Preprints.org

Article

Not peer-reviewed version

Generative Inpainting for Privacy-Preserving Industrial CCTV: Cross-Domain Hard-Hat Compliance Detection on a Temporally Separated Dataset

Remigiusz Baran *, Andrzej Dziech, Natalia Brzezińska, Rafał Gajos

Posted Date: 22 September 2026

Keywords: hard-hat compliance detection; industrial CCTV; object detection; generative image inpainting; image restoration; privacy-preserving image editing; domain gap; temporal split; LaMa; industrial evaluation dataset; AI readiness

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Generative Inpainting for Privacy-Preserving Industrial CCTV: Cross-Domain Hard-Hat Compliance Detection on a Temporally Separated Dataset

Remigiusz Baran [1,*], Andrzej Dziech [2], Natalia Brzezińska [1] and Rafał Gajos [1]

[1] Department of Computer Science, Electronics and Electrical Engineering, Kielce University of Technology, Kielce, Poland
[2] The Institute of Telecommunications and Cybersecurity, AGH University of Krakow, Krakow, Poland
* Correspondence: r.baran@tu.kielce.pl

Abstract

Hard-hat compliance detection is a key component of vision-based industrial safety monitoring. Public hard-hat datasets support model development but rarely reflect real industrial closed-circuit television (CCTV) conditions, such as fixed elevated viewpoints, small object scale, temporal correlation between frames, partial occlusion, and privacy-sensitive content. We address three questions: whether public datasets generalize to industrial CCTV, why temporal separation matters for video-derived data, and whether privacy-preserving preprocessing degrades detection. We introduce the Industrial CCTV-HH Dataset, an AI-ready, privacy-preserving object detection resource built from real surveillance recordings in an operational industrial facility, containing 2413 images and 9608 annotated head-region instances in two classes (head without helmet, head with helmet). A day-level temporal split reduces near-duplicate leakage and yields deployment-oriented evaluation. A preprocessing pipeline uses a generative image-inpainting model (LaMa) to remove on-screen overlays and textual identifiers and adds localized face anonymization while preserving head structure, illustrating generative image editing for privacy-preserving dataset preparation. Cross-domain and mixed-domain experiments with a public source dataset reveal a large domain gap: a model trained only on public data reaches mAP@0.5 $\approx$ 0.97 on public images but only $\approx$ 0.11 on industrial CCTV, whereas adding target-domain samples and using higher input resolution raise industrial-domain mAP@0.5 to $\approx$ 0.90. The privacy-preserving pipeline maintains detection utility while reducing exposure of sensitive information.

Keywords: hard-hat compliance detection; industrial CCTV; object detection; generative image inpainting; image restoration; privacy-preserving image editing; domain gap; temporal split; LaMa; industrial evaluation dataset; AI readiness

1. Introduction

Vision-based monitoring systems are increasingly used to support occupational safety in industrial environments [1,2]. One of the most common applications is hard-hat compliance detection, which identifies whether workers visible in camera footage are wearing protective helmets. Reliable detection can assist human operators and enable automated analysis of surveillance video streams [3].

Most hard-hat detection models are developed and evaluated using public image datasets [4,5]. These datasets are valuable for initial training, transfer learning, and reproducible benchmarking, but they often differ substantially from real industrial CCTV data. Public collections typically contain web images, close-range construction-site photographs, or images captured under relatively favorable conditions. Industrial CCTV footage, in contrast, is characterized by fixed and often elevated viewpoints, small head regions, perspective distortion, partial occlusions, motion blur, background clutter, and varying worker-to-camera distances. Consequently, high accuracy on public benchmarks may not translate into reliable performance in deployment.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

A second challenge arises from the temporal structure of CCTV recordings. Frames extracted from continuous video are strongly correlated, and consecutive frames may be nearly identical. A random image-level split can therefore place near-duplicate samples in both training and test subsets, producing split-related leakage and overly optimistic estimates [6]. Deployment-oriented evaluation should instead enforce temporal separation between training and test data.

A third challenge is privacy. Real surveillance footage contains non-semantic and privacy-sensitive content such as timestamps, camera or hall identifiers, vehicle markings, phone numbers, and visible faces. This information is not required for hard-hat compliance detection and may introduce shortcut cues or privacy risks, so dataset preparation should include privacy-preserving preprocessing and sanitization [7,8]. Methodologically, we address this with a generative image-inpainting model [9], reconstructing sanitized regions rather than masking them; this situates the work at the intersection of generative models for image processing and deployment-oriented detection, applying generative image editing to a downstream task rather than to image synthesis for its own sake. Collectively, overcoming these three hurdles—domain shift, temporal data leakage, and privacy risks—is essential for achieving high AI readiness of industrial video data. Transforming raw, unstructured CCTV footage into an AI-ready dataset requires rigorous spatial sanitization and temporal structuring. This ensures that downstream models learn robust, task-relevant features rather than spurious correlations, while simultaneously complying with data protection standards.

Motivated by these challenges, we frame the study around three research questions:

- **RQ1.** Can privacy-preserving preprocessing be applied without substantially degrading detection performance?
- **RQ2.** Do public hard-hat datasets generalize reliably to real industrial CCTV?
- **RQ3.** Why is temporal separation important when evaluating video-derived CCTV datasets?

To address these questions, we introduce the Industrial CCTV-HH Dataset, a privacy-preserving object detection resource for head-region-based hard-hat compliance detection, created from real CCTV recordings acquired in an operational industrial facility. The dataset contains 2413 images with 9608 annotated instances in two classes, `head_without_helmet` and `head_with_helmet`, and uses a day-level temporal split.

The main contributions of this work are:

1. An anonymized, privacy-preserving industrial CCTV hard-hat evaluation dataset that reflects realistic deployment conditions (fixed viewpoints, small head regions, temporal correlation, occlusion, clutter).
2. A day-level temporal split that groups correlated frames before partitioning and reduces near-duplicate leakage, enabling deployment-oriented evaluation.
3. A privacy-preserving preprocessing pipeline that applies a generative image-inpainting model (LaMa-based on-screen display (OSD)/text removal) together with localized face anonymization, preserving task-relevant head structure—an application of generative image editing to deployment-oriented dataset preparation that ensures the AI readiness of the surveillance footage.
4. A cross-domain evaluation quantifying the domain gap between public hard-hat data and the target industrial CCTV domain.
5. Mixed-domain and scale-aware experiments showing how target-domain data and small-object-oriented training improve industrial-domain detection.

Because the recordings contain privacy-sensitive content from a single operational facility, we do not release the raw imagery; instead, the preprocessing and evaluation code, on-screen-display mask coordinates, face-anonymization parameters, split definitions, and aggregate dataset statistics are available from the corresponding author upon reasonable request, subject to approval by the data provider and a confidentiality agreement, and we frame the contribution as a deployment-oriented evaluation protocol that transfers to other industrial CCTV settings rather than as a public comparison benchmark.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

The remainder of the paper is organized as follows. Section 1.1 reviews related work. Section 2.1 describes the dataset and the privacy-preserving pipeline. Section 2.2 details the experimental setup. Section 3 presents results, Section 4 discusses implications and limitations, and Section 5 concludes.

1.1. Related Work

This section reviews five areas directly tied to the study: hard-hat and personal protective equipment (PPE) detection, public hard-hat datasets, domain gap and cross-domain detection, data leakage in video-derived datasets, and privacy-preserving preprocessing.

1.1.1. Hard-Hat and PPE Detection

Hard-hat and PPE detection have been widely studied as object-detection applications for occupational safety, with recent work spanning multi-class PPE detection supported by dedicated datasets [10] and combined helmet–harness detection on construction sites [11]. One-stage detectors, particularly YOLO-based architectures, are common because of their accuracy–speed trade-off, which suits real-time monitoring [4,12]; the YOLO family has evolved rapidly across successive versions, as summarized in recent reviews [13]. However, many studies evaluate under conditions unlike fixed industrial CCTV (closer views, larger helmets, weakly temporally structured collections), so evaluation on realistic CCTV data remains necessary [5]. Recent work extends PPE detection to surveillance and body-worn camera footage with dedicated benchmark datasets, reinforcing the need for realistic, deployment-oriented evaluation [14].

1.1.2. Public Datasets for Hard-Hat Detection

Public hard-hat datasets support training, transfer learning, and reproducible benchmarking. In this work we use the *Hard Hat Workers* dataset as the public source domain [15]; such datasets may not represent CCTV deployment conditions, differing in viewpoint, scale, image source, background, annotation semantics, and temporal organization, and should therefore be treated as useful source-domain data rather than substitutes for target-domain CCTV data.

