

RETRAINING-FREE OBJECT AUTHORIZATION FOR VIDEO SURVEILLANCE: AN EMPIRICAL STUDY OF EMBEDDING VERIFICATION

S R Sheetal^{1*}, Dr. P. Manikandan², Dr. S. Saravanakumar³

^{1*}Research Scholar, Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bengaluru, India, srsheetal26@gmail.com. Orcid ID: 0009-0002-3335-9194

²Professor, Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bengaluru, India. Orcid ID: 0000-0003-3037-7688

³Associate Professor, Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bengaluru, India.

ABSTRACT

Surveillance detectors report what an object is but not whether it is registered with the site, so an unattended bag and an enrolled bag look identical to the system. This paper studies ISS-RUO, a retraining-free authorization layer in which a YOLOv8 detector localizes objects, a frozen ImageNet-pretrained EfficientNetB0 embeds each region, and the maximum cosine similarity to an enrolled database is thresholded to label the object AUTHORIZED or UNREGISTERED; new objects are enrolled simply by adding images. The layer is evaluated with disjoint enrolment, validation, and test partitions, thresholds tuned on validation data only, and five repeated class splits. On object-centred images (Caltech-256; 20 authorized and 20 unseen classes) it reaches an AUROC of 0.892 ± 0.017 and a balanced accuracy of 0.800 ± 0.017 , comparable to a linear classifier with softmax rejection (AUROC 0.897 ± 0.023) and better than CLIP embeddings (0.841 ± 0.038); a fixed threshold of 0.75 fails (accuracy 0.361), and enrolling one image per class lowers AUROC to 0.769. On multi-object COCO scenes, detector-based localization is essential (AUROC 0.900 versus 0.760 for whole-scene embeddings), but the detector found only about 60% of annotated objects, and when authorization coincides with the detector's own classes its label rule (F1 0.932) far outperforms embedding verification (F1 0.567). On a two-core CPU, processing took about 272 ms per frame, and frame skipping cut CPU time by 78%. The results indicate where embedding-based authorization is useful and where it is not.

KEYWORDS: Video surveillance, object authorization, embedding verification, cosine similarity, object detection, YOLOv8, EfficientNetB0, open-set recognition, frame skipping.

1. INTRODUCTION

Video surveillance has become a core component of the security infrastructure of airports, railway stations, campuses, industrial sites, laboratories, and smart cities. Industry analyses estimated roughly 770 million installed surveillance cameras worldwide in 2019 and projected more than one billion by the end of 2021. Human monitoring at this scale is limited by delayed reaction, operator fatigue, thin staffing at night and at weekends, and the simple fact that no team can watch thousands of streams continuously; [1] automated video analysis is therefore no longer optional but a practical necessity for any site that wants to act on what its cameras record rather than merely archive it.

Deep object detectors have made automated analysis practical. Single-stage detectors such as YOLO, SSD, and RetinaNet and two-stage detectors such as Faster R-CNN localize and classify objects in real time on modern hardware, and YOLO-based surveillance systems have been reported. [2], [3] These detectors, however, answer a fixed question: to which of a predefined set of categories does this object belong? They do not answer the question that security staff actually ask, namely whether a detected object is registered with the site. An unattended bag in an airport hall is correctly recognized as “a bag”, yet the detector cannot tell whether it is an enrolled, authorized item or an unregistered one that warrants attention. The same limitation applies to laptops in a laboratory, equipment on a construction site, or pallets in a warehouse, where the operational question is authorization, not category.

Three deficiencies of the existing approaches make this gap hard to close. First, category-level detectors and classifiers have a closed label set: an object is always assigned to one of the known categories, [4] so authorization cannot be expressed unless every authorized item is a trained class, and adding a new item means collecting labelled data and retraining. Second, methods that do verify identity, namely face verification [5], [6] person re-identification, and open-set recognition, are designed and trained for a specific domain with identity- or category-labelled data, and they are not built to sit behind a general object detector. Third, prototype systems for object recognition in surveillance are frequently evaluated on small, hand-picked image sets in which enrolment and test images overlap, in which the decision threshold is tuned on the reported data, and in which no simple baseline is included; under such conditions a reported accuracy says little about behaviour on unseen data, and the contribution of each pipeline component cannot be isolated.

This paper studies ISS-RUO (Intelligent Surveillance System for Recognition of Unregistered Objects), a pipeline that adds an authorization layer to a standard detector without any additional training. Each object localized by YOLOv8 is cropped and embedded by an ImageNet-pretrained EfficientNetB0, its maximum cosine similarity to an enrolled database of authorized objects is computed, and a threshold chosen on validation data labels the object AUTHORIZED or UNREGISTERED and drives a colour-coded alert. New authorized objects are enrolled by embedding a few images, and

no network is retrained. The detect–embed–threshold recipe is established in verification and re-identification, so the contribution of this paper is not a new learning algorithm but a careful empirical study of how well the recipe works when it is evaluated properly, which of its components matter, and where it fails. The study is organized around four questions: RQ1. How well does a frozen embedding with a validation-tuned cosine threshold separate authorized categories from categories never seen at enrolment, under disjoint partitions and repeated class splits?

RQ2. Which components (detector, embedder, similarity measure, enrolment size) drive the result, and how does the pipeline compare with a linear classifier with softmax rejection, CLIP embeddings, and the detector’s own label?

RQ3. In multi-object scenes, how do localization quality, detector recall, and the detector’s vocabulary affect authorization?

RQ4. How robust is the layer to blur, occlusion, and illumination change, and what does each stage cost on a CPU?

The principal contributions of this paper are summarized as follows.

First, a leakage-free evaluation protocol for enrolment-based object authorization is defined and applied: enrolment, validation, and test images are disjoint, unregistered classes are disjoint between validation and test, the threshold is chosen on validation data only, and every result is repeated over five random class or image splits.

Second, the ISS-RUO pipeline is analysed component by component, with ablations of the detector, the similarity measure, and the class representation, an enrolment-size study, and comparisons against two baselines that need no detector (a linear classifier with softmax rejection and CLIP embeddings).

Third, a scene-level study on COCO val2017 quantifies what localization contributes (AUROC 0.900 with detector crops versus 0.760 with whole-scene embeddings) and what the detector’s limited recall costs end to end, and compares embedding verification with the detector’s own class-label rule.

Fourth, robustness to blur, occlusion, and brightness change is measured together with the CPU cost of every stage, with and without frame skipping, and the scaling of the similarity search to a hundred thousand enrolled embeddings.

Fifth, per-class and per-split analyses, including a similarity matrix among authorized categories and the classes that cause most false alarms, identify concrete failure modes and give practical guidance on when embedding-based authorization is justified and when a simpler detector-label rule is preferable.

The remaining structure of this paper is as follows. The Literature Review section surveys surveillance analytics, object detection, embedding-based recognition, and open-set verification, and identifies the gaps addressed here. The Methodology section describes the ISS-RUO pipeline, its decision rule, and its complexity. The Experimental Setup section specifies the datasets, protocols, variants, metrics, and implementation details. The Results and Discussion section reports the quantitative results, per-class and per-split analyses, ablations, scene-level results, robustness, and computational cost. The Conclusion, Future Enhancements, and Limitations and Ethical Considerations sections close the paper.

LITERATURE REVIEW

Automated analysis of surveillance video has evolved from pixel-level change detection through hand-crafted features and shallow classifiers to deep convolutional detectors and, more recently, to embedding models trained at very large scale. A survey of this literature shows three persistent gaps that define the problem addressed in this paper: object-level authorization against a site-specific list of registered items is rarely treated as a task of its own; verification methods and detectors are developed separately and are seldom evaluated together; and reported results are often obtained on small prototype sets with overlapping partitions and without simple baselines. The development of the field is traced below along four threads.

Classical surveillance analytics.

Early automated surveillance relied on modelling the static background of a scene and flagging pixels that deviate from it. Stauffer and Grimson introduced adaptive per-pixel mixtures of Gaussians that remain a reference point for real-time tracking, and optical flow and frame differencing were used in the same spirit. These methods are cheap but raise many false alarms under illumination changes, camera vibration, and harmless motion, and they say nothing about what a moving object is. Hand-crafted descriptors filled that gap: Dalal and Triggs’s histograms of oriented gradients with a linear SVM became the standard pedestrian detector, Lowe’s scale-invariant keypoints supported object matching, and local binary patterns captured texture. Although these features improved accuracy, they required careful design for each object type and generalized poorly across viewpoint, scale, occlusion, and camera setting.

Deep object detection and its use in surveillance.

The success of deep convolutional networks on large-scale recognition, demonstrated by Krizhevsky et al., removed the need for hand-crafted features. Region-based detectors followed: R-CNN classified region proposals with a convolutional network, Fast R-CNN shared computation across proposals, and Faster R-CNN learned the proposals with a region proposal network, giving accurate but comparatively costly two-stage detection. Single-stage detectors improved the speed–accuracy trade-off: SSD predicted boxes from multi-scale feature maps, YOLO framed detection as a single regression problem, and RetinaNet closed much of the accuracy gap to two-stage detectors with the focal loss. YOLOv8 uses anchor-free heads, decoupled classification and regression branches, and C2f modules[7], and is available in several sizes, of which the smallest (YOLOv8n) suits CPU deployment[3].

The YOLO family has been adapted to many domains, including distant tiny objects (Yu et al.)[8], rotated objects (Zhuang et al.)[9], traffic lights (Qu et al.)[10], and maritime scenes (Mao et al.)[7]. In surveillance, Anish et al.[2] applied YOLO to real-time object detection and tracking, Azatbekuly[3] and Mukhanbet developed an intelligent video surveillance system based on YOLO, and surveys of intelligent video surveillance by Chundi[1] and Bammidi and by Huang and Liu

describe the technology, applications, and system architectures. Systems for checking or recognizing predefined categories of objects in specific settings have also been proposed, for example the object-checking frameworks of Zeng and Guo[4],[11] and the recognition system for personnel and safety equipment on construction sites by Chen and Chen[5]; such systems typically rely on classifiers or detectors trained for the categories of interest.

Two further groups of systems are close in spirit to the present study. Chen et al.[5] match face embeddings against photographs of registered personnel by cosine similarity and label a person as unregistered when the score falls below a threshold, and Sarkar et al. [6] combine thermal-camera object detection with facial-biometric authentication against a stored set of pre-approved users; both authorize people by similarity to enrolled examples, but neither addresses general objects. A separate line of work fuses visible and thermal imagery for detection under difficult conditions, including RGB-thermal pedestrian detection [12], salient object detection in UAV RGB-T image pairs [13], [14], and drone detection by infrared–visible feature-level fusion [15]. In those studies “unregistered” describes image pairs that are spatially misaligned, whereas in the present paper it describes objects that have not been enrolled at the site, so the two problems are different.

Embedding-based recognition and transfer.

Deep networks pretrained on ImageNet produce features that transfer across tasks[16]. ResNet and VGG became common backbones, and EfficientNet introduced compound scaling of depth, width, and resolution to obtain accurate but compact models; its baseline B0 has about 5.3 million parameters including the classifier, far fewer than ResNet-50 or VGG-16. With the classifier removed, global average pooling of the last convolutional stage yields a 1280-dimensional embedding for EfficientNetB0. Frozen embeddings of this kind allow new items to be compared without retraining, which is the property exploited by ISS-RUO. Contrastively trained models such as CLIP and self-supervised models such as DINO provide stronger general-purpose embeddings and are natural alternatives. [17],[18],[19]

Metric learning, verification, and open-set recognition.

Verification and open-set problems ask whether a sample matches something known or belongs to a class never seen in training. Siamese networks and triplet-loss networks[20],[21] such as FaceNet learn embedding spaces in which distance reflects identity, and angular-margin losses such as ArcFace sharpen the separation with cosine-based decision rules. Person re-identification, surveyed by Ye et al., applies the same idea to matching people across cameras. Scheirer et al. formalized open-set recognition[22], and Bendale and Boulton proposed the OpenMax layer to reject unknown classes in deep networks[23]. These methods are trained with identity or category supervision for a specific domain; they are not designed to sit behind a general object detector[24],[25],[26],[27],[28], and the class set of the closed-set variants is fixed at training time.[29]

Evaluation practice.

Two methodological points matter for the present study. A threshold-based decision rule needs an operating point, and Youden’s index, the difference between the true-positive and false-positive rates, is a standard way to select it without favouring either error type. On imbalanced test sets, plain accuracy is misleading and metrics such as balanced accuracy and the Matthews correlation coefficient are more informative. Applying these ideas requires that the threshold be chosen on data that are separate from the test data, a condition that small prototype studies often fail to meet. Table **Error! Reference source not found.** summarizes the surveyed work, and Table **Error! Unknown switch argument.** compares the capability profiles of the method families.

