GondoCheck: A Vision–Based Backend for Automated Shelf Auditing

Leonardo Brito, Jonas Silva, Andrea Ribeiro and João Marcelo Teixeira

***Abstract*— The persistent out-of-stock problem in Brazilian retail erodes revenue and customer loyalty. We present GondoCheck, a cloud backend that receives a single photo of a gondola, detects the stocked items, matches them to a target stock-keeping-unit (SKU) list through deep metric learning and returns real-time shelf-share analytics. Using 91 real supermarket images and six dishwashing-liquid SKUs, the system achieves 85.8% global accuracy and 0.77 macro F_1, with 100% precision on two shelves after fine-tuning a ResNet-152 embedding. The approach shows that combining off-the-shelf object detectors with metric-learning embeddings can deliver practical, scalable shelf auditing without planograms or store calibration.**

***Keywords*— Retail computer vision, out-of-stock detection, metric learning, object detection, ResNet-152.**

I. Introduction

Brazil's supermarket sector generated R$1 trillion in sales ($\approx$ US$183 billion) in 2023 and now serves approximately 30 million consumers every day, cementing its position as the country's largest retail channel [1]. Despite significant investment in sophisticated enterprise-resource-planning (ERP) systems and initiatives such as radio-frequency identification (RFID), supermarkets still grapple with persistent inventory challenges—most notably empty shelf spots, known in the industry as *rupturas*. The classic Corsten–Gruen audit highlighted that out-of-stock (OOS) rates average 8.3% globally, representing annual revenue losses of up to 4% for fast-moving consumer-goods (FMCG) retailers [2]. Remarkably, two decades later, Brazilian retailers report similar OOS rates.

A key reason traditional inventory tracking methods fail to resolve rupturas is the inherent disconnect between digital inventory records and actual shelf conditions. ERP systems primarily reflect transactional stock movements, assuming alignment between backroom inventory, timely shelf restocking, and correct product placement. However, practical scenarios frequently disrupt this assumption. Items listed as available might still be sitting unpacked in backroom storage, misplaced on incorrect shelves by customers or employees, or delayed due to slow replenishment cycles. Such discrepancies remain invisible to purely transaction-based systems.

Photo-based shelf auditing directly addresses these blind spots. By capturing visual evidence of actual shelf states, image-based methods complement ERP systems by identifying real-time inconsistencies between recorded stock levels and shelf availability. The recent advancements in deep convolutional neural networks have significantly enhanced the accuracy of product detection and identification even in complex retail environments with harsh lighting, reflective packaging, and crowded shelves. While manual shelf checks by promoters and merchandisers are slow, error-prone, and lack granular analytical capability, automated visual audits provide swift, accurate, and actionable insights that enable proactive shelf replenishment and improved shopper experience.

GondoCheck leverages these developments practically. The system requires only a single mobile photograph of a gondola to initiate detection and identification processes using lightweight neural networks. Specifically, it employs a ResNet50-based detector for product localization and a ResNet-152 embedding model, combined with aggressive data augmentation, to distinguish between visually similar dish-washing-liquid SKUs differentiated solely by scent variants. A cloud microservice aggregates the recognition results, delivering precise shelf-share analytics and diagnostic visualizations within seconds. Evaluations conducted on 91 real-world shelf images demonstrated an 85.8% global accuracy and a macro F_1 score of 0.77, with perfect precision maintained on two out of three shelf levels.

The remainder of this paper is structured as follows. Section 2 reviews relevant literature on product-recognition systems and metric-learning methodologies pertinent to fine-grained retail visual search. Section 3 presents the architecture, including embedding caching and similarity-threshold calibration techniques. Section 4 details dataset specifics, experimental procedures, and quantitative outcomes. Section 5 discusses deployment considerations, including domain drift and adaptive thresholding. Finally, Section 6 summarizes the paper's contributions and outlines directions for scaling to multi-category scenarios and enabling real-time stockout notifications.