1.1.3. Domain Gap and Cross-Domain Object Detection

Domain gap is a common challenge when models trained on one distribution are applied to another [16,17]. Beyond unsupervised domain adaptation, the gap is also addressed through domain-generalization training that targets unseen domains without access to their data [18], and through generative synthesis of labeled, target-like images, including diffusion- and ControlNet-based generation of synthetic construction imagery [19]. For hard-hat detection, the source domain may be public construction or web imagery and the target domain industrial CCTV frames; differences in viewpoint, scale, illumination, background, and annotation style reduce transferability. Cross-domain evaluation typically compares source-only, target-only, and mixed-domain settings, a structure we follow here.

1.1.4. Data Leakage in Video-Derived Datasets

Datasets derived from video require careful splitting because consecutive frames are highly correlated; random frame-level assignment can place near-duplicates in both training and test subsets, inflating performance [6]. Fixed CCTV cameras observe the same scene over time, so correlated samples should be grouped before partitioning; the same principle has been established for spatially autocorrelated data, where naive random splits likewise inflate apparent accuracy [20]. We implement this via a day-level temporal split.

1.1.5. Generative Inpainting and Privacy-Preserving Preprocessing

Surveillance data may contain information irrelevant to the target task but sensitive in nature: OSD overlays, timestamps, camera identifiers, vehicle markings, phone numbers, and faces. Faces in crowded surveillance scenes can be detected even under substantial occlusion and head-pose variation [21], which motivates explicit anonymization rather than relying on small scale alone.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

Privacy-preserving preprocessing reduces such risks while preserving task utility, commonly via masking, blurring, or learning-based inpainting [7,22]. For detection, preprocessing must be spatially controlled to preserve task-relevant structure. We remove OSD and textual identifiers using LaMa-based inpainting [9] and anonymize faces locally. Generative models for image processing—including generative adversarial networks [23] and, more recently, diffusion models [24,25]—have advanced image synthesis, restoration, and inpainting. LaMa belongs to this family: it is a generative image-inpainting model that reconstructs masked regions using fast Fourier convolutions, and we adopt it as a generative image-editing tool for privacy-preserving sanitization rather than for image synthesis. Diffusion-based inpainting (e.g., RePaint [26]) can achieve high-quality reconstructions but is iterative and computationally expensive; we therefore use LaMa as a fast, single-pass model for sanitizing the entire dataset, consistent with the deployment-oriented focus of this work. Generative approaches to face anonymization have also been proposed, such as GAN-based de-identification [27] and more recent reversible generative anonymization that aims to preserve downstream utility [28]; given the very small face scale in our footage, we instead apply localized blurring that preserves head-region geometry, connecting this study to the broader use of generative models for image processing.

2. Materials and Methods

This section describes the dataset and the privacy-preserving preprocessing pipeline (Section 2.1) as well as the experimental protocol used in the study (Section 2.2).

2.1. The Industrial CCTV-HH Dataset

This section describes the Industrial CCTV-HH Dataset, a privacy-preserving object detection resource created from video surveillance recordings collected in a real, operational industrial facility, and targeting hard-hat compliance detection under realistic CCTV conditions.

The dataset contains 2413 images and 9608 annotated instances in two mutually exclusive classes: `head_without_helmet` (a visible human head without a protective helmet) and `head_with_helmet` (a visible human head wearing a helmet, with the bounding box covering the head together with the helmet rather than the helmet alone). The task is head-region detection for hard-hat compliance only; the system performs no facial recognition, no biometric identification, and no re-identification or tracking of individuals. Unlike many public datasets composed mainly of web or construction-site images, this dataset reflects operational surveillance conditions. The following subsections present data acquisition, annotation protocol, class definitions, temporal split, class distribution, and the privacy-preserving pipeline.

2.1.1. Data Acquisition

The dataset was created from surveillance recordings acquired by fixed-position CCTV cameras in an operational industrial facility during a multi-day recording campaign (Figure 1). The cameras monitored areas where workers and other personnel moved during routine activities. The source material was converted into individual frames, which were subsequently annotated for object detection.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

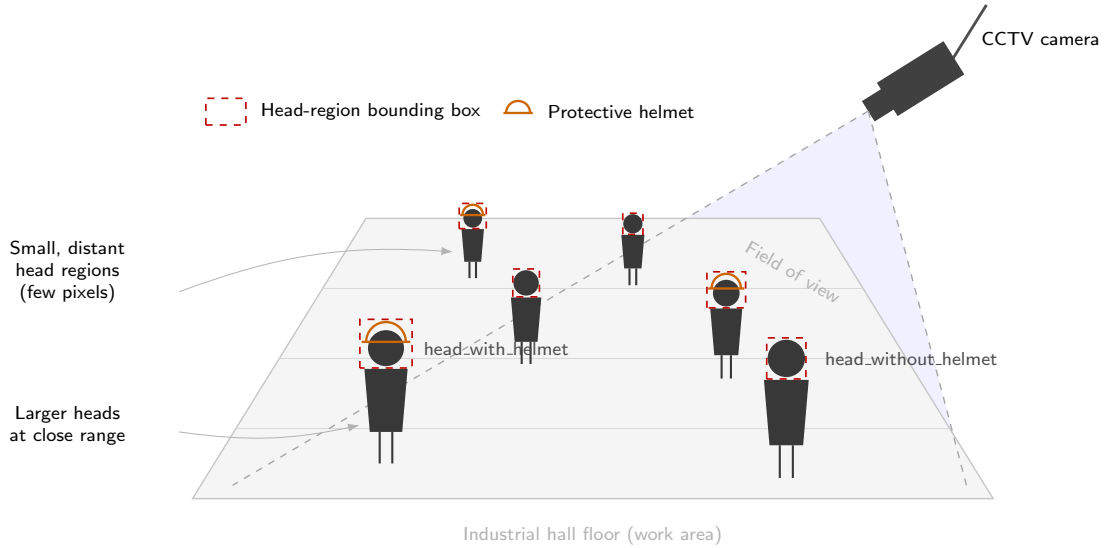


Figure 1. Schematic of the data-acquisition setup: a fixed, elevated CCTV camera observes an industrial hall, so workers appear at a range of distances and head regions are small—especially for distant workers. Head-region bounding boxes enclose the head together with the helmet (`head_with_helmet`) or the bare head (`head_without_helmet`). The figure is a schematic drawing and contains no real surveillance footage.

The dataset is moderate in size; its value lies in the realistic industrial CCTV conditions it represents rather than in scale: fixed viewpoints, small head regions, background clutter, partial occlusions, and temporal dependency between consecutive frames. Acquisition was not performed under controlled laboratory conditions; instead, naturally occurring variation in pose, illumination, visibility, and scene occupancy was retained. Temporal dependency was treated as an intrinsic dataset property and addressed through the day-level split (Section 2.1.4). Before experiments, frames were processed with the privacy-preserving pipeline (Section 2.1.6).

The facility, geographic location, camera and hall identifiers, and exact acquisition dates are intentionally omitted from the core dataset description.

2.1.2. Annotation Protocol

Frames were manually annotated by two authors experienced in computer-vision dataset preparation and safety-related visual inspection. Annotations follow the YOLO format, with each file containing normalized bounding-box coordinates (*class_id*, *x_center*, *y_center*, *width*, *height*). Each instance was assigned to one of the two head-region classes defined in Section 2.1.3.

For `head_without_helmet`, the box covers the visible head without a helmet. For `head_with_helmet`, the box covers the head together with the helmet, not the helmet alone. Partially visible heads were annotated only when helmet status could be reliably determined; severely occluded or ambiguous cases were excluded. Ambiguous cases were resolved by consensus, and final annotations were cross-checked to reduce label inconsistencies, bounding-box inaccuracies, and duplicates. To quantify annotation reliability, both annotators independently labelled the images, and the two label sets were compared at an IoU matching threshold of 0.5. Agreement was high: a box-level F_1 of 0.99 (9502 matched boxes, with only 106 and 91 boxes annotated by a single annotator), a mean IoU of 0.89 for matched boxes, and effectively perfect class-label agreement on matched boxes (raw agreement 100%, Cohen's $\kappa = 1.00$); this perfect class agreement is consistent with the annotation protocol, in which heads whose helmet status could not be reliably determined were excluded, leaving only unambiguous helmet/no-helmet instances. The remaining discrepancies were reconciled by consensus for the final annotations.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

2.1.3. Class Definitions

The dataset uses two mutually exclusive head-region classes because the target application is hard-hat compliance monitoring, where the relevant unit of analysis is the worker's head and its protection status. The task is therefore head-region-based hard-hat compliance detection rather than standalone helmet detection (Table 1).

Table 1. Class definitions used in the Industrial CCTV-HH Dataset.

