

ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ/INFORMATICS AND INFORMATION PROCESSES

DOI: https://doi.org/10.60797/IRJ.2026.170.51 EDN: EQFQVI

COMPUTER-VISION RECOGNITION OF UNSAFE ACTIONS AND PERSONAL PROTECTIVE EQUIPMENT VIOLATIONS FROM A VIDEO STREAM

Research article

Mokshin V.V.[1], Medvedev P.S.[2], Gayfutdinov T.I.[3,] *

[1, 2, 3] Kazan National Research Technical University named after A.N. Tupolev – KAI, Kazan, Russian Federation

* Corresponding author (odie[at]list.ru)

Suggested: 06.06.2026; Accepted: 04.08.2026; Published: 17.08.2026

Abstract

Workplace injury is largely caused by failure to comply with personal protective equipment (PPE) requirements and by violations of work procedures. This paper investigates the use of computer vision methods for the automatic detection of such violations from the video stream of industrial cameras. An information system is proposed and implemented whose core is a fine-tuned YOLOv5 convolutional neural network detector. The system not only records the presence or absence of a safety helmet and a high-visibility vest on a worker but also recognizes a dangerous action — climbing into the bed of a truck without a step ladder — which requires accounting for the spatial context of the scene. The decision logic is formalized as an interpretable safety-violation factor that combines per-object detection confidences, the spatial relations among the worker, the protective items and the vehicle, and the temporal evidence accumulated along object tracks. A Kalman filter is employed for robust object tracking across frames and for smoothing trajectories, and detected violations are automatically compiled into reports. The composition of the training set, the classes of recognized objects, the data preparation and versioning procedure, the training parameters, and the quality-assessment metrics are described. An example of the system operating in an active production workshop is presented, and the limitations of the approach — related primarily to the volume and resolution of the training data — are discussed.

Keywords: computer vision, object detection, YOLOv5, convolutional neural networks, occupational safety, personal protective equipment, video analytics, Kalman filter, violation monitoring, safety-violation factor, industrial safety.

РАСПОЗНАВАНИЕ ОПАСНЫХ ДЕЙСТВИЙ И НАРУШЕНИЙ ПРАВИЛ ИСПОЛЬЗОВАНИЯ СРЕДСТВ ИНДИВИДУАЛЬНОЙ ЗАЩИТЫ С ПОМОЩЬЮ СИСТЕМ КОМПЬЮТЕРНОГО ЗРЕНИЯ ИЗ ВИДЕОТРАНСЛЯЦИИ

Научная статья

Мокшин В.В.[1], Медведев П.С.[2], Гайфутдинов Т.И.[3,] *

[1, 2, 3] Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, Казань, Российская Федерация

* Корреспондирующий автор (odie[at]list.ru)

Предложена: 06.06.2026; Принята: 04.08.2026; Опубликована: 17.08.2026

Аннотация

Травмы на рабочем месте в значительной степени обусловлены несоблюдением требований к средствам индивидуальной защиты (СИЗ) и нарушениями технологических процедур. В данной статье исследуется применение методов компьютерного зрения для автоматического обнаружения таких нарушений на основе видеотрансляции с промышленных камер. Предложена и реализована информационная система, ядром которой является тонко настроенный детектор на основе сверточной нейронной сети YOLOv5. Система не только фиксирует наличие или отсутствие защитной каски и светоотражающего жилета у работника, но и распознает опасное действие — вхождение в кузов грузовика без использования стремянки — что требует учета пространственного контекста сцены. Логика принятия решений формализована в виде интерпретируемого коэффициента нарушения техники безопасности, который объединяет коэффициенты достоверности обнаружения для каждого объекта, пространственные отношения между работником, средствами защиты и транспортным средством, а также временные данные, накопленные по траекториям объектов. Для надёжного отслеживания объектов между кадрами и сглаживания траекторий используется фильтр Калмана, а обнаруженные нарушения автоматически собираются в отчёты. Описываются состав обучающего набора, классы распознаваемых объектов, процедура подготовки данных и управления версиями, параметры обучения, а также метрики оценки качества. Приводится пример работы системы в действующем промышленном цехе, а также обсуждаются ограничения данного подхода, связанные в первую очередь с объемом и разрешением обучающих данных.

Ключевые слова: компьютерное зрение, обнаружение объектов, YOLOv5, сверточные нейронные сети, охрана труда, средства индивидуальной защиты, видеоаналитика, фильтр Калмана, мониторинг нарушений, фактор нарушения техники безопасности, промышленная безопасность.

