

Cascaded detection and classification for annotation-efficient fine-grained brand recognition in video

Rafael Colen de Almeida^{1*} , Rogerio de Oliveira² 

¹ Mackenzie Presbyterian University, Rua da Consolação, 930, 01302-907 São Paulo, SP, Brazil

² Instituto Mauá de Tecnologia, Praça Mauá, 1, 09580-900 São Caetano do Sul, SP, Brazil

* Corresponding author's e-mail: rafaeljbi@gmail.com

ABSTRACT

Fine-grained brand recognition in unstructured video faces high bounding box annotation costs and vulnerability to false positives. This article proposes and evaluates an annotation-efficient cascade architecture for Fine-Grained Visual Categorization (FGVC) in video. The approach couples a generalist detector (YOLO26) for product super-class localization with a specialist classifier (YOLO26-cls) for brand identification over the crops produced by the detector. We compared this cascade model against a single-stage detector under frame-by-frame validation, using 7.070 smartwatch images (Apple, Garmin, and Samsung) from 88 YouTube videos and five superclasses from the public Video E-commerce Retrieval Dataset (VERD). The cascade architecture reduced fine-grained dataset preparation time by approximately 22 times by eliminating manual bounding box annotation. While robustness to false positives proved to be a brand-dependent trade-off, the cascade maintained near-real-time processing throughput. Ultimately, the method's annotation efficiency and capacity for incremental expansion support its viability for video-based retail intelligence applications.

Keywords: computer vision, object detection, fine-grained visual categorization, video analytics.

INTRODUCTION

Video consumption on social media and e-commerce platforms has generated a substantial volume of unstructured visual data. For retail, this material is a source for competitive intelligence metrics such as Share of Voice (SOV), which quantifies the relative visibility of a brand against its competitors in a channel [1]. Operationalizing this content faces the difficulty of identifying specific products in real-world (in-the-wild) scenarios, in which occlusion and visual quality vary markedly [2].

The detection of generic object categories is a well-established task in the literature [3, 4]. By contrast, fine-grained visual categorization (FGVC), which distinguishes visually similar subcategories such as smartwatch models from distinct brands (Apple, Garmin, and Samsung), is a problem of another order [5, 6]. FGVC objects

exhibit high intra-class variance and low inter-class variance, which makes discrimination sensitive to the quality of the learned visual features [6]. In uncontrolled video (in-the-wild, non-professional videos) the problem is aggravated by partial occlusion, motion blur, lighting variation, and extreme camera angles. Smartwatches combine these difficulties in a case of direct commercial interest. The Apple, Garmin, and Samsung brands share a similar dial and strap form factor, which accentuates the low inter-class variance, and they appear frequently in promotional videos and e-commerce reviews, channels in which Share of Voice is the target metric.

The FGVC literature consolidated around benchmarks such as CUB-200-2011 [7], Stanford Cars [8], and FGVC-Aircraft [9], built under controlled conditions, with centered objects, simple backgrounds, and a balanced class distribution [5]. These benchmarks do not model the prolonged

partial absence of classes, which is the dominant pattern in in-the-wild video [2]. This work does not seek to surpass the state of the art in Top-1 Accuracy on controlled datasets. The objective is to evaluate an architecture oriented to the operational bottlenecks of FGVC in video, namely the annotation cost and robustness to false positives in scenarios with imbalanced classes.

The bottleneck is not only technical. Training a single-stage model capable of detecting and classifying brands simultaneously requires manual bounding box annotation for every new brand added to the system. Studies on annotation cost indicate that manually drawing a bounding box takes between 7 and 42 seconds per instance under optimized protocols [10, 11]. For a fine-grained dataset of thousands of images with multiple instances per frame, this cost makes operational scale infeasible without automation mechanisms.

To address these bottlenecks, the proposed method draws on two established principles, multi-stage architectures [12], which decouple tasks to resolve visual ambiguities more efficiently by specializing each stage, and auto-labeling [13], which leverages pre-trained models to reduce human effort. Based on these concepts, this article proposes and evaluates an annotation-efficient cascade architecture for FGVC in video. The solution couples a generalist YOLO26 detector, trained on superclasses, with a specialist YOLO26-cls classifier that operates over the crops produced by the detector. The operational gain comes from building the fine-grained dataset through auto-labeling: the detector extracts and isolates the superclass crops (the smartwatch, in this study), dispensing with manual bounding box annotation at the brand level. To validate the approach, the cascade model is compared against a single-stage detector, trained to detect and classify brands simultaneously in the conventional manner, under a frame-by-frame validation protocol on held-out test videos.

This work makes three main contributions:

- First, evidence is provided that spatial localization (object detection) is crucial before classification in uncontrolled video. It significantly outperforms global frame classification by isolating the object from background noise and preventing false positives.
- Furthermore, annotation efficiency and scalability are addressed by proposing a cascade architecture. This approach reduces the annotation cost of the fine-grained dataset (by roughly 22 times in the experiments) and enables the incremental

expansion of the system to new brands without requiring geometric re-annotation.

- Finally, the operational trade-offs of this approach are characterized. Through a strict frame-by-frame evaluation on uncontrolled videos, it is demonstrated that the cascade's robustness to false positives is a brand-dependent trade-off tied to visual separability, rather than an absolute advantage over single-stage models.

MATERIALS AND METHODS

Methodological procedures

The experimental methodology was organized into four chained stages, in which each stage produces an artifact that feeds the next. Stage 1 trained the base models, a generalist YOLO26 detector and a generalist YOLO26-cls classifier, on five superclasses (Smartwatch, Mixer, Hair Dryer, Grain Grinder, and Blender) of the Video E-commerce Retrieval Dataset (VERD) [2] and validated the detection hypothesis. Stage 2 built the fine-grained smartwatch brand dataset through auto-labeling with the resulting detector and prepared the geometric annotation of this material for the control model. Stage 3 carried out the fine-tuning of the two fine-grained recognition models that would be compared, the specialist classifier and the single-stage detector for control, both from the same fine-grained dataset. Finally, Stage 4 compared, under inference on uncontrolled videos, the two resulting pipelines, the cascade model and the single-stage detector for control. Figure 1 summarizes the four stages. These stages are further described in the *Pipeline architecture* subsection.