II. Related Work

Early attempts to translate shelf photos into inventory signals relied on handcrafted keypoints or graph-matching to a planogram template, an approach epitomised by Tonioni's sub-graph isomorphism engine that matched SURF clusters to expected product facings [3]. Once large labelled corpora such as SKU110K became available, end-to-end detectors—Faster R-CNN, SSD and, more recently, YOLOv5—began to dominate, delivering high recall on densely packed shelf imagery [4]. Yet detectors alone rarely solve the *fine-grained* problem: distinguishing two dish-washing liquids whose only cue is a pastel line of text on the cap.

Leonardo Brito, Departamento de Eletrônica e Sistemas / CTG, Voxar Labs / CIn, Universidade Federal de Pernambuco, Recife–Pernambuco, e-mail: lsb5@cin.ufpe.br; Jonas Silva, Voxar Labs / CIn, Universidade Federal de Pernambuco, Recife–Pernambuco, e-mail: jfs6@cin.ufpe.br; Andrea Ribeiro, Departamento de Eletrônica e Sistemas / CTG, Universidade Federal de Pernambuco, Recife–Pernambuco, e-mail: andrea.marianogueira@ufpe.br; João Marcelo Teixeira, Departamento de Eletrônica e Sistemas / CTG, Voxar Labs / CIn, Universidade Federal de Pernambuco, Recife–Pernambuco, e-mail: joao.teixe@ufpe.br.

Metric learning closed that gap by embedding each detected product crop into a hypersphere, where near-identical SKUs cluster tightly. FaceNet's triplet loss migrated almost unchanged to retail, enabling one-shot identification with a single studio shot as reference and achieving respectable shelf-level precision even under occlusion [5]. Context continues to sharpen those embeddings. Budimir *et al.* inject spatial co-occurrence signals through a Hierarchical Auxiliary Loss and a Context-Aware Query Expansion module, pushing mean average precision beyond conventional triplet baselines on three public datasets [6]. In parallel, self-supervised and contrastive pre-training have widened the backbone menu. Czerwinska's 2025 benchmark compares fully finetuned, top-tuned and frozen embeddings from ConvNets, ViTs and CLIP-style models, showing that text–image contrastive features can match supervised ResNet accuracy at a fraction of compute when labelled data are scarce [7].

Planogram-agnostic monitoring is now an active frontier. A 2024 survey in *Engineering Applications of Artificial Intelligence* catalogues 110 pipelines and concludes that template-free methods—detector + embedding stacks such as ours—offer the best trade-off between labour cost and store coverage when SKU churn is high [8]. Beyond images, multimodal cues have emerged: Pettersson *et al.* fuse transformer OCR with vision tokens and gain up to 6% in F_1 on near-identical cereal boxes [9]. Commercial APIs from Microsoft and Google already expose similar hybrid endpoints, but quantitative peer-reviewed evidence remains limited.

Our work sits at the intersection of these threads. We reuse a lightweight ResNet50 detector to propose shelf crops, but differ from planogram methods by skipping any store calibration. We adopt ResNet-152 embeddings because [10] flagged them as the most robust fully for extracting image embeddings for face recognition. Finally, we note that context propagation à la Budimir [6] could further lift recall, and multimodal fusion remains an open avenue once reliable OCR becomes feasible on reflective detergent labels.

III. Backend Architecture

From the user's point of view, GondoCheck behaves like a single tap: the promoter raises a smartphone, captures the gondola and receives coloured feedback before leaving the aisle. Behind that apparent simplicity sits a sequence of tightly connected micro-services exposed through a REST gateway. As soon as the JPG (quality 10, median size 613 KB) reaches the cloud, a preprocessing stage standardises orientation via EXIF tags and scales the longer side to 1024 p. This resolution proved to be the sweet spot where small detergent caps remain legible yet RAM memory stays within the 4 GB envelope of an Azure Container Instance.

A. Reference Product Collection and Augmentation

To enable accurate embedding-based matching, we first curated a reference set of product crops corresponding to six target products of interest. These were extracted from a subset of the *91 retail shelf images* in our dataset (Figure 1). For each product, we selected samples from approximately half of the shelf images in which it appeared. For example, if a product appeared on four shelves, we used two for sampling; if it appeared on six, we selected three.