Introduction

Workplace injury remains a pressing problem, and a significant proportion of accidents at enterprises is associated with workers neglecting personal protective equipment (PPE). This proportion can be reduced by a system that would monitor PPE use in real-time and record violations. The present work is devoted to the development of such a system.

A substantial share of industrial accidents is preventable and is caused not by equipment failure but by the human factor — a refusal to wear a safety helmet or a high-visibility vest, a breach of the procedure for loading operations, or the neglect of step ladders and guard rails. Traditional monitoring of occupational-safety compliance relies on periodic walk-throughs and the selective review of surveillance recordings; that is, it is episodic in nature and does not allow a response to a violation at the moment it occurs. Meanwhile, the ubiquity of video-surveillance systems and the maturity of deep-learning algorithms create the preconditions for continuous automatic monitoring capable of detecting violations in real time and thereby reducing the risk of injury.

Almost every production process involves the use of PPE, and enterprises build entire sets of measures to ensure occupational-safety compliance. Nevertheless, some workers neglect these requirements and fail to put on a high-visibility vest or a helmet. A typical situation is telling: in order to load a part more quickly, a person climbs into the truck bed straight from the ground instead of positioning a ladder as the rules prescribe. The price of such liberties is injuries and accidents, so detecting them in a timely manner is fundamentally important.

The system is written in Python and uses a convolutional neural network detector of the YOLOv5 family, fine-tuned to the conditions of a specific enterprise by means of transfer learning: weights pretrained on a large publicly available image set are taken as the starting point, after which the network is adapted to the target classes on the enterprise's own data. The choice of the YOLO family is dictated by its combination of high-speed and acceptable accuracy: these models perform well in recognition tasks in manufacturing, transport, and other domains where a decision must be made without delay.

The aim of this work is to investigate computer vision methods as applied to the task of automatically detecting occupational-safety violations and, on this basis, to develop a workable video-analytics system. To achieve this aim, the following tasks were addressed: a comparative analysis of modern object-detection architectures and a justification for the choice of base model; the creation and annotation of a dataset reflecting characteristic situations of a production workshop; fine-tuning of the detector model on the target classes; implementation of a mechanism for tracking objects across frames; organization of the automatic logging of detected violations; and assessment of the quality of the resulting solution.

Most existing systems for monitoring personal protective equipment treat the task as the per-frame detection of isolated objects and reduce a violation to the mere absence of a bounding box, without a formal rule that turns the detector outputs into a decision and without explicit treatment of the temporal instability of those outputs. The scientific novelty of this work is accordingly twofold. First, a formalized model of the safety-violation factor is proposed: the factor is an interpretable scalar variable that maps the raw detector outputs — class labels, bounding boxes and confidence scores — into a graded measure of danger by combining a confidence criterion, the pairwise spatial relations among the worker, the protective items and the vehicle, and a recursive temporal aggregation of the evidence along an object track. The model supplies a reproducible decision rule with explicit, tunable parameters in place of ad hoc heuristics, and it is what enables the system to recognize not only the presence or absence of individual protective items but also a dangerous action defined by the spatial context of the scene — a worker climbing into the bed of a truck without a step ladder. Second, a complete and reproducible technological pipeline is built around this model — from versioned data preparation and detector fine-tuning to track-level violation scoring and the automatic generation of reports — adapted to the conditions of a specific enterprise.

Review of existing approaches

Modern object detectors are commonly divided into two-stage detectors, in which candidate regions are first generated and then classified, and one-stage detectors, which perform localization and classification in a single pass of the network. For real-time systems, processing latency is critical, so it is precisely the comparison of these two approaches that determines the choice of architecture.

Among object-detection approaches, the R-CNN family with its modifications and the YOLO algorithm stand out. The original R-CNN first selects candidate regions by selective search and then classifies each of them. This two-stage scheme rests on a fixed selection algorithm and proves too slow for real-time operation.

Fast R-CNN removes the principal source of latency: the entire image passes through the convolutional network only once, and candidate regions are projected onto a shared feature map by the RoI Pooling operation, so that the convolution for each region is no longer recomputed [1].

Faster R-CNN goes further and abandons slow selective search altogether: candidate regions are now proposed by a trainable region proposal network (RPN) [2], which markedly accelerates both the localization and the classification of objects.

YOLO is among the fastest detectors: the entire scene is analyzed in a single pass of the network, without a separate region-generation stage. The comparison in [3], [4], [5], [6] confirms YOLO's advantage in speed at a comparable recognition quality — and it is precisely this algorithm that underlies the violation-monitoring system being developed.