The proposed approach employed YOLO26 [14] as its backbone, the most recent model of the Ultralytics YOLO family at the time of the study. YOLO26 offers a native classifier variant, YOLO26-cls, which shares the same feature-extraction backbone as the detector. This property unifies detection and classification under a single fine-tuning framework, which is directly relevant to the chained construction of the cascade.

Datasets and data preparation

Dataset: VERD

Stage 1 used a sample of the public VERD, composed of product review videos filmed by real users in uncontrolled environments. VERD

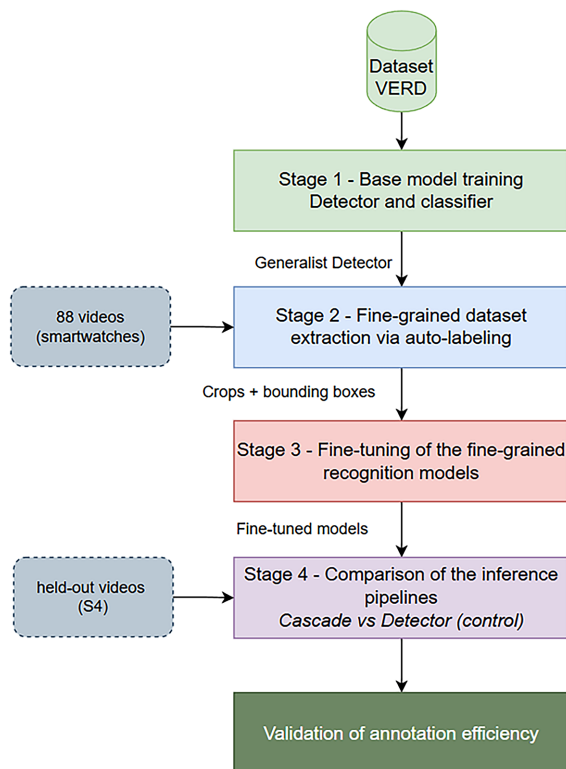


Figure 1. General diagram of the methodology

was selected for its adherence to the in-the-wild scenario, with diversity of backgrounds, camera angles, and lighting conditions that reflect real use on e-commerce platforms. The five superclasses (Smartwatch, Mixer, Hair Dryer, Grain Grinder, and Blender) were annotated in Roboflow, a web-based computer vision annotation platform, with emphasis on the smartwatch class. Figure 2 presents visual examples of the selected classes.

Fine-grained dataset: smartwatches from YouTube

For fine-grained brand recognition, a smartwatch dataset was built from 88 YouTube videos, approximately 30 per brand (Apple Watch, Garmin, and Samsung Galaxy Watch), stratified between promotional videos and reviews from the last five years. The dataset was generated through auto-labeling, applying the generalist detector from Stage 1 in inference over each video and extracting the crops of the smartwatch superclass regions. A subsequent human curation discarded



Figure 2. Examples of the five selected classes from the VERD dataset used to train the first-stage detector

noisy crops, such as severe occlusions, objects partially out of frame, and incorrect detections, without drawing bounding boxes. Table 1 presents the summary of the resulting fine-grained dataset.

Control dataset: annotation for the detector model

For the control detector model, the same set of crops from the fine-grained dataset was re-annotated with explicit bounding boxes in Roboflow, producing a detection dataset with three fine-grained classes. The annotation adopted the guidelines of the MS-COCO protocol [15], with tight bounding boxes limited to the strictly visible pixels of each smartwatch, without inferring areas under occlusion. The cascade pipeline requires only the image label (directory-based classification), whereas the control model requires manual drawing of

bounding boxes. The common origin of the two sets ensures the fairness of the final comparison. Figure 3 juxtaposes the two annotation processes.

Pipeline architecture

Stage 1 – Base model training

A YOLO26 detector and a YOLO26-cls classifier were trained using specialized fine-tuning on five superclasses of VERD. The emphasis fell on the spatial localization of smartwatches, treated as a superclass in the experiments. The objective of this stage was to establish a Region of Interest (ROI) localizer capable of filtering the visual search space for the specialist classifier of the following stage. With the two models in hand, the performance of frame classification was compared with that of object detection in identifying the five

Table 1. Summary of the fine-grained dataset by smartwatch brand

Brand	Number of videos	Number of extracted images	Size of the final dataset
Apple	30	2.264	973 MB
Garmin	30	2.922	1.25 GB
Samsung	28	1.884	896 MB
Total	88	7.070	3.07 GB

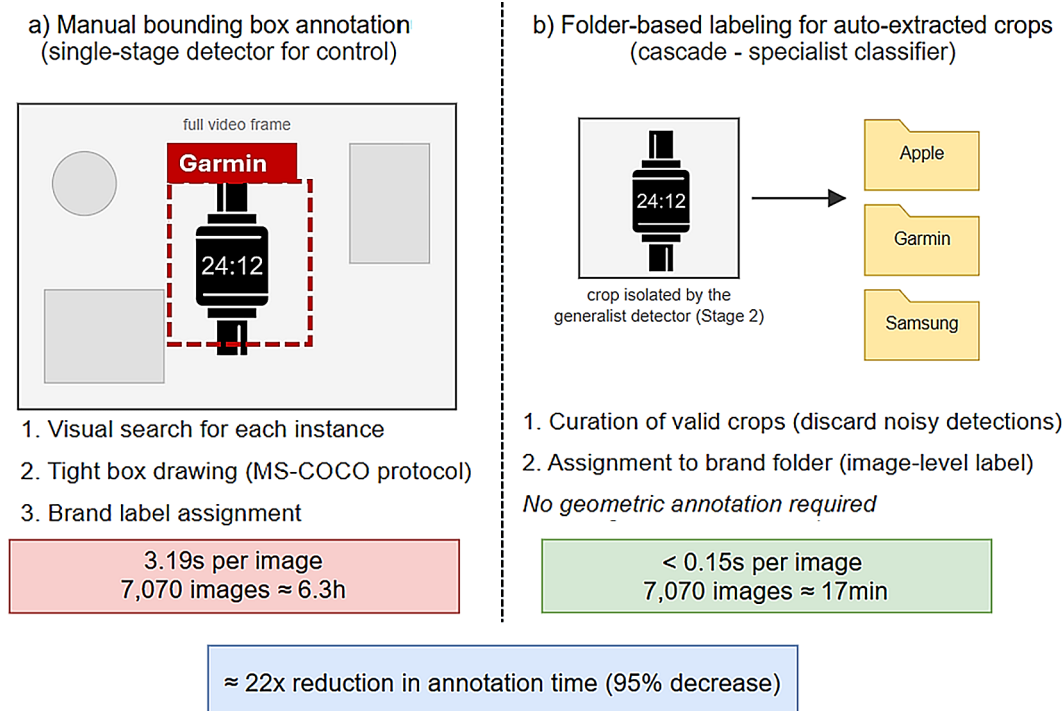


Figure 3. Comparison of the annotation processes for the fine-grained dataset: a) manual bounding box annotation for the single-stage control detector; b) folder-based labeling of crops auto-extracted by the generalist detector for the cascade’s specialist classifier

superclasses, an experiment that grounds the adoption of detection as the basis of the system and whose results are reported in the Results section. Figure 4 shows the detailed flow of the first stage.