Class ID	Class Name	Description
0	`head_without_helmet`	Visible human head without a protective helmet.
1	`head_with_helmet`	Visible human head wearing a protective helmet; the bounding box covers the head region together with the helmet.

2.1.4. Temporal Split

CCTV recordings exhibit strong temporal correlation between consecutive frames, which often share scene, viewpoint, background, and objects, differing only slightly in object position or pose. A random image-level split may place near-duplicates in both training and evaluation subsets, causing split-related leakage and optimistic estimates [6].

To mitigate this, the dataset is divided using a day-level temporal split: all frames from a given recording day are assigned to the same subset. This follows the general principle of grouping correlated samples before partitioning (Table 2).

Table 2. Day-level temporal split of the Industrial CCTV-HH Dataset. Calendar dates are replaced by relative day labels (D1–D4).

Subset	Recording Day(s)	Number of Images
Training	D1–D2	1144
Validation	D3	224
Test	D4	1045
Total	D1–D4	2413

The split was fixed before training and kept unchanged across all experiments. The test set corresponds to a chronologically later recording period than the training and validation sets, approximating a realistic deployment scenario in which a model is applied to later, previously unseen data.

2.1.5. Class Distribution

The dataset contains 9608 annotated instances: 6198 `head_without_helmet` (64.51%) and 3410 `head_with_helmet` (35.49%). The class imbalance is moderate and reflects the natural occurrence of helmet and non-helmet cases in the recorded environment (Table 3).

Table 3. Class distribution across subsets.

Subset	Images	head_without_helmet	head_with_helmet	Total Objects
Training	1144	2642	1664	4306
Validation	224	780	280	1060
Test	1045	2776	1466	4242
Total	2413	6198	3410	9608

Both classes are represented in all subsets. The distribution is not artificially balanced, since the dataset aims to reflect realistic operational conditions; evaluation therefore captures both the visual difficulty of CCTV imagery and the effect of naturally imbalanced class occurrence. The counts in

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

Table 3 correspond to the full annotation set preserved by the LaMa-based E3a variant; the black-mask OSD variants (E2a, E2b) contain slightly fewer boxes due to IoA-based label cleaning (Section 3.1).

2.1.6. Privacy-Preserving Preprocessing

Because the dataset originates from real surveillance recordings, raw frames contained non-semantic artifacts and potentially sensitive information: fixed OSD overlays (timestamps, hall/camera identifiers), scene-level textual identifiers (e.g., phone numbers, vehicle markings), and visible faces (mostly at very small scale). The pipeline comprises three stages: OSD removal, textual-identifier anonymization, and localized face anonymization (Figure 2). Its goal is to remove or obscure information not required for hard-hat compliance detection while preserving scene structure and head-region information.

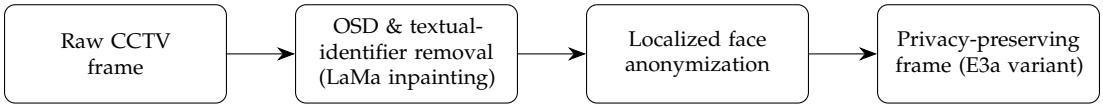


Figure 2. Overview of the privacy-preserving preprocessing pipeline: on-screen-display (OSD) overlays and textual identifiers are removed with a generative image-inpainting model (LaMa), followed by localized face anonymization, yielding the final privacy-preserving variant (E3a) used in all experiments.

OSD Removal and Dataset Sanitization.

Fixed OSD overlays are not semantically related to the targets but may introduce dataset-specific shortcuts; a detector could exploit correlations between OSD content, camera configuration, recording time, or scene identity rather than task-relevant features. OSD removal is therefore an important sanitization step.

Two strategies were considered. The first is *deterministic masking*, replacing fixed rectangular OSD regions with black pixels. Because OSD positions were fixed, two masks were defined in native image coordinates (upper timestamp region, lower hall/camera identifier region) and kept constant. Since black masking may partially cover annotated heads, labels were cleaned afterward: for each box, the intersection-over-area (IoA) with the OSD mask was computed, and boxes whose masked area exceeded a threshold τ were removed. For a bounding box B and an OSD mask region M, the IoA is defined as the fraction of the box area covered by the mask,

$$\mathrm{IoA}(B, M) = \frac{|B \cap M|}{|B|}, \tag{1}$$

so that a box is discarded when $\mathrm{IoA}(B, M) > \tau$. Two thresholds were evaluated: $\tau = 0.30$ (more aggressive) and $\tau = 0.50$ (more conservative). Exact mask coordinates are provided in the preprocessing script for reproducibility.

The second strategy uses *LaMa-based inpainting* [9], a generative image-inpainting model, to remove OSD overlays without introducing black rectangles, reconstructing masked regions for a more natural appearance and avoiding artificial high-contrast areas.

Textual Identifier Anonymization.

Some frames contained potentially identifying scene-level text (e.g., phone numbers, vehicle markings). These were removed using LaMa-based inpainting with binary masks, within the same unified stage as OSD removal. This reduced visible identifying information while preserving surrounding context—important because the targets are small head regions and excessive modification could harm detection.

Face Anonymization.

Visible faces were anonymized to reduce identification risk. Because the task is head-region-based, anonymization obscures facial detail while preserving the head contour. Each face region was slightly expanded, processed with a soft mask to avoid sharp transitions, and Gaussian-blurred within the

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

masked region. This localized strategy was preferred over full-person anonymization to minimize changes to body appearance, context, and head-region geometry. No frames from the industrial recordings are reproduced in this paper, and the dataset itself is not publicly released. In the unlikely event of any restricted sharing, faces would additionally be redacted with an irreversible opaque mask rather than blurred, since blurring and pixelation retain residual information that can be partially inverted by learning-based attacks [29,30].

Faces were generally small: median bounding-box width 11.07 px, 90th percentile 19.48 px. Although low resolution limits identity-related detail, it was not treated as sufficient anonymization by itself; explicit local anonymization was therefore applied in the final variant.

Final Privacy-Preserving Dataset Variant.

Several variants were evaluated: a baseline without OSD sanitization; two deterministic black-mask variants with IoA-based cleaning ($\tau = 0.30$, $\tau = 0.50$); and a LaMa-based inpainting variant. Based on the ablation in Section 3.1, LaMa-based preprocessing was preferred because it removes textual artifacts without introducing black rectangles. The final AI-ready variant, E3a, combines LaMa-based OSD/text removal with localized face anonymization, striking the necessary balance between dataset sanitization and detection utility. Unless stated otherwise, all subsequent experiments use E3a.

2.1.7. Dataset Summary

Table 4 summarizes the main characteristics of the dataset.

Table 4. Summary of the Industrial CCTV-HH Dataset.

Property	Value
Dataset type	Industrial CCTV object detection dataset
Target task	Head-region-based hard-hat compliance detection
Number of images	2413
Number of annotated objects	9608
Classes	`head_without_helmet`, `head_with_helmet`
Annotation format	YOLO
Split strategy	Day-level temporal split
Training images	1144
Validation images	224
Test images	1045
Final dataset variant	E3a
Privacy-preserving preprocessing	LaMa-based OSD/text removal + localized face anonymization

2.2. Experimental Setup

The protocol addresses three objectives aligned with the research questions: assess the impact of sanitization strategies on detection (RQ1), evaluate whether localized face anonymization affects utility (RQ1), and analyze transferability from public hard-hat data to the target industrial CCTV domain (RQ2), under a temporally separated protocol (RQ3).

All experiments use the temporal split from Section 2.1.4. Unless stated otherwise, the E3a variant is used. For fair comparison, the same architecture, input resolution, training protocol, and metrics are used across compared variants. Each experiment is repeated with three random seeds, and results are reported as mean $\pm$ standard deviation.

2.2.1. Detection Model

All experiments use the YOLO11n object detector in the Ultralytics framework [31,32], selected as a lightweight variant suitable for efficient CCTV-oriented monitoring while retaining sufficient localization accuracy for small head regions. The model was initialized from pretrained `yolo11n.pt` weights and fine-tuned per variant to detect the two head-region classes. The same architecture and

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

initialization were used across all comparisons, so performance differences arise from preprocessing, training-data composition, or augmentation rather than architecture. Because the study targets properties of the dataset and the evaluation protocol—the public-to-industrial domain gap, the effect of temporal versus random splitting, and the impact of privacy-preserving preprocessing—rather than a comparison of detector architectures, we deliberately fix a single, modern, lightweight detector as a controlled instrument: holding the architecture, initialization, and training protocol constant ensures that measured differences are attributable to the data and preprocessing rather than to the choice of model. The detector therefore acts as a consistent measurement tool rather than as the object of study, and YOLO11n additionally reflects a realistic deployment choice for resource-constrained CCTV monitoring. The best checkpoint per run was used for test evaluation.