Table 1. Comparative Analysis of Related Work: Paradigm, Domain, Architecture, and Datasets

No.	Authors	Year	Venue	Paradigm	Domain	Architecture	Dataset
1	Stauffer and Grimson	1999	CVPR	Background modelling	Surveillance video	Adaptive per-pixel Gaussian mixture	Surveillance video
2	Dalal and Triggs	2005	CVPR	Hand-crafted detection	Pedestrians	HOG + linear SVM	INRIA person
3	Lowe	2004	IJCV	Hand-crafted features	General objects	Scale-invariant keypoints	—
4	Ojala et al.	2002	TPAMI	Hand-crafted features	Texture	Local binary patterns	Texture databases
5	Krizhevsky et al.	2012	NeurIPS	CNN classification	General	AlexNet	ImageNet
6	Girshick et al.	2014	CVPR	Two-stage detection	General	Region proposals + CNN (R-CNN)	PASCAL VOC, ILSVRC
7	Ren et al.	2017	TPAMI	Two-stage detection	General	Region proposal network (Faster R-CNN)	PASCAL VOC, COCO
8	Liu et al.	2016	ECCV	Single-stage detection	General	Multi-scale feature maps (SSD)	PASCAL VOC, COCO

No.	Authors	Year	Venue	Paradigm	Domain	Architecture	Dataset
9	Redmon et al.	2016	CVPR	Single-stage detection	General	Single regression pass (YOLO)	PASCAL VOC
10	Lin et al.	2017	ICCV	Single-stage detection	General	Focal loss (RetinaNet)	COCO
11	Jocher et al.	2023	Software	Single-stage detection	General	Anchor-free, decoupled head (YOLOv8)	COCO
12	Anish et al.	2023	ICACRS	YOLO application	Surveillance	YOLO detection and tracking	—
13	Azatzbekuly and Mukhanbet	2024	IEEE SIST	YOLO application	Video surveillance	YOLO-based surveillance system	—
14	Zeng and Guo	2025	ICCE	Object checking	Consumer electronics	Multi-category classifiers, designed-augmented data	—
15	Chen and Chen	2024	HNICE M	Recognition system	Construction sites	Personnel and safety-equipment recognition	—
16	Koch et al.	2015	ICML Workshop	Metric learning	One-shot recognition	Siamese CNN	Omniglot
17	Schroff et al.	2015	CVPR	Metric learning	Face verification	Triplet-loss embedding (FaceNet)	LFW, YouTube Faces
18	Deng et al.	2019	CVPR	Metric learning	Face recognition	Additive angular margin loss (ArcFace)	Large face datasets
19	Scheirer et al.	2013	TPAMI	Open-set recognition	General	Formal open-set framework	—
20	Bendale and Boulton	2016	CVPR	Open-set recognition	General	OpenMax layer on a deep network	ImageNet
21	Ye et al.	2022	TPAMI	Survey	Person re-identification	Deep re-identification review	Re-ID benchmarks
22	Tan and Le	2019	ICML	Backbone	General	EfficientNet compound scaling	ImageNet
23	Radford et al.	2021	ICML	Embedding model	General	CLIP contrastive image-text model	400M image-text pairs
24	ISS-RUO (this study)	2026	Under review	Detector embedding verification +	Surveillance	YOLOv8n + EfficientNetB0 + cosine threshold	Caltech-256, COCO val2017

Table I. Capability Profile of Method Families Relevant to Object Authorization

Method family	Localize objects	Embedding matching	Enrols new items without retraining	Rejects unseen categories	Needs domain-specific training	Main limitation for authorization
Hand-crafted detectors (HOG, SIFT, LBP + SVM)	Yes	No	No	No	Yes	Feature design per object type; poor generalization
Two-stage detectors (Faster R-CNN)	Yes	No	No	No	Yes	Fixed label set; higher per-frame cost
Single-stage detectors (YOLO, SSD, RetinaNet, YOLOv8)	Yes	No	No	No	Yes	Fixed label set; no notion of enrolment

Method family	Localizes objects	Embedding matching	Enrols new items without retraining	Rejects unseen categories	Needs domain-specific training	Main limitation for authorization
Open-set deep networks (OpenMax)	No	Partial	No	Yes	Yes	Known classes fixed at training
Metric learning (Siamese, FaceNet, ArcFace)	No	Yes	Yes	Yes	Yes	Trained for a domain (faces, characters); no detector
Person re-identification	No	Yes	Yes	Yes	Yes	Person-specific; trained on re-identification data
General embeddings (CLIP, DINO)	No	Yes	Yes	Partial	No	No localization; threshold must be calibrated
ISS-RUO (this study)	Yes	Yes	Yes	Yes	No	Detector recall and vocabulary; threshold per embedder

The survey and the two tables lead to the following gaps. First, to our knowledge, few studies evaluate a detector-plus-frozen-embedding authorization layer with disjoint enrolment, validation, and test partitions and a validation-tuned threshold, so the true generalization of such systems is rarely known. Second, comparisons against simple alternatives that need no detector, such as a linear classifier with softmax rejection, and against the detector's own label are seldom reported, which makes it difficult to judge when embeddings are worth their cost. Third, robustness to blur, occlusion, and illumination, and the per-stage CPU cost of the whole pipeline, are rarely characterized together with accuracy. The present study addresses all three gaps directly.

METHODOLOGY

ISS-RUO is organized as a short pipeline in which every stage is a pretrained, frozen component and the only site-specific ingredient is the enrolled database. Video frames are passed to a YOLOv8 detector[3],[7],[10] the detected regions are cropped and embedded by EfficientNetB0, each embedding is matched against the enrolled embeddings of authorized objects by cosine similarity, and a threshold chosen on validation data converts the best similarity into an AUTHORIZED or UNREGISTERED decision that drives a colour-coded alert. The overall architecture is illustrated in Fig. **Error! Unknown switch argument.** No stage is trained in this work, so the method has no training procedure, loss function, or optimizer; the study instead concerns the design of the decision rule, the choice of the threshold, and the evaluation of the whole.

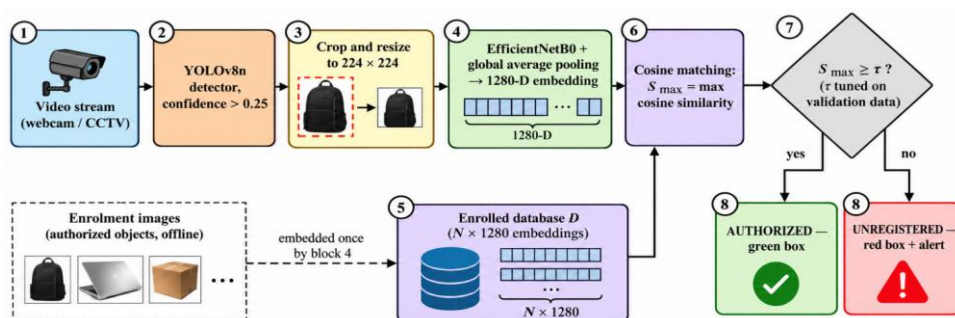


Fig. 1. Overall architecture of the ISS-RUO pipeline. Blocks 1–4 and 6–8 form the online path; block 5, the enrolled database, is filled offline by embedding enrolment images with the same EfficientNetB0 (block 4).

Problem Formulation

Consider a video stream of frames $x_1, x_2, \dots$ and a site that has registered a set of authorized objects belonging to categories C_{auth} . A detector Φ maps a frame x to a set of objects, each described by a bounding box b , a category label ℓ , and a confidence s . The task is to define a decision function $y(o) \in \{\text{AUTHORIZED}, \text{UNREGISTERED}\}$ for every detected object o such that objects that match an enrolled authorized object are accepted and all other objects raise an alert. The site provides only a handful of enrolment images per authorized category; no negative (unregistered) examples are available at enrolment time, and the unregistered objects encountered at run time may belong to categories that have never been seen. The decision function is therefore defined by similarity to the enrolled examples rather than by a trained boundary between known classes, which is what allows new items to be added without retraining.

Enrolment and the Authorized Database

Enrolment consists of collecting a small number of images of each authorized category, embedding them once, and storing the embeddings. Because the embedder is frozen and pretrained, enrolment takes one forward pass per image and involves no gradient computation. With N_c enrolment images for category c , the database is

$$D = \{(F_{c,i}, c) : c \in C_{\text{auth}}, i = 1, \dots, N_c\} \quad (1)$$

where $F_{c,i} \in \mathbb{R}^{1280}$ is the embedding of the i -th enrolment image of category c and $|D| = \sum_c N_c$ is the number of stored embeddings. In the reference configuration $N_c = 5$, so a database for 20 authorized categories holds 100 vectors of 1280 floating-point numbers, about 0.5 MB. Adding a category means appending its embeddings; removing one means deleting them; neither operation touches any model.

YOLOv8 Detection and Object Cropping

YOLOv8 is used as a pretrained detector in its smallest size (YOLOv8n, about 3.2 million parameters), which is the variant that suits CPU deployment[3]. It predicts box coordinates directly with an anchor-free head, and its training objective (distribution focal loss with a CIoU term for boxes, binary cross-entropy for classification, and task-aligned label assignment) is not reproduced here because the detector is used as released. Boxes with confidence above 0.25, the default of the detector's prediction interface, are kept. In the object-centred protocol the highest-confidence box of each image is used; when the detector finds no box, which happened for 34.7% of the Caltech-256 images, the whole image is used so that every image receives a decision. In the scene-level protocol every detection that is not a person is processed. A detected region is cropped from the frame and resized to the network input size:

$$I_o = \text{resize}(\text{crop}(x, b_o), 224 \times 224) \quad (2)$$

where x is the frame and b_o the box of object o . Boxes narrower or shorter than 8 pixels are discarded as degenerate.

EfficientNetB0 Embedding

EfficientNetB0, pretrained on ImageNet, is used in feature-extraction mode. The EfficientNet family is obtained by scaling network depth d , width w , and input resolution r jointly with a single compound coefficient ϕ :

$$\begin{aligned} d &= \alpha^\phi, \quad w = \beta^\phi, \quad r = \gamma^\phi \\ \alpha \cdot \beta^2 \cdot \gamma^2 &\approx 2, \quad \alpha \geq 1, \quad \beta \geq 1, \quad \gamma \geq 1 \end{aligned} \quad (3)$$

where α , β , and γ were found by a small grid search in the original work and take the values $\alpha = 1.2$, $\beta = 1.1$, and $\gamma = 1.15$, for which $\alpha \cdot \beta^2 \cdot \gamma^2 = 1.92 \approx 2$; the baseline model B0 corresponds to $\phi = 0$. The classification head is removed and the global-average-pooled (GAP) output of the last convolutional stage is used as the embedding, which has 1280 dimensions; the backbone without its classifier has about 4 million parameters. Object crops are passed to the network as raw 0–255 pixel values; the first layers of the Keras implementation (rescaling and normalization) apply the standard ImageNet per-channel normalization inside the model, which for an image I corresponds to

$$I_{\text{norm}} = \frac{I / 255 - \mu}{\sigma} \quad (4)$$

with $\mu = (0.485, 0.456, 0.406)$ and $\sigma = (0.229, 0.224, 0.225)$. The embedding of a crop is

$$F = \text{GAP}(f(I_o)) \in \mathbb{R}^{1280} \quad (5)$$

where $f(\cdot)$ is the EfficientNetB0 backbone without its classification layer. Shallow layers capture low-level cues such as edges and colour, and deeper layers capture object structure and category-level semantics. Because global average pooling of activations that are mostly positive produces vectors whose entries are mostly positive, the cosine similarity between embeddings of unrelated images is not close to zero; in the experiments reported below the best-match similarity of unregistered test images averages 0.28, which is one reason why a threshold carried over from another embedder or dataset cannot be trusted.

Cosine Matching and Decision Rule

For a detected-object embedding $F_{\text{det}} \in \mathbb{R}^{1280}$, the similarity to an enrolled embedding B is the cosine of the angle between them,

$$\cos(A, B) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} \quad (6)$$

and the score of the object is the maximum similarity over the database, that is, the similarity to its nearest enrolled neighbour:

$$S_{\text{max}} = \max_{B \in D} \cos(F_{\text{det}}, B) \quad (7)$$

Cosine similarity is invariant to the magnitude of the embedding, is bounded in $[-1, 1]$, costs $O(1280)$ per pair, and assumes no data distribution. Using the maximum rather than an average over a class means that an object is accepted as

soon as it resembles any one enrolled example, which suits categories with diverse appearance but also lets a single unrepresentative enrolment image widen the accepted region. The decision rule is

$$y(F_{\text{det}}) = \begin{cases} \text{AUTHORIZED, } S_{\text{max}} \geq \tau \\ \text{UNREGISTERED, } S_{\text{max}} < \tau \end{cases} \quad (8)$$

Threshold Selection

The threshold τ is the only free parameter of the decision rule, and it is chosen on a held-out validation set that contains both authorized objects and objects of categories that are unregistered but disjoint from the test categories. The value maximizing Youden's index, the true-positive rate minus the false-positive rate, is selected, which is equivalent to maximizing balanced accuracy:

$$\tau = \arg \max_t [\text{TPR}_{\text{val}}(t) - \text{FPR}_{\text{val}}(t)] \quad (9)$$

The test data are never used to choose τ . The choice of Youden's index treats missed authorized objects and false alarms as equally costly; a deployment that weighs the two differently should select a different operating point from the same validation curve, and the threshold must be re-tuned whenever the embedder, the enrolled categories, or the camera setting changes because the similarity scale depends on all three.

Class-Prototype, Softmax-Rejection, and Label-Rule Variants

Three alternative decision rules are used in the ablations and comparisons. In the class-prototype variant each authorized category is represented by the normalized mean of its enrolment embeddings, and the score is the maximum cosine similarity to a prototype:

$$P_c = \frac{m_c}{\|m_c\|}, \quad m_c = \frac{1}{N_c} \sum F_{c,i}, \quad S_{\text{proto}} = \max_c \cos(F_{\text{det}}, P_c) \quad (10)$$

In the softmax-rejection baseline a multinomial logistic-regression classifier is trained on the enrolment embeddings of the authorized categories and an object is accepted when the maximum posterior probability exceeds a validation-tuned threshold:

$$S_{\text{soft}} = \max_c \text{softmax}(W F + b)_c \quad (11)$$

The classifier has 25,620 parameters for 20 categories (1280×20 weights and 20 biases) and must be retrained whenever a category is added. Finally, the detector-label rule needs no embedding at all: a detected object is authorized when the category predicted by the detector belongs to the authorized set,

$$y_{\text{label}}(o) = \text{AUTHORIZED} \Leftrightarrow \ell_o \in C_{\text{auth}} \quad (12)$$

This rule is available only when the authorized categories are among the detector's own classes, but where it applies it is a strong baseline that any embedding-based method must be compared with.