Stage 2 – Fine-grained dataset extraction via auto-labeling

The Stage 1 detector was applied in inference mode over the frames of the 88 smartwatch videos. The crops corresponding to the smartwatch superclass were extracted by code and organized into directories by brand. This auto-labeling strategy [13] dispensed with manual bounding box annotation at the fine-grained level, restricting human intervention to the curation of valid crops and the discarding of noisy detections. For the subsequent training stage, this curated set of crops was formatted in two distinct ways: first, it was kept organized by brand folders for the specialist classifier; second, it was re-annotated with bounding boxes in Roboflow to produce the detection dataset for the control model, as shown in Figure 5.

Stage 3 – Fine-tuning of the fine-grained recognition models

Using the datasets prepared in Stage 2, the specialist classifier was initialized with the weights of the generalist classifier from Stage 1, which ensured learning continuity and faster convergence, and had its fine-tuning carried out exclusively on the fine-grained crop dataset with the three brand classes. Operating over regions already isolated by the detector, the classifier concentrates its discriminative capacity on the fine features of each brand, free from the background noise of the scene. The single-stage

detector for control was trained for the simultaneous detection and classification of the three brands, on the same set of crops re-annotated with bounding boxes. This model represents the conventional approach in literature, characterized by the high cost of geometric annotation, and serves as a reference for predictive performance and latency. The flow of Stage 3 is represented in Figure 6.

Stage 4 – Comparison of the inference pipelines

The two resulting pipelines, the cascade approach and the single-stage detector for control, were evaluated under continuous inference on the held-out validation set S4 (three videos, 59,573 frames), described in the *Inference parameters and evaluation metrics* subsection. Running both models on the exact same sequence of frames allowed for a strict cross-referencing of predictive data to compute Precision, Recall, and F1-Score. Figure 7 compares the two inference pipelines.

Model training and hyperparameter configuration

This subsection presents the training details and hyperparameter configurations for the models developed in Stages 1 and 3. Training runs were standardized at 100 epochs with early stopping (patience of 20 epochs). The data augmentation strategies were customized according to the function of each model. For the detector models (Stages 1 and 3), the input resolution was fixed at 640×640 pixels with a batch size of 16. The transformations included color perturbations in

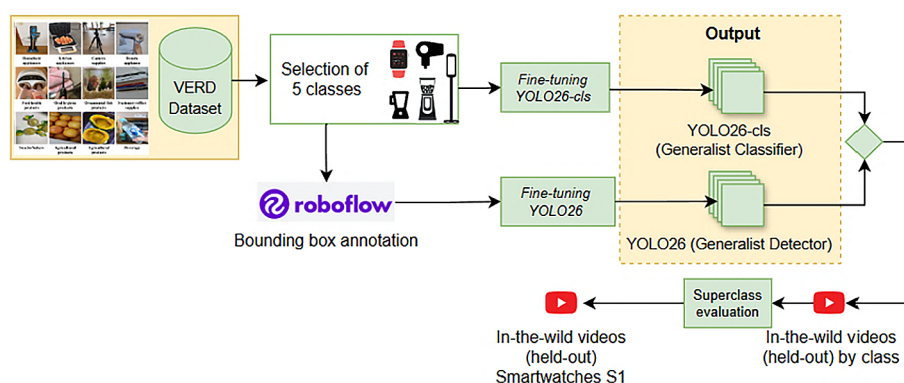


Figure 4. Flow of selection, annotation, and fine-tuning of the generalist detector (YOLO26) and generalist classifier (YOLO26-cls) models

