2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Efficient Real-World Benchmarking for Practical Fine-Grained Product Identification in Retail Robotics for Picking and Stock Taking

Jochen Lindermayr[1], Florian Jordan[1], Cagatay Odabasi[1], Werner Kraus[1], Richard Bormann[1], Marco F. Huber[2]

***Abstract*— The rapid evolution of retail robotics is set to transform in-store operations through advanced automation, spanning vision-based inventory tracking, order picking, packing, and restocking. Yet fine-grained product identification remains a bottleneck: assortments change, packaging evolves, and shelves host thousands of near-duplicates—requiring perception systems that can adapt quickly with minimal setup. This paper targets that gap with two contributions. First, we present a semi-automated, robot-assisted acquisition pipeline that records 3D scene ground truth via iterative placement, projecting it into each image, yielding dense, low-cost annotations at scale. Second, we extend IPA-3D1K with challenging real shelf scenes containing 130 near-duplicate SKUs. While scenes are not paired one-to-one, the same product set appears across synthetic and real images, enabling controlled, object-level sim/real analyses under occlusion, rearrangement, and lighting variation. Using frozen DINOv3 features, our baseline recognition pipeline allows index updates in minutes. We evaluate training-free or fast approaches (kNN and a lightweight classifier head) to assess the capabilities and limitations of this representation in fine-grained retail identification. Experiments show that on the FineGrainedOCR dataset the lightweight head improves over kNN by ∼11 percentage points, narrowing the gap to fully trained models to 1.9–5.3 pp. On IPA-3D1K (1,000 SKUs), synthetic-scene retrieval is strong (Top-1 ≈90%, Top-2 ≈95%), while exact disambiguation among near-duplicates remains challenging. We find that confidence thresholds enable targeted triage during inference, and a neighborhood-based risk signal predicts confusion during training, indicating where specialized modules are most beneficial.**

I. Introduction

Retail robotics represents a rapidly evolving sector poised to revolutionize business operations through advanced automation. This progression in robotics technology enables the automation of a broad spectrum of retail tasks, from vision-based inventory management [1] to customer order picking [2], including packing [3], and restocking processes [4], [5]. Such advancements hold promise for enhancing operational efficiency, mitigating costs, addressing workforce shortages, and improving customer experiences. Additionally, the recent global COVID-19 crisis has underscored the importance of minimizing human contact in retail settings, further emphasizing the role of robotics in facilitating seamless, contactless shopping solutions.

Fig. 1. General physical setup: Two calibrated RGB-D cameras (yellow) mounted on the robot end-effector view at the shelf; the camera pose in the image captures the iterative (image, SKU) pairs. A handheld barcode scanner (blue) records SKUs; a staging area holds upcoming products (white); a controllable studio light varies illumination (green); and already placed objects in the shelf are indicated in purple.

For accurate inventory tracking and for ensuring correct order fulfillment a visual object recognition system with fine-grained distinction capabilities is required. The domain of fine-grained object recognition within retail robotics presents unique challenges, primarily due to the dynamic nature of retail object sets. These sets are characterized by frequent changes in product packaging designs and the continuous evolution of product lines, necessitating adaptable object perception systems. Retail environments typically feature thousands of products, with products from the same brand often bearing similar designs or packaging [6]. This situation introduces significant complexity, requiring object perception systems to manage large, evolving sets of similar-looking items effectively. Fine-grained object recognition holds significant potential also in healthcare inventory tracking [7] due to the shared characteristic with the retail use-case of similarly appearing items, such as medications in nearly

This work has received funding from the German Ministry of Education and Research (BMBF) under grant agreement No 01IS21061A for the Sim4Dexterity project, the German Federal Ministry for Economic Affairs and Climate Action (BMWK) under grant agreement No 01MK20001G for the Knowledge4Retail project and the Baden-Württemberg Ministry of Economic Affairs, Labour and Tourism for the AI Innovation Center "Learning Systems and Cognitive Robotics".

[1]The authors are researchers with the Department of Robot and Assistive Systems, Fraunhofer IPA, 70569 Stuttgart, Germany. <first name>.<last name>@ipa.fraunhofer.de

[2]Department Cyber Cognitive Intelligence (CCI), Fraunhofer IPA, and Institute of Industrial Manufacturing and Management IFF, University of Stuttgart, 70569 Stuttgart, Germany. marco.huber@ieee.org

identical packaging or medical instruments with subtle variations. By accurately distinguishing between these similar objects, fine-grained recognition ensures precise tracking and management of healthcare inventories.

Existing methodologies in object recognition face several hurdles, particularly in retail settings. The necessity for rapid integration of new products into detection systems and the removal of outdated models poses a significant challenge [8].

Despite progress in category-level recognition, there is no standard dataset to evaluate fine-grained identification for all relevant robotics tasks at retail scale. Existing benchmarks either cover few classes [9], [10], use only synthetic data [11], [12], or have only real images [13] without calibration or multi-view pairs. Most do not offer synthetic and real scenes of the same large and similar-looking object set, making sim2real comparability difficult and limiting analysis for training-free, quickly updated recognizers.