Stream Processing, Frame Skipping, and Alerts

On a video stream the detector is the dominant fixed cost, so only every N_{skip} -th frame is processed, with $N_{\text{skip}} = 5$ in the reference configuration. This reduces the number of frames entering the pipeline by a factor $1 - 1/N_{\text{skip}} = 80\%$; the saving in computation depends on the per-frame cost and is measured in the experiments. At a nominal 30 frames/s the interval between processed frames is about 167 ms, during which a briefly visible object can be missed. For every processed frame, each detected object is cropped, embedded, and matched, and the alert module draws a green box labelled AUTHORIZED with the score when $S_{\text{max}} \geq \tau$, or a red box labelled UNREGISTERED with the score when $S_{\text{max}} < \tau$, which also notifies the operator and logs the event with time, frame, and score. The complete procedure is formalized in Algorithm 1.

Algorithm 1: ISS-RUO — Enrolment-Based Object Authorization on a Video Stream

Input: enrolment images of authorized categories, video stream V , threshold τ , skip factor N_{skip} , confidence 0.25

Output: annotated frames with AUTHORIZED / UNREGISTERED labels and logged alerts

Phase 1 — Enrolment (offline)

- 1: for each authorized category c and each enrolment image $I_{c,i}$ do
- 2: $F_{c,i} \leftarrow \text{GAP}(f(\text{resize}(I_{c,i}, 224 \times 224)))$
- 3: end for; store $D = \{(F_{c,i}, c)\}$

Phase 2 — Threshold selection (offline, validation data only)

- 4: compute S_{max} for all validation objects; $\tau \leftarrow \arg \max_t [\text{TPR}_{\text{val}}(t) - \text{FPR}_{\text{val}}(t)]$

Phase 3 — Online verification

- 5: $\text{frame_count} \leftarrow 0$
- 6: for each frame x in V do
- 7: $\text{frame_count} \leftarrow \text{frame_count} + 1$; if $\text{frame_count} \bmod N_{\text{skip}} \neq 0$ then continue
- 8: $O \leftarrow \text{detector}(x)$ with confidence > 0.25
- 9: for each object o in O do

10: $I_o \leftarrow \text{resize}(\text{crop}(x, b_o), 224 \times 224)$; $F_{\text{det}} \leftarrow \text{GAP}(f(I_o))$
11: $S_{\text{max}} \leftarrow \max \text{ over } (F_c, i, c) \in D \text{ of } \cos(F_{\text{det}}, F_c, i)$
12: if $S_{\text{max}} \geq \tau$ then label AUTHORIZED, draw green box
13: else label UNREGISTERED, raise alert, draw red box, log (time, frame, score)
14: end for; display the annotated frame to the operator
15: end for

Complexity: $O(|D| \cdot 1280)$ per detected object for the similarity search (line 11); the detector and the embedder dominate the cost.

Complexity and Memory

For a processed frame that contains n detected objects, the time is approximately

$$T_{\text{frame}} \approx T_{\text{det}} + n \cdot (T_{\text{crop}} + T_{\text{emb}} + T_{\text{search}}(|D|)) \quad (13)$$

where T_{search} grows linearly with $|D|$ and is negligible for small databases, while T_{emb} is a fixed forward pass of the embedder. With frame skipping the average cost per input frame is divided by N_{skip} . Memory for the database is $4 \cdot 1280 \cdot |D|$ bytes in single precision, which is 0.5 MB for the reference database of 100 embeddings and 512 MB for 100,000 embeddings. For very large databases an approximate nearest-neighbour index would be needed; this was not evaluated.

Inference Pipeline and Model Configuration

At inference time the system therefore performs, for each processed frame, one detector pass, one embedder pass per detected object, and one matrix–vector product against the database, and it needs no gradient computation, no training data beyond the enrolment images, and no labels at run time. Table **Error! Unknown switch argument.** summarizes the components, their key settings, and their sizes. None of the parameters of the detector or the embedder is updated by this work; the only trainable object is the logistic-regression head of the softmax-rejection baseline.

Table II. ISS-RUO Architectural Configuration and Component Settings

Component	Architecture	Key settings	Dimension	Parameters / training
Detector	YOLOv8n, anchor-free single-stage	confidence 0.25; default input size 640; person boxes removed in scene tests	—	$\approx 3.2\text{M}$, pretrained, frozen
Crop and resize	Bilinear resizing	minimum box side 8 px; output 224×224	$224 \times 224 \times 3$	None
Embedder	EfficientNetB0 (MBConv), GAP, no classifier	raw 0–255 input; normalization inside the model	1280	$\approx 4\text{M}$, ImageNet, frozen
Enrolled database	Matrix of embeddings	5 images per authorized class (default)	$ D \times 1280$	None
Matcher	Maximum cosine similarity (1-NN)	float precision; exhaustive search	1	None
Decision	Threshold on S_{max}	τ from validation (Youden’s index)	1	One scalar
Baseline: classifier	Multinomial logistic regression	$C = 10$, up to 3000 iterations; softmax rejection	$1280 \rightarrow 20$	25,620, trained on enrolment
Baseline: CLIP	CLIP ViT-B/32 image encoder	frozen; cosine 1-NN	512	Pretrained, frozen
Alert	Colour-coded overlay and log	green AUTHORIZED, red UNREGISTERED	—	None

Experimental Setup

The experimental evaluation of ISS-RUO is designed to answer the four questions posed in the introduction with a protocol that prevents information from leaking between enrolment, threshold selection, and testing. It combines four dimensions: a leakage-free object-centred evaluation on Caltech-256 with repeated class splits; ablations and baseline comparisons; studies of enrolment size and of robustness to perturbations; and a scene-level evaluation on COCO val2017 together with a per-stage CPU timing study. All random processes are seeded, and every reported number is either produced by the code that accompanies this paper or derived arithmetically from its outputs.

Datasets

Two public datasets are used. Caltech-256 contains 30,607 photographs in 256 object categories plus a clutter category, with at least 80 images per category; most images show a single, centred object. It represents the case in which the object

of interest fills the image and the authorized categories are not classes of the detector. The categories named clutter, people, and face were removed for privacy, and only categories with at least 50 images were kept, leaving 254 categories. In each of five random class splits, 20 categories are authorized and 20 further categories are unseen and therefore unregistered, ten used only for validation and ten only for test. For every authorized category, ten enrolment images (the first five are used by default), 20 validation images, and 20 test images are drawn without overlap. Each test set therefore contains 400 authorized and 200 unregistered images, and the five splits together use 6,352 unique images. COCO val2017 contains 5,000 everyday scenes annotated with 80 object categories, and it was not used to train the detector. Person annotations are excluded for privacy, crowd regions are ignored, and objects smaller than 32×32 pixels or narrower than 24 pixels are dropped, leaving 4,689 images and 16,454 annotated objects. Eight categories are authorized (backpack, handbag, suitcase, laptop, cell phone, umbrella, car, and bicycle; 2,471 objects) and all others are unregistered (13,983 objects). Images are split at image level into 20% enrolment, 40% validation, and 40% test images in five random repetitions, and five enrolment crops of at least 48 pixels per authorized category are taken from the ground-truth boxes of enrolment images only. The protocols are summarized in Fig. **Error! Unknown switch argument.** and Table **Error! Unknown switch argument.**

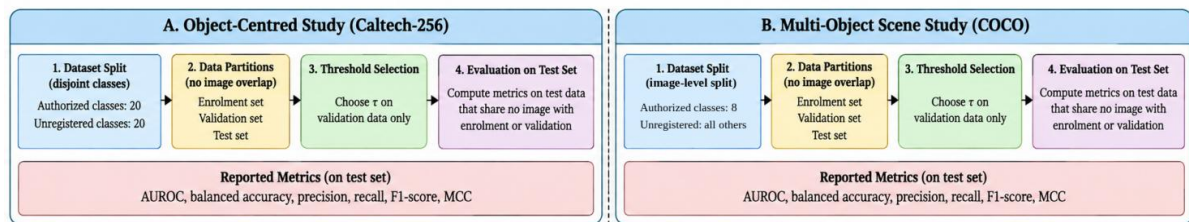


Fig. 2. Evaluation protocols of the two studies: (A) object-centred study on Caltech-256, with 20 authorized and 20 unregistered classes per split, and (B) multi-object scene study on COCO val2017, with eight authorized categories and all other non-person categories unregistered and an image-level split. In both studies the enrolment, validation, and test sets share no image, the threshold τ is chosen on validation data only, and the metrics are computed on the test set; in the object-centred study the unregistered classes are also disjoint between validation and test, and the scene study additionally reports detector recall and end-to-end recall.

Table III. Dataset Characteristics and Experimental Configuration

Property	Caltech-256 (object-centred)	COCO val2017 (scene-level)
Domain	Photographs with one main object	Everyday multi-object scenes
Source size	30,607 images, 256 categories + clutter	5,000 images, 80 categories
Data used	254 categories (clutter, people, face removed; at least 50 images); 6,352 unique images over five splits	4,689 images with a usable non-person object; 16,454 objects
Authorized categories	20 random categories per split	8 fixed: backpack, handbag, suitcase, laptop, cell phone, umbrella, car, bicycle (2,471 objects)
Unregistered categories	20 unseen categories per split: 10 for validation, 10 for test	All other non-person categories (13,983 objects)
Evaluation unit	One image: highest-confidence box, or the whole image if no box is found	One detection matched to an annotated object at $\text{IoU} \geq 0.5$ (10,873 detections)
Enrolment	5 images per authorized class (pool of 10)	5 ground-truth crops (≥ 48 px) per authorized class from enrolment images
Validation	20 images per authorized class and 10 unseen classes $\times$ 20 images (600 images)	40% of the images
Test	400 authorized and 200 unregistered images	40% of the images (about 4,300 detections per split)
Class balance	2 : 1 (authorized : unregistered)	About 15% authorized among evaluated detections
Repetitions	Five random class splits (seeds 0–4)	Five random image splits (seeds 0–4)
Filters and exclusions	Privacy exclusions; leakage assertions on image and class overlap	Objects $\geq 32 \times 32$ px; crowd regions ignored; persons excluded

Compared Variants and Baselines

The full pipeline is compared with five variants, chosen so that each isolates one design decision. Removing the detector (whole image) tests whether localization helps on object-centred images. Replacing cosine similarity with Euclidean

distance tests whether the similarity measure matters. Replacing nearest-neighbour matching by class prototypes tests whether aggregating the enrolment images of a class helps. Replacing EfficientNetB0 by CLIP ViT-B/32 embeddings tests whether a stronger general-purpose embedder helps. Finally, a linear classifier with softmax rejection trained on the enrolment embeddings tests whether a trained decision rule beats the similarity rule; it is the natural closed-set-plus-rejection alternative[4].[11] In the scene-level study the pipeline is compared with three further alternatives: ground-truth boxes instead of detections, a whole-scene embedding assigned to every object (no localization), and the detector class-label rule. Table **Error! Unknown switch argument.** lists all variants.

Table IV. Variants and Baselines: Components, Score, and Training Requirements

Variant	Localization	Embedder	Score	Trained component	New category
ISS-RUO (full)	YOLOv8n crop	EfficientNetB0	Max cosine, 1-NN	None	Append embeddings
No detector	Whole image	EfficientNetB0	Max cosine, 1-NN	None	Append embeddings
Euclidean distance	YOLOv8n crop	EfficientNetB0	Negative minimum Euclidean distance	None	Append embeddings
Class prototype	YOLOv8n crop	EfficientNetB0	Max cosine to class mean	None	Update class mean
Baseline: CLIP	YOLOv8n crop	CLIP ViT-B/32	Max cosine, 1-NN	None	Append embeddings
Baseline: softmax rejection	YOLOv8n crop	EfficientNetB0	Max softmax probability	Logistic regression	Retrain head
Scene: ground-truth boxes	Annotated boxes	EfficientNetB0	Max cosine, 1-NN	None	Append embeddings
Scene: whole-scene embedding	None	EfficientNetB0	Max cosine, 1-NN	None	Append embeddings
Scene: detector label rule	YOLOv8n	None	Predicted class $\in$ authorized set	None	Not possible outside detector classes

Evaluation Metrics

Authorized is the positive class throughout. Because the two test sets are imbalanced (2:1 in the object-centred study and about 15% authorized in the scene study), plain accuracy is reported only for completeness and the primary indicators are the area under the ROC curve (AUROC), which does not depend on the threshold, and balanced accuracy and the Matthews correlation coefficient (MCC), which are computed at the validation-selected threshold. Average precision (AP), precision, recall, and F1 complete the set. In the scene study two further quantities are reported: the detector recall, the fraction of annotated authorized objects that are matched by any detection, and the end-to-end recall, the fraction of all authorized objects in the test scenes that are both detected and accepted,

$$R_{e2e} = \frac{|\{\text{authorized objects with a matched detection and } S_{\max} \geq \tau\}|}{|\text{authorized objects in test scenes}|} \quad (14)$$

so that objects missed by the detector count as misses. Table **Error! Unknown switch argument.** gives the definitions, where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives.