The YOLO family has continued to develop well beyond the version used here. Recent releases move the accuracy–latency frontier through contrasting design philosophies. YOLO12 (2025) adopts an attention-centric architecture, introducing an area-attention mechanism and residual efficient layer-aggregation networks that raise detection accuracy at a comparable parameter budget, at the cost of heavier memory use on the central processor [7]. YOLO26 (2025) instead pursues a deployment-oriented, end-to-end design: it dispenses with non-maximum suppression and with the distribution focal loss and adds training refinements — progressive loss balancing and small-target-aware label assignment — aimed specifically at small objects, while reportedly reducing processor inference time by up to 43% relative to the preceding generation [8]. Because the protective items in the present task are exactly such small and variable objects, these newer architectures are directly relevant; a quantitative comparison with them is given in the discussion below.

Proposed method

3.1. Architecture and training of the YOLOv5 network

The basic YOLOv5 architecture follows the “backbone — neck — head” principle: the backbone extracts image features, the neck aggregates them at several scales, and the head produces the final predictions. Having received an image, the network builds feature maps at several scales and connects the levels to one another via skip connections [9]. The maps are then brought to a common resolution (upsampling) and merged into a shared representation, from which the class and the bounding-box position are determined for each prediction. Redundant boxes that repeatedly cover the same object are eliminated by non-maximum suppression — the box with the higher confidence is retained [10], [11], [12].

YOLOv5 itself is a modern detection model implemented in the PyTorch framework [13]. It is based on a mathematical apparatus that predicts the coordinates and dimensions of object bounding boxes; the hidden layers employ the SiLU activation function, also known as Swish [14].

For each potential box, the network outputs five numbers — the center coordinates x and y, the width, the height, and the confidence (objectness) score [15] — as well as a vector of probabilities that the box belongs to each of the recognized classes. Training is guided by a loss function [16]: the total loss is the sum of three terms — the classification error, the objectness error (the presence of an object), and the localization error [17], [18].

Two engineering decisions noticeably affected the training quality of this model. The first is the Focus input layer, which in YOLOv5 replaced the starting block of several convolutions used in earlier versions [19]. It regroups neighboring pixels into channels, thereby reducing the number of operations and the consumption of video memory and making the forward and backward passes faster with almost no loss in mAP [20]. The second is the revised formula for predicting box centers: the normalized offset now spans not the interval 0…1 but −0.5…1.5 [21]. As a result, a box center can fall exactly on a cell boundary, which improves the detection of objects at the very edges of the frame [22].

3.2. Dataset and violation classes

The training set was formed from surveillance frames typical of a production workshop: workers with and without protective clothing, freight vehicles, loading and unloading operations, background equipment, and stored parts. The following object classes were annotated: worker, safety helmet (helmet), high-visibility vest (vest), truck (kamaz), and a dangerous action — climbing into the truck bed without a step ladder (no_stairs). The violation-detection logic is built on relating these classes: if a tracked worker is not associated with a helmet or vest detection, a PPE-use violation is recorded; and if the worker’s region coincides with the truck bed in the absence of a ladder, a violation of the work procedure is registered.

All images were obtained from the enterprise’s own stationary surveillance cameras rather than from public collections, so the set reproduces exactly the viewing geometry in which the system is intended to operate. The recordings were made by three fixed IP cameras with a native frame resolution of 1920×1080 pixels, mounted at a height of four to six meters and inclined at 25–40 degrees to the horizontal; frames were extracted from the recordings at intervals of at least two seconds and screened for near-duplicates, so that the set does not contain adjacent frames of one and the same scene. The distance from a camera to a worker ranges from about five to twenty-five meters, at which a helmet or a vest occupies from roughly ten to fifty pixels along its larger side in the source frame and becomes smaller still after scaling to the network input — this viewing geometry is the root of the difficulty of the small classes discussed below. The shooting conditions span the day and evening shifts: the workshop is lit by overhead luminaires at all times, daylight from the gates and the window strips is added during the day shift, and the set deliberately includes backlit scenes against open gates, glare on metal surfaces, and shadows cast by the equipment. A typical frame contains from one to four workers, often partially occluded by machinery, stored parts, or the sides of the truck bed; the clothing worn under the vest varies with the season, which adds appearance variability to the worker class.