Hence, we extend the existing IPA-3D1K [14], a dataset comprising 1,000 retail product models with fine-grained classes, with substantial real shelf scenes. The extension preserves breadth in Stock Keeping Unit (SKU) coverage, provides calibrated multi-view captures, and uses the same products as in the synthetic scenes. This enables controlled synthetic-to–real studies and stress tests under occlusion, rearrangement, and light variances.

As data collection at this scale is non-trivial, we developed a semi-automated, robot-assisted acquisition pipeline for fine-grained shelf scenes (see Fig. 1). We combine assisted manual iterative (image, SKU) capture [15], automated SKU-based retrieval of the product model and metadata, and multi-view recording along predefined robot poses under varied lighting. 3D poses estimated in world frame coordinates from iterative captures are then projected into each robot image to yield dense annotations. It reduces annotation cost, increases ground truth correctness, and supports research on view selection in potential active robot vision.

We evaluate three state-of-the-art product identification approaches on the IPA-3D1K dataset [14] (1,000 products) and our real-scene extension to test recognition among highly similar items in shelf scenes. We compare synthetic and real results to quantify sim2real behavior. We also test on a third-party dataset [16] (256 products) tailored for fine-grained retail products to assess generalization beyond our capture domain. Our experiments aim for a system which allows rapid, training-free index updates within minutes and large-scale identification accuracy under occlusion and chaotic placement.

The key contributions of this work are as follows:

- A semi-automated and robot-assisted pipeline for efficient dataset acquisition tailored specifically for retail product scenarios.
- Creation of a novel real-world dataset for fine-grained (active) product identification in retail shelf scenes using the proposed acquisition pipeline.[1]
- Evaluation of three training-free methods for scalable retail object identification, including the recently published DINOv3 [17], which is also integrated into our product identification method, and a sim2real analysis enabled by the newly recorded data.
- Comparative analysis with training-based methods on an alternative fine-grained retail identification dataset [16].

[1]This real-world dataset will be made publicly at `https://ipa-jcl.github.io/ipa-3d1k/real-benchmarking`.

II. Related Work

1) Retail Product Recognition: The works [6], [18] provide a comprehensive overview of the state-of-the-art in product recognition within retail shelves, highlighting several challenges inherent to the domain, like the similarity between products. Similar to [1], [19], they focus on recognizing orderly placed products from front views alone. The survey points out that much of the existing research focuses on products that are neatly arranged, overlooking the common retail scenario of arbitrarily placed items, often misplaced when put back by customers. This gap in consideration underscores the need for methodologies capable of handling the unpredictable nature of product placement within retail environments, thereby extending the applicability of recognition systems to more realistic retail settings. The work in [20] discusses an approach for product identification heavily relying on text recognition, highlighting its advantage of not requiring annotations. However, there are challenges such as decreased effectiveness in cases of blurry images, occlusion, or self-occlusion. [16] supports this finding with text embeddings adding only very small additional information besides the visual appearance in general. Moreover, correcting Optical Character Recognition (OCR) errors proves difficult under these conditions, indicating a need for enhanced robustness in recognition methodologies for reliable product identification. InstanceNet [8] follows a similar motivation of low-effort training of object detectors capable of handling large object sets but needs up to one hour per object. Other works in the retail area focus on detecting free-space between products in shelves [21] or detecting whole colli [22] which has another focus than our work. The work in [2] presents a holistic robot picking system incorporating an object recognition system using SIFT feature points which fails in case of less textured products.

2) Foundation Models on Related Tasks: Qwen2.5-VL is useful as an open-vocabulary localizer or detector as it can return object bounding boxes/points in zero-shot mode making it strong for proposal generation on retail shelves (e.g., “box all bottles”) [23]. However, for fine-grained SKU identification among look-alike products, recent evidence shows Large Vision Language Models (VLMs) including Qwen-style generative paradigms struggle: on the retail-focused MIMEX benchmark, state-of-the-art VLMs achieved unsatisfactory fine-grained zero-shot classification, and broader evaluations on fine-grained tasks report notable limitations [24], [25]. This is consistent with our own experiments using Qwen and GroundingDino. Moreover, Qwen2.5-VL’s available sizes (up to 72B) imply that per-instance reasoning over many proposals in dense scenes is

computationally costly, so applying it to every box is typically impractical [23]. Hybrid retrieval-augmented pipelines are emerging as a more scalable alternative for SKU-level identification [26].

DINOv2 [27] has shown promise in various domains by integrating global context with local image features. This architecture facilitates the encoding of spatial information concerning object parts and semantic categories. Its effectiveness in zero-shot scenarios, where it establishes semantic correspondences without specific training, has proven also successful in object detection [28] or 6-DoF pose estimation [29]. The recently published DINOv3 [17] extends the DINO family into a suite of scalable vision foundation models that produce high-quality dense features and deliver state-of-the-art results across diverse tasks with a frozen backbone. Compared to DINOv2, DINOv3 scales both data and model size. Under a standard linear probing approach, the authors train a linear head on frozen features. For classification they probe the CLS token [17]. In the paper's fine-grained classification evaluation, DINOv3 attains the highest linear-probe accuracy on iNaturalist-2021 (89.8%) and competitive results on iNaturalist-2018 and Places205, while averaging 93.0% over 12 small fine-grained benchmarks [17]. Although these results support the claim that DINOv3 captures fine-grained details, the reported datasets do not include retail product identification. Thus, our setting remains untested in their benchmarks.

