Title：Smart Dust-Free Protective Equipment and Cleanroom Inspection System

Author(s): Shih-Shuan Tai 、Yi-Li Ho

Class: 4nd year of Department of Electronic Engineering

Student ID: D0916225、D0949852

Course: DESIGN AND DEPLOYMENT OF INTELLIGENT VISION SYSTEM

Instructor: Dr. Sze-Teng Liong

Department: Department of Electronic Engineering

Academic Year: Semester 2,2024

Smart Dust-Free Protective Equipment and Cleanroom InspectionSystem

ABSTRACT

In response to the stringent safety requirements of semiconductor cleanrooms, this study aims to develop an advanced dust-free protective equipment inspection system tailored for personnel entering these critical environments. Central to this system is the application of human pose detection techniques, which precisely identify essential body parts such as the head, body, hands, and foot. These detections serve as the foundation for deep learning network architectures that rigorously evaluate the adequacy of protective suits worn by personnel. Additionally, a system is designed to monitor the effectiveness of dust removal and ensure comprehensive coverage during the air shower process, crucial for maintaining impeccable cleanliness standards. The efficacy of the proposed pipeline is substantiated through rigorous validation encompassing comprehensive quantitative and qualitative analyses. Initial trials demonstrate robust performance, with quantitative accuracy rates of 98.42% for the protective equipment inspection system and an error rate of approximately 5% for the air showering process system. These results affirm the system's capability to reliably assess adherence to safety protocols in real-time scenarios. Beyond its immediate application in semiconductor cleanrooms, this adaptable system holds promise for integration into diverse sectors where stringent safety and contamination- free environments are paramount. Future research aims to enhance the system's adaptability to varying operational conditions and expand its functionalities through advancements in real-time feedback mechanisms and integration with edge computing technologies.

Keywords: cleanroom, inspection, deep learning, protective equipment

Table of Content

1. Introduction

The significance of automated detection of protective clothing is increasingly evident, enhancing workplace safety and boosting productivity. This technology is applied across various sectors, including chemical plants [1, 2], construction sites [3, 4], healthcare [5, 6], and others. Automated detection systems ensure compliance with safety standards, reduce the incidence of accidents, and guarantee regulatory adherence. With the continual advancements in technologies such as the Internet of Things (IoT) and artificial intelligence (AI), the performance and accuracy of these systems are constantly improving, making them more intelligent and reliable. This technology plays a significant role in enhancing workplace safety, increasing productivity, and ensuring regulatory compliance.

From Figure 1, it can be observed that the semiconductor industry's output value has surged alongside an increase in demand, resulting in exponential growth in the requirements for production line efficiency and quality. This underscores the necessity of automated recognition of proper cleanroom garment wearing, especially within the semiconductor sector, given its direct impact on production efficiency and product quality. Therefore, ensuring strict adherenceto cleanroom garment protocols among workers emerges as a pressing concern. Neglecting to address instancesof improper garment wearing could lead to heightened production line downtime, increased production costs, and potential ramifications on product quality and market competitiveness.

The primary objective of this research is to develop two inspection systems capable of automatically verifying the proper use of protective equipment by personnel entering cleanrooms, and determining whether personnel have fully undergone the air showering process before entering the cleanroom. The semiconductor industry requires environments of unparalleled purity in its relentless pursuit of perfection. Cleanrooms are meticulously designed to maintain low levels of pollutants such as dust, airborne microbes, aerosol particles, and chemical vapors. Personnel entering these environments are required to wear specialized protective equipment to prevent contamination. However, ensuring compliance with these stringent safety regulations has traditionally been a manual and time-consuming process.

Inspired by recent advancements in technology, this research proposes the development of an intelligent, dust-free protective equipment inspection system by leveraging body keypoint detection and deep learning approaches. This initiative aims to enhance compliance with safety regulations and streamline operations. To achieve this, the system

ORCID(s):