The distribution of the images and of the annotated object instances over the classes is summarized in Table 1. The imbalance visible there is a property of the domain rather than of the sampling: a worker is present in almost every frame and is often not alone; protective items are somewhat rarer, because it is precisely their absence that constitutes a violation of interest; the truck appears only in loading scenes; and the dangerous action no_stairs is rare because the violation itself is rare, and staging it deliberately for the camera was ruled out for safety reasons. The remaining frames without people serve as background examples that reduce false positives. The set was divided into training and validation subsets in a fixed 4 : 1 proportion — 800 and 200 images — with stratification by the rarest class, so that the share of no_stairs instances in the validation subset matches its share in the set as a whole; frames extracted from the same recording were always assigned to the same subset, which prevents near-identical scenes from leaking across the split. Table 1 also makes the connection with the per-class quality metrics explicit: the best-recognized class (kamaz) is a large, low-variability object, whereas the weakest (no_stairs) combines the smallest number of examples with the highest variability of pose and viewpoint.

Table 1 - Composition of the dataset: images and annotated object instances by class

DOI: https://doi.org/10.60797/IRJ.2026.170.51.1

Class	Images with the class	Instances, training	Instances, validation	Share of instances, %
Worker (worker)	962	1684	428	37.0
Helmet (helmet)	874	1198	306	26.3
Vest (vest)	843	1114	283	24.5
Truck (kamaz)	517	414	103	9.1
Climbing without a ladder (no_stairs)	178	142	36	3.1

Class	Images with the class	Instances, training	Instances, validation	Share of instances, %
Total	1000	4552	1156	100.0

Note: as a rule an image contains objects of several classes at once, so the image counts do not sum to the size of the set

Data preparation is split off into a separate step that precedes training [11]. The corresponding script downloads the set locally, validates it, and registers it in the Weights & Biases (W&B) service as a versioned artifact, while also generating a configuration file for training. This approach ensures the reproducibility of experiments: each version of the data and configuration corresponds unambiguously to its own training result. This step does not train the model itself — it only prepares the data and the configuration [3].

The model was trained on the set prepared in this way [12]: the 1000 images described above were used at an input resolution of 416 pixels, training ran for 200 epochs, and it took roughly six hours. To improve generalization, augmentation techniques standard for YOLOv5 were applied to the training subset — the mosaic composition of several images and changes to photometric parameters (brightness, contrast, hue) — which expand the variability of scenes without additional annotation.

3.3. Object tracking and formalization of the safety-violation factor

To track objects robustly from frame to frame and to smooth their trajectories, a Kalman filter was added to the system. It is called linear because its state and measurement models are specified by linear relationships, and each new estimate is constructed recursively — from the previous estimate and the fresh measurement supplied by the detector [18]. Tracking makes it possible to link detections of the same object across consecutive frames, to filter out brief false positives, and to attribute a violation robustly to a specific worker rather than to an individual frame.

We now formalize the violation-detection logic outlined above. At each frame t the detector returns a set of detections; an individual detection is a triple d = (c, b, s) consisting of a class label c, a bounding box b and a confidence score s. We write P with a class superscript for the subsets of detections that are helmets (H), vests (V), the dangerous action no_stairs (A) and trucks (T), and we denote a tracked worker by w. The spatial relation between two boxes is measured by their coverage:

$$\operatorname{cov}\left(b_1, b_2\right) = \frac{\operatorname{area}(b_1 \cap b_2)}{\operatorname{area}(b_1)} \tag{1}$$

which gives the fraction of the area of the first box that lies inside the second; unlike the symmetric intersection-over-union, this asymmetric quantity is the natural measure of a containment relation, such as a small helmet box lying inside a larger worker box. A protective item of class κ — a helmet or a vest — is regarded as belonging to the worker when its detection is sufficiently confident, sufficiently contained in the worker region, and geometrically plausible:

$$a_\kappa(w, p) = \mathbf{1}\left[s_p \geq \tau_\kappa\right] \cdot \mathbf{1}\left[\operatorname{cov}\left(b_p, b_w\right) \geq \theta_\kappa\right] \cdot g_\kappa\left(b_p, b_w\right) \tag{2}$$