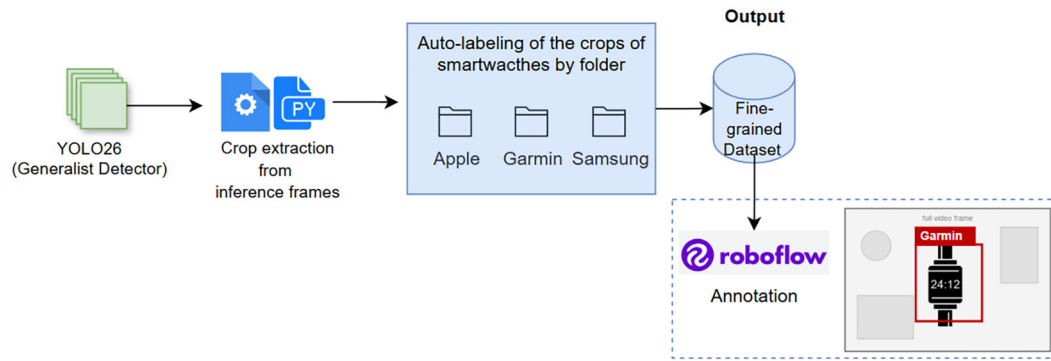


Figure 5. Flow of fine-grained dataset creation via auto-labeling

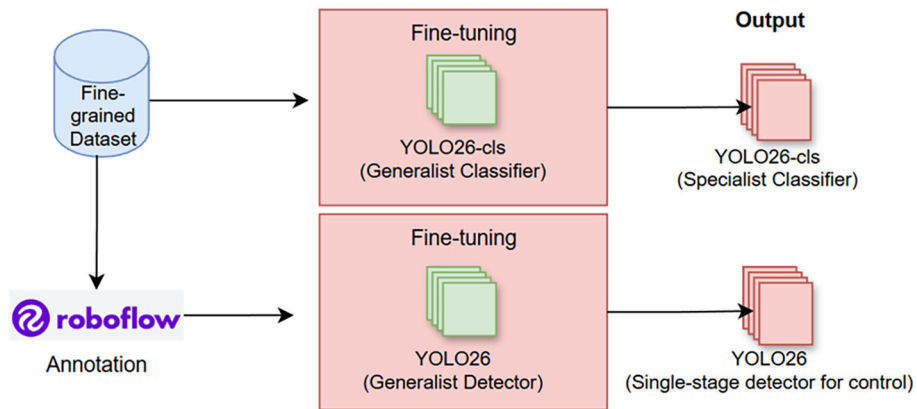


Figure 6. Fine-tuning of the fine-grained recognition models

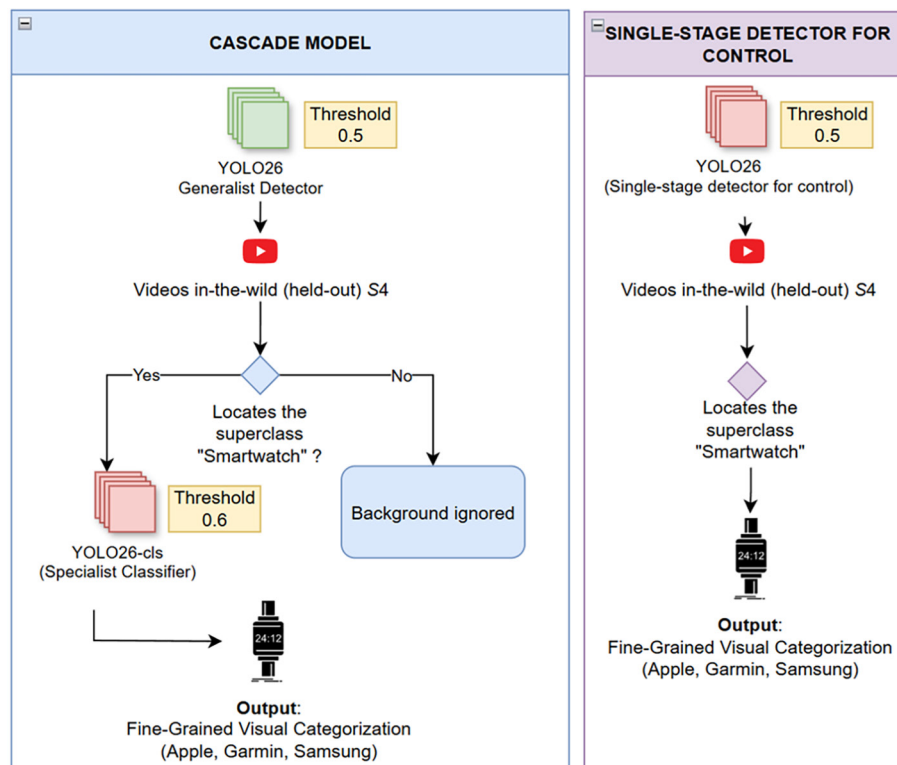


Figure 7. Comparison of the inference pipelines between the cascade approach and the single-stage detector for control

the HSV space, translation, scaling, horizontal flipping, and rotation restricted to $\pm 15^\circ$, in addition to mosaic augmentation, which forces the network to resolve ambiguities in dense visual contexts and variable backgrounds. For the specialist classifier of Stage 3, operating over already isolated crops, the input resolution was set to 320×320 pixels with a batch size of 64. Since the object of study is a wearable device, the spatial augmentation regime was broadened, with a rotation tolerance of $\pm 90^\circ$ and horizontal and vertical flipping, simulating variations of the right, left, raised, and lowered wrist. Spatial scaling remained conservative, so as to preserve the fine-grained features that discriminate between brands, such as physical buttons and watch crowns.

The training runs were executed on different accelerators. The generalist detector of Stage 1 was trained on an NVIDIA T4, whereas the specialist classifier and the control model of Stage 3 were run on an NVIDIA A100. The computational cost analysis is restricted to the comparison between the specialist classifier and the control model (both on the NVIDIA A100), which represents the real marginal cost of expanding the system to new brands.

Inference parameters and evaluation metrics

The evaluation strategy was divided into two approaches. For the fine-tuning validation over approximately balanced sets, the macro average was adopted. For continuous inference over uncontrolled videos, characterized by severe class imbalance, per-class Precision, Recall, and F1-Score were prioritized, alongside absolute counts of false positives (FP) and false negatives (FN). The global aggregation of these videos employed the micro average to assign proportional weight to the frame volume of each video.

The reported metrics follow standard definitions. For each brand class, a prediction is counted as a true positive (TP) when it matches a ground-truth instance of that class in the same frame, as a false positive (FP) when the class is predicted in a frame where it is absent, and as a false negative (FN) when a present instance is not predicted. Precision is $TP/(TP+FP)$, Recall is $TP/(TP+FN)$, and F1-Score is their harmonic mean, $2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall})$. The macro average computes each metric independently per class and then takes the unweighted mean across classes, giving equal weight to

Table 2. Per-class validation metrics – fine-tuning results of the Stage 1 and Stage 3 models

Model	Class	Precision	Recall	mAP@50	mAP@50-95 / Acc
Stage 1 Generalist Detector (YOLO26)	Blender	1.0000	0.9826	0.9950	0.8767
	Grain_grinder	0.9885	1.0000	0.9950	0.8723
	Hair_Dryer	0.9930	0.9725	0.9944	0.8598
	Mixer	0.9927	0.9417	0.9931	0.6657
	Smartwatch	0.9792	0.9869	0.9933	0.7452
	Macro avg	0.9907	0.9767	0.9942	0.8039
Stage 1 Generalist Classifier (YOLO26-clc)	Blender	1.0000	1.0000	–	–
	Grain_grinder	0.9989	1.0000	–	–
	Hair_Dryer	1.0000	1.0000	–	–
	Mixer	1.0000	0.9976	–	–
	Smartwatch	0.9973	0.9973	–	–
	Acc-top1 (all)	–	–	–	99.92%
Stage 3 Specialist Classifier (YOLO26-clc)	Apple	0.9911	0.9888	–	–
	Garmin	0.9795	0.9913	–	–
	Samsung	0.9911	0.9737	–	–
	Acc-top1 (all)	–	–	–	98.61%
Stage 3 Single-stage detector for control (YOLO26)	Apple	0.9780	0.9190	0.9833	0.7584
	Garmin	0.9796	0.9834	0.9943	0.7655
	Samsung	0.9271	0.9450	0.9679	0.6826
	Macro avg	0.9616	0.9491	0.9818	0.7355

every brand regardless of its frequency, whereas the micro average pools TP, FP, and FN across all classes before computing the metrics, so that each frame-level decision contributes equally and frequent classes dominate the aggregate.