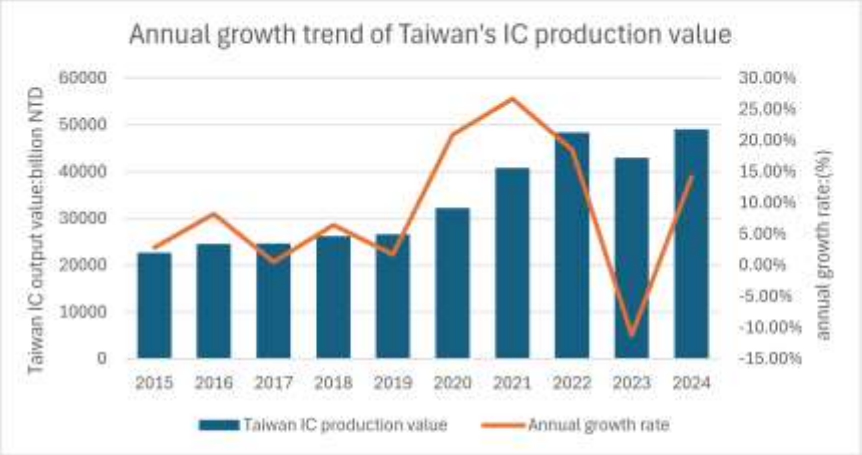


Figure 1: The IC production value and annual growth trend of Taiwan's semiconductor industry

will employ cutting-edge computer vision technology to identify and analyze the protective gear worn by individuals. This approach will not only ensure adherence to safety protocols but also significantly reduce the time and labor traditionally required for manual inspections.

In short, the main contributions of this work are highlighted as follows:

1. Compilation of a varied dataset consisting of 8687 images featuring personnel with protective equipment, focusing on six body parts.

2. Development of a real-time and robust dust-free automated protective equipment recognition system.

3. Implementation of an air shower process verification system using body pose estimation to monitor personnel movement.

4. Through an interactive GUI system, it is able to respond promptly to personnel and situations, thereby preventing contamination incidents in advance.

5. Extensive experimental work and thorough analysis conducted to assess the performance and effectiveness of the developed system.

The remaining structure of the paper is organized as follows. Section 2 conducts a synthesis of content from other relevant papers, including a comparative summary table and addressing the research gap. Section 3 outlines the research process of this system with intuitive insights into the proposed end-to-end protective equipment detection system and the cleanroom inspection system. Section 4 details the setup configurations, experiment settings, and performance metrics used. Subsequently, Section 5 presents and discusses the experimental results, including the challenges and limitations encountered. Finally, Section 6 concludes the paper with important findings highlighted and offers perspectives for future developments.

2. Literature Review

Despite considerable progress in image processing technologies in recent years, the advancement of automated dust-free protective equipment and cleanroom inspection system research has been limited, mainly because of the absence of a dataset for experimental assessment. Section 2.1 provides a succinct overview of prior work on automated personal protective equipment (PPE) analysis, highlighting key findings. Moreover, Section 2.2 identifies a research gap, emphasizing the need for the proposed system.

Table 1
Summary of the existing methods proposed for the automated personal protective equipment inspection system

No.	Ref.	Task	Target	Images	Labels	Model	Accuracy
1	[7]	Detection	Pictor-V3	4727	4	YOLOv3	72.3 %
2	[8]	Detection	CHVG	1699	8	YOLOX-m	89.84%
3	[9]	Detection	D1,2,3	69227	12	YOLOv4-tiny	87.5 %
4	[10]	Classification/ Detection	ODPD	18600	5	RFA-YOLO	88.41 %
5	[11]	Detection	CHVG	1189	8	Faster R-CNN	-
6	[12]	Classification	ReID	6245	4	VGG16	87.33%

2.1. Inspection System for Personal Protective Equipment

Research groups across the globe have adopted automated inspection systems, aiming to utilize advanced technologies to improve the precision and efficiency of wearing protective equipment correctly. This section sheds light on and discusses the applications and innovations of notable studies, emphasizing the significance of automated inspection systems in fostering a more knowledgeable and accurate approach to sustainability and cost-effectiveness. Previously published works are examined in this section, with a concise summary of these articles presented in Table 1.

In the field of object detection for PPE on construction sites, [7] employs one of the object detectors, specifically the YOLOv3 model [13], for real-time object detection. The study conducted experiments using three different methods, each involving various sizes of output layers and target recognition scenarios. Their self-collected dataset, namely Pictor-v3, includes individuals wearing vests, individuals wearing helmets, and both, with a total of 10,193 images. The best model i.e., YOLO-v3-A2 model achieved an average precision of 72.3%. However, the authors acknowledged limitations in their approach, such as the need for high-quality training data and the potential for false positives. Nonetheless, this work underscores the potential of deep learning in enhancing safety in the construction industry.

