

# Deep Learning-Based Object Detection for Fashion, Uniform, and PPE Items: A Systematic Literature Review

Isa Aulia Almadani<sup>1</sup>, Nurcahya Pradana Taufik Prakisyaa<sup>2</sup>

<sup>1,2</sup> Informatics Education Department, Faculty of Teacher Training and Education, Sebelas Maret University, Surakarta, Indonesia

## Received:

June 6, 2026

## Revised:

July 20, 2026

## Accepted:

September 7, 2026

## Published:

October 24, 2026

Corresponding Author:

## Author Name\*:

Nurcahya Pradana Taufik  
Prakisyaa

## Email\*:

nurcahya.ptp@staff.uns.ac.id

## DOI:

10.63158/journalisi.v8i5.1818

© 2026 Journal of Information Systems and Informatics. This open access article is distributed under a (CC-BY License)



**Abstract.** Deep learning-based object detection for clothing and uniform items spans two research streams: fashion and apparel detection for e-commerce, and personal protective equipment (PPE) compliance detection for occupational safety. No prior review has jointly synthesized these streams with a detection-centric scope. This study addresses that gap through a PRISMA 2020-guided systematic literature review of 1,378 records from Scopus as the sole database, identified via two Boolean queries and screened for object detection or instance segmentation of clothing, uniform, or PPE items, yielding 26 empirical studies published between 2021 and 2026. One-stage YOLO-based detectors dominate the corpus (18 of 26 studies, up to 96.9% mAP50 at 7.3 ms inference), a dominance partly attributable to an architectural constraint in the PPE-domain search query; two-stage R-CNN and transformer-based architectures remain comparatively underexplored, the latter represented by only two studies. Dataset scale is markedly asymmetric between public fashion benchmarks (up to 491,895 images) and PPE-specific collections (as few as 1,330 images), though an emerging benchmark-reuse pattern partly offsets this gap within the PPE domain. This review contributes a cross-domain synthesis benchmarked against five prior reviews, an explicit architecture-dataset-metric mapping, a proposed nine-element minimum reporting standard, and a challenge taxonomy that surfaces previously undiscussed ethical and cultural-representation gaps in this literature.

**Keywords:** Fashion Object Detection, PPE Detection, Deep Learning, Systematic Literature Review, Computer Vision

## 1. INTRODUCTION

Automatic recognition of clothing and uniform items has become a foundational visual task across multiple domains. In retail and e-commerce, garment-level detection drives visual product search, automatic catalog enrichment, and image or video-based captioning, motivated by an apparel market valued in the range of USD 1.7–3 trillion depending on forecast methodology, with market analysts projecting continued growth [1]. Detection-based pipelines for sub-category prediction, attribute recognition, and outfit description are now standard in this commercial setting [2], [3], [4], [5]. In the construction and industrial safety domain, the same primitive enables automated personal protective equipment (PPE) compliance monitoring of helmets, vests, gloves, and full uniforms. The construction sector suffered 1,075 fatal injuries in the United States in 2023, more than any other industry, emphasizing this vital need [1], [6], [7], [8], [9]. Clothing detection has various other applications apart from the two major applications mentioned. For example, clothing detection can be utilized for estimating the insulation provided by the clothes for smart thermostats [10]. Whether it be a primary application area or a more niche application, the technical problem that clothing detection presents is the simultaneous need for classification and localization of clothing or uniform instances. This makes clothing detection a part of the broader field of deep learning based object detection.

Before surveying the architectural landscape, the boundaries of "detection" as used in this review should first be set. Object detection is the task of simultaneously localizing an object instance using a bounding box and class label. Instance segmentation goes one step further, creating a mask for each individual instance – preserving instance-level separation, pixel by pixel. These tasks are both within the scope of this review. In contrast, semantic segmentation labels the individual pixels in the image, but not on a per-instance basis, and image classification assigns only one label to the entire image, but not to individual instances of a class, and neither of these tasks is addressed by the research questions that are being asked in this review. Only signaled when accompanied by a qualifying detection or segmentation task, landmark estimation is considered, and predicts keypoint locations instead of bounding regions. Within this context, the review also categorises the three application areas identified in the corpus, namely: fashion &

apparel detection, which is used for commercial applications like visual search and catalog generation; safety & personal protective equipment (PPE) detection, which is used for safety-oriented applications like occupational compliance monitoring; and specialised applications, which use the same detection or segmentation task, but are applied to a much narrower use case, such as clothing-insulation estimation for smart home systems. PPE detection is viewed as a domain separate from general uniform detection, as PPE objects (helmets, vests etc.) are a category of objects regulated and safety-related and have their own datasets, classes and evaluation conventions.

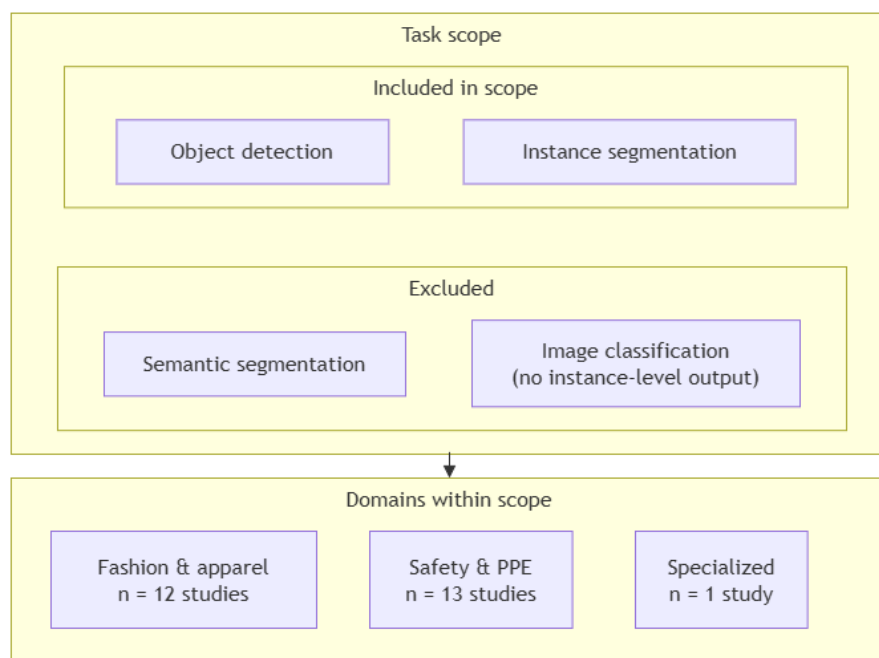


**Figure 1.** Scope of this review: included task types (top left) versus excluded task types (top right), and the three application domains covered within scope (bottom row).

Prior to the advent of deep-learning techniques, handcrafted features were used like Histograms of Oriented Gradients, Scale-Invariant Feature Transform descriptors, color histogram, or texture histogram, along with linear classifiers [11] or [12]. These representations provide a decomposition of the fine-grained, multi-layered, and deformable nature of clothing, as one clothing item can change dramatically from pose, occlusion, and illumination [13], [14]. The move towards data-driven representations was

concretized by the success of AlexNet, which achieved a top-5 error rate of 15.3% on ILSVRC 2012, down from 26.2% for the runner-up, and was formalized as a general paradigm in LeCun [15, 16]. Today there are three families of architecture that dominate the detection of clothing or uniform. The two-stage detectors based on Faster R-CNN and Mask R-CNN achieve high fine-grained accuracy [17, 18]. Various architectures have since been used for instance segmentation with localized attributes, such as Attribute-Mask R-CNN [5] and hierarchical classification with multi-branch hybrids [18]. One-stage detectors focus on real-time inference and represent the largest group in the subset of detectors reviewed. The YOLO family from YOLOv3 to YOLOv11 can be adapted to various applications, like apparel classification, style recognition, two-stage detection, one-stage landmark estimation, and PPE compliance monitoring through domain-specific variants such as MFD-YOLO [6], MSCG-YOLO [23], and YOLO11x [7]. SSD offers dense prediction [24] based on anchors. One alternative architecture is a one-stage model based on RetinaNet which includes focal loss [25] that focuses on robustness to handle class-imbalance. A third family is transformer based, as represented by Fashionformer-style detectors, and was opened by DETR [26]. This evolution can be documented algorithmically in methodological surveys; however, few of these surveys are conducted in the context of specific application demands.

A focused synthesis is long overdue since two parallel research streams have developed in five years with no common crossings. The first is a consumer facing and commerce focussed one, aimed at fashion and apparel visual search, catalog generation, and attribute prediction [2, 3, 4, 5]. The second, safety- and security-oriented, focuses on occupational PPE compliance [6, 8, 27]. Both streams have similar challenges: non-rigid deformation, high occlusion, scale variation, fine-grained inter-class similarity, and class imbalance, and reference one another rather superficially. The architecture of object detectors has also changed rapidly in the same time frame: YOLOv8 was released in January 2023, YOLO11 in September 2024, and the detection paradigm has shifted from fully supervised to semi-supervised training like FC-YOLOv4 [4] has recently become commonplace. This is a timely and necessary consolidation of this rapidly diversifying literature.

Previous reviews address parts of this landscape but do not integrate it. In this survey Mameli et al. [28] reviewed 219 papers from the field of fashion that included five tasks:

object detection & classification, object generation, object knowledge extraction, and object recommendation, where object detection was one of the tasks. Fashion meets computer vision, but without a detection-centric view, is also explored by Cheng et al. [13] in the realms of landmarks, parsing, retrieval, attribute recognition, style learning, synthesis and recommendation. Twenty year retrospective or review of general object detection surveys, namely Zou et al. [14] and Terven et al. [29] of YOLO architectures, provide a thorough documentation of methodology, but lack the level of detail for the clothing/uniform domain in terms of dataset, class taxonomy and evaluation reporting. Vukicevic et al. [27] discuss issues of PPE compliance as an industrial safety concern, but don't relate it to fashion detection. There are still numerous primary studies that are not synthesized [30]. There are four distinct gaps in this fragmented coverage: (1) Fashion detection and Occupational PPE detection are analyzed separately in each of these; (2) Properties of clothing are not treated as architecture and properties of architectures are not analyzed with respect to clothing properties such as non-rigid deformation and fine-grained attribute granularity; (3) The dataset landscape is divided between fashion benchmarks and safety-specific custom datasets; (4) Evaluation reporting is inconsistent across different studies, making it difficult to compare across studies.