The evaluation used two independent sets of in-the-wild videos, ensuring no overlapping frames with the training data. Set S1 (three videos, 26,812 frames), validated the generalist detector trained in Stage 1. Set S4 (used in the Stage 4 comparison, three longer videos, 59,573 frames) served as a continuous inference stress test for comparing the cascade and control models. The S4 videos were specifically selected to model prolonged partial class absence: one video contains all three brands simultaneously, whereas the other two completely lack either Samsung or Garmin.

For the final comparison on the S4 set, asymmetric confidence thresholds were defined to reflect each component's function. The single-stage control model used a standard detection threshold of 0.50. In the cascade, the first-stage generalist detector used 0.50 to favor Recall in capturing the Region of Interest, while the second-stage specialist classifier adopted a stricter 0.60 threshold to favor Precision in defining the brand. This asymmetry enforces the decoupling principle, with the first stage acting as a broad spatial filter and the second as a strict categorical decision-maker.

RESULTS

This section presents the experimental results structured around the three core contributions of this study. First, it empirically validates the spatial decoupling hypothesis by comparing object detection against frame classification. Second, it demonstrates the annotation efficiency and

computational scalability of the cascade approach. Finally, it characterizes the operational trade-offs between the cascade and the control model regarding predictive performance and latency in continuous video inference.

Empirical validation of spatial decoupling (Stage 1)

Before evaluating the continuous inference pipelines, initial tests on balanced validation sets confirmed the convergence of all individual models. Notably, the Stage 1 generalist detector achieved an mAP@50 of 0.9933 for the smart-watch superclass, establishing a solid foundation for the auto-labeling process and the cascade's first stage, while the Stage 3 specialist classifier reached a Top-1 Accuracy of 98.61% (Table 2).

The decisive validation of the spatial decoupling hypothesis, which compares object detection against global frame classification, was conducted on the S1 set, comprising three in-the-wild videos (26,812 frames). As shown in Table 3, the micro-average results reveal a stark contrast in robustness to background noise. The generalist detector maintained an F1-Score of 0.7934 with 1.632 false positives (FP). In contrast, the frame classifier achieved an F1-Score of only 0.5687, generating 12,544 FPs, a difference of nearly an order of magnitude. Per-video results show that the frame classifier can outperform the detector in short videos with near-continuous object presence (Videos 1 and 3). The detector's advantage emerges precisely under prolonged class absence, the dominant condition in Video 2 and in the target application scenario.

This discrepancy is driven by the classifier's context bias, which associates surrounding scene elements with the presence of the product, forcing positive predictions even during prolonged

Table 3. Evaluation metrics of the models trained in Stage 1 with the VERD dataset

Stage 1 metrics	Video 1-S1		Video 2-S1		Video 3-S1		Micro average	
	Detector	Classifier	Detector	Classifier	Detector	Classifier	Detector	Classifier
Accuracy	0.3333	0.8782	0.8666	0.4471	0.3206	0.7735	0.8365	0.4671
Precision	0.9806	0.9136	0.8365	0.4007	0.8213	0.8707	0.8376	0.4289
Recall	0.2457	0.9513	0.8186	0.8379	0.2506	0.8596	0.7536	0.8437
F1-Score	0.3963	0.9321	0.8274	0.5422	0.3840	0.8651	0.7934	0.5687
FP	2	37	1.583	12.397	47	110	1.632	12.544
FN	310	20	1.795	1.604	646	121	2.751	1.745
Total frames	468		25,324		1.020		Sum of frames: 26,812	

periods of class absence (as observed in the longest video, where the classifier alone accumulated 12,397 FPs). By conditioning the prediction on the explicit localization of an instance, object detection isolates the product from background noise. This empirical evidence shows that detection mitigates the failure mode most damaging to the application, justifying its adoption as the system's first stage and as the source of the crops used in the auto-labeling process.

Annotation efficiency and scalability (Stages 2 and 3)

The most direct operational advantage of the cascade architecture lies in the cost of data preparation. An average manual bounding box annotation time of 3.19 seconds per image was measured for the single-stage control model. In contrast, the folder-based labeling of the cascade, which acts upon the crops already isolated by the generalist detector, required less than 0.15 seconds per image. Projected over the 7,070 images of the fine-grained dataset, these rates correspond to approximately 6.3 hours of human effort for the control model versus 17 minutes for the cascade. As detailed in Table 4, this represents a reduction of roughly 22 times (a 95% decrease) in annotation time, achieved by entirely suppressing the geometric step at the fine-grained level.

Beyond the manual effort, the cascade also demonstrated computational efficiency during the fine-tuning phase. Evaluated on the same NVIDIA A100 hardware, the Stage 3 specialist classifier converged in approximately 48 minutes, whereas the control detector required 85 minutes, which corresponds to a 1.8-fold speedup for the cascade. This differential occurs because the classifier operates on pre-isolated crops at a $320 \times$

320 resolution, while the single-stage detector processes complete scenes at 640×640 .

Together, the elimination of geometric annotation and the faster training directly support the scalability of the system. Incorporating a new brand into the cascade requires only the collection of videos, the automated extraction of crops via the first-stage detector, rapid human curation of these folders, and the fine-tuning of the specialist classifier. This incremental expansion model is operationally infeasible for the single-stage approach, which demands a full new round of bounding box drawing and computationally heavier retraining for every new class added to the system.

Operational trade-off characterization (Stage 4)

The comparison between the cascade and control models under continuous video inference (S4 set) reveals that the cascade's robustness to false positives is a brand-dependent trade-off rather than a universal advantage. Evaluated at a fixed operating point (confidence thresholds of 0.50 for the first-stage detector, 0.60 for the specialist classifier, and 0.50 for the control model), neither pipeline dominated across all scenarios (Tables 5, 6, and 7).