On the other hand, a study conducted by [8] centers on the utilization of YOLOX [14]. To validate the proposed pipeline in a testing environment, a new dataset, namely CHVG [15], has been collected. This dataset comprises eight categories representing common and essential equipment found on construction sites, such as safety helmets, vests, safety glasses, body, and headgear. It encompasses a total of 1699 images, which include instances with various environmental conditions like rain, haze, and low-light situations, aimed at simulating real-world environments. The results of the experiments have revealed that the YOLOX-m architecture can achieve the highest average precision (mAP) at 89.84%. However, it is important to note that complex backgrounds may lead to detection errors.

On a related note, [9] introduces an embedded real-time PPE detection system with a specific focus on head, helmets, chest, vests, hand and gloves. The dataset comprises three subsets (i.e., D1, D2, and D3), consisting of a total of 7283 images. To validate the robustness of the methods employed, five different CNN models were utilized (i.e., YOLOv4 [16], YOLOv4-tiny [16], SSD MobileNet V2 [17], CenterNet ResNet-50 V2 [18] and EfficientDet D0 [19]). The experimental results include measurements of recognition time and accuracy. Interestingly, the YOLOv4 architecture achieved the highest accuracy,almost all have reached an accuracy of 90% or higher,especiallythe part related to hands and gloves exceeds the other four models by more than 20%.However, considering the significantly faster recognition time of YOLOv4-tiny (7.6 milliseconds compared to YOLOv4's 30.6 milliseconds) and its comparable accuracy to ResNet-50 V2, the authors opted for YOLOv4-tiny as the operational network for theirsystem.

In a similar vein, [10] utilizes the RFA-YOLO model to recognize protective gear, with a specific focus on helmets and work suits in the context of offshore drilling. Their dataset is extensive, comprising an object detection dataset that includes objects such as a person, helmet, and workwear, a feature classification dataset to assess workwear appropriateness, and a personal protective equipment dataset that encompasses objects like helmets and workwear. Intotal, these three datasets contain 18,600 images. During experimentation, a comparative analysis of YOLO models r anging from V3 to V5 [13, 16, 20] was conducted. This analysis revealed that the RFA-YOLO model achieved optimal performance in helmet and work suit recognition, achieving accuracies of 84.21% and 84.72%, respectively. However, it is important to note that practical applications in offshore environments may encounter complex challenges, which could potentially limit the replicability of these results in real-world offshore settings.

[11] outlined in the document contributes significantly to the field of safety compliance monitoring in construction environments by developing a novel framework that leverages deep learning models (i.e., ResNet50 [21], OSNet [22] and OSNet+BDB [23]) for worker re-identification (ReID) and PPE classification. This work introduces a new loss function, named similarity loss, to enhance the accuracy of worker ReID, and a weighted-class strategy to address the challenge of imbalanced classes in PPE classification. The efficacy of these methods is quantitatively validated through improved accuracies—4% in ReID and 13% in PPE classification—using a real-world construction site dataset. Despite these innovations, the paper recognizes limitations, notably the lack of extensive discussion on the system's performance in varying lighting conditions and across different site layouts, which suggests areas for further exploration and optimization.

Recently, Ahmed et al. [12] explores the application of deep learning for the real-time detection of PPE to enhance worker safety in hazardous environments. The researchers employ Faster Region-based Convolutional Neural Network (Faster RCNN) [24] and YOLOv5 [25] models trained on a specifically developed dataset, the CHVG dataset [15], which includes various classes of PPE such as helmets, vests, and safety glasses. This methodological approach allows for the detection of multiple types of PPE with a high degree of accuracy, achieving a mean average precision (mAP) of 96%, highlighting its potential to significantly improve safety compliance monitoring. Despite its promising results, the study does not delve into the model's performance across different environmental conditions or its scalability, which could impact its utility in diverse real-world scenarios.