**Table 1.** Comparison of this review against prior surveys

| Prior review | Scope                                     | Detection-centric | Fashion domain | PPE domain | Architecture-dataset-metric mapping |
|--------------|-------------------------------------------|-------------------|----------------|------------|-------------------------------------|
| [28]         | Fashion knowledge extraction (5 tasks)    | No                | Yes            | No         | No                                  |
| [13]         | Fashion meets computer vision (7 tasks)   | No                | Yes            | No         | No                                  |
| [14]         | General object detection (20-year survey) | Yes               | No             | No         | No                                  |
| [29]         | YOLO architecture review                  | Yes               | No             | No         | No                                  |

| Prior review | Scope                             | Detection-centric | Fashion domain | PPE domain | Architecture-dataset-metric mapping |
|--------------|-----------------------------------|-------------------|----------------|------------|-------------------------------------|
| [27]         | PPE compliance (industrial)       | Partial           | No             | Yes        | No                                  |
| This review  | Clothing, uniform & PPE detection | Yes               | Yes            | Yes        | Yes                                 |

Therefore, this systematic literature review (SLR) intends to fill these gaps by providing a cross-domain synthesis of deep learning-based object detection for clothing and uniform items. The objectives of this research are to answer the following five research questions (RQs):

- 1) RQ1. What deep learning architectures are most commonly used for object detection of clothing/uniform items, and what are their strengths and weaknesses?
- 2) RQ2. What types of clothing/uniform items are typically detected in object detection studies?
- 3) RQ3. What datasets are used to train and evaluate object detection models for clothing/uniform items, and what are their characteristics?
- 4) RQ4. What evaluation metrics are employed to assess object detection performance for uniform and clothing items?
- 5) RQ5. What are the main challenges and future directions for object detection of uniform and clothing items identified in the literature?

Each of these research questions maps directly onto a corresponding contribution of this review. RQ1, concerning architectural strengths and weaknesses, is addressed through the systematic taxonomy of deep learning architectures presented in Section 3.2. RQ2, concerning the types of items detected, is addressed in Section 3.3. RQ3, concerning dataset characteristics, is addressed through the consolidated catalog of public benchmarks and custom datasets in Section 3.4. RQ4, concerning evaluation metrics, is addressed through the comparative analysis of evaluation practices in Section 3.5, including a proposed minimum reporting set. RQ5, concerning challenges and future

directions, is addressed through the unified challenge taxonomy and research agenda presented in Section 3.6.

To the best of our knowledge, no previous review simultaneously meets all three of the following criteria: detection-centric scope, cross-domain fashion-PPE coverage, and explicit mapping of architecture-dataset-metrics. In comparison to this context, the following are five contributions of the present review. First, it provides a joint synthesis of fashion and PPE detection in a single corpus, which is implemented through a cross domain integration. Second, it has a narrow scope of detection, focusing only on works that conduct localization by means of bounding-box detection or instance segmentation, which differs dramatically from the broader fashion surveys proposed in [13] and [28]. Third, the 2021–2026 range covers the latest trend in the field of architecture, such as YOLOv8, YOLO11, attention-based lightweight models, transformer detectors, and new training approaches. Fourth, the review builds an explicit architecture–task–dataset–metric mapping across all studies included to facilitate like-for-like comparison, which is currently not reflected in any single paper. Finally, it brings together a common challenge taxonomy and research program across deformability, occlusion, fine-grained attributes, domain shift, and reporting standardization.

These innovations result in five tangible contributions: (a) a systematic taxonomy for deep learning architectures for clothing and uniform detection; (b) a consolidated catalogue of public benchmarks and custom datasets; (c) a comparative study of evaluation metrics and reporting practices with a proposed minimum reporting set; (d) identification of open challenges and the establishment of a prioritized research agenda; and (e) guidance for practitioners on the selection of deep learning architectures, datasets and evaluation protocols for e-commerce and PPE compliance applications. The results will be useful to the researchers because they will have a single common resource; to the practitioners in giving evidence-based selection advice; and to the policy makers in providing an evaluation of needs for evaluation standardization.

The remainder of this paper is organized as follows. Section 2 details the review methodology. Section 3 presents the results organized around the five research questions. Section 4 discusses the implications, threats to validity, and future directions. Section 5 concludes.

## 2. METHODS

This study uses the methodology of the Systematic Literature Review (SLR) that follows the PRISMA 2020 guidelines [31] and the systematic review protocol that was proposed by Kitchenham and Charters [32]. The review process involves multiple stages of the activity.

### 2.1. Search Strategy Execution

Scopus was chosen as the sole database, as, unlike a publisher, it serves as an indexing platform, pulling together peer-reviewed content from a variety of sources such as IEEE Xplore, ACM Digital Library, Springer, and Elsevier. As can be seen in the final corpus, this includes IEEE (IJCNN, IEEE Access), IEEE/CVF (CVPRW), ACM (ICDLT), Springer, Elsevier, MDPI, PLOS ONE, and Taylor & Francis – major computer vision venues were therefore not specifically excluded by this decision. However, Scopus does not index arXiv, meaning that preprint-stage contributions in the field of computer vision (where conference and preprint publication often predates or replaces journal publication) could not be included. This is a known limitation of the search strategy, which is further discussed in Section 2.5.

Two combined Boolean search strings were executed simultaneously on 25 July 2026 to collectively retrieve the two main application areas covered in this review: fashion/apparel detection and safety/PPE detection. Both queries were limited to articles and conference papers published in English in the years 2021-2026. A subject area filter was not used as the purpose of this review is to cover multiple disciplines (including computer science, engineering, and textile/apparel) which would be unduly restricted by limiting results to a single Scopus subject area. The first string targeted deep learning-based clothing and apparel detection:

TITLE-ABS-KEY ( ( "deep learning" OR "neural network" OR "machine learning" OR "artificial intelligence" ) AND ( "architecture" OR "model" OR "framework" OR "design" ) AND ( "multi-object detection" OR "object detection" OR "recognition" OR "detection" ) AND ( "clothing" OR "apparel" OR "garment" OR "fashion" ) AND ( "attributes" OR "features" OR "characteristics" OR "properties" ) ) AND PUBYEAR > 2020 AND PUBYEAR < 2027 AND DOCTYPE ( ar OR cp ) AND LANGUAGE ( english )

The second string targeted personal protective equipment and safety-compliance detection:

TITLE-ABS-KEY ( ( "personal protective equipment" OR "PPE" OR "safety helmet" OR "safety vest" OR "hard hat" ) AND ( "detection" OR "compliance" OR "recognition" ) AND ( "construction" OR "industrial" OR "worker" ) AND ( "YOLO" OR "Faster R-CNN" ) ) AND PUBYEAR > 2020 AND PUBYEAR < 2027 AND DOCTYPE ( ar OR cp ) AND LANGUAGE ( english )

The results of both queries were merged and deduplicated prior to screening.

## 2.2. Screening and Eligibility Criteria

A two-stage screening process was conducted to identify relevant studies. The inclusion and exclusion criteria used throughout the screening are presented in Table 2.

**Table 2.** Inclusion and exclusion criteria

| Inclusion Criteria                                                                                                                | Exclusion Criteria                                                                               |
|-----------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
| Peer-reviewed journal article or conference paper with empirical evaluation                                                       | Closed-access only                                                                               |
| Published between 2021 and 2026                                                                                                   | Published before 2021                                                                            |
| Full text available                                                                                                               | Image classification only (without bounding boxes or instance masks)                             |
| Written in English                                                                                                                | Non-English text                                                                                 |
| Primary focus on object detection or instance segmentation of clothing, garment, fashion, attire, uniform, PPE, or wearable items | Non-empirical document types (e.g., editorials, position papers, reviews without new evaluation) |

A search protocol was devised to guarantee reproducibility within the two main application areas described in this review. Two combined Boolean searches were conducted in one search session on 25th July 2026 to simultaneously capture fashion/apparel detection and safety/PPE detection studies. The identification process

produced 1378 records from Scopus. From these, 139 duplicates identified by either query were removed resulting in 1239 records being screened against the criteria in Table 2. Studies not fulfilling the inclusion criteria including those published outside the specified time frame, non-empirical studies, non-English publications, and studies unrelated to object detection of clothing or uniform items were excluded. This step excluded 921 records leaving 318 reports for which retrieval was sought.

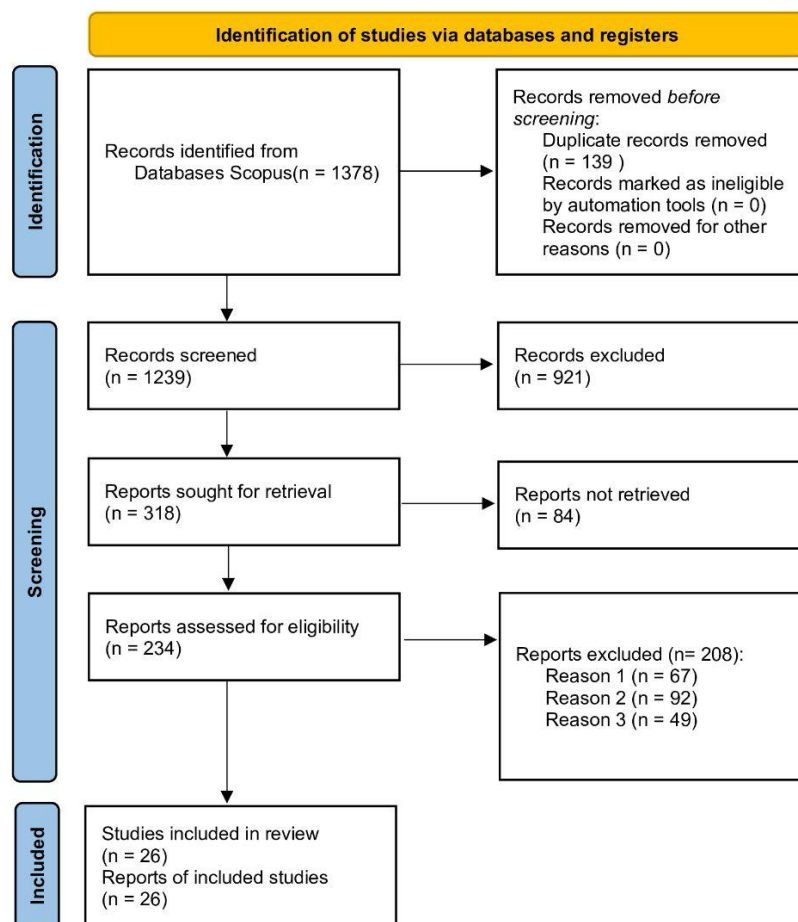


**Figure 2.** PRISMA 2020 flow diagram of the study selection process

Full-text retrieval was made through the “View at Publisher” links provided by Scopus, with the help of institutional access and open access, when available. The retrieval rate of the 318 reports requested was about 73.6%, with 84 reports not being retrieved. The remaining 234 reports were then assessed for each of the criteria in Table 2. During this step, the studies were excluded for three specific reasons: 67 reports were excluded because they were on a wrong computer vision task, such as image and semantic

segmentation or image classification but not including instance-level localization; 92 reports were excluded because they were for a wrong target domain, such as pedestrian detection without clothing attributes, and 49 reports were excluded because they lacked empirical or quantitative evaluation. 208 reports were excluded and a final set of 26 studies was included in the qualitative synthesis. The process of selecting the studies is summarized in the PRISMA flow diagram (Figure 2). Throughout this manuscript, the 26 included studies are cited as references [2]–[8], [10], [19]–[23], and [34]–[46]; all other references (e.g., [1], [9], [11]–[18], [24]–[33], [47]) serve as background or methodological citations and are not part of the empirical corpus.