Table V. Evaluation Metrics: Definitions, Range, and Role

Metric	Definition	Range	Role in this study
AUROC	Area under the curve of TPR against FPR over all thresholds	[0, 1]	Threshold-free separability of authorized and unregistered scores; primary indicator
AP	Area under the precision–recall curve	[0, 1]	Ranking quality when the positive class is rare
Balanced accuracy	$\frac{TP/(TP + FN) + TN/(TN + FP)}{2}$	[0, 1]	Accuracy insensitive to class imbalance; equals Youden’s criterion up to scale
Precision	$\frac{TP}{TP + FP}$	[0, 1]	Share of accepted objects that are truly authorized
Recall	$\frac{TP}{TP + FN}$	[0, 1]	Share of authorized objects that are accepted

Metric	Definition	Range	Role in this study
F1-score	$\frac{2PR}{P + R}$	[0, 1]	Harmonic mean of precision (P) and recall (R)
MCC	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$	[-1, 1]	Single balanced summary of the confusion matrix
Detector recall	Annotated authorized objects matched by any detection / annotated authorized objects	[0, 1]	Upper bound on what any verifier behind the detector can achieve
End-to-end recall	Equation (Error! Unknown switch argument.)	[0, 1]	Recall that counts detector misses

Ablation Study Design

The ablations change one element at a time and leave the rest of the pipeline and the data splits unchanged, so that differences can be attributed to the element. Besides the variants of Table Error! Unknown switch argument., two further sweeps are run on the full pipeline: the number of enrolment images per category is varied over {1, 3, 5, 10}, and six perturbations are applied to the test crops before embedding while the enrolment set and the threshold stay fixed. Table Error! Unknown switch argument. lists the ablations and the question each one answers.

Table VI. Ablation and Sweep Design: Modified Component and Question Addressed

Ablation	Component modified	Substitution	Question addressed
No detector	Localization	Whole image resized to 224×224	Does the detector help on object-centred images?
Euclidean distance	Similarity measure	Negative Euclidean distance	Is angular similarity needed?
Class prototype	Class representation	Mean embedding per class	Does aggregating enrolment images help?
CLIP embedding	Embedder	CLIP ViT-B/32 image encoder	Does a stronger general embedder help?
Softmax rejection	Decision rule	Logistic regression + maximum probability	Does a trained rule beat similarity?
Enrolment size	Images per class	$k = 1, 3, 5, 10$	How many enrolment images are needed?
Perturbations	Test crops	Brightness $\times 0.4$ and $\times 1.6$; blur (radius 3); occlusion 25% and 50%	How fragile is the embedding to image degradation?
Scene variants	Localization and decision	Ground-truth boxes; whole scene; detector label	What do localization and the detector's label contribute?

Implementation Details and Hyperparameters

All components are used as released and all settings were fixed before the experiments and not tuned on test data; the only quantity tuned on data is the threshold τ , and it is tuned on validation data. Embedding extraction used the GPU, and the timing study used the CPU only, with the EfficientNetB0 and the detector forced onto the CPU. Table Error! Unknown switch argument. lists the settings and Table Error! Unknown switch argument. the environment, which was recorded automatically by the experiment code.

Table VII. Implementation and Hyperparameter Configuration

Category	Parameter	Value	Rationale
Detector	Weights and confidence	YOLOv8n; confidence 0.25	Smallest size; library default confidence
Detector	Box matching (scene)	$\text{IoU} \geq 0.5$	Standard detection criterion
Detector	Minimum box side; small-object filter	8 px; area $\geq 32 \times 32$ px and sides ≥ 24 px	Excludes degenerate crops and objects too small to embed
Embedder	Backbone and input	EfficientNetB0, ImageNet weights; 224×224 bilinear, raw 0–255	Native resolution; compact 1280-D embedding
Embedder	Batch size	64 (32 images per detector batch)	Throughput on the GPU

Category	Parameter	Value	Rationale
Matching	Score	Maximum cosine similarity (1-NN)	Simplest similarity rule
Enrolment	Images per class	5 (pool of 10; 1, 3, 5, 10 studied)	Small number a site can realistically provide
Partitions	Per authorized class	20 validation, 20 test images	Disjoint from enrolment
Partitions	Unregistered classes	10 for validation, 10 for test	Disjoint class sets
Threshold	Selection	Youden's index on validation	Balanced error costs
Baseline	Logistic regression	C = 10, up to 3000 iterations	Fixed a priori, not tuned
Baseline	CLIP model	openai/clip-vit-base-patch32	Widely used general embedder
Robustness	Perturbations	Brightness $\times 0.4$, $\times 1.6$; blur radius 3; black-box occlusion 25%, 50% at random position	Simple synthetic degradations
Timing	Stream	300 frames of 640×480 ; 10 warm-up frames; N_skip = 1 and 5	Repeatable synthetic stream
Statistics	Repeats and bootstrap	Seeds 0–4; 2000 bootstrap resamples	Repeated splits; percentile intervals
Reproducibility	Random seeds	Global seed 0; split seeds 0–4	Deterministic splits

Table VIII. Hardware and Software Environment

Component	Specification
CPU	Intel(R) Xeon(R) CPU @ 2.00GHz; 2 physical / 4 logical cores
RAM	31.3 GB
GPU (embedding extraction only)	NVIDIA Tesla T4
Operating system	Linux-6.12.90+-x86_64-with-glibc2.35
Python	3.12.13
TensorFlow / Keras (EfficientNetB0)	2.20.0
PyTorch (CLIP, detector backend)	2.10.0+cu128
Ultralytics (YOLOv8n)	8.4.160
OpenCV	4.13.0
NumPy / SciPy / scikit-learn	2.0.2 / 1.16.3 / 1.6.1

Statistical Protocol

Every experiment is repeated on five random splits, and results are reported as the mean and the sample standard deviation over the splits, which reflects variation in class composition and image sampling. For the primary split (seed 0) 95% percentile bootstrap intervals are computed from 2,000 resamples of the test samples. Comparisons between variants use the same splits, so paired per-split differences are reported; with only five splits no hypothesis tests are performed and no p-values are quoted. The protocol is summarized in Table **Error! Unknown switch argument..**

Table IX. Statistical Protocol

Element	Configuration	Justification
Repetitions	Five random splits (seeds 0–4) for every experiment	Captures variation in class composition and sampling
Summary statistic	Mean $\pm$ sample standard deviation over splits	Transparent spread instead of a single split
Confidence intervals	95% percentile bootstrap, 2000 resamples of test samples (primary split)	Non-parametric; no distributional assumption
Threshold	Chosen on validation data only	Prevents optimistic bias from tuning on test data

Element	Configuration	Justification
Primary metric	AUROC (threshold-free); secondary metrics at the validation-selected threshold	Separates ranking quality from operating-point choice
Variant comparison	Paired per-split differences on identical splits	Removes split-to-split variation from the comparison
Significance testing	Not performed	Five splits are too few for reliable p-values

RESULTS AND DISCUSSION

This section reports the empirical validation of ISS-RUO in the order of the four questions posed in the introduction: leakage-free performance on object-centred images, threshold behaviour and per-class analysis, split-to-split variability, ablations and baselines, enrolment size and robustness, scene-level verification, and computational cost, followed by a summary of what each component contributes and a discussion of when the approach is justified. Unless stated otherwise, values are the mean $\pm$ standard deviation over five random splits, and every threshold is the one chosen on validation data.

Primary Results on Object-Centred Images

Table **Error! Unknown switch argument.** reports the main result and Fig. **Error! Unknown switch argument.** shows the ROC and precision–recall curves and the score distributions of the primary split. Over the five class splits ISS-RUO reaches an AUROC of 0.892 ± 0.017 , an average precision of 0.944 ± 0.008 , a balanced accuracy of 0.800 ± 0.017 , an F1-score of 0.833 ± 0.032 , and an MCC of 0.584 ± 0.043 . Precision (0.901 ± 0.046) exceeds recall (0.782 ± 0.092): the objects that are accepted are usually authorized, but roughly one authorized object in five is rejected, which is expected when five enrolment images must cover the whole appearance range of a category. On the primary split the validation-selected threshold ($\tau = 0.414$) yields 277 true positives, 123 false negatives, 11 false positives, and 189 true negatives, that is, 11 of 200 unregistered images are accepted and 123 of 400 authorized images are rejected. Authorized test images scored 0.499 ± 0.140 (range 0.238–0.861) and unregistered test images 0.280 ± 0.094 (range 0.076–0.615); the two ranges overlap, so no threshold separates them perfectly (Fig. **Error! Unknown switch argument.c**).

Table X. Results of ISS-RUO on Object-Centred Images (Caltech-256): Mean $\pm$ SD over Five Class Splits and 95% Bootstrap Intervals on the Primary Split

Metric	Mean $\pm$ SD, five splits	Primary split [95% CI]
AUROC	0.892 ± 0.017	0.906 [0.881, 0.928]
Average precision	0.944 ± 0.008	0.953 [0.937, 0.966]
Balanced accuracy	0.800 ± 0.017	0.819 [0.791, 0.845]
Accuracy	0.794 ± 0.029	0.777 [0.743, 0.810]
Precision	0.901 ± 0.046	0.962 [0.938, 0.983]
Recall	0.782 ± 0.092	0.693 [0.647, 0.737]
F1-score	0.833 ± 0.032	0.805 [0.772, 0.836]
MCC	0.584 ± 0.043	0.602 [0.546, 0.657]

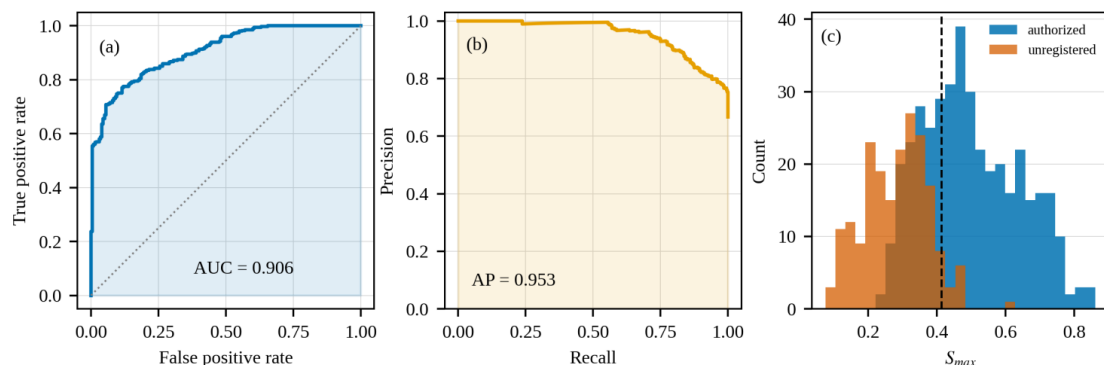


Fig. 3. Primary split (object-centred): (a) ROC curve, (b) precision–recall curve, and (c) distribution of S_{max} for authorized and unregistered test images, with the validation-selected threshold as a dashed line; computed from the 600 per-sample test scores.

Threshold Behaviour

The threshold selected on validation data averaged 0.381 ± 0.037 and ranged from 0.318 to 0.414 across the five splits (Fig. **Error! Unknown switch argument.b**). A fixed threshold of $\tau = 0.75$, which is a natural default for a cosine similarity, is not usable with frozen EfficientNetB0 embeddings[5]: it yields an accuracy of 0.361 ± 0.013 because the

similarities of authorized test images average only 0.50. On the primary split, $\tau = 0.75$ accepts no unregistered image (false-positive rate 0.000) but rejects 96.3% of the authorized ones (Table **Error! Unknown switch argument.**). The threshold that suits an embedder is therefore a property of that embedder: with CLIP embeddings the same fixed value gives an accuracy of 0.791 ± 0.032 . Fig. **Error! Unknown switch argument.**a shows that, on the validation-selected side of the curve, moving τ upwards trades a small reduction in false alarms for a rapidly growing false-negative rate, so an operator who wants fewer false alarms can raise τ at a known cost in missed authorized objects.

Table XI. Threshold Sensitivity on the Primary Split

τ	Validation balanced acc.	Test FPR	Test FNR	Test balanced acc.
0.414(validation-selected)	0.801	0.055	0.307	0.819
0.50	0.721	0.005	0.550	0.723
0.60	0.642	0.005	0.745	0.625
0.70	0.564	0.000	0.890	0.555
0.75 (fixed default)	0.526	0.000	0.963	0.519
0.80	0.509	0.000	0.985	0.507
0.90	0.500	0.000	1.000	0.500

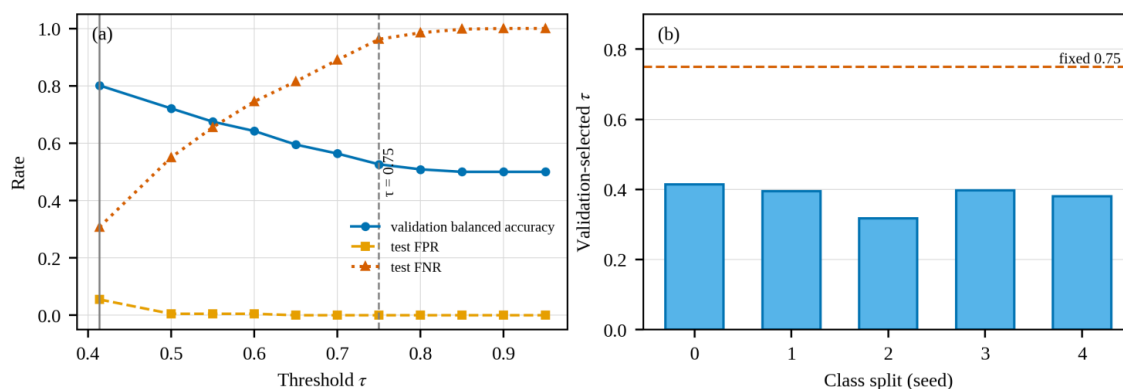


Fig. 4. Threshold analysis: (a) validation balanced accuracy and test false-positive and false-negative rates against τ for the primary split; (b) validation-selected τ for the five class splits compared with the fixed value 0.75.