2.2.2. Training Protocol

For each variant, the model was trained on the corresponding training subset and evaluated on the test subset using the best checkpoint. Within each comparison group, architecture, initialization, input resolution, batch size, epochs, and evaluation procedure were fixed. Each experiment used seeds 0, 1, and 2, with results reported as mean ± standard deviation (Table 5).

Table 5. Training configuration used in the experiments.

Parameter	Value
Model	YOLO11n
Initial weights	`yolo11n.pt`
Input image size	640×640
Batch size	16
Number of epochs	50
Random seeds	0, 1, 2
Deterministic mode	Enabled
Training device	2 GPUs
Evaluation device	1 GPU
Evaluation split	Test
Checkpoint used for evaluation	Best checkpoint

2.2.3. Evaluation Metrics

Detection performance was evaluated on the test subset using mAP@0.5, mAP@0.5:0.95, precision, and recall, computed with the Ultralytics validation procedure on the configured test split. The YOLO fitness score is reported as an auxiliary framework-specific summary, but the main comparisons rely on mAP, precision, and recall. Repeated experiments are reported as mean ± standard deviation across three seeds.

A prediction is matched to a ground-truth box using the intersection over union (IoU). For a predicted box B_p and a ground-truth box B_g,

$$\mathrm{IoU}(B_p, B_g) = \frac{|B_p \cap B_g|}{|B_p \cup B_g|}. \tag{2}$$

A detection is counted as a true positive (TP) when its IoU with a ground-truth box of the same class exceeds the evaluation threshold; otherwise it is a false positive (FP), and undetected ground-truth boxes are false negatives (FN). Precision and recall are then

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{3}$$

The average precision (AP) for a class is the area under its precision–recall curve, and the mean average precision (mAP) is the average of the per-class AP over the C classes:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

$$\mathrm{AP} = \int_0^1 p(r)\,dr, \qquad \mathrm{mAP} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{AP}_c. \tag{4}$$

Here mAP@0.5 uses a fixed IoU threshold of 0.5, following the PASCAL VOC convention [33], whereas mAP@0.5:0.95 averages mAP over ten IoU thresholds from 0.5 to 0.95 in steps of 0.05, following the COCO protocol [34].

2.2.4. Experimental Scenarios

Experiments are organized into three groups, aligned with the research questions:

- **Group 1 (RQ1):** impact of privacy-preserving preprocessing—baseline (no OSD removal), deterministic black masking with IoA cleaning, LaMa inpainting, and localized face anonymization.
- **Group 2 (RQ2):** transferability of public-trained models to the target industrial CCTV domain (domain-gap quantification).
- **Group 3 (RQ2):** mixed-domain and scale-aware training to assess whether combining public and target-domain data improves industrial-domain performance.

All scenarios use the same temporal test subset for industrial-domain evaluation. Reported public/target ratios refer to the training subsets; validation subsets follow the same principle and are used only for model selection.

2.2.5. Implementation Details

Experiments were implemented with Ultralytics YOLO [32] on PyTorch [35]. Training and evaluation ran on an Ubuntu Server workstation with two NVIDIA Quadro RTX 6000 GPUs. Weights & Biases was used for experiment tracking; final metrics were exported to CSV for aggregation and statistical reporting. The hardware and software environment is summarized in Table 6. Although the system CUDA toolkit was 11.5, the PyTorch build used CUDA 11.8 (PyTorch 2.7.1+cu118).

Table 6. Hardware and software environment used in the experiments.

Component	Value
Operating system	Ubuntu 22.04.3 LTS
Kernel	5.15.0-76-generic
CPU	Intel Xeon Platinum 8164 @ 2.00 GHz
RAM	251 GiB
GPUs	2× NVIDIA Quadro RTX 6000
GPU memory	24 GB per GPU
NVIDIA driver	525.125.06
CUDA toolkit	11.5
Python	3.11.13
PyTorch	2.7.1+cu118
PyTorch CUDA	11.8
cuDNN	9.1.0
Ultralytics	8.3.161
OpenCV	4.11.0
NumPy	2.1.2
pandas	2.3.0
Weights & Biases	0.21.0

2.2.6. Public Dataset and Label Harmonization

The public source domain is the *Hard Hat Workers* dataset, version 5 (*raw_HeadHelmetClasses*), selected because it provides the raw, non-augmented images and already excludes the auxiliary `person` class [15]. This version uses a 70/20/10 train/validation/test split and defines two object classes, `head` (a bare human head) and `helmet` (a worker wearing a hard hat). The exported subset comprises 7035 images (4916 train, 1413 validation, 706 test).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

The public dataset was exported in YOLO format. Because the dataset uses a head-region-level protocol with two classes (`head_without_helmet`, `head_with_helmet`), the public labels were harmonized to the same two-class target space using a direct one-to-one mapping: the public class `head` was mapped to `head_without_helmet` (class 0), and the public class `helmet` was mapped to `head_with_helmet` (class 1). No class merging or splitting was required, since the auxiliary `person` class is already absent from version 5, and no public images were discarded during harmonization. It should be noted that the public label name `helmet` refers to the annotation category rather than to the helmet object in isolation: in this dataset the corresponding bounding box encloses the head region together with the helmet, following the same head-region-level convention adopted in the present dataset for `head_with_helmet` (Section 2.1.3). The mapping is therefore semantically consistent, and the renaming serves only to make the class meaning explicit and to align the label space across the two domains. To confirm this, we manually inspected a random sample of public-dataset annotations and verified that the `helmet` boxes enclose the head together with the helmet rather than the helmet alone, in agreement with our head-region-level convention; we found no systematic mismatch that would bias the cross-domain comparison.

The public dataset was used in three settings: public-only, mixed public–target, and balanced public–target training, enabling assessment of both the domain gap and the benefit of target-domain samples.

3. Results

Results are organized in four parts: OSD-sanitization impact (Section 3.1), face-anonymization impact (Section 3.2), cross-domain generalization (Section 3.3), and mixed-domain/scale-aware training (Section 3.4), followed by qualitative analysis (Section 3.8).

3.1. Impact of OSD Sanitization (RQ1)

Four preprocessing variants were compared under the same temporal split, YOLO11n architecture, identical training settings, and three seeds (Table 7).

Table 7. Impact of OSD sanitization on detection performance. Results are reported as mean ± standard deviation across three random seeds.

Variant	Description	mAP@0.5	mAP@0.5:0.95	Precision	Recall	Fitness
E1	Baseline, no OSD sanitization	0.800 ± 0.005	0.448 ± 0.005	0.852 ± 0.030	0.686 ± 0.015	0.483 ± 0.005
E2a	Black OSD masks, IoA $\tau = 0.30$	0.791 ± 0.009	0.435 ± 0.006	0.827 ± 0.062	0.694 ± 0.019	0.471 ± 0.005
E2b	Black OSD masks, IoA $\tau = 0.50$	0.784 ± 0.016	0.432 ± 0.013	0.823 ± 0.047	0.691 ± 0.007	0.467 ± 0.013
E3	LaMa inpainting	0.795 ± 0.003	0.442 ± 0.003	0.855 ± 0.016	0.684 ± 0.000	0.478 ± 0.002

The baseline achieved the highest mean mAP@0.5 and mAP@0.5:0.95 in the in-domain evaluation; however, the differences among variants are small and largely overlap within the across-seed standard deviation, so, given only three repetitions, they should be read as trends rather than as statistically conclusive effects. The two black-mask variants showed the largest mean drop relative to the baseline, which may reflect the black rectangles and the associated annotation cleaning altering image statistics or removing useful information near OSD regions. Among the sanitization methods, LaMa inpainting had the mean closest to the baseline and above both black-mask variants. These observations suggest that the choice of OSD-sanitization method can matter: removing non-semantic artifacts is desirable for dataset hygiene, and LaMa inpainting removes overlays while preserving a natural appearance and avoiding the artificial high-contrast regions introduced by black masking. Within the limits of these measurements, a generative inpainting model thus appears to be a reasonable alternative to deterministic masking for privacy-preserving sanitization, editing sensitive content while preserving image statistics relevant to the downstream detector. The recall standard deviation for E3 was below 0.0005 and is shown as 0.000 after rounding.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

3.2. Impact of Face Anonymization (RQ1)

The LaMa variant without face anonymization (E3) was compared with the final E3a variant, which adds localized Gaussian-blur face anonymization. All other settings were identical (Table 8).

Table 8. Impact of localized face anonymization on detection performance. Results are reported as mean ± standard deviation across three random seeds.