Leveraging insights from these other works on foundation models, our work employs DINOv2 and DINOv3 to address specifically the task of fine-grained retail product recognition for robotics.

3) Retail Datasets: Public datasets relevant to robot vision and retail fall into two groups. **(i) 3D-model–centric:** *YCB* [10] (77 objects) targets manipulation and grasping but has few retail items and limited texture fidelity. *ShapeNet* [30] (51k objects/55 classes) offers many models but often coarse geometry and weak textures, and is largely non-retail. *BlenderKit* [31] provides ∼7k royalty-free 3D assets of general objects. *NVIDIA HOPE* [9] delivers 28 textured, retail-oriented 3D models but with a small object count not suitable for large-set fine-grained retail evaluation. **(ii) Retail 2D imagery:** *StandardSim* [12] contains rendered retail scenes (no 3D assets released), useful for detection but limiting view resampling/physics. *CPGDet-129* [11] offers rendered shelf scenes with boxes/masks but no product meshes. *SKU110K* [13] and *RP2K* [19] provide large-scale real 2D product images for dense detection/fine-grained classification, yet lack 3D models. *FineGrainOCR* [16] contains high-resolution 2D images of conveyor-belt objects (real-only, random train/test splits; no 3D). *MIMEX* [24] provides crops of hand–product interactions with fine-grained classes (real-only; no 3D).

IPA-3D1K [14] addresses these gaps by focusing on 1,000 retail products with textured 3D models and many near-duplicates, enabling recognition and physics-based manipulation in simulation. Still, the amount of real scenes and their variance is comparably low.

4) (Semi-)Automated Instance Annotation: Semi-automated annotation reduces manual effort in datset generation but often trades accuracy for assumptions. Tracker-based pipelines initialize poses and propagate masks over time, which can drift and degrade masks [32]. RFID- and multi-view systems add identity and geometric constraints but require objects to be fitted with RFID-tags or complex calibration [33], [34]. CAD-alignment methods achieve high pixel fidelity by registering product models to depth [35]. [15] presents a semi-autoamted method for change region detection and segmentation in iterative image captures.

In our retail-shelf setting, we adopt and combine the last two methods: product 3D models are selected by Barcode-SKU and combined with calibrated RGB-D captures, world-space ground-truth aggregation, and visibility-aware projection. This removes dependence on tracking or RFID and yields dense per-instance labels with consistent identities and poses.

III. Dataset Creation Pipeline

This chapter details the physical setup and the data-collection pipeline used to generate per-instance annotations for shelf scenes. The overall pipeline is shown in Fig. 2 which focuses on the steps *Scene Setup* and *Post-processing*.

A. General physical recording setup

We are leveraging an existing robot cell built for feasibility and experimental evaluation on retail robotic handling tasks. In the following, we explain the complete physical recording setup as shown in Fig. 1. Colored ellipses in the figure mark the key components:

- Yellow: two calibrated RGB-D cameras mounted on the robot end effector. The robot and camera pose seen in the image provides a perpendicular, far view onto the shelf which is the reference viewpoint used to capture the iterative *(image, SKU)* tuples during the human-in-the-loop scene setup. All sensor intrinsics and extrinsics are known.
- Blue: handheld barcode scanner used by the operator to read the SKU of each product before placing on the shelf.
- White: staging area containing the next products to be placed.
- Green: controllable studio light used to vary illumination.
- Purple: objects that have already been scanned and placed on the shelf.

We use two RGB-D cameras mounted on the robot arm, providing synchronized color and depth from known robot poses. The Ensenso-C (industrial-grade stereo camera with fixed optics) delivers stable, high-quality depth, while the OAK-D Pro (consumer-grade active stereo camera) offers configurable lens positions to target products within the current view frustum at different focus distances across different robot arm poses. Both sensors are calibrated (intrinsics, extrinsics) for consistent fusion of geometry and appearance