Similarity Structure and Class-Level Analysis

Fig. **Error! Unknown switch argument.** shows the mean cosine similarity between test images of each authorized category and the enrolment images of every category on the primary split. The mean intra-class similarity (diagonal) is 0.341 and the mean inter-class similarity (off-diagonal) is 0.079, with a maximum of 0.205 (the three highest pairs are mussels against bat (0.205), tennis-court against windmill (0.190), and giraffe against bat (0.188)), so different authorized categories are well separated in the embedding space, while images of the same category are only moderately similar to their enrolled examples. These are averages over all enrolment images of a category; the score used for the decision, $S_{\max}$, takes the best-matching enrolment image and is therefore higher (0.499 on average for authorized test images).

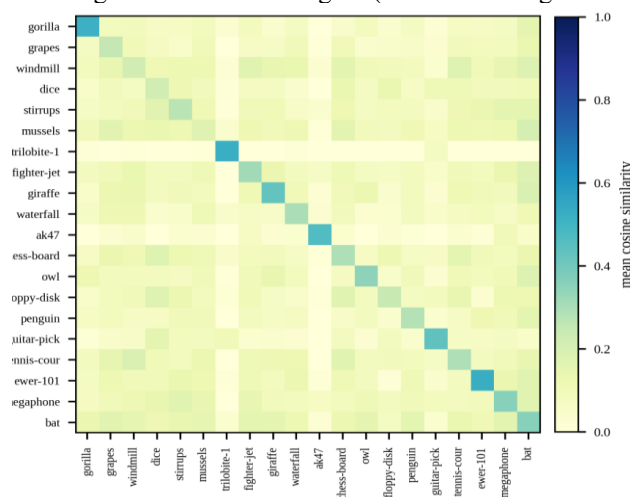


Fig. 5. Mean cosine similarity between test images of each authorized category (rows) and the enrolment images of every category (columns), primary split; the diagonal is intra-class similarity.

Table **Error! Unknown switch argument.** and Fig. **Error! Unknown switch argument.** break the primary result down by class. Recall of authorized categories ranges from 0.15 (dice, floppy disk, and mussels) to 1.00 (ewer and gorilla); five categories have a recall of at most 0.40 (dice, floppy disk, mussels, stirrups, and windmill), and they are also the categories with the lowest mean scores, which indicates that their five enrolment images do not cover the variation among their test images. Among the ten unseen test categories the false-positive rate is at most 0.15 (basketball hoop) and three categories produce no false alarm at all, so false alarms are rare and are spread over several categories rather than concentrated in one. Poor recall on a few categories, rather than confusion with unseen categories, is thus the main source of error at the validation-selected threshold.

Table XII. Per-Class Results on the Primary Split at the Validation-Selected Threshold (20 Test Images per Class)

Group	Category	Mean Smax $\pm$ SD	Rate at τ
Authorized	gorilla	0.600 $\pm$ 0.088	recall 1.00
Authorized	ewer-101	0.609 $\pm$ 0.086	recall 1.00
Authorized	giraffe	0.578 $\pm$ 0.094	recall 0.90
Authorized	megaphone	0.556 $\pm$ 0.122	recall 0.90
Authorized	trilobite-101	0.690 $\pm$ 0.136	recall 0.90
Authorized	owl	0.466 $\pm$ 0.062	recall 0.85
Authorized	tennis-court	0.503 $\pm$ 0.105	recall 0.85
Authorized	ak47	0.615 $\pm$ 0.154	recall 0.85
Authorized	waterfall	0.504 $\pm$ 0.112	recall 0.80
Authorized	penguin	0.490 $\pm$ 0.082	recall 0.80
Authorized	bat	0.462 $\pm$ 0.074	recall 0.80
Authorized	guitar-pick	0.612 $\pm$ 0.186	recall 0.80
Authorized	grapes	0.477 $\pm$ 0.110	recall 0.75
Authorized	chess-board	0.506 $\pm$ 0.138	recall 0.75
Authorized	fighter-jet	0.443 $\pm$ 0.067	recall 0.65
Authorized	windmill	0.395 $\pm$ 0.081	recall 0.40
Authorized	stirrups	0.400 $\pm$ 0.071	recall 0.40
Authorized	dice	0.360 $\pm$ 0.075	recall 0.15
Authorized	floppy-disk	0.381 $\pm$ 0.102	recall 0.15
Authorized	mussels	0.336 $\pm$ 0.057	recall 0.15
Unseen (test)	basketball-hoop	0.346 $\pm$ 0.094	FPR 0.15
Unseen (test)	coffin	0.357 $\pm$ 0.048	FPR 0.10
Unseen (test)	tricycle	0.280 $\pm$ 0.094	FPR 0.10
Unseen (test)	yarmulke	0.260 $\pm$ 0.082	FPR 0.05
Unseen (test)	ketch-101	0.183 $\pm$ 0.084	FPR 0.05
Unseen (test)	fire-truck	0.183 $\pm$ 0.093	FPR 0.05
Unseen (test)	welding-mask	0.344 $\pm$ 0.056	FPR 0.05
Unseen (test)	cake	0.308 $\pm$ 0.048	FPR 0.00
Unseen (test)	spoon	0.301 $\pm$ 0.052	FPR 0.00
Unseen (test)	tomato	0.241 $\pm$ 0.057	FPR 0.00

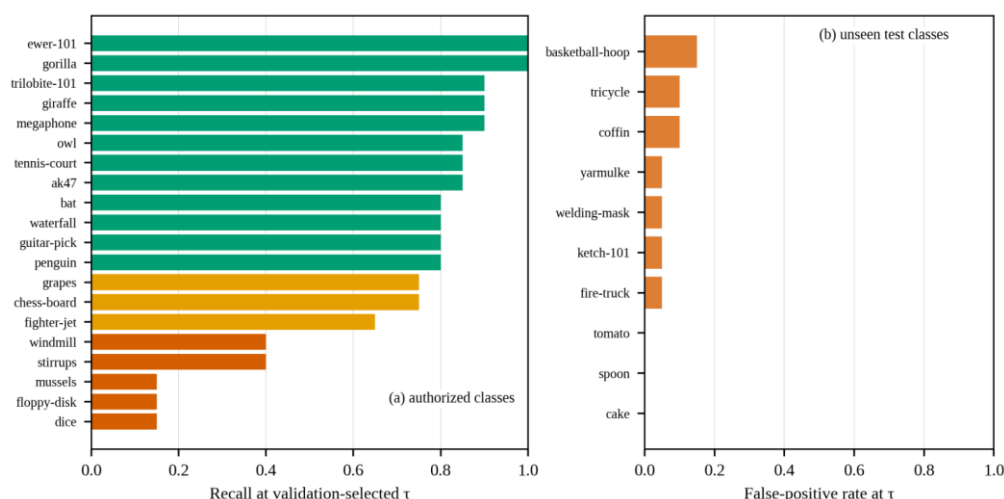


Fig. 6. Per-class behaviour on the primary split: (a) recall of each authorized category (bar colour: green, recall of at least 0.8; orange, 0.5 to 0.8; red, below 0.5) and (b) false-positive rate of each unseen test category at the validation-selected threshold.

Split-to-Split Variability and Paired Comparisons

Table **Error! Unknown switch argument.** and Fig. **Error! Unknown switch argument.** show that the operating-point metrics vary much more between splits than the threshold-free AUROC. AUROC stays between 0.869 and 0.912 (SD 0.017), whereas recall ranges from 0.693 to 0.935 (SD 0.092) and specificity from 0.665 to 0.945 (SD 0.108). The reason is that the validation-selected threshold differs between splits (0.318 to 0.414), and a lower threshold moves the operating point towards higher recall and lower specificity along an almost unchanged ROC curve; split 2, for example, selects the lowest threshold (0.318) and shows the highest recall (0.935) and the lowest specificity (0.665). Results at a single operating point should therefore be read together with their spread across splits, and a deployment should select the operating point on validation data from its own site.

Table XIII. ISS-RUO Results for Each Class Split

Split (seed)	τ	AUROC	Balanced acc.	Recall	Specificity	F1-score	MCC
0	0.414	0.906	0.819	0.693	0.945	0.805	0.602
1	0.395	0.869	0.775	0.780	0.770	0.823	0.529
2	0.318	0.912	0.800	0.935	0.665	0.889	0.641
3	0.397	0.891	0.814	0.738	0.890	0.823	0.593
4	0.381	0.882	0.792	0.765	0.820	0.825	0.557
Mean $\pm$ SD	0.381 $\pm$ 0.037	0.892 $\pm$ 0.017	0.800 $\pm$ 0.017	0.782 $\pm$ 0.092	0.818 $\pm$ 0.108	0.833 $\pm$ 0.032	0.584 $\pm$ 0.043

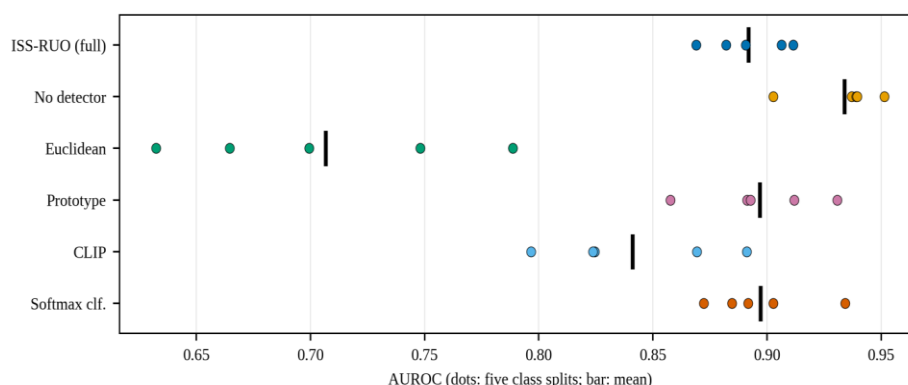


Fig. 7. AUROC of ISS-RUO and its variants on each of the five class splits (dots) with the mean (bar).

Because all variants are evaluated on identical splits, differences can be judged split by split (Table **Error! Unknown switch argument.**). Removing the detector improves AUROC on 5 of 5 splits (mean +0.042); Euclidean distance is worse on all five splits (mean -0.185); CLIP embeddings are worse on all five (mean -0.051); class prototypes (mean +0.005) and the softmax-rejection classifier (mean +0.005) differ from ISS-RUO by less than the split-to-split standard deviation and do not win on every split.

Table XIV. Paired Per-Split Differences Relative to Full ISS-RUO (Positive = Variant Better)

Variant vs. full ISS-RUO	Δ AUROC, mean $\pm$ SD	Range over splits	Splits better (AUROC)	Δ Balanced acc.	Δ F1
No detector (whole image)	$+0.042 \pm 0.022$	+0.021 to +0.070	5 of 5	+0.042	+0.048
Euclidean distance	-0.185 ± 0.052	-0.250 to -0.123	0 of 5	-0.145	-0.282
Class prototype	$+0.005 \pm 0.019$	-0.015 to +0.030	3 of 5	+0.015	-0.004
CLIP embedding	-0.051 ± 0.022	-0.072 to -0.020	0 of 5	-0.026	-0.030
Softmax rejection	$+0.005 \pm 0.011$	-0.006 to +0.023	3 of 5	+0.019	-0.004

Ablations and Baselines

Table Error! Unknown switch argument., Fig. Error! Unknown switch argument., and Fig. Error! Unknown switch argument. summarize the ablations and baselines. Four observations stand out. First, cosine similarity is clearly better than Euclidean distance (AUROC 0.892 ± 0.017 versus 0.707 ± 0.063), which suggests that the direction of an embedding is more informative than its magnitude. Second, class prototypes (0.897 ± 0.027) and nearest-neighbour matching perform alike, so aggregating five enrolment images into one prototype neither helps nor hurts. Third, the linear classifier with softmax rejection reaches essentially the same AUROC as ISS-RUO (0.897 ± 0.023 versus 0.892 ± 0.017); the embedding approach therefore does not outperform a simple trained baseline in accuracy, and its advantage lies in needing no training when a category is added. Fourth, the CLIP embedding scores lower than EfficientNetB0 (0.841 ± 0.038) under the same decision rule, so a larger general-purpose embedder is not automatically better for this task. Removing the detector improved every metric (whole-image AUROC 0.934 ± 0.018). Caltech-256 images are already object-centred, YOLOv8n found a box in only 65.3% of them, and its boxes can crop away context; these are plausible explanations that were not tested individually, but the empirical conclusion is clear: on object-centred images the detector does not help.