2.2. Research Gap

While recent advancements in automated PPE detection have shown promising results across various industries, several research gaps persist that limit the broader applicability and effectiveness of these technologies. First, there is a notable deficiency in the adaptability of current models to diverse and complex environmental conditions typical of real-world scenarios. Furthermore, there is a scarcity of comprehensive datasets that encompass the personnel entering the dust-room, which is crucial for training models to recognize and adapt to diverse situations effectively. Additionally, there is a lack of an end-to-end inspection system, encompassing both protective suit inspection and air showering procedure identification. By focusing on these aspects, our work not only contributes significantly to the development of a dust-free workplace but also aims to promote a comprehensive safety strategy in other similar industrial settings.

3. Proposed Method

The primary aim of this study is to ensure the correct wearing of cleanroom suits and the execution of cleanliness procedures upon entering the air shower room. To verify the suitability of the suit for each individual, the body is divided into six distinct parts. Deep learning models, enhanced by a transfer learning strategy, are then employed to analyze each part separately. Furthermore, upon entering the air shower room, both the frontal and rear body parts of personnel are recognized to confirm the completion of cleanliness actions.

Concretely, the process is divided into four main steps: (a) data preparation: data collection, body parts defining, and regions of interest extraction; (b) CNN model development: to train six deep learning models to address each of the six body parts independently; (c) protective equipment verification: to evaluate the suitability of the protective suit worn; (d) self-rotation identification: to ensure the front and rear of each individual are captured to finalize the air showering process. To enhance understanding of the inspection procedure, Figure 2 presents a flowchart diagram with illustrative examples.

3.1. Data Preparation

To develop a robust and effective protective equipment inspection system, constructing a comprehensive dataset is a critical first step. Images acquired from five different individuals cover a complete range from head to toe, capturing the full scope of body variations. One of the notable challenges addressed by this approach is the variation in body size among individuals. To overcome this issue, a keypoint landmark detector, specifically the MediaPipe algorithm [26], is utilized for its skeleton-based approach to precisely locate each joint across the body. This method facilitates the detection of key points on various parts of the body, including the face, shoulders, elbows, wrists, hands, hips, knees, ankles, and foot, as illustrated in Figure 3(a). It is worth noting that the landmarks of the body can be detected accurately, both with and without wearing the protective suit. Thus, the landmark detector serves as an intuitive tool to ensure the entire body is captured within the image frame, despite these variations, thereby guaranteeing a comprehensive analysis of protective equipment fit and placement on diverse body types.