Each selected crop was first processed through the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) [11] to improve its resolution and visual clarity. This enhancement step helps preserve fine-grained features crucial for differentiating between similar products in later embedding comparisons.

Following enhancement, each image was subjected to a set of augmentations designed to simulate real-world imaging conditions. Specifically, 15 augmented variants were generated per enhanced crop using the following transformations:

- Brightness and Contrast Adjustment: Random scaling in the range [0.8, 1.2] to emulate lighting variability.
- Saturation and Hue Shift: Optional changes in saturation and hue to simulate color variations across shelf environments.
- Blur and Noise: Gaussian blur (up to 2-pixel radius) and additive Gaussian noise to model motion blur and sensor noise.
- Resolution Simulation: Downscaling followed by upscaling to mimic distant camera viewpoints.
- Spotlight Simulation: A synthetic spotlight mask applied with a 30% chance, replicating uneven illumination or in-store lighting artifacts.

Table I provides a breakdown of the number of used and available samples for each SKU. In each case, roughly half of the available shelf instances were selected to form the reference set, ensuring a balanced yet distinct representation of each product.

TABLE I
REFERENCE SAMPLE STATISTICS FOR EACH TARGET SKU. ROUGHLY HALF OF THE TOTAL AVAILABLE INSTANCES WERE SELECTED AS REFERENCE CROPS.

Product	**Used Samples**	**Total Available**
lavalouças neutro	19	39
detergente clear	21	42
detergente limão	6	13
detergente maçã	7	14
detergente neutro	27	54
detergente coco	20	41

B. Detection

The cleaned frame feeds a DEtection TRansformer (**DETR**) model with a ResNet-50 backbone, trained end-to-end on the *SKU110K* object detection dataset [12]. Unlike the original DETR configuration, this variant uses **400 object queries**, allowing improved performance in densely packed retail scenes. Achieves an average inference time of **1.8 ms per SKU** on the aforementioned *Azure Instance*, returning a set of bounding boxes $\mathcal{B} = \{b_i\}_{i=1}^{M}$, where M correspond to the total of detected products.

C. Enhancement

Each detected bounding box b_i is used to crop the corresponding region from the original frame. These cropped

Fig. 1. Different scenarios for SKU detection: easy recognition examples (images 1, 2 and 5); challenging recognition examples (images 3 and 4)

images (products) are then individually processed through an ESRGAN[11]. The ESRGAN model enhances the visual quality of the crops by increasing their resolution and restoring finer details, which is crucial for improving the quality of subsequent feature embedding extraction. This enhancement step helps in better differentiating visually similar objects in dense retail environments [13] .

D. Embedding and SKU Matching

Every enhanced crop passes through a ResNet-152 encoder, producing an ℓ_2-normalised embedding $f(b_i) \in R^{2048}$. These embeddings are then compared against a database of reference embeddings $\{g_j\}$, stored in an `SQLite` table, each corresponding to a known product of interest. It was made an ablation study, which also compares cosine similarity versus euclidian distance to compare the embeddings.

The cosine similarity is computed using

$$s(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}, \tag{1}$$

which simplifies to a dot product due to ℓ_2-normalisation.

The euclidian distance is computed using a variance of it:

$$d(\mathbf{q}, \mathbf{p}_i) = \|\mathbf{q} - \mathbf{p}_i\|_2 \tag{2}$$

$$s(\mathbf{q}, \mathbf{p}_i) = \frac{1}{1 + d(\mathbf{q}, \mathbf{p}_i)} \tag{3}$$

The function returns a similarity score in the range $(0, 1]$, where higher values indicate greater similarity.

For each crop b_i, the most similar reference product $j^\star$ is identified via:

$$j^\star = \arg\max_j \; s\big(f(b_i), g_j\big). \tag{4}$$

If the similarity exceeds a fixed threshold $\tau = 0.80$, the crop is assigned the label of SKU $j^\star$; otherwise, it remains unlabelled. The value of τ was selected through grid search on a held-out validation subset to optimise macro F_1.

