on a small representative dataset, assigns the labels. The approach proved more advantageous than simple metric comparison in Foundation Model embedding spaces. The ensemble of Hopfield heads was successfully applied to both ground truth annotations and annotations generated by the Segment Proposer. While detectors trained purely on pipeline-generated annotations remain functional, the observed 0.15 mAP gap to expert-corrected annotations makes a brief human review advisable for final deployment. We applied our approach to datasets gathered at previous RoboCup competitions, which we make publicly available together with the code.

While the Labeler Heads require only a small dataset to train, they must still be re-trained for the objects newly published at each competition. A promising direction is a "vocabulary" of representatives that is extended at each competition with previously unseen instances, reducing re-training overhead and the need for a new representative dataset beyond the instances not yet covered. Our experiments also show the importance of a reliable Segment Proposer, as it accounts for at least 0.2 mAP loss compared to ground truth segmentations. Future work could focus on models with better grounding and on a learned filter that

Table 7: Pipeline vs. Ground Truth performance by dataset.

	Pipeline			Ground Truth		
City	Precision	Recall	mAP	Precision	Recall	mAP
Bonn	.754 ±.004	.835 ±.005	.693 ±.004	.914 ±.007	.943 ±.004	.797 ±.004
Bordeaux	.768 ±.007	.584 ±.002	.422 ±.002	.867 ±.006	.913 ±.001	.698 ±.001
Eindhoven	.788 ±.003	.831 ±.005	.671 ±.004	.949 ±.005	.955 ±.003	.813 ±.002
Kassel	.668 ±.004	.826 ±.004	.703 ±.001	.962 ±.003	.968 ±.002	.855 ±.007
Cologne	.725 ±.005	.822 ±.002	.646 ±.001	.908 ±.002	.944 ±.003	.766 ±.002
Nuernberg	.687 ±.009	.818 ±.003	.663 ±.003	.869 ±.006	.956 ±.004	.811 ±.004
Salvador	.804 ±.006	.837 ±.001	.639 ±.002	.927 ±.005	.950 ±.003	.776 ±.004
Final	**.742±.005**	**.793±.003**	**.634±.003**	**.914±.005**	**.947±.003**	**.788±.003**

Storing Groceries (RCGO 2026) Clean the Table (RC 2025) Storing Groceries (RC 2025)

Fig. 6: Examples of the tasks that the robot performed during the RoboCup@Home competitions. Our labeling approach allowed for efficient perception model training that resulted in successful grasping in multiple tasks.

prevents out-of-distribution segments from reaching the Ensemble model, complementing the current confidence-based rejection. A direct comparison against modern open-vocabulary detectors [7, 21] on the full competition object set, as well as ablations of the Foundation Model and Hopfield bank count, are further directions.

Another promising next step is to extend our approach to video, enabling more efficient data capture: our method could label a few frames, and the labels could be propagated through the remaining frames with video object segmentation methods such as SAM 2 [18].

Acknowledgment

The authors would like to thank Jan Nogga for his support in the data collection process and the preparation of the models.

References

[1] N. Carion et al. "End-to-end object detection with transformers". In: *European Conference on Computer Vision (ECCV)*. Springer. 2020, pp. 213–229.

[2] CVAT.ai Corporation. *Computer Vision Annotation Tool (CVAT)*. Version v2.4.3. Apr. 2023. DOI: 10.5281/zenodo.7863887.

[3] A. Gouda et al. "Learning Embeddings with Centroid Triplet Loss for Object Identification in Robotic Grasping". In: *IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE. 2024, pp. 3577–3583.

[4] G. Jocher et al. *Ultralytics YOLOv5 v7.0 – SOTA Realtime Instance Segmentation*. Version v7.0. 2022. DOI: 10.5281/zenodo.7347926.

[5] A. Kirillov et al. "Segment anything". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 4015–4026.

[6] F. Li et al. "Mask DINO: Towards a unified transformer-based framework for object detection and segmentation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 3041–3050.

[7] S. Liu et al. "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection". In: *European Conference on Computer Vision (ECCV)*. Springer. 2024, pp. 38–55.

[8] N. Massouh, L. Brigato, and L. Iocchi. "RoboCup@Home-Objects: Benchmarking object recognition for home robots". In: *Robot World Cup*. Springer, 2019, pp. 397–407.

[9] M. Matamoros et al. "RoboCup@Home: Summarizing achievements in over eleven years of competition". In: *2018 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. IEEE. 2018, pp. 186–191.

[10] R. Memmesheimer et al. "Adaptive Domestic Service Robotics through Foundation Models for Perception, Interaction, and Action". In: *2nd German Robotics Conference (GRC)*. 2026.

[11] R. Memmesheimer et al. "NimbRo@Home 2023 Open Platform League Team Description". In: *RoboCup@Home 2023, Team Description Papers*. 2023.

[12] R. Memmesheimer et al. "RoboCup@Home 2024 OPL winner NimbRo: Anthropomorphic service robots using foundation models for perception and planning". In: *Robot World Cup*. Springer, 2024, pp. 515–527.

[13] D. D. R. Meneghetti et al. *Annotated image dataset of household objects from the RoboFEI@Home team*. 2020. DOI: 10.21227/7wxn-n828.

[14] T. Novkovic et al. "CLUBS: An RGB-D dataset with cluttered box scenes containing household objects". In: *The International Journal of Robotics Research* 38.14 (2019), pp. 1538–1548.

[15] B. Pätzold, J. Nogga, and S. Behnke. "Leveraging vision-language models for open-vocabulary instance segmentation and tracking". In: *IEEE Robotics and Automation Letters* (2025).

[16] A. Radford et al. "Learning transferable visual models from natural language supervision". In: *International Conference on Machine Learning (ICML)*. PmLR. 2021, pp. 8748–8763.

[17] H. Ramsauer et al. "Hopfield networks is all you need". In: *International Conference on Learning Representations (ICLR)*. 2021.

[18] N. Ravi et al. "SAM 2: Segment Anything in Images and Videos". In: *arXiv preprint arXiv:2408.00714* (2024).

[19] J. Shang et al. "Theia: Distilling diverse vision foundation models for robot learning". In: *Conference on Robot Learning (CoRL)*. 2024.

[20] S. Tyree et al. "6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2022, pp. 13081–13088.

[21] X. Zhou et al. "Detecting twenty-thousand classes using image-level supervision". In: *European Conference on Computer Vision (ECCV)*. Springer. 2022, pp. 350–368.