In Video 1-S4 (all brands present), the single-stage control model achieved a higher micro F1-Score (0.8376) than the cascade (0.6597). This disparity was largely driven by the Apple class, where the cascade produced 4.388 false positives compared to the control's 83. Frame-by-frame inspection revealed that isolating the crop deprives the specialist classifier of scene context, leading to two main failure modes. The first is the persistence of prediction after total occlusion, which accounted for 76.9% of pure background

Table 4. Comparative annotation cost, literature vs. measurements of this work

Metric / stage	Literature (worst case)	Literature (best case)	Detector for control (this paper)	Cascade (this paper)
Annotation type	Manual bounding box	<i>Extreme clicking</i>	Manual bounding box	Classification by folder
Reported time per image	25.5 to 42.3 s	7.4 s	3.19 s	< 0.15 s
Images in the fine-grained dataset	–	–	7,070	7,070
Estimated time (7,070 imgs)	~83 h	~15 h	~6.3 h (measured)	17 min (measured)
Necessity of bounding boxes	Yes	Yes	Yes	No
Reduction vs the detector for control model (this paper)	–	–	1× (reference)	~22×
Source	[11]	[10]	Measured (Roboflow)	Measured

Table 5. Frame-by-frame evaluation metrics – Video 1-S4 (all classes present)

Metrics	Apple		Garmin		Samsung		Micro average	
	Cascade	Control	Cascade	Control	Cascade	Control	Cascade	Control
Accuracy	0.6499	0.9398	0.9602	0.9405	0.9146	0.9286	0.8415	0.9363
Precision	0.2943	0.9527	0.9463	0.8407	0.9586	0.8602	0.5879	0.8744
Recall	0.7463	0.6823	0.8215	0.8191	0.7081	0.8720	0.7515	0.8037
F1-Score	0.4221	0.7952	0.8795	0.8298	0.8145	0.8661	0.6597	0.8376
FP	4.388	83	118	393	116	537	4.622	1.013
FN	622	779	452	458	1.106	485	2.180	1.722
Total frames	14,343 frames							

Table 6. Frame-by-frame evaluation metrics – Video 2-S4 (Samsung absent)

Metrics	Apple		Garmin		Samsung		Micro average	
	Cascade	Control	Cascade	Control	Cascade	Control	Cascade	Control
Accuracy	0.7396	0.9062	0.9812	0.8821	0.9909	0.9671	0.9039	0.9185
Precision	0.5193	0.9154	0.8649	0.4975	-	-	0.5889	0.6845
Recall	0.8534	0.7300	0.9953	0.6514	-	-	0.8955	0.6954
F1-Score	0.6457	0.8123	0.9255	0.5642	-	-	0.7105	0.6899
FP	3.196	273	265	1.121	133	479	3.594	1.873
FN	593	1.092	8	594	-	-	601	1.686
Total frames	14,754 frames							

Table 7. Frame-by-frame evaluation metrics – Video 3-S4 (Garmin absent)

Metrics	Apple		Garmin		Samsung		Micro average	
	Cascade	Control	Cascade	Control	Cascade	Control	Cascade	Control
Accuracy	0.7219	0.7276	0.7548	0.813	0.9108	0.8971	0.7958	0.8126
Precision	0.1673	0.2995	-	-	0.989	0.9818	0.4622	0.5053
Recall	0.0609	0.1616	-	-	0.7491	0.7144	0.4779	0.4966
F1-Score	0.0893	0.2099	-	-	0.8525	0.8271	0.4699	0.5009
FP	1.892	2.360	6.837	5.214	80	127	8.809	7.701
FN	5.863	5.234	-	-	2.408	2.741	8.271	7.975
Total frames	30,476 frames							

false positives. The second is inter-class confusion, in which degraded crops of Garmin or Samsung were mistakenly classified as Apple. By contrast, the single-stage control resolves these ambiguities using the global context of the scene.

However, this same decoupling mechanism proved advantageous for visually distinct brands. In Video 2-S4 (Samsung absent), the cascade outperformed the control (F1-Score 0.7105 vs. 0.6899), effectively suppressing false positives for the absent Samsung (133 FPs vs. 479 for the control). Conversely, in Video 3-S4 (Garmin absent), the cascade's inter-class confusion resurfaced, generating more false positives

for the absent Garmin (6.837) than the control (5.214). These results demonstrate that the cascade's ability to suppress false positives under class absence depends heavily on the visual separability of the brand.

Finally, this architectural decoupling dictates the system's latency, as summarized in Table 8. While the control model maintained a stable throughput of approximately 51 to 53 FPS regardless of frame content, the cascade's throughput was asymmetric. In empty frames, where the second stage is not triggered, the cascade matched the control at around 50 FPS. In frames with positive detections, the sequential queue of extracting,

Table 8. FPS comparison between the models

Video	Model	Average FPS (Global)	Average FPS (Empty Frames)	Average FPS (With Detection)	Minimum FPS (worst scenario)
Video 1-S4	Single-stage detector for control	50.76	50.86	50.68	24.88
	Cascade	33.00	50.03	26.16	3.59
Video 2-S4	Single-stage detector for control	52.89	52.98	52.76	24.21
	Cascade	38.24	49.96	30.22	13.71
Video 3-S4	Single-stage detector for control	51.28	51.54	50.89	14.31
	Cascade	38.53	50.10	30.61	9.05

resizing, and re-evaluating the crop dropped the throughput to the 26–30 FPS range, reaching lower values under multiple simultaneous detections. Nevertheless, the cascade maintained an average global throughput of 33 to 38 FPS across the videos, remaining fully compatible with near-real-time commercial video analysis.

DISCUSSION

The comparison between object detection and global frame classification indicates that spatial localization is a structural requirement for FGVC in uncontrolled video rather than an implementation choice. The frame classifier learned to associate surrounding scene elements with the presence of the product, a context bias that generated false positives precisely during the prolonged periods of class absence that dominate in-the-wild footage. By conditioning the prediction on the explicit localization of an instance, the detector suppresses this failure mode at its origin. This finding also aligns the system design with the target application, since metrics built on exposure depend on registering absence as reliably as presence. The adoption of YOLO26 as the backbone is further supported by recent architectural benchmarks of next-generation real-time detectors, which confirm its favorable accuracy-latency profile relative to earlier YOLO generations [16].

The elimination of manual bounding box annotation at the brand level is the most direct operational result of the study. Human intervention was restricted to the curation of crops already isolated by the generalist detector, which removes the spatial search and instance verification steps that concentrate most of the cost in full-scene annotation protocols [10]. The gain extends to training, since the specialist classifier operates

over small crops instead of complete scenes and converges faster on equivalent hardware. From the standpoint of scalability, the modularity of the cascade is its main structural advantage. Incorporating a new brand requires only the collection of videos, the application of the existing detector, the curation of the resulting crops, and the fine-tuning of the specialist classifier, while the generalist detector remains unchanged. In the single-stage approach, each new class demands a further round of geometric annotation and the complete retraining of the model, which makes incremental expansion progressively more expensive.