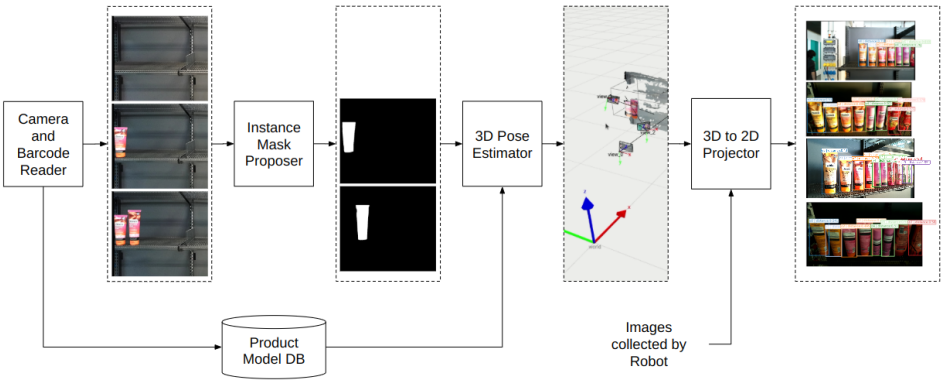

Fig. 2. System overview of the iterative semi-automated shelf-annotation pipeline: A human operator scans a product with a handheld barcode reader to get the SKU, places the product in the scene, and then captures an image by a static calibrated camera, yielding an (image, SKU) pair per iteration. An Instance Mask Proposer segments candidate regions in an image by comparing it to its predecessor image. The scanned SKU indexes the corresponding product model and metadata, which are passed to a 3D pose estimator that aligns the model to the candidate image region and returns the object's 6-DoF pose in world coordinates. These poses are accumulated as scene-level 3D ground-truth (GT). A 3D→2D projector then renders this 3D GT scene back into 2D image space for a second image stream collected by a robot arm equipped with calibrated color cameras following a predefined sequence of poses. The projections produce dense, per-instance annotations on the robot-captured images. Stacking results yields a dataset of fully annotated scene images with known identity for each visible product.

along predefined robot trajectories. For the iterative (image, SKU) pair acquisition, we use the OAK-D Pro. The lens values were optimized before data collection for each robot pose to ensure optimal focus and hence visibility of fine details and readable text.

B. Scene setup phase

As proposed in [15], a human operator iteratively assembles the shelf scene according to the following steps:

1) The operator scans a product from the staging area with the handheld barcode reader to get the SKU.
2) The product is placed in the shelf at the desired position and orientation.
3) A reference RGB-D image is captured from the perpendicular, far viewpoint, producing an *(image, SKU)* pair.
4) Based on images of two consecutive pairs, the *Instance Mask Proposer* [15] outputs a candidate region. For each *(image, SKU)* pair, the SKU retrieves the corresponding product model and metadata from the product database. Using the product model and the region, we estimate the object's 6-DoF pose in world coordinates. Aggregating these poses over all placed products yields a 3D scene representation that serves as ground truth in world space.

Automating product selection via the SKU is essential, as manual choices are error-prone, especially for product sets with highly similar-looking products. Using the SKU to index the product model and metadata increases annotation consistency and, consequently, the quality of the resulting ground-truth data. For the *Instance Mask Proposer* module, we use the approach presented in [15] where it was evaluated in detail. It is crucial for clean ground truth data to rely on the proposed mask as otherwise due to similar looking products pose estimation may choose the wrong product in the scene.

C. Robot data collection

After the scene is assembled, a UR10e robot arm equipped with calibrated cameras records images from a predefined sequence of camera poses. The trajectory covers multiple viewing angles and distances (near, far, frontal, oblique from above, and oblique from the side) to ensure that changes in product appearance due to perspective and scale are well captured. To increase variability, we also capture under different lighting conditions: first with the studio light switched off, then with it switched on to intentionally introduce specular reflections and illumination diversity. To further enrich the variance, we repeat all the above steps after applying physical rotations of the products around their respective upwards vector to capture viewpoints which would not have been seen otherwise as these surface pointed towards the shelf backside in earlier image recordings.

D. Post-processing

From this point the pipeline follows the 3D→2D projection path which is defined as follows:

1) For every image frame captured by the robot, the known intrinsics and the per-pose extrinsics from hand–eye calibration are used by the 3D→2D projector to render the 3D scene ground truth into 2D image space. The resulting projections provide dense per-instance annotations (masks, bounding boxes, and labels).
2) Visibility is enforced by depth-aware rendering so that only the pixels visible from a given robot viewpoint are kept. Frames are discarded when a projected instance is more than 50% outside the image to avoid ambiguous, heavily truncated annotations. Occlusion handling is performed per viewpoint and product to ensure that annotations reflect the true visible extent of each object.

This procedure yields an automatically annotated dataset

with known identity and pose for every product across diverse viewpoints and lighting conditions.

IV. Dataset of Real Images

Using the pipeline described in Sec. III, we collected a new dataset of real product images captured on a retail shelf. As the primary goal is fine-grained product identification, we selected roughly 130 SKUs from the original dataset [14] that proved difficult to distinguish in earlier work, yielding a hard-case benchmark with many near-duplicate appearances for which also 3D models and simulated image data is available. Refrigerated products were excluded because they could not be stored from the previous dataset creation period.

The dataset comprises 10 base shelf scenes that we iteratively annotated following the procedure in Sec. III. To increase intra-class variability without altering global layout, we produced additional scene variants by rotating products about their vertical (up-)axis in place, thereby changing visual appearance while keeping positions and hence ground truth data fixed. Randomly selected examples at frontal view before product rotation are shown in Fig. 3. It can be seen that nearly every product has a near-duplicate.