Variant	Description	mAP@0.5	mAP@0.5:0.95	Precision	Recall	Fitness
E3	LaMa inpainting	0.795 ± 0.003	0.442 ± 0.003	0.855 ± 0.016	0.684 ± 0.000	0.478 ± 0.002
E3a	LaMa inpainting + face anonymization	0.798 ± 0.012	0.444 ± 0.007	0.835 ± 0.034	0.700 ± 0.001	0.480 ± 0.007

Localized face anonymization did not substantially degrade performance: mean mAP@0.5 and mAP@0.5:0.95 remained comparable, with differences within seed-level variability. E3a showed slightly higher mean recall and lower mean precision, indicating a modest shift in the precision–recall trade-off; given only three repetitions, this should be read as a trend rather than a statistically conclusive effect. E3a was selected as the final privacy-preserving variant for the cross-domain and mixed-domain experiments, improving privacy properties while maintaining utility.

3.3. Cross-Domain Generalization from Public Data to Industrial CCTV (RQ2)

To quantify the domain gap, a model trained only on the public dataset was evaluated on both the public and the target industrial CCTV test subsets (source-only scenario, no target-domain samples in training).

Using the multi-seed results (Table 9), the source-only model achieved very high performance on the public test set (mAP@0.5 $= 0.9737 \pm 0.0003$) but collapsed on the target industrial CCTV test set (mAP@0.5 $= 0.1056 \pm 0.0234$; mAP@0.5:0.95 $= 0.0221 \pm 0.0064$). This collapse is consistent across all three seeds (target standard deviation 0.0234) and was produced with exactly the same evaluation pipeline, class mapping, and confidence settings as the public-test evaluation; qualitative inspection of the source-only predictions confirmed that the errors correspond to genuinely missed or mislocalized heads under industrial CCTV conditions rather than to a labelling or thresholding artifact. This large drop indicates a strong domain gap: public hard-hat data is useful for initial training and benchmarking but is insufficient as a standalone training source for deployment in realistic industrial CCTV. This directly answers RQ2 in the negative.

3.4. Mixed-Domain Training and Scale Adaptation (RQ2)

To assess whether target-domain samples improve industrial-domain performance, we evaluated source-only, target-only, full mixed supervision, domain-balanced training, a higher-resolution balanced variant, and a scale-aware augmentation variant (Table 9).

Table 9. Cross-domain and mixed-domain training results. Results are reported as mean ± standard deviation across three random seeds.

ID	Training Setting	Public/ Target	Input Size	Aug.	Public mAP@0.5	Public mAP@.5:.95	Target mAP@0.5	Target mAP@.5:.95
A	Source only	100/0	640	Default	0.9737 ± 0.0003	0.6925 ± 0.0013	0.1056 ± 0.0234	0.0221 ± 0.0064
D	Target only	0/100	640	Default	0.0069 ± 0.0009	0.0016 ± 0.0000	0.7978 ± 0.0120	0.4442 ± 0.0069
B1	Full mixed supervision	81.1/18.9	640	Default	0.9742 ± 0.0019	0.6921 ± 0.0023	0.6609 ± 0.0088	0.3320 ± 0.0028
B2	Domain-balanced	50/50	640	Default	0.9607 ± 0.0012	0.6609 ± 0.0094	0.7076 ± 0.0175	0.3634 ± 0.0010
C1	Higher-resolution balanced	50/50	960	Default	0.9656 ± 0.0013	0.6702 ± 0.0016	0.8978 ± 0.0037	0.5316 ± 0.0081
C2	Scale + mosaic balanced	50/50	640	[1]	0.9638 ± 0.0033	0.6619 ± 0.0100	0.6810 ± 0.0136	0.3457 ± 0.0066

[1] Augmentation for C2: scale = 0.8, mosaic = 1.0. Input size is given in pixels (square input). "Public/Target" is the training-data ratio (%).

Including target-domain samples substantially improved target-domain performance while preserving public-domain performance. Relative to source-only (target mAP@0.5 = 0.1056), full mixed

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

supervision (B1) raised target mAP@0.5 to 0.6609, and domain-balanced training (B2) further to 0.7076, indicating that a higher proportion of target-domain samples benefits industrial CCTV performance. The target-only model (D) reached 0.7978 on the target test set but failed almost entirely on the public set (0.0069), confirming that the two domains are not interchangeable. Notably, at the same 640 × 640 resolution the target-only model attains the highest target accuracy among the 640-pixel settings, so mixing public data does not maximize target accuracy by itself; rather, mixed supervision is the appropriate choice when a single model must remain accurate on *both* domains, since it retains near-source-level public performance that the target-only model loses entirely. The additional gain of the balanced model at higher resolution (C1) is therefore attributable to input resolution rather than to mixing as such.

The higher-resolution balanced variant (C1, 960 × 960) achieved the best target-domain results overall (mAP@0.5 = 0.8978; mAP@0.5:0.95 = 0.5316), substantially exceeding the 640 × 640 balanced variant. This supports the hypothesis that input resolution is a key factor for the small head regions characteristic of CCTV imagery, in line with the recognized difficulty of small-object detection [36]. The scale-and-mosaic augmentation variant (C2; mosaic augmentation following [37]) at 640 × 640 reached target mAP@0.5 = 0.6810, comparable to B2 and below C1, suggesting that, in this setting, increased input resolution was more effective than scale/mosaic augmentation at the base resolution. Public-domain performance remained stable across all mixed variants (mAP@0.5 between 0.9607 and 0.9742), showing that adding target-domain data did not harm public-benchmark performance.

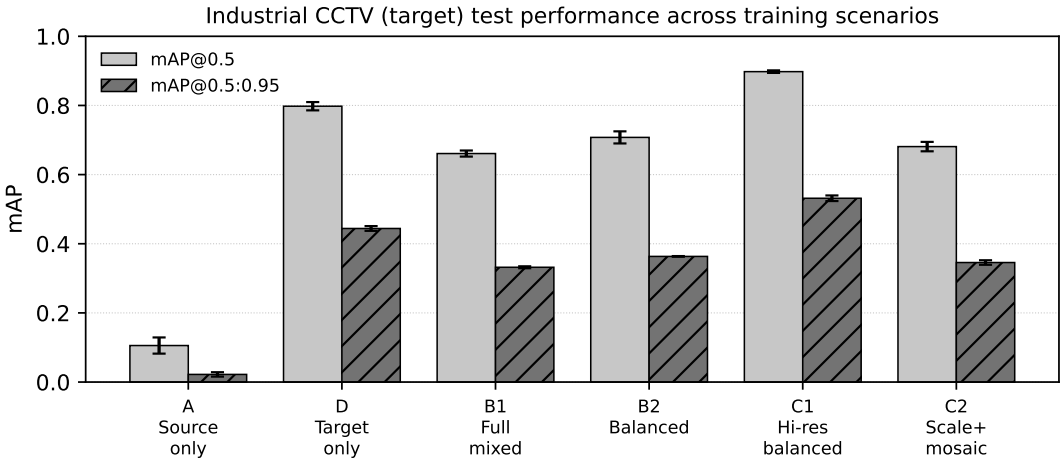


Figure 3. Industrial CCTV (target-domain) test performance across the cross-domain and mixed-domain training scenarios of Table 9. Bars show mAP@0.5 and mAP@0.5:0.95 (mean over three seeds; error bars denote standard deviation). The source-only model (A) collapses on the target domain, whereas adding target-domain data (B1, B2) and, especially, higher input resolution (C1) substantially improves performance.

3.5. Effect of the Data Splitting Strategy: Random vs. Temporal (RQ3)

To empirically assess the impact of the splitting strategy motivated in Section 2.1.4, we compare the day-level temporal split against a frame-level random split. Both settings use the same images, model, and training configuration (E3a variant, YOLO11n, 640 × 640, three seeds), and the per-subset image counts are kept identical (1144/224/1045); they differ only in how frames are assigned to the training, validation, and test subsets, which isolates the effect of temporal separation. For the random setting, three independent random splits (one per seed) are generated to avoid dependence on a single partition.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

Table 10. Effect are reported as mean ± standard deviation across three random seeds.

Split Strategy	mAP@0.5	mAP@0.5:0.95	Precision	Recall
Random (frame-level)	0.856 ± 0.009	0.483 ± 0.008	0.899 ± 0.005	0.760 ± 0.009
Temporal (day-level, ours)	0.798 ± 0.012	0.444 ± 0.007	0.835 ± 0.034	0.700 ± 0.001

As expected, the random split yields higher apparent performance than the temporal split, with a difference of 0.058 mAP@0.5 and 0.039 mAP@0.5:0.95 (and consistently higher precision and recall), because near-duplicate frames from the same recording day can appear in both training and test subsets. The day-level temporal split removes this near-duplicate leakage and therefore provides a more conservative, deployment-oriented estimate of detection performance.