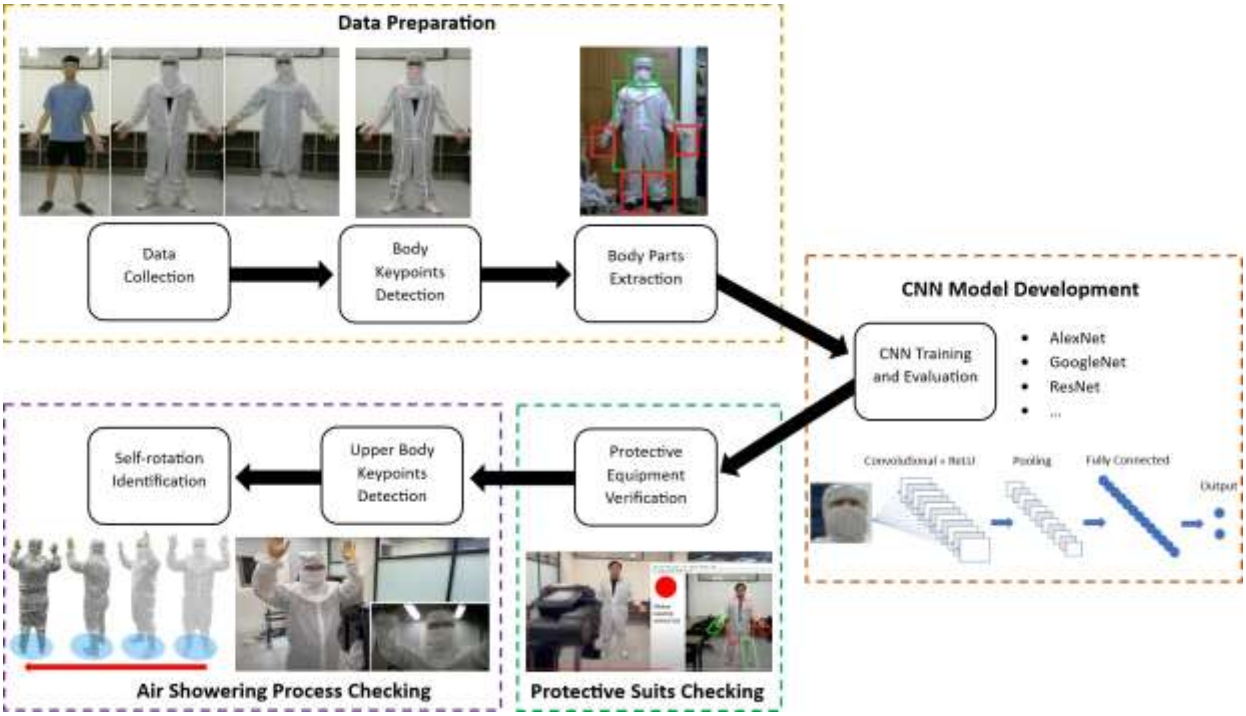

Data Preparation
Data Collection
Body Keypoints Detection
Body Parts Extraction
CNN Model Development
CNN Training and Evaluation
AlexNet
GoogleNet
ResNet
...
Convolutional + ReLU
Pooling
Fully Connected
Output
Self-rotation Identification
Upper Body Keypoints Detection
Protective Equipment Verification
Air Showering Process Checking
Protective Suits Checking

Figure 2: The flow chart of the proposed pipeline comprises four main components: (a) data preparation; (b) CNN model development; (c) protective suits checking; and (d) air showering process checking

Following the initial detection, the body is further segmented into six distinct parts: the head, body, left hand palm, right hand palm, left feet, and right feet. This segmentation, detailed in Figure 3(b), enables a more focused analysis of each body part. By dividing the body in this manner, the system can conduct targeted evaluations of the protective equipment worn on each specific body region. This detailed approach enhances the inspection system's precision, enabling it to more effectively identify and address potential issues with equipment fit or placement.

3.2. CNN Model Development

Upon completing data collection, training each of the six body parts separately, particularly with a transfer learning strategy, accelerates the training speed and enhances efficiency. Transfer learning leverages knowledge from a related task (previously learned by the model on a dataset, namely ImageNet) and applies it to a new task (e.g., binary classification for "pass" or "fail" in equipment inspection). This strategy significantly reduces the required volume of data and shortens the training duration. The visualization of the binary classification for each body part, along with their respective passing criteria, is provided in Figure 4.

The robustness of this approach is affirmed by experimenting with a variety of CNN architectures (i.e., VGG-19[27], ResNet-18 [21], Inception-V3 [28], and GoogLeNet [28]), each with its unique design and fine-tuning requirements. This diversity ensures that the selected models are best suited for the task at hand, taking into account the specific features and challenges of protective equipment inspection. The four architectures experimented with are briefly described below:

- VGG-19:
  It utilizes a deep architecture with repeated blocks of convolutional and max pooling layers, followed by fully connected layers. The simplicity and depth of VGG make it excellent for learning hierarchical features, with the depth contributing to the network's strong performance on image recognition tasks.

- ResNet-18:
  It introduces residual connections to facilitate the training of much deeper networks. These connections allow