In total, we gathered about 7,000 cropped product images from two cameras across varied viewpoints and illumination. This combination of challenging classes, controlled layout changes, and multi-camera capture is intended to stress-test fine-grained recognition methods in realistic retail shelf environments.

Fig. 3. Randomly selected examples from different shelf setups before random rotation of products. Nearly every product has a near-duplicate product.

V. Experiments

Our objective is efficient recognition of *fine-grained* products in *large, dynamically changing* retail environments. We deliberately avoid monolithic end-to-end retraining by decoupling representation learning from classification, thereby improving scalability and facilitating the rapid onboarding of new classes.

We study fine-grained product identification with two complementary regimes:

1) a training-free k-nearest-neighbor (kNN) retrieval classifier operating on frozen DINOv3 embeddings, and
2) a lightweight training-based classifier (a shallow linear head) trained on the same frozen embeddings. This combines the value of modern foundation features with fine-tuning, while keeping compute and maintenance overhead minimal.

We examine two ways to construct the template set at setup time to be used at inference:

1) Real template images from the FineGrainOCR dataset [16] (real images only).
2) Synthetic template images rendered from 3D product models (IPA-3D1K dataset) [14].

Embedding and indexing: All real (FineGrainOCR) or rendered (IPA-3D1K) template images are encoded with a frozen DINOv3 backbone to produce feature vector embeddings. We ℓ_2-normalize them and populate a vector database with them. For large-scale, low-latency search we use Faiss [36], which provides efficient approximate nearest-neighbor indices and scalable cosine-similarity retrieval. During experiments, we quantified the runtime performance of each system module: For the setup phase on an *RTX2080* graphics card, rendering 34 views per object took 2.73 minutes, and embedding all of these views took just 19.3 milliseconds.

Inference: At test time, segmented product crops are embedded with the same DINOv3 encoder and queried against the template index using approximate kNN with cosine similarity. The system returns the Top-k neighbors, their cosine scores, and associated metadata (e.g., SKU, class ID and template image info like rendering perspective (IPA-3D1K) or real image name (FineGrainOCR). Exemplary qualitative Top-k retrievals are shown in Fig. 4. The kNN vector database loading time at robot startup takes in mean 12 seconds to load all data for 1000 products. Embedding per scene is measured at 0.26 seconds, and database retrieval at negligible 0.3 milliseconds. These results highlight our system's efficiency, balancing speed with accuracy effectively which allows scene inference on a robot during operation.

Protocol and metrics: We evaluate multi-class classification with Top-1 and other Top-k accuracies. For the kNN classifier, the predicted ranking is obtained by sorting per-class similarity scores. All methods use the official test split. No images from the test set are used for training, feature

Fig. 4. Top-k results (right) sorted by similarity score from top left to bottom right for a given scene image crop of a segmented object proposal (left). Besides views of the correct object (green underline) we also get views of similar-looking objects (note that the different dark blue object renderings do not belong to the same object - they are near-duplicates).

fitting, or hyperparameter selection.

A. FineGrainOCR Dataset

Dataset: FineGrainOCR [16] comprises 256 retail product categories captured with a high-resolution camera while products move on a conveyor belt. We follow the official train/test split provided by the accompanying paper. Both split sets contain randomly sampled real images only, and neither synthetic data nor 3D models of the products are available. Because FineGrainOCR is real-only and provides no 3D assets, we do not perform synthetic augmentation or 3D mesh model based rendering on this dataset. This contrasts with our retail 3D settings where matched real+synthetic views of the *same* object identities can be leveraged.

Product Identification Methods: **kNN retrieval classifier:** We build a vector database from all training images by extracting DINOv3 features (frozen backbone; ℓ_2-normalized) as embeddings. At test time we compute cosine similarities to all training embeddings, retrieve the top-k neighbors. **Linear classification head (clsHead):** On the same frozen DINOv3 features, we train a small network with one hidden layer of dimension 256 and a final softmax layer using cross-entropy. Input size is defined by the DINO embedding vectors and output size by the number of products in the dataset. We use the Adam Optimizer, learning rate $1e{-}3$, weight decay 0.05, batch size 4096, up to 200 epochs with early stopping on a 20% validation split taken from the training set on which the top-1 validation accuracy is evaluated during training. Because training is performed on embeddings rather than raw images, the required training time is only 2.5 up to 5 minutes on a mobile PC with a RTX2080 GPU, depending on convergence. **Published baseline:** For reference and comparison to a fully learned method, we also report the *ResNet-50* results published by the dataset authors.

Preprocessing: Product crops are segmented from the whole images using SAM and the crop images are resized such that the longer side is 512 px as recommended by the authors of the dataset.