Recent foundation models offer an alternative route to annotation efficiency. Open-vocabulary detectors such as Grounding DINO [17] and real-time variants such as YOLO-World [18] can localize objects from textual prompts without task-specific training and have been increasingly adopted as automatic annotators in dataset construction pipelines. However, these models remain limited in fine-grained discrimination. Recent evaluations show that open-vocabulary detectors struggle to distinguish visually similar subcategories that share a common form factor, which is precisely the regime of brand-level recognition [19]. The cascade proposed here is complementary rather than competing with this trend, since its first stage is architecture-agnostic, it means that an open-vocabulary localizer could replace the fine-tuned YOLO26 detector as the crop source while preserving the specialist classifier and the annotation efficiency property. This positions the cascade as a pragmatic middle ground between fully supervised single-stage pipelines, whose geometric annotation cost grows with each new class, and zero-shot foundation models, whose fine-grained brand discrimination remains an open problem.

The predictive evaluation situates these gains within a clear trade-off. With all classes present, the single-stage control model benefited from correlating the global context of the scene with the fine-grained texture of the brands, a capacity that the cascade, operating over isolated crops, does not reproduce. Under partial class absence the relationship inverted in a brand-dependent manner. The cascade suppressed false positives effectively for the visually distinct Samsung, but remained a target of inter-class confusion for the absent Garmin. Its robustness to false positives should therefore be read as conditioned on the visual separability of each brand rather than as an absolute property. The latency profile follows the same logic. The cascade saves resources in empty frames, in which the second stage is not triggered, and imposes additional operations of crop extraction, resizing, and a second inference in frames with positive detections. Figure 8 shows the cascade model running on a frame of uncontrolled video. The resulting throughput remains compatible with near-real-time processing, which configures a favorable balance against the gains in annotation cost and scalability. This applied orientation is consistent with recent YOLO-based systems reported in this journal, in which detection models are deployed to extract operational metrics from imagery acquired in unstructured environments [20].

Implications for share of voice measurement

The conservative bias of the second stage, which prefers to omit a detection rather than commit a false positive, is well suited to competitive intelligence metrics such as Share of Voice, in

which the false detection of a competing brand corrupts the indicator directly. As a proof of concept, both pipelines operated end to end over a video stream, generating a structured per-frame log with the predicted brand, the confidence, and the box coordinates. Figure 8 shows the cascade model running on a frame of uncontrolled video. This log translates the exposure time of each brand into an auditable record, the primitive from which Share of Voice is computed. A production deployment would still require per-class calibration, with brand-specific thresholds or the reintroduction of some scene context in the second stage.

Limitations and future directions

Several directions can extend this work. The ground truth was produced by a single annotator, and independent review would strengthen inter-annotator agreement, although the frame-by-frame granularity and the volume of more than eighty thousand validated frames mitigate this point. The reported metrics correspond to a single final run per architecture, and multiple runs with independent seeds would add variance estimates to the comparison. The stress test did not include a video with the exclusive absence of Apple, the brand that concentrated the most expressive errors. By analogy with the Garmin-absent case, in which the cascade produced more false positives than the control, and given the tendency of the specialist classifier to default to Apple under degraded crops, an Apple-absent video would be expected to elicit the worst case for the cascade, with Apple precision collapsing and the cascade underperforming the single-stage control. Covering the absence of each

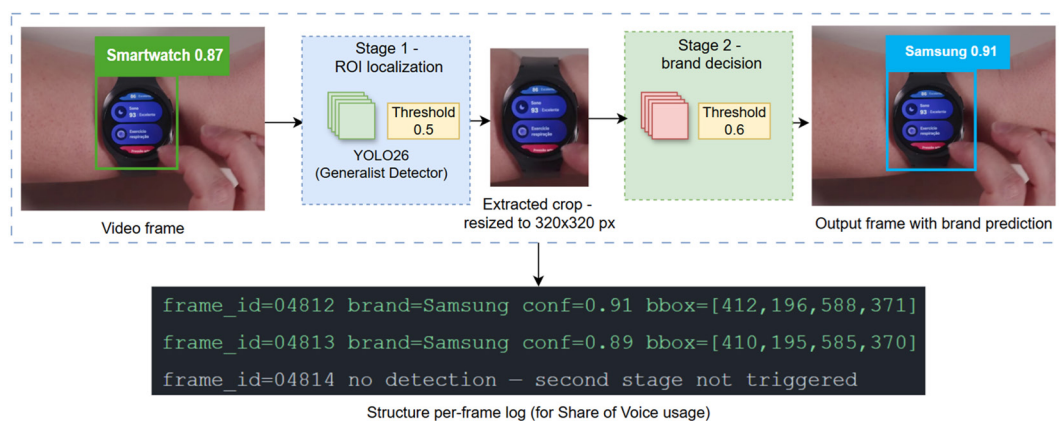


Figure 8. Cascade model in execution

class systematically would complete the characterization of the trade-off. Retaining a small context margin around each crop, rather than the tight bounding box used in this study, is a further low-cost candidate to reduce the inter-class confusion observed for Apple, since recent work on object-centric cropping indicates that some surrounding context aids fine-grained classification while its complete removal can degrade it [21]. Any such margin would itself require calibration, since excessive padding reintroduces the background bias that the cascade was designed to suppress. Finally, per-brand calibration in the second stage and the empirical evaluation of the open-vocabulary localizers discussed above, or of Transformer-based architectures, in the first stage are natural candidates to improve the predictive side of the cascade without sacrificing its annotation efficiency. Because the first stage acts purely as a spatial filter and the reported latency is dominated by the second stage, a re-parameterized real-time detector such as YOLO-World [18] or YOLOE [22] could remove the manual superclass training while preserving the near-real-time throughput, whereas a Transformer-based localizer such as Grounding DINO [17] would trade latency for zero-shot flexibility. Training the first stage nonetheless remains valuable for superclasses that are rarer or defined by subtle shape differences that a text prompt cannot easily express.