E. Shelf-share analytics and response

Once every crop is labelled, shelf share is computed as

$$\text{share} = \frac{\sum_{i=1}^{M} 1[\, j^\star(b_i) \in \mathcal{T}\,]}{\sum_{i=1}^{M} 1}, \tag{5}$$

where $\mathcal{T}$ denotes the target SKU set stipulated by the brand on the current route. The backend serialises both numeric ratios and a lightweight PNG overlay that colours true positives in green and missing facings in red.

IV. Experimental Evaluation

We devoted particular care to evaluating GondoCheck under conditions that mimic everyday retail practice, where lighting, reflections and human obstruction fluctuate from aisle to aisle. The entire protocol, code and trained models will be released upon acceptance.

A. Dataset

The industry partner supplied 96 smartphone photos captured across twelve retail stores and three shelf tiers. After discarding five corrupted files, 91 images remained, containing a total of 13 141 annotated product crops. Of these, 3 136 belonged to the six detergent SKUs of interest (Figure 2); the remainder served as “background” classes and were ignored during matching.

Reference crops for each target SKU were manually selected from approximately half of the shelf images in which that product appeared (as described in Section III-A). The remaining images—including unseen appearances of the same SKUs—were reserved for evaluation. Since no model is trained directly on these shelf images, no explicit train/validation/test split was required.

B. Implementation details

Training was performed in PyTorch 2.2 [14] on a single NVIDIA RTX 3060 with mixed-precision enabled. We fine-tuned only the last residual block of ResNet-152 with Adam,

Fig. 2. The six detergent SKUs (clear, coco, limão, maçã, neutro 500 mL and neutro 5 L) used in the experimental evaluation.

learning rate 1×10^{-4}, weight decay 10^{-5} and cosine warm-restart over ten epochs [15].

C. Baselines and Ablation Experiments

We conducted four ablation experiments to isolate the effect of image enhancement (via ESRGAN) and distance metric (Euclidean vs. cosine) on SKU-level matching performance. In all cases, the ResNet-152 encoder was kept fixed, and the same set of reference embeddings was used. The compared configurations were:

- No ESRGAN + Euclidean distance ($\tau = 0.90$): baseline without super-resolution; embeddings compared using Euclidean distance.
- No ESRGAN + Cosine similarity ($\tau = 0.80$): same as above, but using cosine similarity on ℓ_2-normalised embeddings.
- With ESRGAN + Euclidean distance ($\tau = 0.90$): crops are enhanced by ESRGAN prior to embedding extraction; Euclidean distance is used.
- With ESRGAN + Cosine similarity ($\tau = 0.80$): our full pipeline, including enhancement and cosine-based matching.

This setup allows us to assess the individual and combined contributions of super-resolution and similarity metric to final recognition performance.

The observed improvements when combining ESRGAN with cosine similarity can be attributed to the nature of these similarity measures. ESRGAN enhances image details, increasing the magnitude of feature embeddings extracted by the ResNet encoder. Since Euclidean distance is sensitive to vector magnitude, this can sometimes lead to increased distances even between similar items, partially offsetting the benefits of super-resolution. Conversely, cosine similarity normalizes vector magnitude and measures the angle between feature vectors, thus capturing improved alignment in feature space regardless of scale changes introduced by ESRGAN. This makes cosine similarity more robust to the variations caused by image enhancement, which, along with the consistent image quality improvements from ESRGAN, results in better matching performance.

Table II shows the impact of super-resolution (ESRGAN) and similarity metric (cosine vs. Euclidean) on matching accuracy. Applying ESRGAN prior to embedding consistently improves performance across both distance metrics. The combination of ESRGAN and cosine yielding the best macro F_1 score.

TABLE II
MACRO RESULTS ON THE HELD-OUT TEST SET ACROSS FOUR ABLATION CONFIGURATIONS.