Findings: The results are shown in Tab. I and Fig. 5. The following aspects are observed:

- **DINOv3 vs. DINOv2 on this fine-grained retail dataset:** At $k{=}1$, DINOv3 improves over DINOv2 by $+5.4$ pp for kNN (82.2 vs. 76.8) and $+2.6$ pp for the *clsHead* (91.8 vs. 89.2).
- **Classifier head vs. kNN (same backbone):** At $k{=}1$, adding the lightweight classification head yields $+12.4$ pp on DINOv2 (89.2 vs. 76.8) and $+9.6$ pp on DINOv3 (91.8 vs. 82.2) — ~ 11 pp on average.
- **Gap to fully trained models:** Our best zero-/low-shot variant (*DINOv3_clsHead*, 91.8%) is only 1.9–5.3 pp behind fully trained CNNs/ConvNeXts (93.7–97.1%).
- **Plausibility vs. prior DINOv3 reports:** Absolute accuracies and the shape of the Top-k curves are in the expected range for DINO-style features on fine-grained recognition, consistent with results reported by the DINOv3 authors [17]. This confirms that the previously observed behavior also holds in the retail domain.
- **Parameter vs. accuracy:** With frozen DINO features and a small classifier head, we achieve performance close to a fully trained model on this dataset while only fine-tuning the output layer, resulting in significantly reduced training cost and preserving full flexibility and scalability for retail applications. The training effort scales inversely with the number of learnable parameters given in Tab. I.

TABLE I

TOP-1 CLASSIFICATION ACCURACY USING IMAGE SIZE 512 PX WITH 400 MAX SAMPLES PER CLASS ON THE TEST SET OF FINEGRAINOCR. BEST METHOD PER CATEGORY IS HIGHLIGHTED IN **BOLD**. OVERALL BEST IS UNDERLINED.

Method	Accuracy (%)	Learnable Param.
MobileNetV3-Small	93.7	1.74 M
MobileNetV3-Large	95.8	4.45 M
ResNet18	96.1	11.32 M
ResNet50	**97.1**	24.08 M
ConvNeXt-Tiny	96.4	28.03 M
ConvNeXt-Large	96.3	196.86 M
Ours: DINOv2_kNN	76.8	0 M
Ours: DINOv2_clsHead	89.2	0.53 M
Ours: DINOv3_kNN	82.2	0 M
Ours: DINOv3_clsHead	**91.8**	0.33 M

B. IPA-3D1K Dataset

Dataset: The dataset in [14] comprises 1,000 shelf-ready retail product 3D models, each defining a distinct object class. Models span diverse physical attributes (size, weight, shape) and packaging variations (material, design), sourced from eight retailers and 300 brands. Packaging types include cuboid boxes, cans, bottles, jars, foil bars, beverage cartons, and bags/pouches, yielding varied visual cues such as logos and symbols. To counter low inter-class variance, the set deliberately includes look-alike products from different

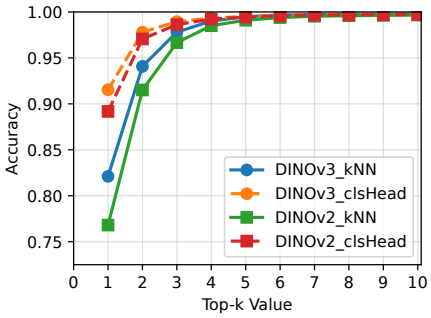

Fig. 5. Top-k accuracy given the top-k value for the FineGrainOCR test image dataset using image size 512 px with max samples per class set to 400.

classes (e.g., food and hair-care variants), enabling fine-grained classification of subtle product differences. [14]

The synthetic test set consists of rendered shelf scenes with retail objects in chaotic layouts to mimic real stores, where customer interactions disrupt order. This introduces intra-scene variation in class, pose, distance, and occlusion. Using BlenderProc [37] with physics, we varied shelf types, materials, backgrounds, lighting, and systematically sampled camera viewpoints to ensure realistic, robust coverage.

The real test set is based on the recorded dataset presented in this work (see Sec. IV). Please note, that the real dataset only contains roughly 130 hard product classes.

Product Identification Methods: We use the same methods for kNN and linear class head as presented for the other dataset above.

Findings: The results for different configurations and datasets subsets (synthetic, real) are shown in Fig. 6. The following is observed:

- **DINOv3 on a large, fine-grained label space:** Despite the huge number of classes, DINOv3 remains strong: on synthetic scenes both *kNN* and *clsHead* reach around $\sim$90% at k=1 and approach $\sim$95% already at k=2. This shows that DINOv3 is good at correctly identifying the group of similar products.
- ***clsHead* vs. kNN:** The *clsHead* underperforms on this dataset, even scoring below *kNN* using the same backbone. This likely indicates overfitting to the rendered templates, which exhibit lower intra-class variance compared to real product images.
- **Real vs. synthetic difficulty:** Besides the fact of real images including real environment difficulties, the real dataset is substantially harder than FinegrainedOCR and the full IPA-3D1K synthetic set, as it contains only "hard" cases. When we filter the synthetic evaluation results to only the product identities present in the real image dataset, DINOv3 kNN Top-1 accuracy drops to **78%**, confirming the increased difficulty for these products.