### 2.3. Quality Assessment

The quality of the included 26 studies was evaluated using a five-item criterion adapted from the Critical Appraisal Skills Program (CASP) checklist [33]. The five criteria are shown below:

- 1) Q1. Clarity of research objective – whether the included study clearly states the aim or research question.
- 2) Q2. Appropriateness of methodology – whether the selected deep learning architecture and experimental design are appropriate for the stated objective.
- 3) Q3. Adequacy of data source description – whether the dataset (its size, source, class distribution, and annotation type) is adequately described.
- 4) Q4. Relevance to the review's research questions (RQs) – whether the included study reports relevant information that directly answers any of the five RQs of this SLR.
- 5) Q5. Practical or scientific applicability – whether the included study discusses the practical application, limitations, and/or contribution to the field.

Each criterion was rated as 'Yes', 'No' or 'Unclear' (insufficient information to rate). The studies were classified according to the number of 'Yes' ratings as High (all five criteria met), Moderate (three to four) or Low (two or fewer). The decision to require all five criteria for a high rating was made to ensure that the appraisal distinguished between studies, rather than giving the same rating to all.

The assessment was conducted independently by two reviewers. The first reviewer scored all 26 studies against the checklist; a second reviewer, without sight of the first

reviewer's ratings, independently scored the same studies using an explicit set of decision rules for each criterion. To reduce ambiguity, the two methodology-related criteria were anchored to reported evidence: Q2 required detection-specific metrics (mAP, AP, or IoU) rather than proxy measures such as overall accuracy or caption-quality scores, and Q5 required operational evidence such as inference speed, model size, or an explicit statement of limitations. Item-level agreement between the two reviewers was 90.8% across the 130 judgements (26 studies  $\times$  5 criteria), with a Cohen's kappa of 0.31. The kappa value is depressed by the high prevalence of "Yes" ratings — particularly for Q4, on which the reviewers agreed unanimously as a consequence of the relevance-based inclusion criteria — a known limitation of chance-corrected agreement measures under skewed marginal distributions. Disagreements were concentrated in Q2 (appropriateness of methodology), the most interpretive criterion, and were resolved through discussion against the documented evidence to reach the consensus ratings reported in Table 3. The appraisal yielded 17 studies rated High, eight rated Moderate, and one rated Low. The Moderate and Low ratings arise primarily from two recurring gaps: reliance on proxy metrics that do not directly evidence detection performance (Q2), and incomplete data-source description such as unquantified custom datasets or unreported class counts (Q3). Because the corpus had already passed a two-stage screening and eligibility process, no study fell below the Low threshold on more than three criteria; the quality assessment was used to inform interpretation of the findings and was not applied as an exclusion criterion. Table 3 documents the per-criterion ratings and a justification note for each study.

**Table 3.** Quality appraisal of the included studies (N = 26)

| Study | Q1  | Q2  | Q3  | Q4  | Q5  | Rating | Justification                                                                                   |
|-------|-----|-----|-----|-----|-----|--------|-------------------------------------------------------------------------------------------------|
| [5]   | Yes | Yes | Yes | Yes | Yes | High   | FashionFail — 2,495 images, 22 classes, semi-automatic bbox+mask; multi-architecture comparison |
| [2]   | Yes | Yes | Yes | Yes | Yes | High   | OIDV4 + custom apparel; mAP/precision/recall/F1; generalisation gap acknowledged                |

| Study | Q1  | Q2      | Q3      | Q4  | Q5      | Rating   | Justification                                                                                                  |
|-------|-----|---------|---------|-----|---------|----------|----------------------------------------------------------------------------------------------------------------|
| [4]   | Yes | Yes     | Yes     | Yes | Yes     | High     | Shopee 4,116 images; standard detection metrics + 15.6 ms inference                                            |
| [19]  | Yes | Yes     | Yes     | Yes | Yes     | High     | Mallzee 5,343 images + DeepFashion2; multi-task ablation                                                       |
| [10]  | Yes | Unclear | Yes     | Yes | Yes     | Moderate | Q2 unclear: only overall accuracy reported, no mAP/AP/IoU to verify localisation                               |
| [3]   | Yes | No      | Yes     | Yes | Unclear | Moderate | Q2 not met: detection stage evaluated only with BLEU (caption metric), no detection metric; Q5 unsupported     |
| [20]  | Yes | Yes     | Yes     | Yes | Yes     | High     | 4,455→5,141 images, 5 style classes; AP/mAP/F1 + model size + FPS                                              |
| [34]  | Yes | Unclear | Unclear | Yes | Yes     | Moderate | Q2 unclear: only Top-3 accuracy; Q3 unclear: FashionItem quantified only as ">300,000", provenance unspecified |
| [7]   | Yes | Yes     | Yes     | Yes | Yes     | High     | CHVG 1,699 images, 8 classes; mAP50 + 7.3 ms inference                                                         |
| [6]   | Yes | Yes     | Unclear | Yes | Yes     | Moderate | Q3 unclear: image count of SODA subset not reported                                                            |
| [23]  | Yes | Yes     | Yes     | Yes | Yes     | High     | Private 4,408 + SHWD 7,581; full metric set + 114 FPS cross-validation                                         |
| [8]   | Yes | Yes     | Yes     | Yes | Yes     | High     | CHV 1,330 images/9,209 instances, defined split; exemplary dataset description                                 |

| Study | Q1      | Q2      | Q3      | Q4  | Q5      | Rating   | Justification                                                                                                                   |
|-------|---------|---------|---------|-----|---------|----------|---------------------------------------------------------------------------------------------------------------------------------|
| [21]  | Yes     | Yes     | Yes     | Yes | Yes     | High     | CCP 2,098 images; ablation vs baseline; 92.98% mAP at 64 FPS                                                                    |
| [22]  | Yes     | Yes     | Yes     | Yes | Yes     | High     | DeepFashion2 fully specified; mAP + 42 ms inference                                                                             |
| [35]  | Yes     | Yes     | Yes     | Yes | Yes     | High     | Custom 11,000 images/88,725 labels, 4 compliance classes incl. negatives                                                        |
| [36]  | Yes     | Yes     | Yes     | Yes | Yes     | High     | CHVG 1,699 images, 8 classes; two-stage Fast R-CNN counterpoint                                                                 |
| [37]  | Unclear | No      | Yes     | Yes | Unclear | Low      | Aim conflates segmentation with retrieval; no detection/segmentation metric (only Hamming/Euclidean); applicability unsupported |
| [38]  | Yes     | Yes     | Unclear | Yes | Unclear | Moderate | Q3 unclear: unspecified DeepFashion2 subset; Q5 unclear: very low AP, no runtime/deployment discussion                          |
| [39]  | Yes     | Yes     | Yes     | Yes | Yes     | High     | Garment40K 140k+40k, 15 subclasses, bbox+captions; parameter-efficient (1.3M)                                                   |
| [40]  | Yes     | Yes     | Unclear | Yes | Yes     | Moderate | Q3 unclear: custom 10,700-image set reported without provenance/class distribution                                              |
| [41]  | Yes     | Yes     | Yes     | Yes | Yes     | High     | Helmet/mask 8,000 + identity 7,800/9,850; dual-stage evaluation                                                                 |
| [42]  | Yes     | Unclear | Yes     | Yes | Yes     | Moderate | Q2 unclear: overall accuracy + FPS only, no mAP/per-class AP for RFA-YOLO                                                       |

| Study | Q1  | Q2      | Q3      | Q4  | Q5  | Rating   | Justification                                                                                                                |
|-------|-----|---------|---------|-----|-----|----------|------------------------------------------------------------------------------------------------------------------------------|
| [43]  | Yes | Yes     | Yes     | Yes | Yes | High     | SHWD 7,581 images; ablation with precision/recall/mAP/FPS/params                                                             |
| [44]  | Yes | Yes     | Yes     | Yes | Yes | High     | CHVG 1,699 images (source dataset); mAP/FPS across YOLOX scales                                                              |
| [45]  | Yes | Yes     | Yes     | Yes | Yes | High     | MakeML 5,000 images, 60:20:20 split; full metric set                                                                         |
| [46]  | Yes | Unclear | Unclear | Yes | Yes | Moderate | Q2 unclear: only precision/FPR/miss-rate, no mAP/IoU; Q3 unclear: class taxonomy not applicable (region-intersection method) |

#### 2.4. Data Extraction and Thematic Analysis

The 26 studies included were systematically coded and a set of characteristics identified across each study. The deep learning architecture, the backbone network, the name and description of the dataset, the evaluation metrics, the clothing or uniform objects identified, the main findings, and challenges or future directions were extracted. All studies used a standardised form for data extraction. Data extracted were then analysed thematically using an approach advocated by Braun and Clarke [47] which focuses on the identification of common patterns and themes within the selected articles. Research studies with similar characteristics were classified into the major theme categories based on the five research questions. For instance, in the context of RQ1, the studies were grouped together according to the category of architecture (e.g., YOLO family, two-stage detectors, transformer-based detectors). Likewise, datasets and assessment metrics were aligned with RQ3 and RQ4, and challenges and future directions were combined for RQ5. The thematic synthesis allowed for a structured comparison of the architecture strengths, the characteristics of the datasets, and reported performance metrics in both the fashion and uniform/PPE fields.

## 2.5. Limitations

There are some limitations to this review that need to be acknowledged. Firstly, database search was carried out using only one database – Scopus. While Scopus offers broad coverage of computer vision and deep learning literature due to its inclusion of peer-reviewed publications from IEEE Xplore, ACM Digital Library, Springer and Elsevier, it does not include arXiv, and relevant contributions in the preprint stage could not be included in this review. This is especially pertinent for computer vision, since the field favours publishing in conferences and preprint platforms such as arXiv.

Second, the search was limited to articles and conference papers published in English only and between 2021 and 2026, so that important foundational work published prior to 2021 was not synthesised in this review, though it was referred to in the introduction. Third, although full-text retrieval succeeded for the majority of sought reports (a retrieval rate of approximately 73.6%), 84 of 318 reports could not be retrieved through Scopus's "View at Publisher" links, institutional access, and open-access channels; interlibrary loan and direct author contact were not pursued. This remains a residual source of potential selection bias.

Fourth, the Boolean query targeting the safety/PPE domain required the presence of "YOLO" or "Faster R-CNN" as a mandatory term, reflecting the architectural families most commonly reported in this literature at the time of query design. This design choice means the search could not surface PPE studies employing other architectures (e.g., RetinaNet, SSD, DETR, EfficientDet) unless they also referenced YOLO or Faster R-CNN. Consequently, the pronounced YOLO-family dominance observed within the safety/PPE domain (Section 3.2) should be interpreted with this search constraint in mind, and may not fully represent the architectural diversity of the broader PPE detection literature. This asymmetry also means the two domain-specific queries differ in maturity and construction, which may partly contribute to the observed balance between the fashion/apparel and safety/PPE domains.

Fifth, the quality appraisal was based on the data extracted into the master extraction table rather than on independent re-reading of all 26 full texts by the second reviewer. While the extraction fields were verified against the primary sources, criteria that depend on narrative content not fully captured in the table — such as an explicit limitations

discussion – were rated conservatively as "Unclear" where the information could not be confirmed. In addition, item-level agreement, though high at 90.8%, corresponds to a modest Cohen's kappa (0.31) owing to the skewed distribution of ratings; the consensus ratings in Table 3 should therefore be read as an informative appraisal rather than a definitive quality ranking. Sixth, the review focused exclusively on object detection and instance segmentation studies; related tasks such as image classification, landmark estimation, or clothing parsing without bounding-box localization were excluded by design.