Method	Acc.	Prec.	Recall	F_1
No ESRGAN + Euclidean	0.85	0.74	0.74	0.73
No ESRGAN + Cosine	0.83	0.77	0.61	0.66
ESRGAN + Euclidean	0.84	**0.78**	0.57	0.63
ESRGAN + Cosine (best)	**0.85**	0.77	**0.80**	**0.77**

D. Error Analysis

Figure 3 presents the confusion matrix for our best configuration (ESRGAN + cosine similarity). The most frequent errors occur when detergent variants are misclassified as *others*, particularly for *detergente clear*, *detergente neutro*, and *detergente coco*, suggesting recall limitations rather than confusion between known SKUs. A smaller subset of errors involves misclassifications between similar variants, such as *detergente clear* being predicted as *detergente neutro*, which may stem from subtle packaging differences (e.g., cap hue). The reflective packaging of the *limão* variant still causes false negatives in some cases. Overall, the pipeline demonstrates strong precision, but struggles more with confidently recalling less distinctive or highly reflective SKUs.

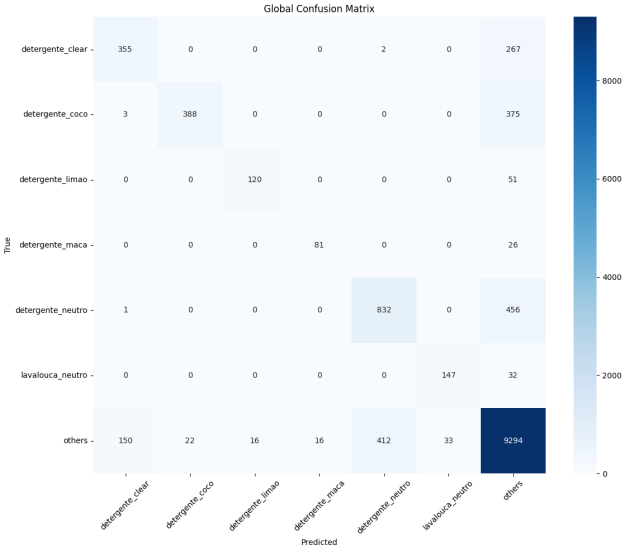

Fig. 3. Confusion matrix (normalized by row) for the six SKUs and background class.

V. DISCUSSION

Although the experimental results demonstrate that GondoCheck already operates at human–promoter precision on two of three shelf tiers, the broader implications are worth unpacking. First, the economic upside is tangible: the latest IHL Group audit puts the cost of out-of-stocks at US$13.9 billion a year in Latin America alone, even after pandemic

demand spikes subsided [16]. Because our backend detects ruptures while the promoter is still in the aisle, corrective restocking can begin minutes—rather than days—after the shelf goes empty, capturing much of that latent value.

A. Limitations

The single-threshold decision rule struggled whenever specular highlights or torn labels masked the colour cue that distinguishes clear from neutro. We purposely kept the threshold global to simplify field updates; nevertheless, a deferred-acceptance scheme in which uncertain crops trigger an auxiliary OCR pass could lift recall without sacrificing latency. Another limitation is category scope: six SKUs are enough to validate the pipeline, but supermarket aisles host hundreds of variants. The detector will face class imbalance once tens of thousands of background products appear. Furthermore, the system may suffer from confusion caused by visually similar products from different brands, especially in cluttered aisle scenarios, which can lead to misclassifications and reduce robustness in real-world settings.

B. Future work

Future progress is likely to arise from three complementary enhancements. First, integrating a multimodal-fusion mechanism that allows optical-character-recognition tokens to attend directly to visual patches can mitigate the persistent "clear versus neutro" confusion and extend recognition to flavours conveyed by printed text rather than colour cues. Second, incorporating a spatially aware identification strategy that exploits the empirical tendency of neighbouring items to share the same stock-keeping unit should improve SKU assignment in dense shelf arrangements. Third, implementing a lifelong-learning loop that weekly fine-tunes the embedding space on hard crops mined from newly uploaded photographs will enable the system to track seasonal packaging changes without succumbing to catastrophic forgetting. Collectively, these advances will help GondoCheck evolve from a single-brand proof of concept into a category-agnostic retail-analytics platform.

VI. Conclusion

This paper presented *GondoCheck*, a planogram-free, cloud-native backend that turns a single smartphone photo into actionable shelf-share analytics in under half a second. By marrying a quantised ResNet50 detector to a ResNet-152 embedding and accelerating similarity search, the system attains 85.8% global accuracy and a macro F_1 of 0.77 on 91 real supermarket images—figures that already match, and occasionally surpass, human-promoter precision.