C. General Findings

During the experiments we have observed several findings valid for both datasets:

- **Confusion-set retrieval vs. exact identification:** DINOv3 features are highly effective for rapidly retrieving a compact *confusion set* (high Top-k). However, distinguishing the exact product within this set remains challenging, particularly for near-duplicates. Incorporating additional cues (e.g., alternative viewpoints, OCR, or size information) can improve disambiguation. The generic pretraining of DINOv3 appears to encode both visual and semantic attributes, occasionally causing confusion when products share similar shapes, colors, or brand identity. For example, round, blue items from the same brand were often misclassified among each other. Although DINOv3's broad training scope enhances its versatility, it also makes distinguishing items with subtle differences challenging, underscoring the importance of fine-grained recognition in retail.
- **Early Top-k strength:** Top-2 accuracy is already high on both dataset, and performance increases sharply for small k, largely saturating by $k\leq$10. This indicates that considering the top 10 predictions is typically sufficient for effective retrieval.
- **Confidence-driven triage:** Model prediction probabilities provide a useful trigger for follow-up actions. Empirically, predictions with softmax confidence below $\sim$ 0.80 should be flagged for detailed review.
- **Predicting failure cases from neighborhood structure at train time:** For each sample, we examine its nearest neighbor of the same class and analyze that neighbor's k=5 to predict whether a similarly-viewed image of the same class is likely to be misclassified during inference. The "problem/no-problem" prediction achieves a precision of 0.99 and recall of 0.759, providing a basis for designing specialized strategies to handle these challenging cases.
- **Robustness across domains:** The neighborhood-based risk signal generalizes across both datasets and across synthetic as well as real data.

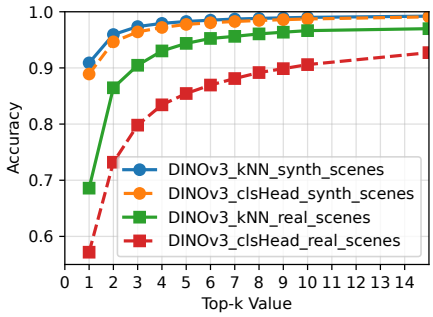

Fig. 6. Top-k accuracy given the top-k value for the synthetic and real test images in the IPA-3D1K dataset.

VI. Conclusions

We introduced a semi-automated, robot-assisted data acquisition pipeline and used it to create a hard real-world benchmark ($\sim$ 130 near-duplicate SKUs) for fine-grained

retail product identification, being released upon publication. The recently published DINOv3 was used to evaluate the performance of training-free methods for fine-grained retail product identification on two domain-specific datasets. With frozen DINOv3 features plus a lightweight classifier head, we attain strong, scalable performance: on FineGrainOCR (256 classes) the head adds ~ 11 pp over kNN's accuracy, narrowing the gap to fully trained models to 1.9–5.3 pp.

On IPA-3D1K, despite much bigger with 1000 product classes, DINOv3 excels at rapid confusion-set retrieval (Synthetic Top-1 $\sim 90\%$; Top-2 $\sim 95\%$), but the exact identification among near-duplicates remains a challenge. On our hard subset of real images, kNN Top-1 drops to even 78%, highlighting the need for methods focusing on details. The experiments show that confidence thresholds enable targeted triage during inference and a simple neighborhood risk signal (precision 0.99; recall 0.759 in experiments) predicts confusion cases reliably already during training which enables additional specialist methods to solve it. Future work includes active multi-view acquisition and fusion of OCR or geometry information to resolve near-duplicates.

REFERENCES

[1] M. Beetz et al., "Robots collecting data: Modelling stores," in *Robotics for Intralogistics in Supermarkets and Retail Stores*, Springer, 2022, pp. 41–64.

[2] R. Bormann, B. de Brito, J. Lindermayr, M. Omainska, and M. Patel, "Towards Automated Order Picking Robots for Warehouses and Retail," in *Proceedings of the 12th International Conference on Computer Vision Systems (ICVS)*, 2019.

[3] T. Nickel, R. Bormann, and K. O. Arras, "Multi-heuristic robotic bin packing of regular and irregular objects," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 10 730–10 736.

[4] A. Cavallo et al., "Robotic clerks: Autonomous shelf refilling," in *Robotics for Intralogistics in Supermarkets and Retail Stores*, Springer, 2022, pp. 137–170.

[5] L. Villani, C. Natale, M. Beetz, and B. Siciliano, *Robotics for Intralogistics in Supermarkets and Retail Stores*. Springer Nature, 2022, vol. 148.

[6] B. Santra and D. P. Mukherjee, "A comprehensive survey on computer vision based approaches for automatic identification of products in retail store," *Image and Vision Computing*, vol. 86, pp. 45–63, 2019.

[7] J. Lindermayr, C. Odabasi, T. Wohlleber, and B. Graf, "Multi-modal visual withdrawal detection for inventory management on a robotic care cart," in *ISR 2020; 52th International Symposium on Robotics*, VDE, 2020, pp. 1–6.

[8] R. Bormann, X. Wang, M. Völk, K. Kleeberger, and J. Lindermayr, "Real-time Instance Detection with Fast Incremental Learning," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 056–13 063.