### 3. RESULTS AND DISCUSSION

#### 3.1. Data Reporting

A structured data extraction template was established to collect relevant information from each included study. Specifically, deep learning model / architecture and backbone, dataset name and characteristics, type of clothing/ uniform detected, evaluation metrics, key findings, and challenges or future directions were extracted. Such information was correspondent to the five research questions (RQs) addressed in this review. Eventually, a total of 26 empirical studies were included in the qualitative synthesis, which were published between 2021 and 2026, with the majority of them appeared in 2022 (n=9) and 2023 (n=7), suggesting the period of most active model design and development, as shown in Table 4.

**Table 4.** Summary of Included Studies

| Ref | Year | Architecture                                                        | Dataset                        | Images     | Classes | Best Result                                 |
|-----|------|---------------------------------------------------------------------|--------------------------------|------------|---------|---------------------------------------------|
| [5] | 2024 | Transformer<br>(Fashionformer,<br>Attribute-Mask R-<br>CNN, Facere) | FashionFail                    | 2,495      | 22      | 93.4% weighted<br>mAP                       |
| [2] | 2022 | One-stage: YOLOv3-<br>SPP                                           | OIDV4 and<br>custom<br>apparel | 6,000; 500 | 6; 5    | 98.4% mAP<br>(custom); 42.4%<br>mAP (OIDV4) |

|      |      |                                              |                            |                         |      |                                |
|------|------|----------------------------------------------|----------------------------|-------------------------|------|--------------------------------|
| [4]  | 2022 | One-stage: FC-YOLOv4                         | Shopee                     | 4,116                   | 5    | 96.01% mAP; 15.6 ms            |
| [19] | 2022 | Two-stage: Faster R-CNN with EfficientNet-B4 | Mallzee and DeepFashion2   | 5,343                   | 15   | 97.57% Top-5 accuracy          |
| [10] | 2022 | One-stage: YOLOv3, Tiny YOLOv3               | Custom                     | 12,000                  | 8    | At least 90% accuracy          |
| [3]  | 2024 | One-stage: YOLOv8 with CNN encoder-decoder   | Custom                     | 863, augmented to 5,411 | 26   | BLEU-1 0.983                   |
| [20] | 2022 | One-stage: YOLOv5s                           | DeepFashion subset and web | 5,141                   | 5    | 90-99% AP per style            |
| [34] | 2023 | Two-stage: Faster R-CNN (upstream YOLO)      | FashionItem                | Over 300,000            | 33   | 94.56% Top-3 accuracy          |
| [7]  | 2026 | One-stage: YOLO11x                           | CHVG                       | 1,699                   | 8    | 96.9% mAP50; 7.3 ms            |
| [6]  | 2025 | One-stage: MFD-YOLO                          | SODA and CHV               | Not reported            | 3    | 74.4% mAP@0.5                  |
| [23] | 2025 | One-stage: MSCG-YOLO                         | Private and SHWD           | 4,408; 7,581            | 4; 2 | 91.6% AP50; 114 FPS            |
| [8]  | 2021 | One-stage: YOLOv3, v4, v5                    | CHV                        | 1,330                   | 6    | 86.55% mAP (v5x); 52 FPS (v5s) |
| [21] | 2021 | One-stage: YOLOv4-TPD                        | CCP                        | 2,098                   | 3    | 92.98% mAP; 64 FPS             |
| [22] | 2021 | One-stage: EfficientDet                      | DeepFashion2               | 491,895                 | 13   | 0.686 mAP; 42 ms               |

|      |      |                                            |                            |                 |    |                                                  |
|------|------|--------------------------------------------|----------------------------|-----------------|----|--------------------------------------------------|
| [35] | 2022 | One-stage: YOLOv3, v4, v7                  | Custom PPE                 | 11,000          | 4  | 97% mAP; 25 FPS                                  |
| [36] | 2023 | Two-stage: Fast R-CNN                      | CHVG                       | 1,699           | 8  | 96% mAP                                          |
| [37] | 2023 | Two-stage: Mask R-CNN                      | DeepFashion2               | 491,895         | 13 | Hamming and Euclidean distance                   |
| [38] | 2022 | Two-stage: Faster R-CNN and Mask R-CNN     | DeepFashion2 subset (wild) | Subset          | 13 | 21.05% AP (Mask R-CNN); 15.12% AP (Faster R-CNN) |
| [39] | 2023 | Transformer: Grounding DINO (improved)     | Garment40K                 | 40,000; 140,000 | 15 | 43.5 AP with 1.3M parameters                     |
| [40] | 2023 | One-stage: YOLO_CA                         | Custom helmet              | 10,700          | 2  | 94.8% mAP; 134 FPS                               |
| [41] | 2023 | One-stage: YOLOv5 with Facenet and SVM     | Kaggle and LFW-based       | 8,000           | 3  | 0.954 mAP; 95.4% identity accuracy               |
| [42] | 2023 | One-stage: RFA-YOLO with Random Forest     | ODPD                       | 10,000          | 4  | 93.1% accuracy; 13 FPS                           |
| [43] | 2024 | One-stage: YOLOv5 with SPPM, FAM, Soft-NMS | SHWD                       | 7,581           | 2  | 96.1% mAP; 97.3% precision                       |
| [44] | 2022 | One-stage: YOLOX-m                         | CHVG                       | 1,699           | 8  | 89.84% mAP                                       |
| [45] | 2022 | One-stage: YOLOv5x                         | MakeML Hard Hat            | 5,000           | 2  | 92.44% mAP; 45 FPS                               |

|      |      |                                    |        |                                  |                           |                                                   |
|------|------|------------------------------------|--------|----------------------------------|---------------------------|---------------------------------------------------|
| [46] | 2021 | One-stage: YOLOv4<br>with OpenPose | Custom | 8,000,<br>augmented<br>to 40,000 | N/A<br>(region-<br>based) | 13.7% false-<br>positive rate;<br>15.8% miss rate |
|------|------|------------------------------------|--------|----------------------------------|---------------------------|---------------------------------------------------|

In terms of application domains, the included studies address three distinct areas: the majority address the safety and PPE compliance domain ( $n=13$ ), which reflects an urgent practical need for automated monitoring in industrial and construction settings. A close second is the fashion and apparel domain ( $n=12$ ), driven by the interest in e-commerce and visual search. One paper addresses a niche use case of estimating clothing insulation for smart thermostats. Thus, the spectrum of deep learning applications for clothing and uniform items is broad, but the safety-driven and commercial-driven research appear to be in different tracks. In terms of research questions' coverage, all 26 included studies covered the entire set of five RQs defined in this review, thus providing a solid basis for the subsequent thematic synthesis.

### 3.2. Deep Learning Architectures

The analysis of the 26 included studies shows a clear picture of the different deep learning architectures used in clothing and uniform object detection. Three families of architectures were found in the corpus: one-stage detectors (the YOLO family), two-stage detectors (the R-CNN family), and transformer-based detectors as an emerging third family. A summary of the detected architectures is shown in Figure 3.

One-stage: YOLO family (18 studies, 69.2%)    Transformer-based (2 studies, 7.7%)  
 Two-stage: R-CNN family (5 studies, 19.2%)    Other one-stage (1 study, 3.8%)

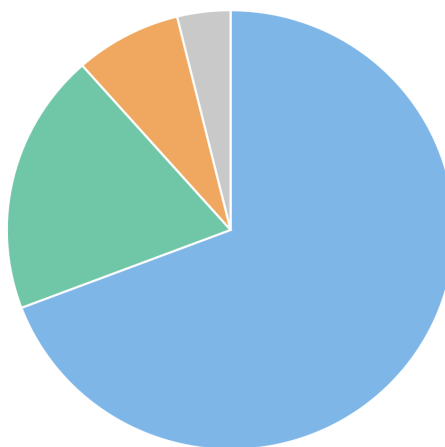


**Figure 3.** Distribution of deep learning architecture families across 26 included studies

One-stage detectors are the most popular detector type, being used in 19 of 26 papers, of which 18 used the YOLO family (YOLOv3 to YOLO11) in both PPE and fashion areas. Some of the key performance metrics are: YOLOv5x published 86.55% mAP at 52 FPS [8]; YOLOv7 published 97.29% mAP for PPE compliance [35]; YOLO11x published 96.9% mAP50 at 7.3 ms inference [7]. The newly added PPE studies report accuracies between 92.4% to 96.7% mAP [44], [39], [42] and scale- and edge-specific variants like MFD-YOLO [6] and MSCG-YOLO [23] achieve even better performance. In addition to YOLO, EfficientDet can detect clothing and estimate the corresponding landmarks simultaneously in 42ms per frame [22] which is the only one-stage architecture in the corpus other than YOLO. The primary strength of one-stage architectures is real-time deployability; their main drawbacks are the challenge of detecting small objects, being easily confused by visually similar objects [8] and their limited precision in localization as compared to two-stage detectors [4]. The accuracy numbers reported in this paper cannot be directly compared between different studies because they are based on different data sets, different class taxonomies, different IoU thresholds, and different hardware setups (such as the GPU model used and the batch size). These are not consistently reported across the corpus (Section 3.5).

Overall, the two-stage detectors (R-CNN family) emerged from five studies, with a higher localization accuracy achieved by separating region-proposal and classification from each other. The faster R-CNN algorithm achieved an mAP of 80.6% on the Mallzee fashion dataset [19] and Fast R-CNN had an mAP of 96% on eight PPE classes in the CHVG dataset [35]. In three of these works [5, 37, 38], they adopted Mask R-CNN and it achieved about 20% higher AP than Faster R-CNN on DeepFashion2 wild images [38] while suffering from the problem of class imbalance [37] or distributional shift in context-free studio images [5]. The fifth one [34] uses a YOLO-based person detector followed by Faster R-CNN in the pipeline, showing a combination of the R-CNN family in a multi-stage fashion for 33 garment subcategories. A major advantage of two-stage detectors is their fine-grain localization, which is ideal for recognizing fashion attributes; however, their key drawback is inference latency, where the original R-CNN took about 47 seconds per image [36] and is therefore not suitable for real-time surveillance deployments without dedicated hardware.

Transformer-based detectors are mentioned in two studies and employ self-attention mechanisms to enable global context modeling capabilities that traditional convolutional approaches lack. Fashionformers achieved state-of-the-art results on Fashionpedia-val but performed significantly worse on out-of-distribution studio images [5], while a refined Grounding DINO with adapter tuning achieved 43.5 AP on Garment40K using only 1.3 million trainable parameters while avoiding catastrophic forgetting [39]. The primary advantage of this family is open-set generalization through language-guided inference; common limitations include excessive computational demands, domain shift sensitivity, and language ambiguity leading to erroneous body part identification as clothing items [39].

### 3.3. Types of Clothing/Uniform Items

The 26 included studies target three broad item categories that mirror the application domains identified in Section 3.1: general clothing and apparel items, safety and occupational PPE items, and specialized niche items. A summary of the distribution is presented in Table 5.