CONCLUSIONS

This study proposed and evaluated an annotation-efficient cascade architecture for fine-grained brand recognition in uncontrolled video. The main findings can be summarized as follows:

- Spatial localization is a structural requirement for FGVC in uncontrolled video: conditioning the prediction on the explicit localization of an instance reduced background-induced false positives by nearly an order of magnitude compared to global frame classification (1.632 vs. 12,544 FPs).
- The cascade reduced fine-grained dataset preparation time by approximately 22 times (6.3 hours to 17 minutes for 7,070 images) by replacing manual bounding box annotation with folder-based labeling of auto-extracted crops, and accelerated fine-tuning by 1.8 times on equivalent hardware.

- Robustness to false positives under class absence is brand-dependent rather than absolute: the cascade effectively suppressed false positives for the visually distinct Samsung, but remained subject to inter-class confusion for the absent Garmin due to the loss of scene context.
- The incremental scalability of the cascade, integrating new brands using only auto-labeled crops, without geometric re-annotation or full retraining, resolves the primary operational bottleneck of fine-grained video analytics.
- Operating at a global average of 33 to 38 FPS, the method remains compatible with near-real-time extraction of retail intelligence metrics, such as Share of Voice, from commercial videos.

Taken together, these results confirm the technical viability of annotation-efficient cascades for video-based retail intelligence, provided that per-brand visual separability is taken into account when interpreting robustness under class absence.

Acknowledgements

The authors thank the VERD team for making the Video E-commerce Retrieval Dataset publicly available for research. The YOLO26 fine-tuning scripts and the auto-labeling pipeline used in this research are publicly available at <https://github.com/rafaeljbi/annotation-efficient-fine-grained-brand-recognition-in-video>. During the preparation of this manuscript, generative artificial intelligence tools were used exclusively for translation and language editing. The authors reviewed and edited the resulting text and take full responsibility for the content of this publication.

REFERENCES

1. Farris PW, Bendle NT, Pfeifer PE, Reibstein DJ. Marketing metrics: the manager's guide to measuring marketing performance. 4th ed. Pearson Education; 2021.
2. Lee G, Jo W, Choi Y. VERD: emergence of product-based video e-commerce retrieval dataset from user's perspective. *Sensors*. 2023; 23(1): 513. <https://doi.org/10.3390/s23010513>
3. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016. p. 779–788. <https://doi.org/10.1109/CVPR.2016.91>

4. Terven J, Córdova-Esparza DM, Romero-González JA. A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS. *Mach Learn Knowl Extr.* 2023; 5(4): 1680–1716. <https://doi.org/10.3390/make5040083>
5. Wei XS, Song YZ, Mac Aodha O, Wu J, Peng Y, Tang J, et al. Fine-grained image analysis with deep learning: a survey. *IEEE Trans Pattern Anal Mach Intell.* 2022; 44(12): 8927–8948. <https://doi.org/10.1109/TPAMI.2021.3126648>
6. Xie Y, Gong Q, Luan X, Yan J, Zhang J. A survey of fine-grained visual categorization based on deep learning. *J Syst Eng Electron.* 2023; 35(6): 1337–1356. <https://doi.org/10.23919/JSEE.2022.000155>
7. Wah C, Branson S, Welinder P, Perona P, Belongie S. The Caltech-UCSD Birds-200-2011 dataset. California Institute of Technology; 2011. Technical Report CNS-TR-2011-001.
8. Krause J, Stark M, Deng J, Fei-Fei L. 3D object representations for fine-grained categorization. In: *Proceedings of the IEEE ICCV Workshop on 3D Representation and Recognition*; 2013. <https://doi.org/10.1109/ICCVW.2013.77>
9. Maji S, Rahtu E, Kannala J, Blaschko M, Vedaldi A. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*; 2013. <https://doi.org/10.48550/arXiv.1306.5151>
10. Papadopoulos DP, Uijlings JRR, Keller F, Ferrari V. Extreme clicking for efficient object annotation. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; 2017. p. 4940–4949. <https://doi.org/10.1109/ICCV.2017.528>
11. CVAT.ai. Image annotation guide: annotation types, times, and best practices [Internet]. 2026 [cited 2026 Mar]. Available from: <https://www.cvat.ai/resources/blog/image-annotation-guide>
12. Cai Z, Vasconcelos N. Cascade R-CNN: delving into high quality object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2018. p. 6154–6162. <https://doi.org/10.1109/CVPR.2018.00644>
13. Zhou X, Jiang Y, Wang Y, Liu J, Lin S, Bai S. Rethinking pseudo labels for semi-supervised object detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence*; 2022; 36(3): 3580–3588. <https://doi.org/10.1609/aaai.v36i3.20268>
14. Jocher G, Qiu J, Liu M, Lyu S, Akyon FC, Kalfaoglu ME. Ultralytics YOLO26: unified real-time end-to-end vision models. *arXiv preprint*; 2026. <https://doi.org/10.48550/arXiv.2606.03748>
15. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: *Computer Vision – ECCV 2014*. Springer International Publishing; 2014. p. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
16. Chakrabarty S. YOLO26: an analysis of NMS-free end-to-end framework for real-time object detection. *arXiv preprint arXiv:2601.12882*; 2026. <https://doi.org/10.48550/arXiv.2601.12882>
17. Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, et al. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In: *Computer Vision – ECCV 2024*. Springer; 2024. <https://doi.org/10.48550/arXiv.2303.05499>
18. Cheng T, Song L, Ge Y, Liu W, Wang X, Shan Y. YOLO-World: real-time open-vocabulary object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2024. <https://doi.org/10.48550/arXiv.2401.17270>
19. Bianchi L, Carrara F, Messina N, Gennaro C, Falchi F. The devil is in the fine-grained details: evaluating open-vocabulary object detectors for fine-grained understanding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2024. <https://doi.org/10.48550/arXiv.2311.17518>
20. Al-Abbas AR, Gierz Ł, Falih BS, Al-Sammarraie MAJ. Application of you only look once algorithms to crop production management using unmanned aerial vehicles and computer vision systems. *Adv Sci Technol Res J.* 2026; 20(1): 1–13. <https://doi.org/10.12913/22998624/207678>
21. Abdali A, Boguslawski B, Drumetz L, Gripon V. Object-centric cropping for visual few-shot classification. *arXiv preprint arXiv:2508.00218*; 2025. <https://doi.org/10.48550/arXiv.2508.00218>
22. Wang A, Liu L, Chen H, Lin Z, Han J, Ding G. YOLOE: real-time seeing anything. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2025. <https://doi.org/10.48550/arXiv.2503.07465>