Here τ and θ are the per-class confidence and coverage thresholds, and the geometric-consistency term equals one when the centroid of the item lies in the expected sub-region of the worker box — the upper part for a helmet, the torso region for a vest — and zero otherwise. The presence of a helmet and of a vest for the worker at frame t is then the best associated detection of each:

$$h_t(w) = \max_{p \in P_t^H} a_H(w, p), \quad v_t(w) = \max_{p \in P_t^V} a_V(w, p) \tag{3}$$

with the maximum over an empty set defined as zero, so that a zero value signals a missing item. The dangerous action is contextual: a no_stairs detection is counted only when it is at once attached to the worker and co-located with a vehicle, which expresses the requirement that the worker be climbing into a truck bed rather than merely standing beside it:

$$n_t(w) = \max_{q \in P_t^A}\left(\mathbf{1}\left[s_q \geq \tau_A\right] \cdot \mathbf{1}\left[\operatorname{cov}\left(b_q, b_w\right) \geq \theta_A\right] \cdot \max_{k \in P_t^T} \mathbf{1}\left[\operatorname{cov}\left(b_q, b_k\right) \geq \theta_{\text{bed}}\right]\right) \tag{4}$$

where τ and θ are again confidence and coverage thresholds; a value of one marks a procedural violation associated with the worker at that frame. These indicators are combined into an instantaneous safety-violation factor for the worker at frame t:

$$\varphi_t(w) = \lambda_H\left(1 - h_t(w)\right) + \lambda_V\left(1 - v_t(w)\right) + \lambda_A n_t(w), \ \lambda_H + \lambda_V + \lambda_A = 1 \tag{5}$$

in which the non-negative severity weights λ express the relative danger of a missing helmet, a missing vest and an unsafe climb, normalized to sum to one, so that the factor lies between zero and one. It is therefore an interpretable, graded measure: zero when the worker is fully compliant, and approaching one as several severe violations occur together.

A decision taken on a single frame would inherit the instability of the detector, whose safety-critical classes attain only moderate confidence, as the per-class metrics show (Table 2). The track produced by the Kalman filter links the detections of one worker across successive frames and so lets the evidence be accumulated rather than judged frame by frame. The factor is aggregated along the track by an exponential moving average, written in the same recursive form as the filter itself:

$$\Phi_t(w) = \gamma \Phi_{t-1}(w) + (1 - \gamma)\varphi_t(w), \ \gamma \in [0, 1) \tag{6}$$

where the smoothing coefficient γ governs the memory of the estimate: larger values suppress brief false detections more strongly but react more slowly to a genuine violation. A violation is registered when the aggregated factor exceeds a decision threshold:

$$V_t(w) = \mathbf{1}\left[\Phi_t(w) \geq \Theta\right] \tag{7}$$

optionally with the additional condition that this hold over a minimum number of consecutive frames, which removes residual flicker.

The formulation makes the behavior of the system explicit and reproducible. Its only free parameters — the confidence and coverage thresholds, the severity weights, the smoothing coefficient γ and the decision threshold Θ — each carry a clear operational meaning, and in this work they were fixed on the validation subset: the confidence thresholds at the values that maximized the per-class F-measure, and the smoothing and decision thresholds so as to balance missed violations against false alarms on the held-out recordings. Because the factor is a single scalar computed from quantities that the detector already produces, the same decision rule transfers without modification to a stronger detector; only the confidence thresholds need to be recalibrated.

Experimental evaluation

4.1. Quality-assessment metrics

The detector's quality was assessed using metrics common in detection tasks. Precision indicates the proportion of correct detections among all detections, recall indicates the proportion of detected violations among all those present, and their generalization is the mean average precision (mAP), computed at a box-overlap threshold of IoU = 0.5 (mAP@0.5) and averaged over the range of thresholds 0.5…0.95 (mAP@0.5:0.95). The system's suitability for real-time operation is characterized by the number of frames processed per second (FPS). The values of these metrics by class are given in Table 2.

Table 2 - Detection quality metrics by class

DOI: https://doi.org/10.60797/IRJ.2026.170.51.2

Class	Precision	Recall	mAP@0.5
Helmet (helmet)	0.79	0.68	0.72
Vest (vest)	0.77	0.66	0.70
Truck (kamaz)	0.92	0.89	0.91
Climbing without a ladder (no_stairs)	0.71	0.58	0.63
All classes	0.80	0.70	0.74

4.2. Results in a production environment

The system's performance was tested on video recordings made in an operating production workshop. An example of a recognized scene is shown in Figure 1: the system highlights workers wearing high-visibility vests and helmets, identifies the truck, and interprets climbing into its bed without a step ladder as a violation.