Table XV. Ablations and Baselines on Object-Centred Images: Mean $\pm$ SD over Five Class Splits

Variant	AUROC	AP	Balanced acc.	F1-score	MCC
ISS-RUO (YOLO crop, EfficientNetB0, cosine 1-NN)	0.892 ± 0.017	0.944 ± 0.008	0.800 ± 0.017	0.833 ± 0.032	0.584 ± 0.043
Ablation: no detector (whole image)	0.934 ± 0.018	0.967 ± 0.009	0.842 ± 0.028	0.881 ± 0.025	0.671 ± 0.054
Ablation: Euclidean distance	0.707 ± 0.063	0.852 ± 0.037	0.655 ± 0.049	0.551 ± 0.071	0.327 ± 0.102
Ablation: class-prototype cosine	0.897 ± 0.027	0.949 ± 0.015	0.815 ± 0.037	0.829 ± 0.032	0.600 ± 0.069
Baseline: CLIP ViT-B/32 embedding, cosine	0.841 ± 0.038	0.894 ± 0.028	0.773 ± 0.020	0.803 ± 0.024	0.525 ± 0.030
Baseline: linear classifier with softmax rejection	0.897 ± 0.023	0.952 ± 0.011	0.819 ± 0.021	0.829 ± 0.019	0.604 ± 0.038

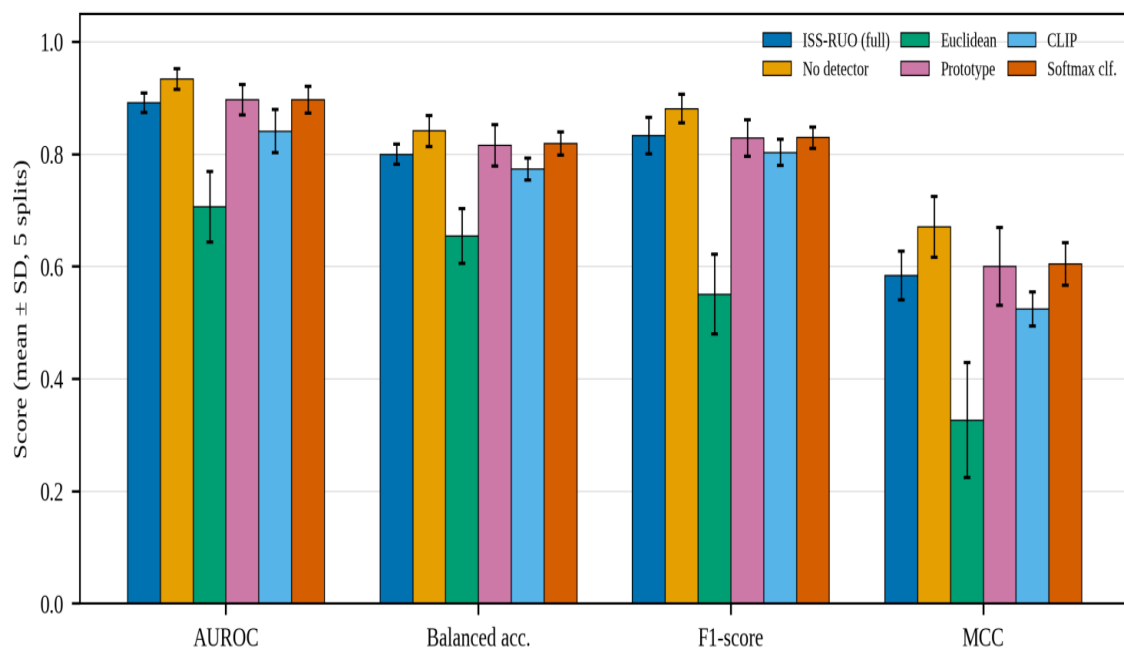


Fig. 8. Ablations and baselines on object-centred images: AUROC, balanced accuracy, F1-score, and MCC (mean $\pm$ SD over five class splits).

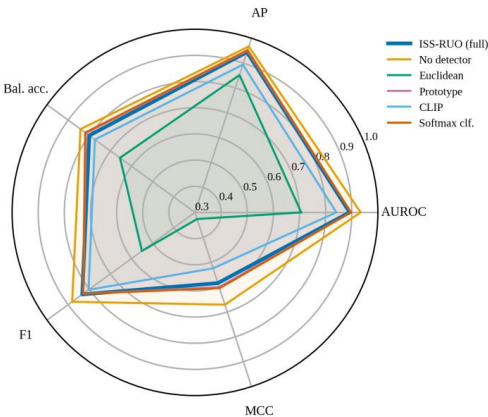


Fig. 9. Multi-metric profile of ISS-RUO and its variants (mean over five class splits) across AUROC, average precision, balanced accuracy, F1-score, and MCC.

Enrolment Size and Robustness

Table **Error! Unknown switch argument.** and Fig. **Error! Unknown switch argument.**a show that the number of enrolment images matters most between one and five images: AUROC rises from 0.769 with one image per category to 0.863 with three, 0.892 with five, and 0.903 with ten, so five images capture most of the achievable gain. Robustness was probed by perturbing the test crops of the primary split before embedding while the enrolment set and the threshold stayed fixed (Table **Error! Unknown switch argument.** and Fig. **Error! Unknown switch argument.**b). Darkening or brightening changes AUROC by about 0.02, whereas blur with radius 3 and occlusion of half of the object reduce AUROC to 0.76 and 0.68, and blur raises the false-positive rate from 0.055 to 0.180. Occlusion at 50% mainly lowers the true-positive rate (0.215), that is, occluded authorized objects are flagged as unregistered, which in a security setting is the safer error but produces many false alarms. Synthetic perturbations are only a proxy for real degradation, so these numbers indicate direction and rough size rather than expected field performance.

Table XVI. Effect Of The Number Of Enrolment Images Per Category (Mean $\pm$ Sd Over Five Class Splits)

Images per class	AUROC	Balanced accuracy	F1-score
1	0.769 $\pm$ 0.059	0.682 $\pm$ 0.057	0.755 $\pm$ 0.047
3	0.863 $\pm$ 0.020	0.766 $\pm$ 0.007	0.787 $\pm$ 0.056
5	0.892 $\pm$ 0.017	0.800 $\pm$ 0.017	0.833 $\pm$ 0.032
10	0.903 $\pm$ 0.018	0.805 $\pm$ 0.038	0.853 $\pm$ 0.016

Table XVII. Robustness of ISS-RUO to Perturbations of the Test Crops (Primary Split, Fixed Enrolment and Threshold)

Condition	AUROC	Balanced acc.	TPR	FPR
clean	0.906	0.819	0.693	0.055
dark (x0.4)	0.883	0.796	0.642	0.050
bright (x1.6)	0.882	0.784	0.637	0.070
blur (r=3)	0.757	0.675	0.530	0.180
occlusion 25%	0.844	0.745	0.580	0.090
occlusion 50%	0.679	0.583	0.215	0.050

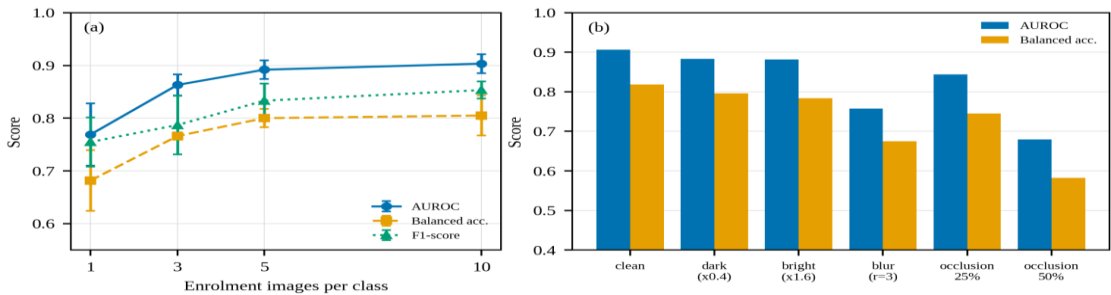


Fig. 10. (a) AUROC, balanced accuracy, and F1-score against the number of enrolment images per category (mean $\pm$ SD over five class splits); (b) AUROC and balanced accuracy under perturbations of the test crops (primary split).

Scene-Level Verification

Table **Error! Unknown switch argument.** and Fig. **Error! Unknown switch argument.** report the COCO scene study, in which 10,873 detections matched annotated objects. Localization is essential in scenes: ISS-RUO with detector crops reaches an AUROC of 0.900 against 0.760 when the whole-scene embedding is assigned to every object, because a single embedding of a multi-object scene cannot say which object is authorized. The detector, however, matched only 59.2% of annotated authorized objects and 60.5% of the other annotated objects (small, occluded, or hard-to-detect objects), so the end-to-end recall of authorized objects is 0.485: about half of all authorized objects are never verified because the detector does not find them. Precision is low (0.438) because only about 15% of the evaluated detections are authorized, so even a modest false-positive rate produces many false alarms in absolute terms. The ground-truth-box variant scores lower than detector crops (AUROC 0.848) because it evaluates every annotated object, including those the detector never found (small and heavily occluded ones), so it is not an upper bound on the same units; the comparison is informative about difficulty, not about the detector's box quality.

Table XVIII. Scene-Level Verification on COCO val2017: Mean $\pm$ SD over Five Image Splits

Variant	AUROC	Balanced acc.	Precision	Recall	F1-score	MCC	End-to-end recall
ISS-RUO (YOLO crops, EfficientNetB0, cosine)	0.900 $\pm$ 0.023	0.815 $\pm$ 0.020	0.438 $\pm$ 0.080	0.826 $\pm$ 0.044	0.567 $\pm$ 0.060	0.502 $\pm$ 0.059	0.485 $\pm$ 0.020
Ground-truth boxes (all annotated objects)	0.848 $\pm$ 0.024	0.764 $\pm$ 0.024	0.363 $\pm$ 0.047	0.771 $\pm$ 0.027	0.492 $\pm$ 0.044	0.406 $\pm$ 0.048	—
No localization: whole-scene embedding	0.760 $\pm$ 0.021	0.690 $\pm$ 0.023	0.272 $\pm$ 0.022	0.718 $\pm$ 0.069	0.393 $\pm$ 0.021	0.277 $\pm$ 0.031	—
Baseline: detector class-label rule	—	0.966 $\pm$ 0.004	0.919 $\pm$ 0.012	0.946 $\pm$ 0.006	0.932 $\pm$ 0.009	0.920 $\pm$ 0.011	—

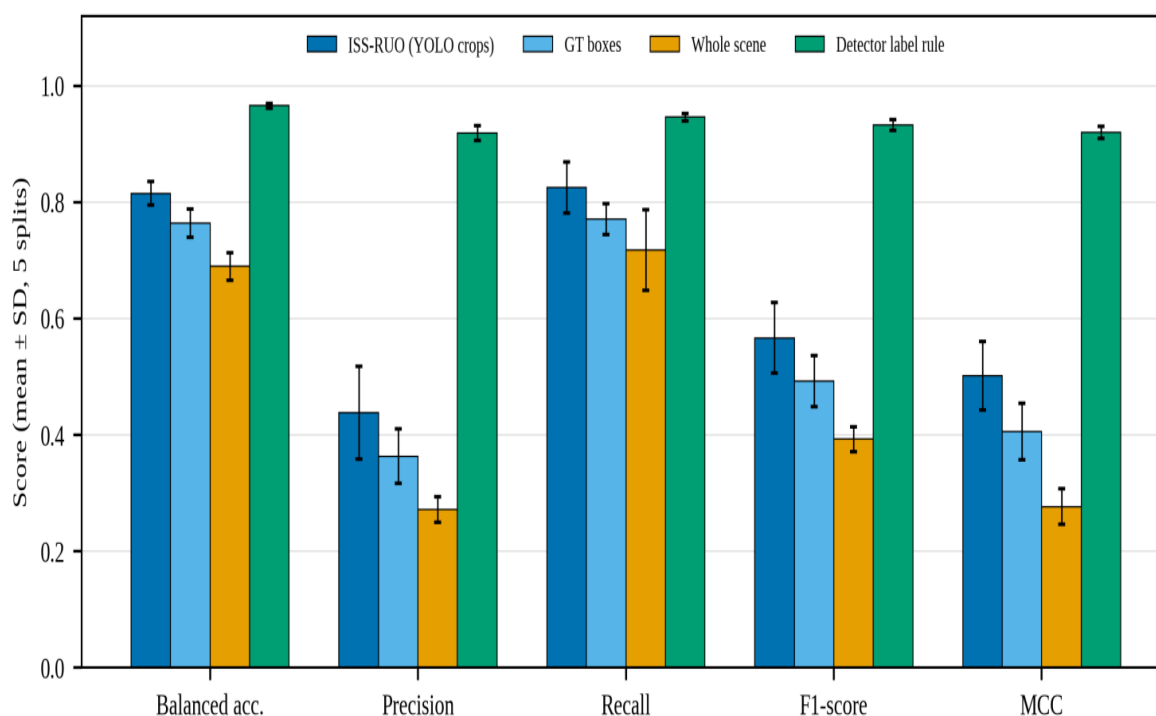


Fig. 11. Scene-level results on COCO val2017: balanced accuracy, precision, recall, F1-score, and MCC of ISS-RUO, ground-truth boxes, whole-scene embedding, and the detector class-label rule (mean $\pm$ SD over five image splits).