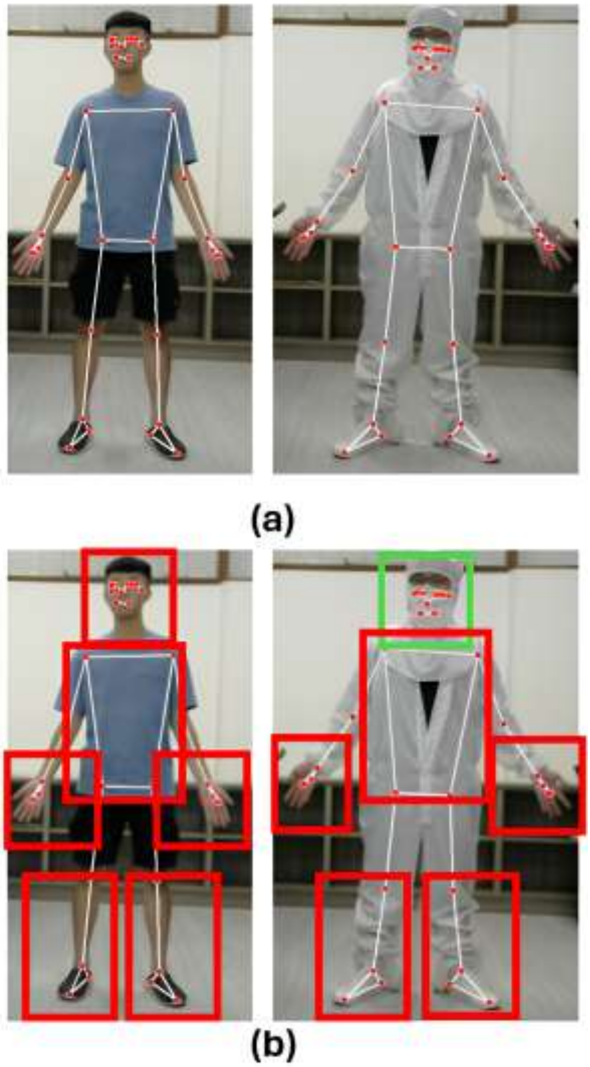


Figure 3: (a) The landmarks are detected accurately both with and without wearing a protective suit. (b) The illustration of the body partition that consists of six distinct parts: the head, body, left hand palm, right hand palm, left feet, and right feet.

gradients to flow through the network more effectively, solving the vanishing gradient problem and enabling the network to learn more complex features without a significant increase in training difficulty.

- Inception-V3:
  It employs modules that perform several convolutions in parallel, merging their outputs. This architecture allows the network to adapt to various scales and dimensions of the input data, making it highly efficient in recognizing patterns across different sizes and resolutions.

- GoogLeNet:
  It incorporates a similar inception module concept, optimizing the network for computational efficiency by reducing the dimensionality of the convolutions. This makes it particularly suitable for environments where computing resources are limited.

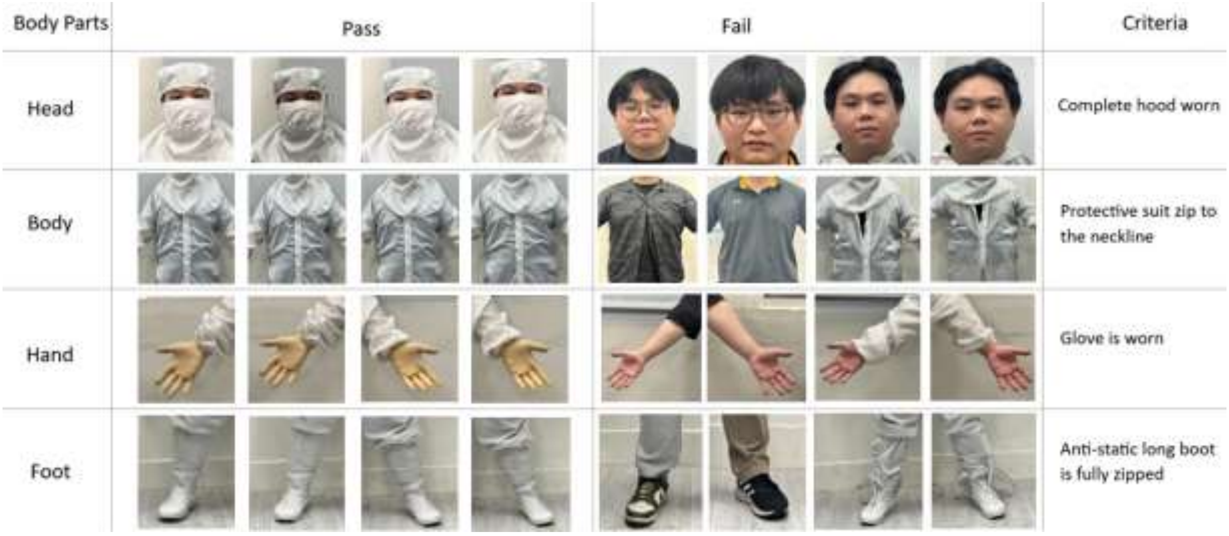


Figure 4: Example of binary classification for each body part, including descriptions of their respective criteria

To ensure fair and reliable performance evaluation, a five-fold cross-validation strategy is employed. This method ensures that each test image is evaluated once, allowing for a comprehensive assessment of the model's effectiveness across different data subsets. It should be noted that all body parts are first resized to a fixed size (i.e., 200×200 or 150×225) to accommodate the training and testing processes of the deep learning models.