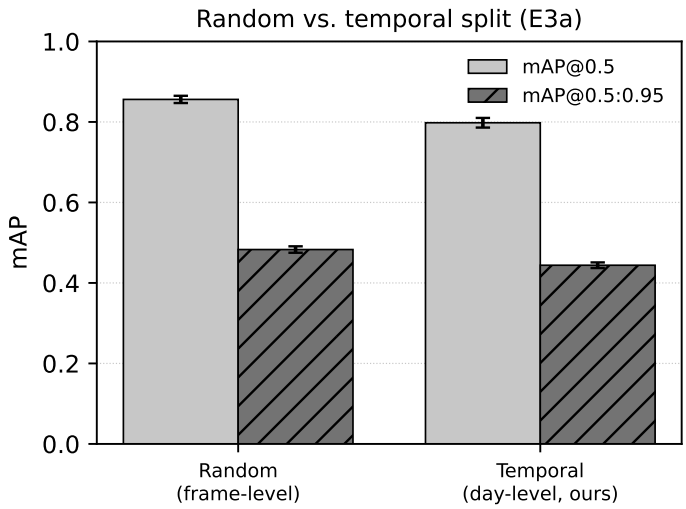


Figure 4. Effect of the data splitting strategy (E3a variant). A frame-level random split yields higher apparent performance than the day-level temporal split, reflecting near-duplicate leakage between training and test subsets. Bars show mean over three seeds; error bars denote standard deviation.

3.6. Detection Performance by Object Size

Because head regions in industrial CCTV are predominantly small, we analyze detection performance stratified by object size; small objects are widely recognized as a challenging regime for modern detectors [36]. We adopt the standard COCO area ranges [34], where small, medium, and large objects correspond to bounding-box areas below 32^2, between 32^2 and 96^2, and above 96^2 pixels, respectively. The analysis is performed post-hoc on the test-set predictions of the higher-resolution C1 model (Section 3.4), without additional training, and is evaluated using the COCO protocol (pycocotools). Results are reported as mean ± standard deviation across the three random seeds.

Table 11. Detection performance stratified by object size (COCO area ranges) for the higher-resolution C1 model on the industrial CCTV test set, evaluated with the COCO protocol. Results are mean ± standard deviation across three random seeds.

Object Size (Area)	GT Instances	AP@0.5:0.95
Small ($<32^2$ px)	3164	0.453 ± 0.010
Medium (32^2–96^2 px)	1078	0.626 ± 0.026
Large ($>96^2$ px)	0	n/a
All	4242	0.503 ± 0.012

The test set contains no large head regions: all annotated heads fall in the small (3164 instances) or medium (1078 instances) COCO ranges, which directly reflects the small-object nature of industrial

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

CCTV. Detection accuracy increases with object size, from AP@0.5:0.95 = 0.453 for small heads to 0.626 for medium heads, so the smallest regions are the dominant source of error. This is consistent with the improvement obtained at higher input resolution (variant C1, Section 3.4) and explains why input resolution is a key factor for CCTV-oriented hard-hat compliance detection. Because these values are computed with the COCO evaluation protocol, the absolute AP differs slightly from the Ultralytics mAP reported in Table 9; the size-stratified results should therefore be read as a relative comparison across object sizes rather than as directly comparable to Table 9.

3.7. Deployment-Oriented Efficiency

Since the dataset targets real industrial CCTV monitoring, we report the inference efficiency of the deployed configuration. Efficiency is summarized by the number of parameters, the number of floating-point operations (FLOPs), the on-disk model size, the per-image latency, and the throughput in frames per second (FPS). All measurements use the same input resolution as the corresponding training configuration and a fixed inference device, with a warmup phase preceding timing.

Table 12. Deployment-oriented efficiency of the detection configurations. Latency and FPS are measured on a single NVIDIA Quadro RTX 6000 GPU at batch size 1 after a warmup phase; values are averaged over 200 runs. Efficiency metrics are architecture-dependent and are therefore reported for a single representative checkpoint per configuration.

Model	Input Size	Params (M)	FLOPs (G)	Size (MB)	Latency (ms)	FPS
YOLO11n (E3a)	640×640	2.59	6.44	5.20	8.45 ± 0.40	118.3
YOLO11n (C1)	960×960	2.59	14.49	5.24	9.58 ± 0.35	104.4

The results confirm that both configurations operate well above real-time requirements on the target hardware. The reference 640-pixel model (E3a) processes surveillance frames at 118.3 FPS (8.45 ms per image), while the higher-resolution 960-pixel variant (C1), which achieved the best target-domain accuracy (Section 3.4), runs at 104.4 FPS (9.58 ms). Increasing the input resolution from 640 to 960 raises FLOPs from 6.44 G to 14.49 G and reduces throughput by about 12%, while leaving the parameter count (2.59 M) and model size (≈5.2 MB) essentially unchanged. This modest latency cost is acceptable given the substantial accuracy improvement on small CCTV head regions, confirming that the configuration is suitable for deployment-oriented hard-hat compliance monitoring on industrial CCTV streams.

3.8. Qualitative Analysis

Beyond the aggregate metrics, we summarize the qualitative behavior observed on the target test subset. Correct detections spanned small head regions, partially occluded workers, and workers at varying distances and viewpoints. The most common failure modes involved very small heads, partial occlusion, motion blur, and ambiguous helmet boundaries: false positives occasionally appeared on background structures or high-contrast objects resembling head or helmet shapes, whereas false negatives occurred mainly under strong occlusion, large camera distance, or poor local contrast. These observations are consistent with the quantitative results and underline the difficulty of applying hard-hat detection to industrial CCTV. Illustrative image examples from the industrial recordings are not reproduced here, as the underlying surveillance footage cannot be shared (Data Availability Statement).

4. Discussion

The results support three findings aligned with the research questions. First (RQ1), privacy-preserving preprocessing can be applied without substantial degradation when image structure is preserved, as in the LaMa-based, face-anonymized E3a variant. Second (RQ2), models trained only on public hard-hat data transfer poorly to the target industrial CCTV domain despite excellent

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

public-benchmark performance. Third (RQ2), combining public and target-domain data substantially improves industrial-domain performance, so public datasets are valuable for pretraining or mixed supervision but insufficient as standalone training sources. The temporal protocol (RQ3) underlies all of these conclusions by ensuring deployment-oriented evaluation.

4.1. Impact of Privacy-Preserving Preprocessing (RQ1)

The pipeline reduces non-semantic and potentially identifying information while maintaining utility. Deterministic black masking reduced OSD visibility but introduced artificial high-contrast rectangles and required annotation cleaning for partially covered boxes; the lower performance of black-mask variants suggests these artifacts and label modifications can harm the detector. LaMa-based inpainting offered a better compromise, removing overlays and textual identifiers without black rectangles and preserving a natural appearance. Localized face anonymization did not substantially degrade performance, indicating that privacy-related modifications can be compatible with detection utility when spatially localized and designed to preserve task-relevant structure such as the head contour. This agrees with prior evidence that face obfuscation has only a limited effect on the accuracy of downstream visual-recognition models [7]. More broadly, these results position generative image inpainting as a practical tool for achieving AI readiness in privacy-preserving dataset preparation: a generative model edits task-irrelevant, sensitive content while preserving the structures a downstream detector relies on. This complements the more common use of generative models for image synthesis and restoration, here applied to deployment-oriented industrial monitoring rather than to content creation. Beyond inpainting, generative super-resolution offers a complementary route to recover detail in small regions [38]; here we rely on higher native input resolution rather than synthesized detail, but combining the two is a promising direction, as demonstrated by joint super-resolution and detection pipelines for small targets in industrial inspection [39].

4.2. Domain Gap Between Public Datasets and Industrial CCTV (RQ2)

A central finding is the large gap between public hard-hat imagery and the target industrial CCTV domain. The public-only model performed well on the public test set but degraded sharply on the target test set, so strong public-benchmark performance does not imply reliable industrial deployment. This sharp cross-dataset drop is consistent with the long-documented phenomenon of dataset bias, whereby detectors tuned on one benchmark generalize poorly to images drawn from a different distribution [40]. The gap stems from domain-specific factors: public datasets often contain web sources, construction-site photographs, or close-range views, whereas the target domain comprises fixed-camera CCTV frames with small, partially occluded heads seen from elevated or oblique viewpoints against complex backgrounds. Public data remains useful as source-domain data, especially for initial training and mixed supervision, but target-domain samples are necessary for deployment-oriented performance.