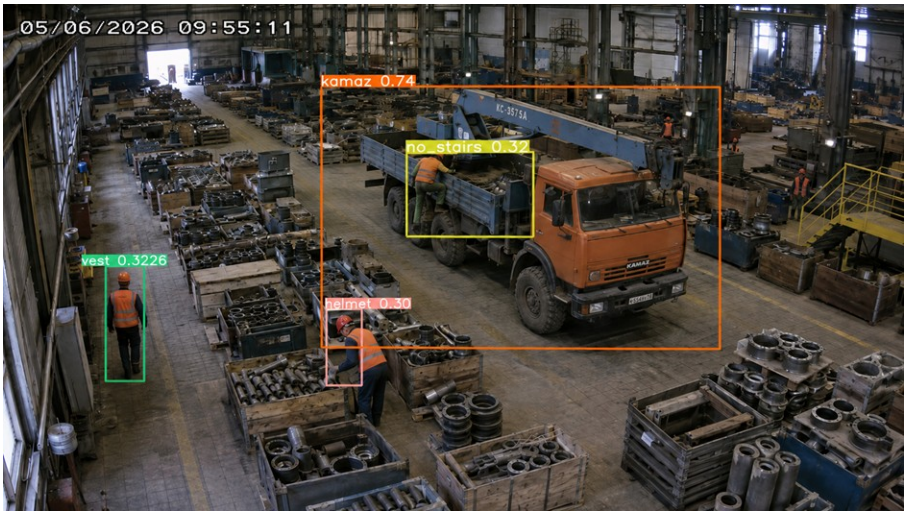

Figure 1 - Detection of a violation: climbing into a KamAZ truck bed without a ladder

DOI: https://doi.org/10.60797/IRJ.2026.170.51.3

In the frame shown, the truck is detected with a confidence of 0.74, whereas the dangerous action (no_stairs), the high-visibility vest, and the helmet received noticeably lower scores — about 0.30…0.32. Such a gap is to be expected: a large, high-contrast, weakly variable object (the truck) is recognized more confidently than small classes or classes that vary in appearance — items of equipment and a person's posture — especially at a distance from the camera and under partial occlusion by other objects.

The low confidence values for the classes that are central to occupational safety point to the main limitation of the current version — the moderate size of the training set (on the order of a thousand images) and the comparatively low input resolution

(416 pixels). Both restrict the network's ability to distinguish small details of equipment in a complex, object-dense scene. Improvements in quality can be expected from expanding and balancing the dataset across classes, increasing the input resolution, and applying more capacious network configurations while maintaining acceptable performance.

Object tracking with a Kalman filter smooths the results across frames and reduces the influence of individual missed detections: even if a helmet or vest is not recognized in a particular frame, a stable worker track prevents the false registration of a violation, while the accumulation of evidence over a series of frames, conversely, increases the reliability of recording an actual violation.

Having recorded a violation, the system automatically generates a report that includes the frame from the moment of the violation and its type — the absence of a helmet or vest, climbing into a truck bed without a ladder, and the like. The result of recognition thus becomes not merely an annotated frame but a document suitable for further review, which can be used by the enterprise's occupational-safety service.

When working with a video stream, processing is expected in a near-real-time mode, with an average performance on the order of 30 frames per second.

4.3. Comparison with modern detector architectures

The moderate accuracy of the safety-critical classes reported above is in part a property of the YOLOv5 detector adopted for this study, and it is therefore useful to set that choice against the current state of the art. Table 3 collects the detection accuracy reported in the literature for the original YOLOv5 and for two recent detectors — the attention-centric YOLO12 and the end-to-end YOLO26 — on the public COCO benchmark.

Table 3 - Detection accuracy reported in the literature on the COCO val2017 benchmark

DOI: https://doi.org/10.60797/IRJ.2026.170.51.4

Detector	Year	Detection paradigm	mAP@0.5:0.95, %
YOLOv5-s	2020	anchor-based CNN, NMS	37.4
YOLOv5-m	2020	anchor-based CNN, NMS	45.4
YOLO12-s	2025	attention-centric, NMS	48.0
YOLO12-m	2025	attention-centric, NMS	52.5
YOLO26-s	2025	end-to-end, NMS-free	≈47.2
YOLO26-m	2025	end-to-end, NMS-free	≈51.5

Note: mAP@0.5:0.95, 640-pixel input; the YOLO26 figures are approximate values reported by the model developers

Two observations follow. First, the figures in Table 3 are not directly comparable with those of Table 2: they are obtained on a different dataset of eighty general-purpose classes and at a stricter, averaged-IoU threshold, whereas Table 2 reports mAP@0.5 on roughly a thousand domain-specific images. The comparison, therefore, concerns the relative capability of the architectures, not the accuracy attainable on the present task. Second, within that comparison, the advantage of the newer models is large and systematic: at a comparable model size the small YOLO12 model exceeds the original YOLOv5 baseline by more than 10 percentage points of mAP (48.0 against 37.4), and the small and medium YOLO26 models reach a similar level while additionally dispensing with non-maximum suppression and with the distribution focal loss and reducing processor inference time by up to 43% [7], [8].