3.3. Protective Equipment Verification

To facilitate real-time detection of protective equipment, which relies on the deep learning models developed in Section 3.2, an interactive interface has been implemented. Specifically, this integrated interface operates in four phases: (a) standby mode: to detect the presence of individuals; (b) posture adjustment: to ensure the visibility of all six human body parts; (c) image acquisition: to capture a full-body image; (d) protective equipment recognition: to verify whether the appropriate suit is correctly worn, with each of the six body parts being extracted and recognized.

To enhance the clarity of the protective equipment verification stage, a flowchart accompanied by illustrative GUI screenshots is presented in Figure 5. The function of each phase is detailed below:

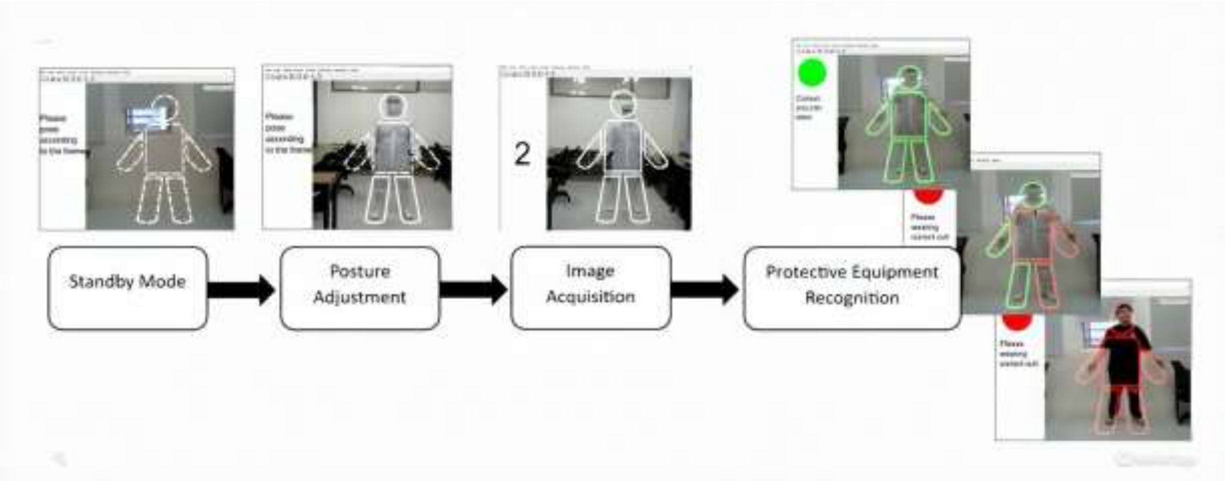


Figure 5: The protective equipment verification process, accompanied by illustrative GUI screenshots

1. Standby mode:
   In the absence of individuals within the detection area, the system engages in continuous monitoring through a camera, maintaining a standby state. Simultaneously, the landmark detector (i.e., the MediaPipe algorithm) is employed to ascertain the entry of personnel into the detection area and to compute their appearance time based on the frame rate. Upon meeting the predetermined duration of appearance, the system transitions to the subsequent phase.

2. Posture adjustment:
   Upon detection of body movement, a human-shaped outline is designed to guide the person to present the predefined pose on the screen, delineating the six body segments: the head, body, left and right hands, and left and right feet. Individuals are required to align their posture with the outline to activate the ensuing detection phase. Initially displayed as dashed lines, these outlines indicate misalignment of body parts with the specified locations. Correct positioning transforms the outlines to solid lines, signaling alignment to the user.

3. Image acquisition:
   Once the correct posture is detected, a brief period is allocated for acquiring image data of the individual for further recognition. To reduce background interference, landmarks are utilized to individually extract the partial image based on the six body segments.

4. Protective equipment recognition:
   Finally, the GUI displays the protective equipment verification result for each body part. Upon deriving the final recognition results for each segment, the human-shaped outline displayed on the screen will signify the correctness of each part's attire in distinct colors. A green display denotes correct attire, whereas red signifies incorrect. For successful recognition, all six segments must be verified in green.