**Table 5.** Distribution of Detected Item Types Across Included Studies

| Theme                 | Item Category    | Representative Items                                     | Studies (n) |
|-----------------------|------------------|----------------------------------------------------------|-------------|
| Fashion & Apparel     | General clothing | Tops, bottoms, full-body garments, footwear, accessories | 12          |
| Safety & Occupational | PPE / uniforms   | Helmets, safety vests, glasses, gloves, masks            | 13          |
| Specialized/Niche     | Domain-specific  | Clothing-insulation garments                             | 1           |

Fashion and apparel studies cover the most extended range of item categories, including tops, bottoms, upper wear, footwear, and accessories at both coarse and fine-grained levels. On the coarse level, studies such as Lee & Lin [20] and Thwe et al. [4] target categories of jackets, tops, pants, skirts, and dresses. Whereas Papadopoulos et al. [19]

and Amin et al. [34] predict more fine-grained subcategories encompassing blazers, blouses, hoodies, jumpsuits, suits, and footwear, including boots, heels, and accessories like belts, bags, or jewelry. The most granular taxonomy appears in Agarwal et al. [3] with 26 fashion classes including culturally specific items such as sarees and kurtas, and in Ma & Xu [39] with 15 subclasses under two macro-categories of pants and tops. Studies using the DeepFashion2 benchmark consistently adopt its standard 13-class taxonomy covering sleeve-length and silhouette variants of tops, outerwear, dresses, and bottoms [22], [37], [38]. A recurring characteristic of this theme is the deliberate exclusion of fine-grained part-level attributes — such as sleeves, pockets, and closures — from detection scope in favor of main garment boundaries reflecting a design choice to maintain tractable class taxonomies [37].

The safety and PPE studies converge around a smaller but still operationally important set of items focused on the detection of helmets, safety vests, or safety glasses, in the majority of the thirteen PPE studies. Some studies have added more categories, such as gloves, masks, safety shoes and face shields, and Wang et al. [8] added the worker role sub-classification of helmets based on color categorization, including four colors: blue, red, white, and yellow, to make worker role identification possible beyond simple compliance detection [7] [36]. In the new studies, Kwak & Kim [41] are able to detect both helmet and mask compliance, as well as worker ID, while Ji et al. detect PPE in a combination of persons-helmet-workwear for offshore drilling platforms, demonstrating the widening of the domain of PPE detection into more specialized operating environments. In PPE studies, the detected item set is always accompanied by a "person" or "head" class, which is used as an anchor for association-based compliance checking, and the negative classes like "no helmet" and "no vest" are also included to allow the direct detection of the class. This item set is specific, as opposed to fashion studies, because the application is safety-critical, with real-life implications in cases of misclassification.

The only specialized study, Medina et al. [10], categorizes garments based on their thermal insulation properties into eight classes for use with smart thermostats, regrouping items not by style but by heat retention value. This study highlights that the scope of use cases within the domain of clothing can range from aesthetic to non-visual, functional

properties, though its dearth within the corpus points to its nascent status within the domain of fashion and PPE.

### 3.4. Datasets

The datasets used across the 26 included studies reflect the structural divide between application domains, ranging from large-scale public fashion benchmarks to purpose-built safety and industrial collections. Three themes emerge from the dataset landscape: large-scale public fashion benchmarks, reused public PPE benchmarks, and self-curated or specialized collections. A summary is presented in Figure 4.

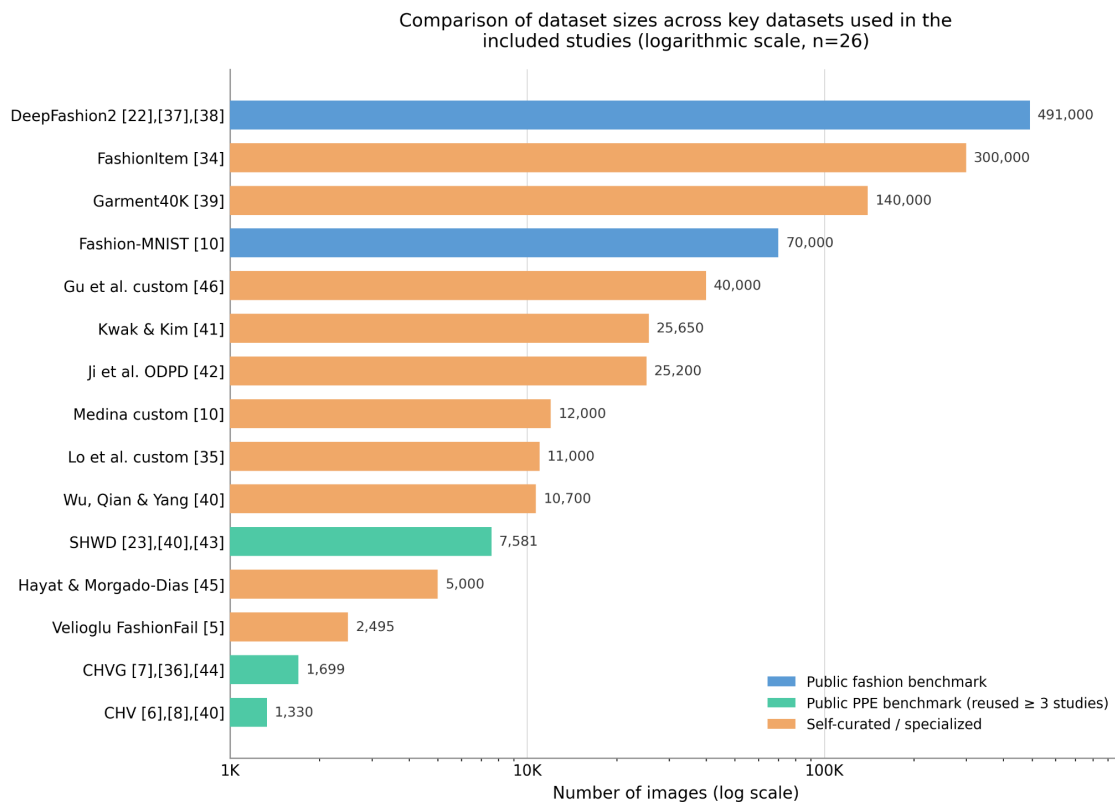


**Figure 4.** Comparison of dataset sizes (number of images) across key datasets used in the included studies (logarithmic scale)

The most popular public fashion benchmarks for fashion and apparel studies are DeepFashion2 [22], [37], [38]. DeepFashion2 comprises about 491,895 images of 13 different clothing categories and is well-annotated with bounding boxes, segmentation masks, landmarks, scale, occlusion levels, and viewpoint information. Kim et al. [22] claim that they have 191,961 images for training and 32,153 images for validation. Medina et al.

[10] used fashion images for 70,000 low-resolution images of grayscale fashion products across 10 categories, which were evaluated as a classification baseline, and were explicitly rejected in favor of a custom YOLO-based detection dataset, because they failed to support multi-garment location.

While safety and PPE studies are building on a limited number of public 'safety benchmarks' instead of creating their own. CHVG is a new benchmark that spans 8 PPE classes from Ferdous & Ahsan [44] and was later used by two additional studies in this corpus [35,7] for benchmark consolidation over PPE. As for the primary training, Tong et al. [6] and Wu, Qian, & Yang [39] both used the CHV (9209 annotated instances) 1,330-image data set for training and cross-validation, respectively. Su & Gao [43] used a 7,581-image public set of helmets as the main data set, while Lin, Zhang, & Chen [23] and Wu, Qian, & Yang [40] used it as a cross validation set. The three studies using each of these approaches (CHV, CHVG, and SHWD) indicate a growing standardisation, albeit still informal, in the PPE detection literature which is not parallel or in any way matched by the disjointed nature of the various approaches to fashion-benchmarking.

There are several studies that introduce custom datasets to provide flexibility in data sets that may fall outside the current public benchmarks. The latter three are all large scale open set datasets: Ma & Xu [39] built Garment40K, with more than 140,000 images divided into 40,000 garment sets, each portraying a garment from multiple angles; Amin et al. [34] created FashionItem, which includes over 300,000 images across 33 subclasses and 150 attribute classes, sourced from film and television; and Velioglu et al. [5] gathered FashionFail, comprising high-resolution clean studio images and clearly set out to document the distributional problem between in-the-wild training data and context-free product photography, with 2,495 pairs of images. Lo et al. [35] created a large custom dataset of 11,000 images and 88,725 bounding-box labels with negative classes, such as 'no hard hat', for direct violation detection; Wu, Qian, & Yang [40] built a helmet dataset of 10,700 detection images; Ji et al. [42] added the Offshore Drilling Platform Dataset (ODPD) with 10,000 detection images, two sets for identity verification: 7,800 and 9,850 images respectively, and 2,000 evaluation images specifically designed for offshore operational scenarios; Gu et al. [46] enlarged an initial 8,000-image dataset to 40,000 images using a dataset augmentation technique; and Kwak & Kim [41] extended an original 8,000-image dataset with two separate identity-verification datasets of 7,800 images

and 9,850 images and a 2,000-image evaluation set designed to represent the operational context of offshore drilling. Medina et al. [10] is the only dedicated study to developing a thermal-insulation dataset for 12,000 images and adding 2,000 images. Both sets of self-curated datasets are targeted to address specific deployment gaps and have relatively smaller scales, and are less reproducible between studies than the reused public datasets of PPE described above.

### 3.5. Evaluation Metrics

The evaluation metrics employed across the 26 included studies reflect two complementary priorities: accuracy-oriented metrics that quantify detection and classification quality, and efficiency-oriented metrics that assess computational suitability for real-time deployment. A summary of metric usage across studies is presented in Table 6.

**Table 6.** Evaluation Metrics Employed Across Included Studies

| Theme             | Metric                        | Description                                 | Studies                                                                                                |
|-------------------|-------------------------------|---------------------------------------------|--------------------------------------------------------------------------------------------------------|
| Accuracy-oriented | mAP / mAP@0.5 / mAP@0.5:0.95  | Primary detection accuracy measure          | [2], [4], [7], [8], [19], [20], [21], [22], [23], [35], [36], [38], [39], [40], [41], [43], [44], [45] |
| Accuracy-oriented | Precision / Recall / F1-score | Classification correctness and completeness | [2], [4], [7], [8], [10], [20], [21], [23], [35], [36], [39], [40], [43], [46]                         |
| Accuracy-oriented | AP / AP50 / AP75              | Per-class or threshold-specific precision   | [4], [19], [22], [23], [38], [39]                                                                      |
| Accuracy-oriented | IoU / GIoU                    | Bounding box overlap quality                | [2], [4], [8], [21], [22], [38]                                                                        |
| Accuracy-oriented | Top-K Accuracy / Recall       | Classification ranking metric               | [19], [34]                                                                                             |
| Accuracy-oriented | Task-specific accuracy        | Identity or PPE-combination correctness     | [41], [42]                                                                                             |

| Theme               | Metric                       | Description                             | Studies                                                  |
|---------------------|------------------------------|-----------------------------------------|----------------------------------------------------------|
| Efficiency-oriented | FPS / Inference time (ms)    | Real-time deployability                 | [4], [7], [20], [23], [35], [36], [40], [42], [43], [45] |
| Efficiency-oriented | Parameters / FLOPs / GFLOPs  | Model complexity and computational cost | [23]                                                     |
| Specialized         | BLEU score                   | Caption generation quality              | [3]                                                      |
| Specialized         | Hamming / Euclidean distance | Texture similarity measure              | [37]                                                     |
| Specialized         | Zero-shot Transfer AP        | Open-set generalization                 | [39]                                                     |

In the corpus, the most common measurement method is the Mean Average Precision (mAP) metric, used in 18 of 26 studies, and used as the threshold to measure the performance of the architectures for detection over the studies and data sets. Some studies use a fixed IoU threshold for calculating the mAP, i.e., mAP@0.5 or AP50 where the same means the percentage of successfully localized detections at IoU 50% [7], [8], [36], while mAP at various IoU thresholds requires a stricter assessment of localization using varying thresholds, including 0.5:0.95 [19, 23, 38]. In 14 studies the mAP is consistently reported together with complementary metrics like precision, recall and F1-score. Precision and recall metrics serve as a way to assess the quality of detection and link false positive and false negative results and are important in PPE compliance set-ups where omission of a helmet detection has direct safety implications [7], [34], [35]. Intersection over Union (IoU) is used as a criterion in 6 studies both on its own and as a part of GloU loss in bounding box regression (optimisation) [2],[8]. For fine-grained classification tasks with more than one possible class for each image, a smaller set of narrowly-tuned fashion research uses Top-K accuracy and Top-K recall [19, 34]. Two of the new PPE studies give an accuracy value for the task instead of mAP: worker identity verification [41] and multi-class PPE-combination correctness on an offshore platform [42].