These properties bear directly on the limitation identified earlier. The weak classes in this study are the small protective items, and the small-target-aware training of YOLO26 together with the higher accuracy of YOLO12 address exactly that weakness, while the faster, suppression-free inference of YOLO26 suits the near-real-time, edge-oriented deployment intended here. Replacing the detector with one of these architectures and re-evaluating it on the enterprise dataset under the protocol of Table 2 is consequently the most promising next step; it is left to future work because it requires re-annotation and re-training on the target classes rather than the reuse of published weights. It must be stressed that the values in Table 3 are reproduced from the literature and were not measured on the data of this study, and that the decision model formalized above is independent of the detector and would carry over unchanged to any of these networks.

Conclusion

This work investigated computer vision methods as applied to the detection of occupational-safety violations and implemented an information system that analyzes the video stream from cameras, monitors safety compliance, and compiles violations into reports. Its core is the YOLOv5 detector model, fine-tuned for the needs of a specific enterprise [23].

The principal contribution of the work is a formalized model of the safety-violation factor — an interpretable scalar that maps the detector outputs into a graded, track-aggregated decision through explicit spatial and temporal criteria — which replaces ad hoc heuristics with a reproducible rule governed by a small set of tunable parameters. It is this model that allows the system to recognize not only the presence of personal protective equipment but also a dangerous action defined by the spatial context of the scene. A further distinctive feature is the integrity of the technological pipeline: versioned data preparation, detector fine-tuning, track-level violation scoring with a Kalman filter, and the automatic logging of violations.

The results obtained confirm the fundamental viability of the approach and, at the same time, outline directions for its development. The main reserve for improving quality is associated with increasing the volume and diversity of the annotated data, raising the input resolution, and fine-tuning the decision thresholds; the addition of new violation classes and the integration of the system with the alerting facilities already in operation at the enterprise also appear promising. A comparison

with current architectures indicates that the clearest route to higher accuracy on the small, safety-critical classes is to replace the YOLOv5 detector with a more recent network — the attention-centric YOLO12 or the end-to-end, small-target-oriented YOLO26 — and to re-evaluate it on the enterprise dataset; since the proposed safety-violation factor is defined over generic detector outputs, such a substitution leaves the decision model itself unchanged.

The practical significance of the work is that the proposed solution can be embedded into existing video-surveillance infrastructure and applied across a wide variety of industries — from metalworking and construction to transport and warehouse logistics — where compliance with occupational-safety requirements is critically important.

Финансирование

Данная работа/публикация выполнена за счет гранта, предоставленного Академией наук Республики Татарстан образовательным организациям высшего образования, научным и иным организациям на поддержку планов развития кадрового потенциала в части стимулирования их научных и научно-педагогических работников к защите докторских диссертаций и выполнению научно-исследовательских работ (соглашение №15/2025-ПД-КАИ от 22.12.2025).

Funding

This work/publication was funded by a grant from the Academy of Sciences of the Republic of Tatarstan provided to higher education institutions, scientific and other organizations to support human resource development plans in terms of encouraging their research and academic staff to defend doctoral dissertations and conduct research activities. (Agreement No. 15/2025-PD-KAI dated 22 December 2025).

Конфликт интересов

Не указан.

Conflict of Interest

None declared.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы на английском языке / References in English

1. Girshick R. Fast R-CNN / R. Girshick // 2015 IEEE International Conference on Computer Vision (ICCV). — 2015. — P. 1440–1448. — DOI: 10.1109/ICCV.2015.169.

2. Ren S. Faster R-CNN: Towards real-time object detection with region proposal networks / S. Ren, K. He, R. Girshick [et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2017. — № 39 (6). — P. 1137–1149. — DOI: 10.1109/TPAMI.2016.2577031.

3. Redmon J. You only look once: Unified, real-time object detection / J. Redmon, S. Divvala, R. Girshick [et al.] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) — 2016. — P. 779–788. — DOI: 10.1109/CVPR.2016.91.