4.3. Importance of Temporal Separation (RQ3)

The day-level temporal split is essential to the validity of the evaluation. Random image-level splits can overestimate performance when frames come from continuous recordings, because visually similar frames may appear in both training and test sets. By assigning all frames from a recording day to a single subset and using a chronologically later test period, the protocol reduces near-duplicate leakage and better approximates deployment, where models are applied to later, unseen data. Our results therefore suggest that temporal (group-wise) splitting is a more appropriate default than random splitting for video-derived CCTV detection datasets. The empirical comparison of random versus temporal splitting is reported in Section 3.5.

4.4. Practical Implications for Industrial Safety Monitoring

For industrial safety monitoring, these findings have three practical implications. First, models should be evaluated on data reflecting the intended deployment domain, using temporal splitting for

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

realistic estimates, since strong public-benchmark performance may not hold under industrial CCTV conditions. Second, privacy-preserving preprocessing should be an integral part of dataset preparation, ensuring the AI readiness of the data by improving suitability for research and collaboration while reducing exposure of sensitive content. Third, public and target-domain data are complementary: public data supports initial training and general representations, but target-domain data is necessary for reliable industrial monitoring. The dominance of small head regions in the error analysis, and the accuracy gains obtained at higher input resolution, are consistent with recent industrial PPE-monitoring detectors that likewise identify small object scale as the principal difficulty and address it through resolution- and scale-oriented designs [41].

4.5. Limitations

This study has several limitations. First, the dataset is moderate in size; its value lies in realistic CCTV character, temporal split, and privacy-preserving preparation rather than scale. Second, recordings span a limited number of days in a single industrial environment, so the dataset may not capture all variation in lighting, season, camera placement, clothing, helmet appearance, and operating conditions; future work should extend to more periods, cameras, and facilities. Because these recordings also cannot be shared publicly due to privacy and contractual constraints, the present work is best read as a deployment-oriented study on real industrial CCTV rather than as a public comparison benchmark; to support reproducibility, the evaluation protocol, preprocessing and evaluation code, mask coordinates, anonymization parameters, and aggregate statistics are available from the corresponding author upon reasonable request, subject to approval by the data provider and a confidentiality agreement, in place of the raw images. Third, by design the experiments use a single, fixed detector (YOLO11n) so that the architecture acts as a controlled variable; this isolates the effects of data and preprocessing that are central to our research questions. The absolute performance values are nonetheless expected to vary with other detectors and model sizes, and confirming that the observed relative trends—the domain gap, the temporal-split effect, and the resolution sensitivity—transfer across architectures is left to future work. In addition, the higher-resolution configuration (C1) was evaluated only in the domain-balanced setting; we did not train a target-only model at 960×960, so the respective contributions of input resolution and training-data composition to the best target-domain result cannot be fully disentangled. Fourth, the privacy-preserving pipeline reduces visible sensitive information but is not a formal anonymization guarantee in a legal sense; recent work shows that some anonymization methods can be partially reversible [30], so broader sharing may require additional legal and organizational review. Finally, the study covers only two head-region classes; other PPE elements (vests, gloves, eye protection) and person-level tracking are out of scope.

5. Conclusions

We presented the Industrial CCTV-HH Dataset, a privacy-preserving object detection resource for deployment-oriented hard-hat compliance evaluation under realistic industrial surveillance conditions, and organized the study around three research questions. The dataset contains 2413 images and 9608 annotated instances in two head-region classes, with a day-level temporal split that reduces split-related leakage and provides a stricter evaluation in which the test data originate from a later recording period than the training data.

We proposed and evaluated a privacy-preserving preprocessing pipeline that removes OSD overlays and textual identifiers via a generative image-inpainting model (LaMa) and applies localized face anonymization. This pipeline preserved detection utility while improving privacy properties and dataset hygiene; the E3a variant was selected as the reference AI-ready configuration (RQ1). The study thus illustrates how generative image editing can be integrated into a deployment-oriented detection pipeline to produce AI-ready surveillance data—editing sensitive content while preserving task-relevant structure—rather than being used solely for image synthesis. Cross-domain experiments showed that models trained only on public hard-hat data generalize poorly to the target industrial CCTV domain despite high public-benchmark performance, while mixed-domain training with target-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

domain samples substantially improved performance (RQ2). The temporal protocol was central to obtaining deployment-relevant estimates (RQ3).

Public datasets and industrial CCTV data should therefore be treated as complementary rather than interchangeable: public datasets are valuable for initial training and benchmarking, but deployment-oriented evaluation requires domain-specific, temporally separated, and privacy-aware datasets. Future work will extend the dataset to additional cameras, recording periods, facilities, and PPE categories, and will investigate alternative detector architectures and further deployment optimization.

Author Contributions: Conceptualization, R.B. and A.D.; methodology, R.B. and N.B.; software, R.G.; validation, N.B. and R.G.; formal analysis, N.B.; investigation, R.B., A.D., N.B. and R.G.; resources, R.B.; data curation, N.B. and R.G.; writing—original draft preparation, R.G. and N.B.; writing—review and editing, R.B., A.D. and N.B.; visualization, R.G.; supervision, R.B. and A.D.; project administration, N.B. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: This study did not involve any medical or experimental intervention on humans or animals. It analyzed pre-existing closed-circuit television (CCTV) recordings that had been lawfully collected by Altrad Babcock Europe S.A. for the routine occupational health and safety monitoring of its own premises and were subsequently made available to the authors for research purposes. The study addresses head-region detection for hard-hat compliance only and performs no facial recognition, no biometric identification, and no re-identification or tracking of individuals. No new data were recorded from human participants for the purpose of this study, and all personal and potentially identifying content was removed or obscured prior to analysis by means of the privacy-preserving preprocessing pipeline described in Section 2.1.6. The underlying CCTV monitoring was operated in accordance with the applicable workplace-monitoring legislation (in particular Article 22^2 of the Polish Labour Code), and, as the processed data are anonymized and do not permit the identification of individuals, they fall outside the scope of personal data within the meaning of the General Data Protection Regulation (Regulation (EU) 2016/679, Recital 26). On this basis, ethical review and approval were not required for this study; any supporting documentation can be provided by the corresponding author upon reasonable request.

Informed Consent Statement: The CCTV recordings were collected within the framework of the INRED project no. POIR.01.01.01-00-0170/17 (implemented with the support of the Polish National Centre for Research and Development), in the context of Altrad Babcock Europe S.A. lawful workplace safety monitoring, for which the operator is responsible under the applicable labour and data-protection regulations. Because all individuals appearing in the footage were anonymized—faces were blurred and on-screen and scene-level identifiers were removed (Section 2.1.6)—no person is identifiable in the dataset or in any published image, and informed consent for the publication of identifiable personal data was therefore not applicable. Use of the recordings for research purposes was authorized by Altrad Babcock Europe S.A. in accordance with the INRED project agreement concluded between Altrad Babcock Europe S.A. (formerly Doosan Babcock Energy Polska S.A.) and the AGH University of Krakow.

Data Availability Statement: The dataset analyzed in this study was derived from real industrial surveillance recordings and contains privacy-sensitive content. Due to privacy, confidentiality, and contractual restrictions, the raw recordings and the full image dataset cannot be made publicly available. The preprocessing scripts (including OSD mask coordinates, inpainting configuration, and face-anonymization parameters), the dataset configuration files, and the evaluation protocol are available from the corresponding author upon reasonable request, subject to approval by the data provider and a confidentiality agreement. The public source-domain dataset (Hard Hat Workers) is publicly available from its original provider. These released artifacts are intended to make the evaluation protocol reproducible on other industrial CCTV data.

Acknowledgments: This work was supported by the Polish National Centre for Research and Development under the Smart Growth Operational Programme, INRED project no. POIR.01.01.01-00-0170/17. The authors want also to address their special thanks to Altrad Babcock Europe S.A. for providing access to the surveillance recordings used to construct the dataset under consideration.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

Conflicts of Interest: The industrial partner provided access to anonymized recordings but had no role in the study design, in the collection, analysis, or interpretation of data and in the writing of the manuscript, or in the decision to publish the result.