The efficiency results can be found across ten studies, but continue to tightly focus within safety and PPE, where real-time inference is in direct service of the operations. The efficiency measures are in terms of Frames Per Second (FPS) and per-image inference time in milliseconds, from 52 FPS for the smallest YOLOv5s [8] to 134 FPS for the coordinate-attention-based YOLO\_CA architecture [40] that already reports significant improvements over the past few years in terms of throughput. To facilitate multi-dimensional assessment of the speed-accuracy-complexity trade-off, Lin, Zhang, & Chen [23] report parameter count, FLOPs and GFLOPs as well as accuracy metrics in this corpus. One interesting observation is that e-commerce studies in fashion do not capture the latency requirements of a visual search system as well, likely due to the lower latency requirements of shopping compared to a continuous surveillance system.

Three studies use evaluation metrics that do not follow the standard conventions for detection evaluation, but are task-specific. As the pipeline goes beyond detecting clothing objects, Agarwal et al. [3] report BLEU scores as the main metric for assessing the quality of the caption of the created clothing images. In a detection-plus-retrieval pipeline, Vararu et al. [37] quantify the similarity of clothing textures by comparing the colors using Hamming distance and comparing the Local Binary Pattern histograms using Euclidean distance. Ma and Xu [39] introduce Zero-Shot Transfer AP to quantify the ability to recognize clothing categories not seen during fine-tuning. The specific metrics used in the literature, along with task-specific accuracy measures developed by the newly added offshore and identity-verification studies [41, 42] highlight that evaluation protocols are often tailored to the specific formulation of the task, instead of standardized across studies, as discussed in Section 3.7.

### 3.6. Challenges & Future Directions

The challenges and future directions identified across the 26 included studies reveal four recurring themes that cut across both the fashion and safety domains: visual and geometric challenges, data and annotation constraints, model reliability and generalization issues, and emerging future directions. A summary is presented in Table 7.

**Table 7.** Challenges and Future Directions Across Included Studies

| Theme              | Sub-challenges / Directions                                                    | Representative Studies                                          |
|--------------------|--------------------------------------------------------------------------------|-----------------------------------------------------------------|
| Visual & Geometric | Occlusion, deformation, scale variation, viewpoint, inter-class similarity     | [4],[10],[19],[21],[22],[23],[34],[35],[36],[38],[40],[43],[46] |
| Data & Annotation  | Class imbalance, limited annotated data, lack of standardized benchmarks       | [3],[4],[7],[19],[37],[38]                                      |
| Model Reliability  | Small object detection, false positives, domain shift, catastrophic forgetting | [5],[8],[20],[39],[46]                                          |

Visual and geometric challenges dominate in the corpus, impacting both fashion and PPE. Occlusion is the most common challenge that arises when parts of garments or PPE are hidden by body parts, hair, overlapping items or objects in the environment [21], [35], [38]. This is of particular concern in wild fashion images where complex poses hide substantial amounts of clothing [4], [19]. Non-rigid deformations further exacerbate the problem: clothing items tend to be highly deformed in different poses, viewpoints and body shapes, leading to high intra-class variances in appearance that hurt model consistency [34], [22]. Among the newly incorporated PPE studies, this challenge extends to attention-based feature refinement: Wu, Qian, & Yang [40] and Su & Gao [43] both introduce architectural modifications – coordinate attention and a dilated-convolution feature alignment module, respectively – explicitly to improve detection of small or partially occluded PPE items at construction-site scale. High inter-class visual similarity is an additional source of error, particularly between visually adjacent categories such as skirts and short dresses [4], [34], shirts and T-shirts [10], and green T-shirts misidentified as safety vests [8], reflecting the fundamental difficulty of fine-grained discrimination in clothing detection.

For the training and evaluation of the detection models, there is a structural constraint in the form of data and annotation constraints that affect the corpus. The most consistently reported data-level problem is class imbalance; in safety datasets, the number of persons and helmets is much greater than the number of gloves and glasses

[7] and, in fashion benchmarks, categories like sling and short-sleeved outerwear have fewer instances than other categories, but the models still perform poorly on those instances, despite their high overall mAP scores [38]. Because of the lack of large-scale annotated data sources, especially in culturally specific and specialized application areas, for instance, culturally specific apparel taxonomy [3] data sources, researchers have to use custom data sources, where there are distributional gaps between the training and deployment environments. This has been noted as one of the key hurdles to a direct comparison across studies in previous reviews of this literature [28] and this is confirmed by the range of image sources used, degree of granularity in the clothing annotations, clothing taxonomy, and evaluation protocols seen in the datasets analysed in this review (Section 3.4). A parallel concern is the cost of annotations: A segmentation mask for each pixel for instance segmentation models requires orders of magnitude more time to obtain than a bounding box label, making the size of the segmentation capable datasets small [5].

Architectural choices and deployment conditions give rise to the concepts of model reliability and generalization. Small object detection is still a significant challenge in one-stage approaches, especially YOLO-based models, as they are unable to accurately detect the location of PPE items at a long distance or in unfavorable environmental conditions. One of the most important failure modes of transformer-based models is domain shift: Velioglu et al. [5] show that Fashionformer is state-of-the-art on the in-the-wild Fashionpedia images, but suffers significantly on clean e-commerce studio images in which the human body lacks context. This indicates that transformer models are sensitive to distributional difference, which will hinder their cross-domain usefulness. The catastrophic forgetting problem has been particularly addressed in transformer-based open-set detection by Ma & Xu [39] because the model was found to forget its general visual representations learned during pre-training when training on clothing data. Another reliability concern are false positives due to background clutter, such as faces being detected as helmet-related when privacy blurring is used, which lowers the detection accuracy of helmets by around 7% [8]. This trade-off is also explicitly quantified in the newly incorporated studies, which report a false-positive rate of 13.7% and a miss rate of 15.8% for their pose-aware head-region localization approach; that is, reliability constraints also exist in architectures explicitly designed to address these issues.

The research priorities that emerged from the corpus are narrowed down across the four interconnected research directions. First, low resource and data-efficient learning paradigms such as semi-supervised learning, for instance FC-YOLOv4 [4] are suggested for annotation scarce scenarios without compromising the model performance. Second, transformer-vision-language integration is shown as a key architectural frontier; transformer-based parameter-efficient fine-tuning methods have been studied [39] and open-set detection approaches that leverage language-based inference to generalize to unseen clothing categories [38]. Third, edge and embedded deployment is a common practical challenge in the PPE domain and studies suggest model compression via pruning [6] and quantization [7] to be able to perform real-time inference on resource constrained surveillance hardware. Fourth, expansion and standardization of datasets is always called for, such as creating cross domain benchmarks across fashion and PPE domains, and collecting data across underrepresented cultural and geographic contexts [3], [19], [34].

### **3.7. Discussion**

This section interprets the findings presented across the five result dimensions — architectures, item types, datasets, evaluation metrics, and challenges — to provide a coherent understanding of the current state of deep learning-based object detection for clothing and uniform items, and to situate these findings within the broader context of the field.

#### **1) Architectural convergence and its implications**

The data show an overwhelming numerical dominance of one-stage YOLO-based architectures, appearing in 19 of 26 studies, with the YOLO family alone accounting for 18. This dominance is most pronounced within the safety/PPE domain (16 of 18 YOLO-family studies), and should be interpreted alongside a specific search-design constraint: as noted in Section 2.5, the Boolean query targeting the safety/PPE domain required "YOLO" or "Faster R-CNN" as a mandatory term. Consequently, the corpus could not surface PPE studies employing other architectures such as RetinaNet, SSD, or DETR unless they also referenced YOLO or Faster R-CNN. We interpret this to mean that the observed YOLO dominance within the PPE domain reflects, at least in part, the search protocol rather than an unconstrained finding about the field's true architectural diversity; the fashion/apparel domain, whose query carried no such architectural

restriction, shows comparatively greater architectural variety (transformer-based, two-stage, and single non-YOLO one-stage representation). Two-stage detectors, despite slower inference, are reported to achieve superior localization precision and fine-grained classification in this corpus [19], [38]. Transformer-based detectors are represented by only two studies [5], [39]; given this small sample, claims regarding their generalizability or superiority over other families cannot be substantiated by this corpus and should be treated as preliminary observations rather than established findings.

## **2) Fragmentation between fashion and safety research streams**

A fundamental finding of this review is the existence of two largely non-intersecting research threads, one focused on the development of techniques for fashion and apparel with a commercial e-commerce driver ( $n = 12$ ), and the one focused on the development of techniques for safety and PPE with an occupational compliance driver ( $n = 13$ ). Though sharing the technical challenges of occlusion, scale variation, non-rigid deformation, and inter-class similarity (Section 3.6), these threads cite each other only sparingly within this corpus. This division is reflected in the body of datasets (Section 3.4), where DeepFashion2 (approximately 491,000 images [22], [37], [38]) and FashionItem (more than

300,000 images [34]) vastly outnumber the largest PPE dataset in the corpus (40,000 images after augmentation [46]), giving the latter a systematic disadvantage in terms of scale. This imbalance is partially compensated for by the emerging pattern of benchmark re-use specific to the PPE domain: CHV, CHVG, and SHWD are each used in three studies in this corpus, a level of convergence that has no equivalent in the largely self-curated and fragmented fashion domain. We view this pattern as informal standardization within the PPE thread rather than coordinated benchmark selection at the level of the research community, as none of the studies comment on the subject in any capacity. Overall, the citation pattern between the two threads and the scale-driven division in dataset availability point towards the same conclusion: the lack of cohesion. Architectural innovations targeting scale variation in the domain of construction site surveillance [40], [43] might be productively applied to the detection of fine-grained accessories in fashion benchmarks, and the insights from the annotation of large-scale fashion datasets might enable more granular compliance-focused PPE systems, but this review does not identify any empirical studies that investigate this possibility. Future work should productively

address this state of affairs through the identification of unified cross-domain benchmarks and the conduction of cross-domain experiments.