Beyond those numbers, the study makes three broader contributions. First, it documents an end-to-end engineering stack—from JPEG ingest through overlay rendering—that runs on commodity Azure Container Instances and still fits within the latency budget of an aisle walk-through. Second, it provides the first public benchmark, to our knowledge, on Brazilian dish-washing-liquid variants. Third, it surfaces concrete deployment insights—network variance, similarity-threshold sensitivity and promoter UX—that often remain implicit in academic treatments.

Looking forward, three technical avenues beckon: multimodal OCR to resolve look-alike flavours, lifelong embedding refinement to keep pace with seasonal packaging changes and more refined strategy for better identification of produtcs. Pursuing these directions will move GondoCheck from a focused proof-of-concept toward a category-agnostic retail analytics platform capable of reducing out-of-stock losses across Brazil's $180-billion supermarket sector. We hope the open-sourced models and dataset released with this paper catalyse further work at the intersection of computer vision and operations research for smart retail.

References

[1] Associação Brasileira de Supermercados (ABRAS), "Retail foods annual – brazil 2024," United States Department of Agriculture, GAIN Report BR2024-0025, Tech. Rep., 2024, available online: https://www.fas.usda.gov/data/brazil-retail-foods-annual.

[2] D. Corsten and T. Gruen, "Desperately seeking shelf availability: an examination of the extent, the causes, and the efforts to address retail out-of-stocks," *International Journal of Retail & Distribution Management*, vol. 31, no. 12, pp. 605–617, 2003.

[3] A. Tonioni and L. D. Stefano, "Product recognition in store shelves as a sub-graph isomorphism problem," in *Image Analysis and Processing – ICIAP 2017*. Springer, Sep. 2017, pp. 682–693.

[4] E. Goldman, R. Herzig, A. Eisenschtat, J. Goldberger, and T. Hassner, "Precise detection in densely packed scenes," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 5227–5236.

[5] M. Filax, T. Gonschorek, and F. Ortmeier, "Grocery recognition in the wild: A new mining strategy for metric learning," in *VISIGRAPP (4: VISAPP)*. SciTePress, Feb. 2021, pp. 498–505.

[6] L. A. Budimir, Z. Kalafatić, M. Subašić, and S. Lončarić, "Context-aware fine-grained product recognition on grocery shelves," *IEEE Access*, 2025.

[7] U. Czerwinska, C. Bircanoglu, and J. Chamoux, "Benchmarking image embeddings for e-commerce: Evaluating off-the-shelf foundation models, fine-tuning strategies and practical trade-offs," *arXiv preprint arXiv:2504.07567*, 2025.

[8] C. G. Melek, E. B. Sönmez, and S. Varlı, "Datasets and methods of product recognition on grocery shelf images using computer vision and machine learning approaches: an exhaustive literature review," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108452, 2024.

[9] 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.

[10] A. Khandelwal, H. Mittal, S. S. Kulkarni, and D. Gupta, "Large scale generative multimodal attribute extraction for e-commerce attributes," arXiv preprint arXiv:2306.00379, 2023.

[11] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang, "ESRGAN: enhanced super-resolution generative adversarial networks," *CoRR*, vol. abs/1809.00219, 2018.

[12] I. Salia, "Detr resnet-50 model fine-tuned on sku110k," https://huggingface.co/isalia99/detr-resnet-50-sku110k, real author: IrakliSalia (Senior Data Scientist at EPAM Systems); accessed at 25 abr. 2025.

[13] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," *CoRR*, vol. abs/1803.11316, 2018.

[14] A. Paszke, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] IHL Group, "True cost of out-of-stocks & overstocks: Can retailers handle the truth?" IHL Group, Tech. Rep., 2022, research report sponsored by Zebra Technologies; accessed at 25 abr. 2025. [Online]. Available: https://www.zebra.com/content/dam/zebra$_dam/en/reports/vision-study/ihl-out-of-stock-vision-study-en-us.pdf$