[9] Y. Lin, J. Tremblay, S. Tyree, P. A. Vela, and S. Birchfield, "Multi-view Fusion for Multi-level Robotic Scene Understanding," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 6817–6824.

[10] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The YCB object and Model set: Towards common benchmarks for manipulation research," in *2015 International Conference on Advanced Robotics (ICAR)*, 2015.

[11] P. Fulop, *CPGDet-129*, `https://github.com/neurolaboratories/reshelf-detection`, 2022. Accessed: Feb. 27, 2023.

[12] C. Mata, N. Locascio, M. A. Sheikh, K. Kihara, and D. Fischetti, "StandardSim: A Synthetic Dataset for Retail Environments," in *International Conference on Image Analysis and Processing*, Springer, 2022, pp. 65–76.

[13] E. Goldman et al., "Precise detection in densely packed scenes.," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] J. Lindermayr et al., "IPA-3D1K: A large Retail 3D Model Dataset for Robot Picking," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[15] F. Jordan, J. Lindermayr, R. Bormann, and M. F. Huber, "Low-effort iterative dataset generation pipeline for unknown object instance segmentation," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2025, pp. 18 612–18 619.

[16] T. Pettersson, M. Riveiro, and T. Löfström, "Multimodal fine-grained grocery product recognition using image and ocr text," *Machine Vision and Applications*, vol. 35, no. 4, p. 79, 2024.

[17] O. Siméoni et al., *DINOv3*, 2025. arXiv: `2508.10104 [cs.CV]`.

[18] V. Guimarães, J. Nascimento, P. Viana, and P. Carvalho, "A review of recent advances and challenges in grocery label detection and recognition," *Applied Sciences*, vol. 13, no. 5, 2023.

[19] J. Peng, C. Xiao, and Y. Li, "RP2K: A Large-Scale Retail Product Dataset for Fine-Grained Image Classification," *arXiv e-prints*, arXiv–2006, 2020.

[20] B. Santra, A. K. Shaw, and D. P. Mukherjee, "Part-based annotation-free fine-grained classification of images of retail products," *Pattern Recognition*, vol. 121, p. 108 257, 2022.

[21] F. Šikić, Z. Kalafatić, M. Subašić, and S. Lončarić, "Enhanced out-of-stock detection in retail shelf images based on deep learning," *Sensors*, vol. 24, no. 2, 2024.

[22] M. Völk, K. Kleeberger, W. Kraus, and R. Bormann, "Towards packaging unit detection for automated palletizing tasks," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2023, pp. 8310–8317.

[23] S. Bai et al., *Qwen2.5-vl technical report*, 2025. arXiv: `2502.13923 [cs.CV]`. [Online]. Available: `https://arxiv.org/abs/2502.13923`.

[24] A. O. Tur et al., "Exploring fine-grained retail product discrimination with zero-shot object classification using vision-language models," *CoRR*, vol. abs/2409.14963, 2024.

[25] H.-T. Yu, X.-S. Wei, Y. Peng, and S. Belongie, *Benchmarking large vision-language models on fine-grained image tasks: A comprehensive evaluation*, 2025. arXiv: `2504.14988 [cs.CV]`. [Online]. Available: `https://arxiv.org/abs/2504.14988`.

[26] B. Lamm and J. Keuper, *A visual rag pipeline for few-shot fine-grained product classification*, 2025. arXiv: `2504.11838 [cs.CV]`. [Online]. Available: `https://arxiv.org/abs/2504.11838`.

[27] M. Oquab et al., *Dinov2: Learning robust visual features without supervision*, 2023.

[28] V. N. Nguyen, T. Groueix, G. Ponimatkin, V. Lepetit, and T. Hodan, "Cnos: A strong baseline for cad-based novel object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2134–2140.

[29] V. N. Nguyen, T. Groueix, M. Salzmann, and V. Lepetit, "Gigapose: Fast and robust novel object pose estimation via one correspondence," *arXiv preprint arXiv:2311.14155*, 2023.

[30] A. X. Chang et al., "ShapeNet: An Information-Rich 3D Model Repository," Stanford University — Princeton University — Toyota Technological Institute at Chicago, Tech. Rep. arXiv:1512.03012 [cs.GR], 2015.

[31] BlenderKit, *BlenderKit*, `https://www.blenderkit.com/`, 2023. Accessed: Feb. 27, 2023.

[32] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," 2020.

[33] E. Sie and D. Vasisht, "Rf-annotate: Automatic rf-supervised image annotation of common objects in context," 2022.

[34] X. Chen, H. Zhang, Z. Yu, S. Lewis, and O. C. Jenkins, "Progress-labeller: Visual data stream annotation for training object-centric 3d perception," 2022.

[35] M. Suchi, B. Neuberger, A. Salykov, J.-B. Weibel, T. Patten, and M. Vincze, "3d-dat: 3d-dataset annotation toolkit for robotic vision," 2023.

[36] M. Douze et al., "The faiss library," 2024. arXiv: `2401.08281 [cs.LG]`.

[37] M. Denninger et al., "BlenderProc," *arXiv preprint 1911.01911*, 2019.