### **3) Ethical and privacy considerations in PPE surveillance**

The safety/PPE domain also involves ethical issues beyond the accuracy of the detection. The lack of systematic discussion regarding ethics of meeting worker privacy protection and system reliability, as seen by Wang et al. [8] which shows that worker privacy protection affects the accuracy in detecting helmets by around 7%, is another direct trade-off seldom discussed in the reviewed studies. This is more evident in Kwak & Kim [41] where the system does not only detect compliance with PPE, but also verifies the identity of the individual wearing it as well, so additional issues around consent and surveillance are not explored when evaluating the system. The 13 studies in this corpus do not mention the concept of worker consent, data governance, or being subject to a surveillance-ethics framework. We note this as a gap the corpus reveals rather than a claim that such engagement is absent from the broader PPE detection literature.

### **4) Cultural and geographic bias in dataset composition**

The majority of these data sets are based on dress conventions and construction site conditions which are culturally and geographically diverse but not explicitly indicated. A notable exception is Agarwal et al. [3] where the culturally specific garment categories of sarees and kurtas have been taken up in a 26-class taxonomy. In none of the PPE studies in this corpus is reference made to differences in worker attire, standards, or site conditions between the countries from which the data in these studies is from. Ecologically that isn't necessarily a sign of cultural or geographic under-representativeness of the field as a whole; it simply means that it is not featured in the studies that were analyzed.

### **5) Evaluation inconsistency and its consequences**

The diversity of evaluation protocols in this corpus — spanning mAP@0.5, COCO-style mAP@0.5:0.95, Top-K accuracy, and task-specific metrics such as BLEU and Zero-Shot Transfer AP (Section 3.5) — makes direct cross-study comparison unreliable. Studies reporting mAP@0.5 alone report values exceeding 96% [7], [8], [36] that are not numerically comparable to studies reporting the stricter mAP@0.5:0.95 [19], [23], [38]. To

improve comparability in future work, we propose the following minimum reporting standard, presented as a checklist in Table 8.

**Table 8.** Proposed minimum reporting checklist for object detection studies in clothing, uniform, and PPE domains

| Reporting Element | What to Report                                                         |
|-------------------|------------------------------------------------------------------------|
| mAP@0.5           | Mean average precision at IoU 0.5                                      |
| mAP@0.5:0.95      | COCO-style mean average precision over IoU 0.5 to 0.95                 |
| Class-wise AP     | Average precision reported per individual class                        |
| Inference speed   | Frames per second or per-image latency in milliseconds                 |
| Model size        | Parameter count and FLOPs or GFLOPs                                    |
| Annotation type   | Bounding box, instance mask, or landmark, and the annotation tool used |
| Hardware          | GPU model and memory used for training and inference                   |
| Dataset scale     | Total image count and number of classes                                |
| Dataset split     | Train, validation, and test partition sizes                            |

Applying this checklist to the present corpus reveals a substantial reporting gap: no study reports all nine elements, and Lin, Zhang, & Chen [23] come closest with seven. Most studies report only mAP@0.5 and dataset scale, while class-wise AP, model size, and dataset split are reported by only a handful. This divergence between common practice and the information required for rigorous cross-study comparison is precisely what the checklist is intended to close.

## 6) Toward integrated and adaptive detection systems

Taken together, these findings suggest architecture selection should be matched to deployment context rather than defaulting to a single family. Architecture selection should be matched to deployment context, as synthesized in Table 9.

**Table 9.** Practitioner guidance for architecture selection by deployment context

| Application context                              | Suggested architecture family         | Basis                                                                                                |
|--------------------------------------------------|---------------------------------------|------------------------------------------------------------------------------------------------------|
| Real-time PPE compliance monitoring              | One-stage (YOLO family)               | Real-time inference demonstrated across 18 studies [7], [40]                                         |
| Fine-grained fashion attribute recognition       | Two-stage (R-CNN family)              | Superior localization precision reported in [19], [38]                                               |
| Open-set / unseen-category detection             | Transformer-based                     | Demonstrated by [39], though evidence is limited to n=2 studies and should be treated as preliminary |
| Worker identity & compliance (privacy-sensitive) | One-stage with explicit ethics review | Raised by [41]; no corpus study addresses governance                                                 |

This guidance is derived from patterns observed within the 26 studies synthesized here and should be read as a starting point for practitioner decision-making, not a substitute for context-specific evaluation. The development of unified cross-domain benchmarks and the minimum reporting standard proposed above are prerequisites for this field to mature from a collection of domain-specific systems into a coherent, comparable body of knowledge. Table 10 consolidates the key findings for each research question, synthesizing the detailed analysis presented across the preceding subsections.

**Table 10.** Summary of Findings by Research Question

| RQ  | Research Question                                                                  | Key Findings                                                                                                                                                                                                                                                                                                                                               | Supporting Studies                                                                                                                                                                                             | Sect. |
|-----|------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|
| RQ1 | Deep learning architectures most commonly used, and their strengths and weaknesses | One-stage detectors dominate (19 of 26 studies), with the YOLO family only in 18 studies (YOLOv3 to YOLO11). Two-stage R-CNN detectors are used in 5 studies, providing higher localization accuracy with the cost of inference latency. Two transformer-based detectors are used in 2 studies, which show good open-set generalization, but require large | YOLO family: [2], [3], [4], [6], [7], [8], [10], [20], [21], [23], [35], [40], [41], [42], [43], [44], [45], [46]; R-CNN Family: [5], [19], [34], [36], [37], [38]; Transformer: [5], [39]; EfficientDet: [22] | 3.2   |

| RQ  | Research Question                                                    | Key Findings                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Supporting Studies                                                                                                                                                                                    | Sect. |
|-----|----------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|
|     |                                                                      | computational resources and are sensitive to domain shifts. Only one non-YOLO one-stage architecture, EfficientDet, is used in this field.                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                       |       |
| RQ2 | Types of clothing and uniform items typically detected               | Two dominant areas of study:<br>Fashion and apparel (12 studies) spanning clothing items like shirts, dresses, and catalogs of multiple types; and Safety and PPE (13 studies) covering helmets, vests, gloves, and masks. One paper focuses on estimating clothing insulation for smart thermostats. The task of detecting PPE is considered as a parallel but not a subset of detecting generic uniforms.                                                                                                                                                                                       | Fashion and apparel: [2], [3], [4], [5], [19], [20], [21], [22], [34], [37], [38], [39]; Safety and PPE: [6], [7], [8], [23], [35], [36], [40], [41], [42], [43], [44], [45], [46]; Specialized: [10] | 3.3   |
| RQ3 | Datasets used for training and evaluation, and their characteristics | The contribution contains Datasets that are split into large-scale public benchmarks (e.g., DeepFashion2 with 491,895 images; SHWD; CHVG) and small custom collections. The contribution has a significantly asymmetrical amount of data with the number of the PPE collections varying from 1,330 images to 491,895 in public fashion benchmarks. There is an emerging pattern of benchmark reuse (CHV, CHVG, SHWD are reused across PPE studies), which partially mitigates the lack of large-scale PPE benchmarks. No standardized benchmark exists, which impairs comparisons across studies. | Public benchmarks: [7], [8], [19], [22], [23], [36], [37], [38], [43], [44], [45]; Custom datasets: [2], [3], [4], [5], [10], [20], [21], [34], [35], [39], [40], [41], [42], [46]                    | 3.4   |
| RQ4 | Evaluation metrics                                                   | Mean Average Precision (mAP) is dominant, used in 18 of 26 studies.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | mAP: [2], [4], [7], [8], [19], [20], [21], [22],                                                                                                                                                      | 3.5   |

| RQ  | Research Question                                                  | Key Findings                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | Supporting Studies                                                                                                                                                                                                                                                  | Sect.       |
|-----|--------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|
|     | employed to assess detection performance                           | Precision, recall, and F1-score accompany mAP in 14 studies; IoU or GloU appears in 6 studies. Efficiency metrics (FPS, inference time) appear in 10 studies, concentrated in the PPE domain. Specialized metrics appear where the task departs from standard detection: BLEU for caption generation [3], Hamming and Euclidean distance for texture similarity [37], and Zero-Shot Transfer AP for open-set generalization [39]. Reporting is inconsistent, motivating a proposed minimum reporting standard.                                              | [23], [35], [36], [38], [39], [40], [41], [43], [44], [45]; Precision/recall/F1: [2], [4], [7], [8], [10], [20], [21], [23], [35], [36], [39], [40], [43], [46]; Efficiency: [4], [7], [20], [23], [35], [36], [40], [42], [43], [45]; Specialized: [3], [37], [39] |             |
| RQ5 | Main challenges and future directions identified in the literature | Recurring challenges: non-rigid deformation and fine-grained inter-class similarity; occlusion and small-object detection; domain shift between training and deployment; class imbalance (rare categories reaching near-zero precision); annotation cost for pixel-level masks; and inconsistent evaluation reporting. The review further surfaces previously undiscussed ethical and cultural-representation gaps. Future directions include cross-domain benchmarks, data-efficient and open-set learning, and standardized multi-dimensional evaluation. | Deformation and inter-class similarity: [4], [8], [10], [34]; Occlusion and small objects: [40], [43], [46]; Domain shift: [5], [39]; Class imbalance: [7], [37], [38]; Annotation cost: [5]; Ethical and cultural gaps: [3], [8], [41]                             | 3.6,<br>3.7 |

#### 4. CONCLUSION

This systematic literature review synthesized 26 empirical studies (2021–2026) on deep learning-based object detection for fashion, uniform, and PPE items. The synthesis shows that the field has converged on one-stage detectors — the YOLO family alone accounts for 18 of the 26 studies — which favor real-time inference over the stronger localization offered by the less common two-stage and the still-underexplored transformer-based architectures, while the detected items divide into two parallel domains, fashion and apparel and safety and PPE, rather than a single unified problem. This division extends to the data: reported datasets are markedly asymmetric, ranging from large public fashion benchmarks to small, self-curated PPE collections, with an emerging benchmark-reuse pattern only partially offsetting the scarcity of standardized PPE data, and it extends to evaluation, where mean Average Precision dominates but inconsistent reporting of efficiency, dataset, and per-class measures limits cross-study comparison — a gap this review addresses through a proposed nine-element reporting checklist. Persistent challenges of class imbalance, occlusion, small-object detection, and domain shift remain largely unresolved, and the most significant open directions are cross-domain knowledge transfer between the fashion and PPE streams and the ethical and cultural-representation concerns that current studies leave largely unaddressed. Beyond answering these questions, this review contributes a cross-domain synthesis of two previously disconnected research streams, an explicit architecture–dataset–metric mapping, a nine-element minimum reporting standard, and a challenge taxonomy surfacing previously undiscussed ethical gaps; its findings should be read in light of its reliance on Scopus alone and an architecturally constrained PPE query, and future work should validate the proposed taxonomy through broader multi-database searches, pursue unified cross-domain benchmarks and transfer-learning experiments, and explicitly address the representational gaps identified here.

#### ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Cucuk Wawan Budiyo for his institutional support and for providing the research facilities that made this study possible.