The most important result of the scene study is the detector class-label rule: it reaches an F1-score of 0.932 and an MCC of 0.920, far above the 0.567 and 0.502 of embedding verification, because the eight authorized categories are themselves detector classes and the detector's own label is a highly accurate authorization signal for them. This does not make embedding verification useless, but it fixes its domain: it is justified where the authorized items are not classes the detector can name, and it is dominated by the label rule where they are. Fig. **Error! Unknown switch argument.** shows the ROC curve, the precision–recall curve, and the score distributions of ISS-RUO on the first image split (AUROC 0.909, AP

0.689); the overlap of the two score distributions explains the large number of false alarms among the many unregistered detections.

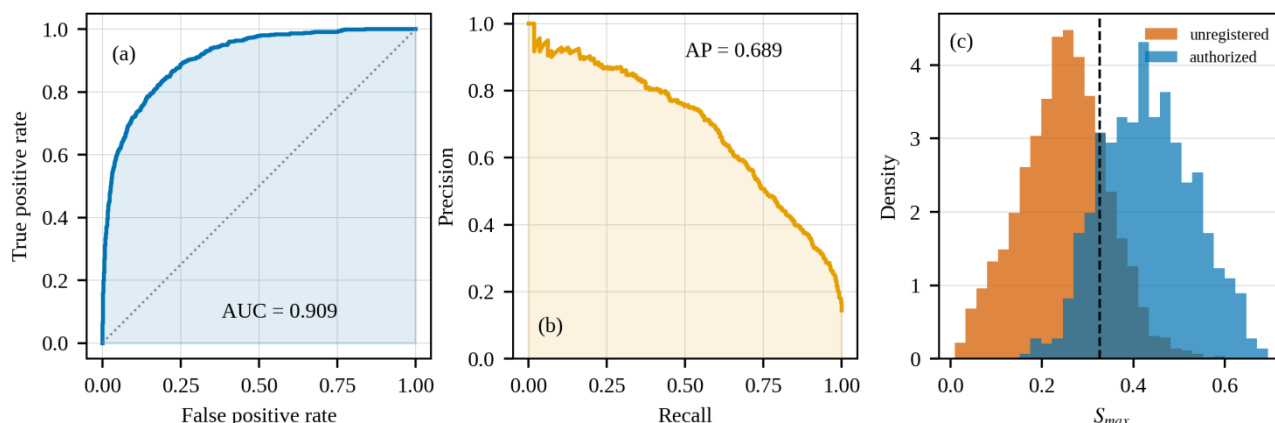


Fig. 12. ISS-RUO on COCO val2017, first image split: (a) ROC curve, (b) precision–recall curve, and (c) density of S_{max} for authorized and unregistered detections with the validation-selected threshold (dashed).

Table **Error! Unknown switch argument.** breaks the first image split down by category. Recall varies from 0.78 for umbrellas to 1.00 for bicycles across the eight authorized categories, so no category is systematically missed. False alarms are dominated by categories that are visually or functionally close to an authorized category or that occur very frequently: chairs cause the most (71 false alarms), whereas trucks, motorcycles, and televisions have the highest rates (0.60, 0.60, and 0.68) and resemble the authorized cars, bicycles, and laptops. This is an observation about the categories, not a tested explanation, but it shows that the unregistered class that matters most operationally is the one that looks like an authorized one, and that a stricter threshold or additional enrolment images of the confusable categories would be needed to control it. Example decisions are shown in Fig. **Error! Unknown switch argument.**.

Table XIX. Scene-Level Results by Category on the First Image Split (Authorized Categories and the Ten Unregistered Categories with Most False Alarms)

Group	Category	Matched detections	Mean S_{max}	Rate at τ	False alarms
Authorized	bicycle	71	0.529	TPR 1.00	—
Authorized	suitcase	47	0.405	TPR 0.87	—
Authorized	handbag	39	0.397	TPR 0.85	—
Authorized	car	272	0.432	TPR 0.83	—
Authorized	cell phone	32	0.412	TPR 0.81	—
Authorized	backpack	26	0.383	TPR 0.81	—
Authorized	laptop	62	0.429	TPR 0.79	—
Authorized	umbrella	69	0.405	TPR 0.78	—
Unregistered	chair	229	0.295	FPR 0.31	71
Unregistered	truck	91	0.358	FPR 0.60	55
Unregistered	motorcycle	90	0.341	FPR 0.60	54
Unregistered	tv	72	0.365	FPR 0.68	49
Unregistered	surfboard	57	0.327	FPR 0.53	30
Unregistered	couch	80	0.307	FPR 0.35	28
Unregistered	skateboard	40	0.350	FPR 0.60	24
Unregistered	bottle	105	0.270	FPR 0.21	22
Unregistered	dining table	140	0.253	FPR 0.15	21
Unregistered	baseball glove	26	0.359	FPR 0.77	20

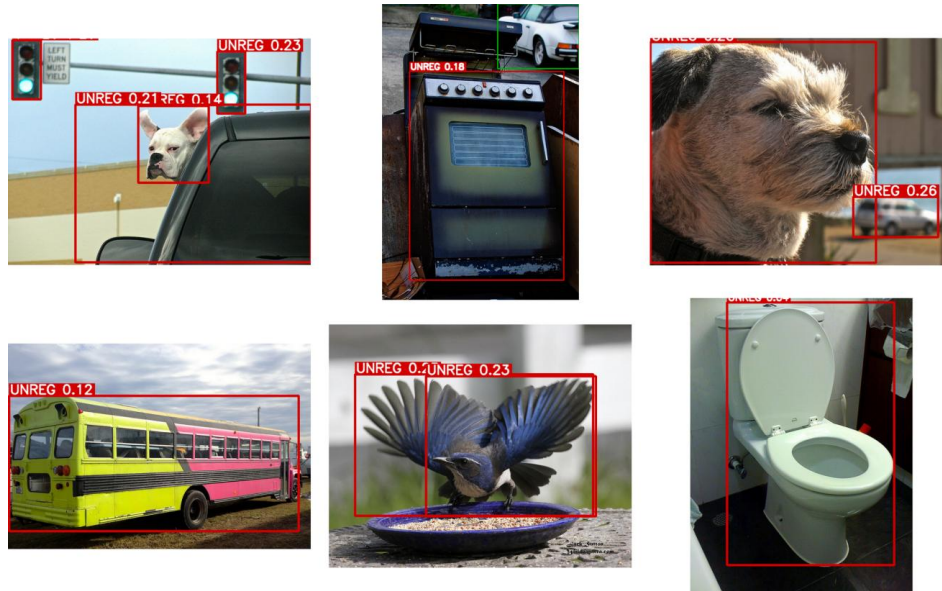


Fig. 13. Example COCO test scenes with ISS-RUO decisions (green: AUTH, red: UNREG, with the similarity score). The top row shows scenes that contain authorized objects, the bottom row scenes with unregistered objects only; some authorized objects (for example the car in the top-right scene) fall below the threshold and are shown as UNREG. Scenes containing people were excluded from this figure.

Computational Cost

Table **Error! Unknown switch argument.** and Fig. **Error! Unknown switch argument.** give the per-stage time on the CPU for 300 frames of 640×480 pixels. YOLOv8n takes 67 ms per frame and the EfficientNetB0 embedding dominates with 204 ms on average (312 ms median), because frames without a detected object skip the embedding stage and frames with several objects embed each of them; similarity search is negligible (0.1 ms). The total is 272 ± 182 ms per processed frame, so about 3.7 frames/s can be processed, which is not real-time for a 30 frames/s stream. Processing all 300 frames took 81.6 s (151.8 CPU-seconds), against 17.6 s (33.5 CPU-seconds) with $N_{\text{skip}} = 5$, reductions of 78% in wall time and 78% in CPU time; average CPU utilization was about 48% in both cases. Keeping pace with a 30 frames/s stream at this per-frame cost would require $N_{\text{skip}} \geq 9$, that is, one processed frame every 300 ms. The search cost grows linearly with database size, from 0.13 ms at 1,000 to 1.26 ms at 10,000 and 18.2 ms at 100,000 enrolled embeddings (Fig. **Error! Unknown switch argument.**b), so exhaustive search remains practical for databases of several thousand embeddings.

Table XX. Per-Stage CPU Time per Processed Frame in Milliseconds (300 Frames of 640×480 Pixels, Two-Core CPU)

Stage	$N_{\text{skip}} = 1$, mean $\pm$ SD	Median	95th percentile	$N_{\text{skip}} = 5$, mean $\pm$ SD
YOLOv8n detection	66.7 ± 8.9	65.8	71.2	65.7 ± 8.6
Crop and resize	1.0 ± 1.7	0.4	3.9	1.1 ± 1.5
EfficientNetB0 embedding	204.1 ± 181.0	311.7	439.5	226.4 ± 191.1
Similarity search	0.1 ± 0.2	0.1	0.2	0.1 ± 0.1
Total per processed frame	271.9 ± 182.4	378.2	520.0	293.3 ± 193.2

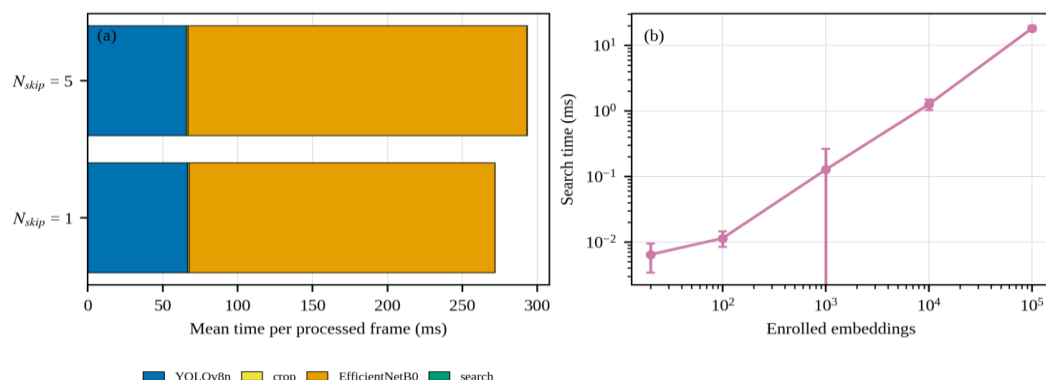


Fig. 14. Computational cost: (a) mean time per processed frame by stage for $N_{\text{skip}} = 1$ and 5; (b) similarity-search time against the number of enrolled embeddings (mean $\pm$ SD of 200 repetitions).

Summary of Component Contributions

Table **Error! Unknown switch argument.** consolidates what each component of the pipeline contributes, as measured in the experiments, and what follows for practice.

Table XXI. Summary of Component Contributions: Finding, Evidence, and Practical Implication

Component or setting	Finding	Evidence	Implication
Detector on object-centred images	Lowers AUROC (0.892 with crops versus 0.934 without); box found in 65.3% of images	Tables Error! Unknown switch argument. , Error! Unknown switch argument. ; worse on all 5 splits	Embed the whole image when the object fills the frame
Detector in scenes	Essential for localization (AUROC 0.900 versus 0.760) but finds only about 60% of objects	Table Error! Unknown switch argument.	Detector recall bounds end-to-end recall (0.485)
Similarity measure	Cosine far better than Euclidean (AUROC 0.892 versus 0.707)	Table Error! Unknown switch argument.	Use angular similarity for GAP embeddings
Class representation	Prototype equals 1-NN (0.897 versus 0.892)	Tables Error! Unknown switch argument. , Error! Unknown switch argument.	Either is adequate for five images per class
Embedder and trained baseline	CLIP lower (0.841); softmax classifier equal (0.897)	Table Error! Unknown switch argument.	Embedding rule matches a trained rule without training
Enrolment size	AUROC 0.769 (1 image) to 0.892 (5) and 0.903 (10)	Table Error! Unknown switch argument.	Enrol at least three to five images per class
Threshold	Validation-selected 0.381 ± 0.037 ; fixed 0.75 gives accuracy 0.361	Table Error! Unknown switch argument. , Fig. Error! Unknown switch argument.	Tune τ per embedder and site on validation data
Robustness	Blur and 50% occlusion lower AUROC to 0.76 and 0.68	Table Error! Unknown switch argument.	Add image-quality control or occlusion handling
Detector label rule	F1 0.932 versus 0.567 for embeddings when categories are detector classes	Table Error! Unknown switch argument.	Use embeddings only for categories outside the detector vocabulary
Computational cost	272 ms per processed frame; frame skipping saves 78% CPU time	Table Error! Unknown switch argument.	Use $N_{\text{skip}} \geq 9$ on this CPU for a 30 frames/s stream

DISCUSSION

Taken together, the two studies show where an enrolment-based authorization layer helps and where it does not. On object-centred images, a frozen embedding with a validation-tuned cosine threshold separates authorized categories from unseen ones moderately well (AUROC about 0.89) and matches a trained linear classifier with softmax rejection, while needing no training when a category is added and no detector at all: the whole-image variant is even better. In multi-object scenes a detector is indispensable for localization, but it also caps performance, since it found only about 60% of annotated objects and its vocabulary is fixed. When the authorized categories are among the detector's own classes, the detector's label is a much stronger authorization signal than embedding similarity, so embedding verification should not be used for them.

The practical niche of ISS-RUO is therefore narrower than the initial intuition suggests but real: authorization of items that a general detector cannot name, in settings where retraining is impractical and a few enrolment images per item can be collected, such as site-specific equipment or bags that are registered by photograph[5]. In such a setting the object-centred results are the more relevant ones, they indicate a moderate but usable separability, and they show that the operating point must be chosen on site-specific validation data, that at least three to five enrolment images per item are needed, and that blur and heavy occlusion will produce false alarms. The instance-level version of the problem, in which a specific registered object must be distinguished from visually similar ones of the same category, was not evaluated and is the most important open question.