4. Redmon J. YOLO9000: Better, faster, stronger / J. Redmon, A. Farhadi // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2017. — P. 6517–6525. — DOI: 10.1109/CVPR.2017.690.

5. Redmon J. YOLOv3: An incremental improvement / J. Redmon, A. Farhadi // arXiv. — 2018. — DOI: 10.48550/arXiv.1804.02767.

6. Wang Z. Fast personal protective equipment detection for real construction sites using deep learning approaches / Z. Wang, Y. Wu, L. Yang [et al.] // Sensors. — 2021. — № 21 (10). — P. 3478. — DOI: 10.3390/s21103478.

7. Tian Y. YOLOv12: Attention-centric real-time object detectors / Y. Tian, Q. Ye, D. Doermann // arXiv. — 2025. — DOI: 10.48550/arXiv.2502.12524.

8. Kalfaoglu M.E. Ultralytics YOLO26: Unified real-time end-to-end vision models / M.E. Kalfaoglu // arXiv. — 2026. — DOI: 10.48550/arXiv.2606.03748.

9. Lin T.-Y. Feature pyramid networks for object detection / T.-Y. Lin, P. Dollár, R. Girshick [et al.] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2017. — P. 936–944. IEEE. — DOI: 10.1109/CVPR.2017.106.

10. Bodla N. Soft-NMS — Improving object detection with one line of code / N. Bodla, B. Singh, R. Chellappa [et al.] // 2017 IEEE International Conference on Computer Vision (ICCV) — 2017. — P. 5562–5570. — DOI: 10.1109/ICCV.2017.593.

11. Liu W. SSD: Single shot multibox detector / W. Liu, D. Anguelov, D. Erhan [et al.] // Computer Vision – ECCV 2016. — 2016. — P. 21–37. — DOI: 10.1007/978-3-319-46448-0_2.

12. Wang C.-Y. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors / C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao // 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — 2023. — P. 7464–7475. — DOI: 10.1109/CVPR52729.2023.00721.

13. Jocher G. YOLOv5 by Ultralytics (Version 7.0) / G. Jocher // Zenodo. — 2020. — DOI: 10.5281/zenodo.3908559.

14. Elfwing S. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning / S. Elfwing, E. Uchibe, K. Doya // Neural Networks. — 2018. — № 107. — P. 3–11. — DOI: 10.1016/j.neunet.2017.12.012.

15. Bochkovskiy A. YOLOv4: Optimal speed and accuracy of object detection / A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao // arXiv. — 2020. — DOI: 10.48550/arXiv.2004.10934.

16. Lin T.-Y. Focal loss for dense object detection / T.-Y. Lin, P. Goyal, R. Girshick [et al.] // 2017 IEEE International Conference on Computer Vision (ICCV) — 2017. — P. 2980–2988. — DOI: 10.1109/ICCV.2017.324.

17. Rezatofighi H. Generalized intersection over union: A metric and a loss for bounding box regression / H. Rezatofighi, N. Tsoi, J. Gwak [et al.] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) — 2019. — P. 658–666. — DOI: 10.1109/CVPR.2019.00075.

18. Kalman R.E. A new approach to linear filtering and prediction problems / R.E. Kalman // Journal of Basic Engineering, — 1960. — № 82 (1). — P. 35–45. — DOI: 10.1115/1.3662552.

19. Terven J. A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS / J. Terven, D.-M. Córdova-Esparza, J.-A. Romero-González // Machine Learning and Knowledge Extraction. — 2023. — № 5 (4). — P. 1680–1716. — DOI: 10.3390/make5040083.

20. Lin T.-Y. Microsoft COCO: Common objects in context / T.-Y. Lin, M. Maire, S. Belongie [et al.] // Computer Vision – ECCV 2014. — 2014. — P. 740–755. — DOI: 10.1007/978-3-319-10602-1_48.

21. Wang C.-Y. Scaled-YOLOv4: Scaling cross stage partial network / C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — 2021. — P. 13029–13038. — DOI: 10.1109/CVPR46437.2021.01283.

22. Hussain M. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection / M. Hussain // Machines. — 2023. — № 11 (7). — P. 677. — DOI: 10.3390/machines11070677.

23. Nath N.D. Deep learning for site safety: Real-time detection of personal protective equipment / N.D. Nath, A.H. Behzadan, S.G. Paal // Automation in Construction, — 2020. — № 112. — Art. 103085. — DOI: 10.1016/j.autcon.2020.103085.