Abbreviations

The following abbreviations are used in this manuscript:

CCTV	Closed-Circuit Television
PPE	Personal Protective Equipment
OSD	On-Screen Display
mAP	Mean Average Precision
IoU	Intersection over Union
IoA	Intersection over Area
YOLO	You Only Look Once
LaMa	Large Mask Inpainting

References

1. Lee, J.; Lee, S. Construction Site Safety Management: A Computer Vision and Deep Learning Approach. *Sensors* **2023**, *23*, 944. https://doi.org/10.3390/s23020944.
2. Mohy, A.; Bassioni, H.; Elgendi, E.; Hassan, T. Innovations in Safety Management for Construction Sites: The Role of Deep Learning and Computer Vision Techniques. *Constr. Innov.* **2024**, *26*, 551–578. https://doi.org/10.1108/CI-04-2023-0062.
3. Dziech, A.; Białas, J.; Glowacz, A.; Korus, P.; Leszczuk, M.; Matiolański, A.; Baran, R. Overview of Recent Advances in CCTV Processing Chain in the INDECT and INSIGMA Projects. In Proceedings of the 2013 International Conference on Systems, Signals and Image Processing (IWSSIP). IEEE, 2013.
4. Hayat, A.; Morgado-Dias, F. Deep Learning-Based Automatic Safety Helmet Detection System for Construction Safety. *Appl. Sci.* **2022**, *12*, 8268. https://doi.org/10.3390/app12168268.
5. Elesawy, A.; Mohammed Abdelkader, E.; Osman, H. A Detailed Comparative Analysis of You Only Look Once-Based Architectures for the Detection of Personal Protective Equipment on Construction Sites. *Eng* **2024**, *5*, 347–366. https://doi.org/10.3390/eng5010019.
6. Kapoor, S.; Narayanan, A. Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. *Patterns* **2023**, *4*, 100804. https://doi.org/10.1016/j.patter.2023.100804.
7. Yang, K.; Yau, J.; Fei-Fei, L.; Deng, J.; Russakovsky, O. A Study of Face Obfuscation in ImageNet. In Proceedings of the 39th International Conference on Machine Learning (ICML); Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; Sabato, S., Eds., PMLR, Baltimore, MD, USA, 17–23 July 2022; Vol. 162, pp. 25313–25330.
8. Gaikwad, B.; Karmakar, A. Real-Time Distributed Video Analytics for Privacy-Aware Person Search. *Comput. Vis. Image Underst.* **2023**, *234*, 103749. https://doi.org/10.1016/j.cviu.2023.103749.
9. Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; Lempitsky, V. Resolution-Robust Large Mask Inpainting with Fourier Convolutions. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 2149–2159. https://doi.org/10.1109/WACV51458.2022.00323.
10. Liu, Y.; Wang, J. Personal Protective Equipment Detection for Construction Workers: A Novel Dataset and Enhanced YOLOv5 Approach. *IEEE Access* **2024**, *12*, 47338–47358. https://doi.org/10.1109/ACCESS.2024.3382817.
11. Mo, Y.; Xu, S.; Ao, S.; Cai, N.; Zhou, S.; Zhou, Y. Safety Helmet and Harness Detection on Construction Sites Based on Deep Learning. *Signal Image Video Process.* **2026**, *20*, 41. https://doi.org/10.1007/s11760-025-05076-6.
12. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. https://doi.org/10.1109/CVPR.2016.91.
13. Ali, M.; Zhang, Z. The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection. *Computers* **2024**, *13*, 336. https://doi.org/10.3390/computers13120336.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

14. Wang, S. Automated Non-PPE Detection on Construction Sites Using YOLOv10 and Transformer Architectures for Surveillance and Body-Worn Cameras with Benchmark Datasets. *Sci. Rep.* **2025**, *15*, 27043. https://doi.org/10.1038/s41598-025-12468-8.
15. Northeastern University—China. Hard Hat Workers Dataset, version 5 (raw_HeadHelmetClasses). J. Nelson workspace, Roboflow Universe; Public Domain license. Available online: https://universe.roboflow.com/joseph-nelson/hard-hat-workers/dataset/5, 2019. (accessed on 8 June 2026), https://doi.org/10.7910/DVN/7CBGOS.
16. Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; Van Gool, L. Domain Adaptive Faster R-CNN for Object Detection in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3339–3348. https://doi.org/10.1109/CVPR.2018.00352.
17. Oza, P.; Sindagi, V.; Vibashan, V.; Patel, V. Unsupervised Domain Adaptation of Object Detectors: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 4018–4040. https://doi.org/10.1109/TPAMI.2022.3217046.
18. Lee, W.; Hong, D.; Lim, H.; Myung, H. Object-Aware Domain Generalization for Object Detection. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). AAAI Press, 2024, Vol. 38, pp. 2947–2955. https://doi.org/10.1609/aaai.v38i4.28076.
19. Davari, S.; Kim, D.; Tohidifar, A. ControlNet-Based Domain Adaptation for Synthetic Construction Images via Graphical Simulation and Generative AI. *Autom. Constr.* **2025**, *180*, 106562. https://doi.org/10.1016/j.autcon.2025.106562.
20. Salazar, J.; Garland, L.; Ochoa, J.; Pyrcz, M. Fair Train-Test Split in Machine Learning: Mitigating Spatial Autocorrelation for Improved Prediction Accuracy. *J. Pet. Sci. Eng.* **2022**, *209*, 109885. https://doi.org/10.1016/j.petrol.2021.109885.
21. Kian Ara, R.; Matiolański, A.; Grega, M.; Dziech, A.; Baran, R. Efficient Face Detection Based Crowd Density Estimation Using Convolutional Neural Networks and an Improved Sliding Window Strategy. *Int. J. Appl. Math. Comput. Sci.* **2023**, *33*, 7–20. https://doi.org/10.34768/amcs-2023-0001.
22. Xiang, H.; Zou, Q.; Nawaz, M.; Huang, X.; Zhang, F.; Yu, H. Deep Learning for Image Inpainting: A Survey. *Pattern Recognit.* **2023**, *134*, 109046. https://doi.org/10.1016/j.patcog.2022.109046.
23. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. https://doi.org/10.1145/3422622.
24. Croitoru, F.A.; Hondru, V.; Ionescu, R.; Shah, M. Diffusion Models in Vision: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10850–10869. https://doi.org/10.1109/TPAMI.2023.3261988.
25. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042.
26. Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; Van Gool, L. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11451–11461. https://doi.org/10.1109/CVPR52688.2022.01117.
27. Hukkelås, H.; Mester, R.; Lindseth, F. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. In Proceedings of the International Symposium on Visual Computing (ISVC), Lake Tahoe, NV, USA, 7–9 October 2019; pp. 565–578. https://doi.org/10.1007/978-3-030-33720-9_44.
28. Wang, R.; Gu, L.; Li, J.; Wang, J.; Sun, J.; Wan, W. GenRAN: GenFusion-Guided Reversible Anonymization Network for Face Privacy Preserving. *Inf. Fusion* **2025**, *121*, 103120. https://doi.org/10.1016/j.inffus.2025.103120.
29. McPherson, R.; Shokri, R.; Shmatikov, V. Defeating Image Obfuscation with Deep Learning. *arXiv* **2016**. arXiv:1609.00408, https://doi.org/10.48550/arXiv.1609.00408.
30. Todt, J.; Hanisch, S.; Strufe, T. Fantômas: Understanding Face Anonymization Reversibility. *Proc. Priv. Enhancing Technol.* **2024**, *2024*, 24–43. https://doi.org/10.56553/popets-2024-0105.
31. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**. arXiv:2410.17725, https://doi.org/10.48550/arXiv.2410.17725.
32. Jocher, G.; Qiu, J. Ultralytics YOLO11, version 11.0.0. Ultralytics. Available online: https://github.com/ultralytics/ultralytics, 2024. (accessed on 8 June 2026).
33. Everingham, M.; Van Gool, L.; Williams, C.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. https://doi.org/10.1007/s11263-009-0275-4.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2026 doi:10.20944/preprints202609.1884.v1

34. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.
35. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2019. https://doi.org/10.48550/arXiv.1912.01703.
36. Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; Han, J. Towards Large-Scale Small Object Detection: Survey and Benchmarks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**. https://doi.org/10.1109/TPAMI.2023.3290594.
37. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**. arXiv:2004.10934, https://doi.org/10.48550/arXiv.2004.10934.
38. Wang, X.; Xie, L.; Dong, C.; Shan, Y. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 1905–1914. https://doi.org/10.1109/ICCVW54120.2021.00217.
39. Falahatnejad, S.; Karami, A.; Nezamabadi-pour, H. PTSRDet: End-to-End Super-Resolution and Object-Detection Approach for Small Defect Detection of Power Transmission Lines. *Comput. Electr. Eng.* **2025**, *124*, 110374. https://doi.org/10.1016/j.compeleceng.2025.110374.
40. Torralba, A.; Efros, A. Unbiased Look at Dataset Bias. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347.
41. Li, Y.; Hu, J.; Li, J.; Tong, X. OAM-YOLO: A Real-Time Small Object Detection Framework for PPE Compliance Monitoring in Industrial Environments. *Process Saf. Environ. Prot.* **2025**, *204*, 108058. https://doi.org/10.1016/j.psep.2025.108058.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.