## REFERENCES

- [1] C. Zhang, X. Ji, and L. Cai, "Clothing Recommendation with Multimodal Feature Fusion: Price Sensitivity and Personalization Optimization," *Appl. Sci.*, vol. 15, no. 8, p. 4591, Apr. 2025, doi: 10.3390/app15084591.
- [2] s. K. Addagarla And A. Amalanathan, "Single Stage Deep Transfer Learning Model For Apparel Detection And Classification For E-Commerce," *Int. J. Electron. Commer. Stud.*, vol. 13, no. 1, p. 69, Sep. 2021, doi: 10.7903/ijecs.1953.
- [3] G. Agarwal, K. Jindal, A. Chowdhury, V. K. Singh, and A. Pal, "Image and Video Captioning for Apparels Using Deep Learning," *IEEE Access*, vol. 12, pp. 113138–113150, 2024, doi: 10.1109/ACCESS.2024.3443422.
- [4] Y. Thwe, N. Jongsawat, and A. Tungkasthan, "A Semi-Supervised Learning Approach for Automatic Detection and Fashion Product Category Prediction with Small Training Dataset Using FC-YOLOv4," *Appl. Sci.*, vol. 12, no. 16, p. 8068, Aug. 2022, doi: 10.3390/app12168068.
- [5] R. Velioglu, R. Chan, and B. Hammer, "FashionFail: Addressing Failure Cases in Fashion Object Detection and Segmentation," in *2024 International Joint Conference on Neural Networks (IJCNN)*, Yokohama, Japan: IEEE, Jun. 2024, pp. 1–8. doi: 10.1109/IJCNN60899.2024.10651287.
- [6] B. Tong, G. Li, X. Bu, Y. Wang, and X. Yu, "A deep learning-based algorithm for the detection of personal protective equipment," *PLOS One*, vol. 20, no. 5, p. e0322115, May 2025, doi: 10.1371/journal.pone.0322115.
- [7] A. Rahman *et al.*, "PPE-EYE: A Deep Learning Approach to Personal Protective Equipment Compliance Detection," *Computers*, vol. 15, no. 1, p. 45, Jan. 2026, doi: 10.3390/computers15010045.
- [8] Z. Wang, Y. Wu, L. Yang, A. Thirunavukarasu, C. Evison, and Y. Zhao, "Fast Personal Protective Equipment Detection for Real Construction Sites Using Deep Learning Approaches," *Sensors*, vol. 21, no. 10, p. 3478, May 2021, doi: 10.3390/s21103478.
- [9] U.S. Bureau of Labor Statistics, "National Census of Fatal Occupational Injuries in 2023," U.S. Department of Labor, USDL-24-2564, Dec. 2024. Accessed: May 08, 2026. [Online]. Available: [https://www.bls.gov/news.release/archives/cfoi\\_12192024.pdf](https://www.bls.gov/news.release/archives/cfoi_12192024.pdf)
- [10] A. Medina, J. I. Méndez, P. Ponce, T. Peffer, A. Meier, and A. Molina, "Using Deep Learning in Real-Time for Clothing Classification with Connected Thermostats," *Energies*, vol. 15, no. 5, p. 1811, Mar. 2022, doi: 10.3390/en15051811.

- [11] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94.
- [12] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA: IEEE, 2005, pp. 886–893. doi: 10.1109/CVPR.2005.177.
- [13] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu, "Fashion Meets Computer Vision: A Survey," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–41, May 2022, doi: 10.1145/3447239.
- [14] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023, doi: 10.1109/JPROC.2023.3238524.
- [15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
- [16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
- [18] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 2980–2988. doi: 10.1109/ICCV.2017.322.
- [19] S.-I. Papadopoulos *et al.*, "Attentive Hierarchical Label Sharing for Enhanced Garment and Attribute Classification of Fashion Imagery," in *Recommender Systems in Fashion and Retail*, vol. 830, N. Dokoohaki, S. Jaradat, H. J. Corona Pampín, and R. Shirvany, Eds., in *Lecture Notes in Electrical Engineering*, vol. 830, Cham: Springer International Publishing, 2022, pp. 95–115. doi: 10.1007/978-3-030-94016-4\_7.
- [20] Y.-H. Chang and Y.-Y. Zhang, "Deep Learning for Clothing Style Recognition Using YOLOv5," *Micromachines*, vol. 13, no. 10, p. 1678, Oct. 2022, doi: 10.3390/mi13101678.
- [21] C.-H. Lee and C.-W. Lin, "A Two-Phase Fashion Apparel Detection Method Based on YOLOv4," *Appl. Sci.*, vol. 11, no. 9, p. 3782, Apr. 2021, doi: 10.3390/app11093782.

- [22] H. J. Kim, D. H. Lee, A. Niaz, C. Y. Kim, A. A. Memon, and K. N. Choi, "Multiple-Clothing Detection and Fashion Landmark Estimation Using a Single-Stage Detector," *IEEE Access*, vol. 9, pp. 11694–11704, 2021, doi: 10.1109/ACCESS.2021.3051424.
- [23] C. Lin, Y. Zhang, and G. Chen, "Intelligent detection of safety helmets and reflective vests based on deep learning," *J. Real-Time Image Process.*, vol. 22, no. 1, p. 5, Feb. 2025, doi: 10.1007/s11554-024-01573-y.
- [24] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in *Computer Vision – ECCV 2016*, vol. 9905, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., in Lecture Notes in Computer Science, vol. 9905, Cham: Springer International Publishing, 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0\_2.
- [25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 2999–3007. doi: 10.1109/ICCV.2017.324.
- [26] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Computer Vision – ECCV 2020*, vol. 12346, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., in Lecture Notes in Computer Science, vol. 12346, Cham: Springer International Publishing, 2020, pp. 213–229. doi: 10.1007/978-3-030-58452-8\_13.
- [27] A. M. Vukicevic, M. Petrovic, P. Milosevic, A. Peulic, K. Jovanovic, and A. Novakovic, "A systematic review of computer vision-based personal protective equipment compliance in industry practice: advancements, challenges and future directions," *Artif. Intell. Rev.*, vol. 57, no. 12, p. 319, Oct. 2024, doi: 10.1007/s10462-024-10978-x.
- [28] M. Mameli, M. Paolanti, R. Pietrini, G. Pazzaglia, E. Frontoni, and P. Zingaretti, "Deep Learning Approaches for Fashion Knowledge Extraction From Social Media: A Review," *IEEE Access*, vol. 10, pp. 1545–1576, 2022, doi: 10.1109/ACCESS.2021.3137893.
- [29] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Mach. Learn. Knowl. Extr.*, vol. 5, no. 4, pp. 1680–1716, Nov. 2023, doi: 10.3390/make5040083.
- [30] V. S. K. Delhi, R. Sankarlal, and A. Thomas, "Detection of Personal Protective Equipment (PPE) Compliance on Construction Site Using Computer Vision Based Deep Learning Techniques," *Front. Built Environ.*, vol. 6, p. 136, Sep. 2020, doi: 10.3389/fbuil.2020.00136.

- [31] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.
- [32] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," vol. 2, Jan. 2007.
- [33] Critical Appraisal Skills Programme, "CASP checklist: for systematic reviews." CASP UK, 2024. Accessed: May 05, 2026. [Online]. Available: <https://casp-uk.net/casp-checklists/CASP-checklist-systematic-reviews-checklist-2024.pdf>
- [34] M. S. Amin, C. Wang, and S. Jabeen, "Fashion sub-categories and attributes prediction model using deep learning," *Vis. Comput.*, vol. 39, no. 9, pp. 3851–3864, Sep. 2023, doi: 10.1007/s00371-022-02520-3.
- [35] J.-H. Lo, L.-K. Lin, and C.-C. Hung, "Real-Time Personal Protective Equipment Compliance Detection Based on Deep Learning Algorithm," *Sustainability*, vol. 15, no. 1, p. 391, Dec. 2022, doi: 10.3390/su15010391.
- [36] M. I. B. Ahmed *et al.*, "Personal Protective Equipment Detection: A Deep-Learning-Based Sustainable Approach," *Sustainability*, vol. 15, no. 18, p. 13990, Sep. 2023, doi: 10.3390/su151813990.
- [37] C. Vararu, C. Simionescu, and A. Iftene, "Integrated Approach for Clothing Detection and Comparison using Structural Shape Detection and Texture Analysis," *Procedia Comput. Sci.*, vol. 225, pp. 2546–2555, 2023, doi: 10.1016/j.procs.2023.10.246.
- [38] R. I. Jadhav, P. Stynes, P. Pathak, R. Haque, and M. Hasanuzzaman, "An Instance Segmentation Model to Categorize Clothes from Wild Fashion Images," in *Proceedings of the 2022 6th International Conference on Deep Learning Technologies*, Xi'an China: ACM, Jul. 2022, pp. 75–83. doi: 10.1145/3556677.3556690.
- [39] B. Ma and W. Xu, "Efficient Fine Tuning for Fashion Object Detection," *Sensors*, vol. 23, no. 13, p. 6083, Jul. 2023, doi: 10.3390/s23136083.
- [40] X. Wu, S. Qian, and M. Yang, "Detection of Safety Helmet-Wearing Based on the YOLO\_CA Model," *Comput. Mater. Contin.*, vol. 77, no. 3, pp. 3349–3366, 2023, doi: 10.32604/cmc.2023.043671.
- [41] N. Kwak and D. Kim, "Detection of Worker's Safety Helmet and Mask and Identification of Worker Using Deeplearning," *Comput. Mater. Contin.*, vol. 75, no. 1, pp. 1671–1686, 2023, doi: 10.32604/cmc.2023.035762.
- [42] X. Ji, F. Gong, X. Yuan, and N. Wang, "A high-performance framework for personal protective equipment detection on the offshore drilling platform," *Complex Intell. Syst.*, vol. 9, no. 5, pp. 5637–5652, Oct. 2023, doi: 10.1007/s40747-023-01028-0.

- [43] J. Su and Z. Gao, "Detection of Safety Helmet Wearing Based on Improved YOLO v5," in *Proceedings of the 2024 4th International Conference on Big Data, Artificial Intelligence and Risk Management*, Chengdu Guangdong China: ACM, Nov. 2024, pp. 938–943. doi: 10.1145/3718751.3718904.
- [44] Md. Ferdous and Sk. Md. M. Ahsan, "PPE detector: a YOLO-based architecture to detect personal protective equipment (PPE) for construction sites," *PeerJ Comput. Sci.*, vol. 8, p. e999, Jun. 2022, doi: 10.7717/peerj-cs.999.
- [45] A. Hayat and F. Morgado-Dias, "Deep Learning-Based Automatic Safety Helmet Detection System for Construction Safety," *Appl. Sci.*, vol. 12, no. 16, p. 8268, Aug. 2022, doi: 10.3390/app12168268.
- [46] Y. Gu, Y. Wang, L. Shi, N. Li, L. Zhuang, and S. Xu, "Automatic detection of safety helmet wearing based on head region location," *IET Image Process.*, vol. 15, no. 11, pp. 2441–2453, Sep. 2021, doi: 10.1049/ipr2.12231.
- [47] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qual. Res. Psychol.*, vol. 3, no. 2, pp. 77–101, Jan. 2006, doi: 10.1191/1478088706qp063oa.