Several threats to validity should be kept in mind. Both datasets consist of still photographs, not surveillance video, and no temporal information such as tracking was used. Authorization was evaluated at category level. Only one detector size and one frozen embedder were tested, and five splits do not support significance tests. The per-class and per-category explanations offered above are observations, not tested causes. Finally, the robustness tests used synthetic perturbations and the timing study a synthetic stream on a two-core cloud CPU.

CONCLUSION

This paper studied ISS-RUO, a retraining-free authorization layer that combines a YOLOv8 detector[3], frozen EfficientNetB0 embeddings, and a validation-tuned cosine threshold, and evaluated it with disjoint enrolment, validation, and test partitions on object-centred images and on multi-object scenes. On unseen object categories the layer reaches an AUROC of about 0.89 with balanced accuracy of 0.80, matches a linear classifier with softmax rejection, outperforms CLIP embeddings and Euclidean matching, and needs only a few enrolment images per category, with no training when categories are added. A fixed threshold of 0.75 fails, and the validation-selected threshold, which lies near 0.38, must be re-tuned for every embedder and site.

The study also delimits the approach. On object-centred images the detector lowers performance, in scenes it finds only about 60% of the objects and so limits end-to-end recall to about half, blur and heavy occlusion cause many false alarms, and, when the authorized categories are among the detector's own classes, the detector's label is a far stronger authorization signal than embedding similarity. On a small two-core CPU the pipeline processes about four frames per second, so real-time use on a 30 frames/s stream requires processing roughly every ninth frame. Enrolment-based embedding verification is therefore justified where authorization concerns items that a general detector cannot name and retraining is impractical; elsewhere simpler rules are preferable. The results establish these conditions empirically and provide a leakage-free protocol against which stronger methods can be compared.

Future Enhancements

Several directions follow from the results. The foremost is instance-level authorization, in which a specific registered object must be distinguished from visually similar objects of the same category; this is the setting in which embeddings can add information that a category label cannot, and it requires datasets with repeated captures of individual objects. Embeddings trained explicitly for verification, and stronger self-supervised or vision-language features, should be tested in place of the frozen ImageNet embedder, together with detectors of higher recall or class-agnostic proposal methods that remove the dependence on a fixed detector vocabulary. Robustness to blur and occlusion might be improved by multi-crop or part-based embeddings and by aggregating decisions over several frames of the same tracked object, which would also exploit the temporal information that this study did not use.

On the systems side, quantized or distilled embedders and approximate nearest-neighbour indexes would reduce the per-frame cost and extend the approach to large databases and edge devices, and adaptive thresholds calibrated on-line from operator feedback would address the dependence of the operating point on the site. Evaluations should move from still photographs to real surveillance video with human-verified ground truth, include user studies with security operators, and report false-alarm rates per hour of video, which is the quantity that determines whether such a system is acceptable in operation.

LIMITATIONS AND ETHICAL CONSIDERATIONS

Caltech-256 and COCO are collections of still photographs, not surveillance video, and no temporal information such as tracking is used. Authorization is evaluated at category level, not instance level; only YOLOv8n and one frozen embedder were tested; and five splits allow no significance testing. In the scene study, unmatched detections are ignored, so false alarms on unannotated objects are not counted, and person authorization was not evaluated. Timing was measured on a two-core cloud CPU with 640×480 frames assembled from still images, not on edge hardware or camera streams, and the robustness tests use synthetic perturbations of crops, not real illumination or occlusion changes.

ISS-RUO processes video that can contain identifiable people, and its logging step stores frame captures of unregistered events. Categories named people or face were removed from the object-centred data, person annotations were excluded from the scene study, and the example scenes shown contain no people. Any real deployment must follow applicable data-protection law (for example, India's Digital Personal Data Protection Act, 2023) and requires notice and consent where applicable, a defined retention period for logs, redaction of faces and licence plates in stored frames, and human review of alerts before action is taken against individuals. The false-alarm behaviour reported here should be taken into account before any automated action is linked to an alert.

Code and Data Availability

The datasets used in this study, Caltech-256 and COCO, are publicly available from their respective official sources. The experimental code, configuration files, class-split definitions, per-sample evaluation scores, and data required to reproduce the reported figures and results will be made available in a public repository upon publication. Until the repository is publicly accessible, these materials are available from the corresponding author upon reasonable request.

REFERENCES

1. V. Chundi, J. Bammidi, A. Pegallapati, Y. Parnandi, A. Reddithala, and S. K. Moru, "Intelligent Video Surveillance Systems," in 2021 International Carnahan Conference on Security Technology (ICCST), Hatfield, United Kingdom: IEEE, Oct. 2021, pp. 1–5. doi: 10.1109/ICCST49569.2021.9717400.
2. A. Anish, S. R. A. H. Malini, and T. Archana, "Enhancing Surveillance Systems with YOLO Algorithm for Real-Time Object Detection and Tracking," in 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India: IEEE, Dec. 2023, pp. 1254–1257. doi: 10.1109/ICACRS58579.2023.10404710.
3. N. Azatbekuly, A. Mukhanbet, and S. D. Bekele, "Development of an Intelligent Video Surveillance System Based on YOLO Algorithm," in 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST), Astana, Kazakhstan: IEEE, May 2024, pp. 498–503. doi: 10.1109/SIST61555.2024.10629617.

4. Y.-C. Zeng and J.-M. Guo, "Object Checking Framework Using a Set of Multi-Category Classifiers with Designed-Augmented Data," in 2025 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA: IEEE, Jan. 2025, pp. 1–6. doi: 10.1109/ICCE63647.2025.10929906.
5. Y.-J. Chen, J.-H. Chen, and R. R. Vicerra, "Design and Development of a Recognition System for Personnel and Safety Equipment in Construction Sites," in 2024 IEEE 16th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Baguio, Philippines: IEEE, Nov. 2024, pp. 1–5. doi: 10.1109/HNICEM64917.2024.11258748.
6. S. Sarkar, T. Bin Khayer, N. H. Kisan, and M. N. Uddin, "Thermogram-based Regions With Convolutional Neural Network (RCNN) and Facial Biometrics for Safe Driving," in 2023 3rd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), Dhaka, Bangladesh: IEEE, Jan. 2023, pp. 207–211. doi: 10.1109/ICREST57604.2023.10070063.
7. Z. Mao, B. Li, R. Zhang, and Q. Fei, "DE-YOLO: Detail-Enhanced Maritime Object Detection Algorithm Based on YOLOv8," in 2025 37th Chinese Control and Decision Conference (CCDC), Xiamen, China: IEEE, May 2025, pp. 3605–3610. doi: 10.1109/CCDC65474.2025.11090244.
8. X. Yu, T. W. Kuan, Y. Zhang, and T. Yan, "YOLO v5 for SDSB Distant Tiny Object Detection," in 2022 10th International Conference on Orange Technology (ICOT), Shanghai, China: IEEE, Nov. 2022, pp. 1–4. doi: 10.1109/ICOT56925.2022.10008164.
9. Z. Zhuang, P. Liu, D. Xu, and J. Cheng, "YOLO-KED: A Novel Framework for Rotated Object Detection in Complex Environments," in ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India: IEEE, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10888504.
10. X. Qu, Y. Zheng, Y. Zhou, and Z. Su, "YOLO v8_CAT: Enhancing Small Object Detection in Traffic Light Recognition with Combined Attention Mechanism," in 2024 10th International Conference on Computer and Communications (ICCC), Chengdu, China: IEEE, Dec. 2024, pp. 706–710. doi: 10.1109/ICCC62609.2024.10941967.
11. Y.-C. Zeng and J.-M. Guo, "Implement of Dual Recognition Processes for Comprehensive Object Recognition and Checking," in 2024 IEEE 13th Global Conference on Consumer Electronics (GCCE), Kitakyushu, Japan: IEEE, Oct. 2024, pp. 789–792. doi: 10.1109/GCCE62371.2024.10760795.
12. C. Tian, Z. Zhou, Y. Huang, G. Li, and Z. He, "Cross-Modality Proposal-Guided Feature Mining for Unregistered RGB-Thermal Pedestrian Detection," *IEEE Trans. Multimed.*, vol. 26, pp. 6449–6461, 2024, doi: 10.1109/TMM.2024.3350926.
13. K. Song, H. Wen, X. Xue, L. Huang, Y. Ji, and Y. Yan, "Modality Registration and Object Search Framework for UAV-Based Unregistered RGB-T Image Salient Object Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023, doi: 10.1109/TGRS.2023.3332179.
14. M. Cui, X. Wen, and Q. Yan, "FBDNet: Frequency-Domain Fusion and Bidirectional Decoding Network for Unregistered RGB-T Salient Object Detection," in 2025 7th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI), Hangzhou, China: IEEE, Nov. 2025, pp. 992–996. doi: 10.1109/RICAI68060.2025.11385061.
15. Z. Ye, Y. Peng, B. Han, H. Hao, and W. Liu, "Unmanned Aerial Vehicle Target Detection Algorithm Based on Infrared Visible Light Feature Level Fusion," in 2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE), Guangzhou, China: IEEE, May 2024, pp. 1–5. doi: 10.1109/CISCE62493.2024.10653241.
16. L. Stanescu and C. Stoica-Spahi, "AI-Based Evaluation of Transfer Learning Strategies for Robust Histopathology Image Classification," in 2025 11th International Conference on Computer and Communications (ICCC), Chengdu, China: IEEE, Dec. 2025, pp. 1196–1201. doi: 10.1109/ICCC68654.2025.11437918.
17. G. Sumbul, M. Muller, and B. Demir, "A Novel Self-Supervised Cross-Modal Image Retrieval Method in Remote Sensing," in 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France: IEEE, Oct. 2022, pp. 2426–2430. doi: 10.1109/ICIP46576.2022.9897475.
18. A. Soni, Anju, and A. Kumar, "Deep Learning-Based Image Retrieval: Addressing the Semantic Gap for Accurate Content-Based Image Retrieval," in 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India: IEEE, Aug. 2023, pp. 1512–1517. doi: 10.1109/ICCPCT58313.2023.10245342.
19. W. Wu, J. Ye, J. Chen, D. Chen, and R. Wang, "Multi-OS: Multimodal OOD Synthesis Enhances Out-of-Distribution Detection for Vision-Language Models," in ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain: IEEE, May 2026, pp. 12737–12741. doi: 10.1109/ICASSP55912.2026.11464872.
20. H. Guanghui, G. Hemantha Kumar, V. N. Manjunath Aradhya, and Punithavalli, "A Few-Shot Learning for HAR Using 3D Siamese Network," in 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India: IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/ICCCNT61001.2024.10725743.
21. Q. Huang and Q. Xie, "Research on Segmentation and Recognition of Oracle Topographies Based on Deep Learning and Image Processing," in 2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China: IEEE, May 2024, pp. 1303–1308. doi: 10.1109/ICETCI61221.2024.10594137.
22. J. Huang, D. Prijatelj, J. Dulay, and W. Scheirer, "Measuring Human Perception to Improve Open Set Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 11382–11389, Sep. 2023, doi: 10.1109/TPAMI.2023.3270772.
23. M. Vendramini, H. Oliveira, A. Machado, and J. A. D. Santos, "Opening Deep Neural Networks With Generative Models," in 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA: IEEE, Sep. 2021, pp. 1314–1318. doi: 10.1109/ICIP42928.2021.9506672.

24. D. K. Singh et al., “New Objects on the Road? No Problem, We’ll Learn Them Too,” in 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan: IEEE, Oct. 2022, pp. 1972–1978. doi: 10.1109/IROS47612.2022.9981886.
25. J. Zhao, P. Wang, W. Du, Y. Zhou, R. Yao, and A. E. Saddik, “Causal Unbiased Feature Intervention for Open World Object Detection,” *IEEE Trans. Multimed.*, pp. 1–11, 2026, doi: 10.1109/TMM.2026.3705513.
26. J. Cao, J. Li, S. Wang, J. Gao, R. Wang, and Z. Zhang, “DT-ProtoYOLO: Mitigating Overconfidence in Open-Set Object Detection via Dual-Linear Thresholding and Prototype Regularization,” in 2026 41st Youth Academic Annual Conference of Chinese Association of Automation (YAC), Changsha, China: IEEE, May 2026, pp. 1863–1870. doi: 10.1109/YAC71005.2026.11615451.
27. K. Oksuz, T. Joy, and P. K. Dokania, “Towards Building Self-Aware Object Detectors via Reliable Uncertainty Quantification and Calibration,” in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 9263–9274. doi: 10.1109/CVPR52729.2023.00894.
28. J. Xu, M. Kovatsch, and S. Lucia, “Open Set Recognition for Machinery Fault Diagnosis,” in 2021 IEEE 19th International Conference on Industrial Informatics (INDIN), Palma de Mallorca, Spain: IEEE, Jul. 2021, pp. 1–7. doi: 10.1109/INDIN45523.2021.9557572.
29. S. S. Gaddipati, T. Lakshmi, N. Krishnan, S. Govindarajan, A. Hadi, and S. K. R. Katta., “An Imbalance-Resilient Deep Classification Framework for Reliable Analytics in Real-World Data Streams,” in 2026 International Conference on Intelligent and Sustainable AI Systems (ICOSAAS), Bali, Indonesia: IEEE, Jul. 2026, pp. 1268–1275. doi: 10.1109/ICOSAAS68663.2026.11648676