3.4. Air Showering Process Checking

Following the verification of appropriate attire for personnel before entry into the cleanroom, as outlined in Section 3.3, the subsequent phase involves ensuring that personnel complete the entire air showering process. This dust removal procedure is conducted in a designated space measuring 80cm×100cm, as portrayed in Figure 6(a). Personnel are required to first raise both hands above the head, facing the air outlet for three seconds, then execute a180° rotation, maintaining this posture for another three seconds to finalize the cleaning.

Given the spatial limitations, only the upper half of the body is captured. A 120° wide-angle lens, mounted on a webcam placed strategically about 30cm from the individual, ensures comprehensive coverage of the upper body. Simple trigonometric calculations are found to be adequate for capturing the air showering movement of individuals ranging in height from 150cm∼180cm.

The MediaPipe algorithm is utilized initially to detect the human skeleton and subsequently to identify self-rotation movements. This involves analyzing the landmarks of the palms (landmarks 15 and 16) and the y-coordinate of the highest point on the head to verify that both hands are elevated above the head. Furthermore, landmarks 11 and 12, located at the shoulder ends, are considered for observing body rotation movements. Changes in the x-axis coordinates of these points signal a body turn. To enhance clarity on body movement detection, Figure 6 (b) offers an illustrative example of the self-rotation motion.

Should any aspect of an individual's actions deviate from the required standards, the system issues a voice alert and resets that phase to guarantee the thorough completion of the cleaning actions. This configuration not only captures movement actions efficiently but also ensures accurate measurements and analyses within the confined space. Moreover, this system is specifically designed to validate the air showering process, not only boosts the system's response time but also significantly enhances its adaptability to varying environments. The inclusion of voice prompts for corrective action represents an intuitive feedback mechanism, further underscoring the system's user-oriented design. This approach ensures compliance with hygiene protocols, reinforcing the cleanroom's integrity and safety standards.

4. Experiment Setup

This section details the process of collecting and distributing a database of images, designing model training and testing, setting up experiments with detailed configurations for training deep learning models, and the performance metrics used to evaluate the models.

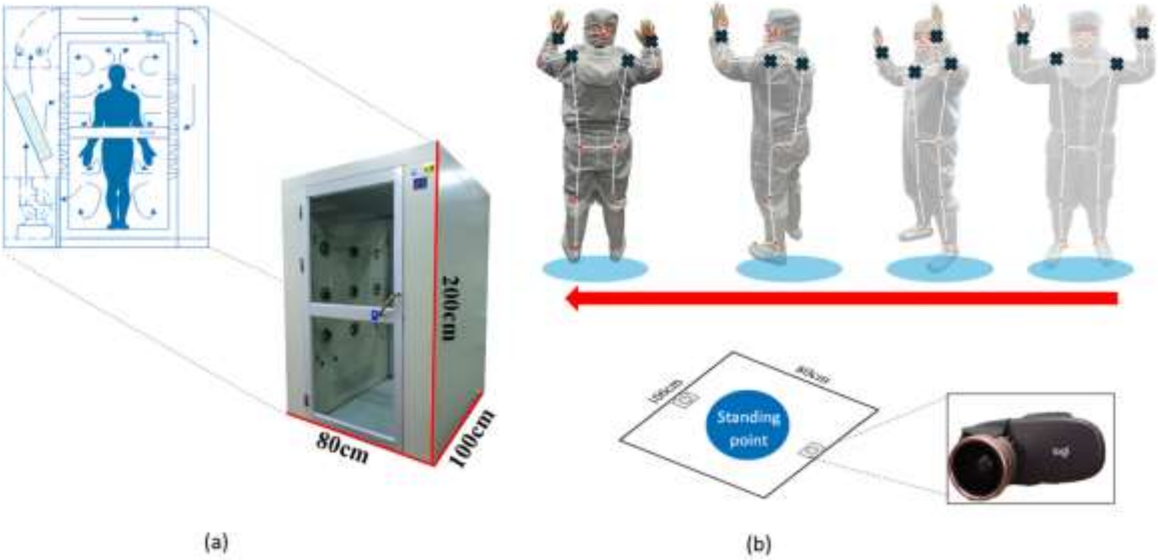


Figure 6: (a) Illustration of the cleanroom air shower with dimensions of 80cm×100cm; (b) Utilization of a landmark detector to detect the hand-raising and self-rotation movements of personnel.

4.1. Experimental Dataset

The dataset used in this experiment comprises a total of 8687 images, covering six body parts: head, body, both hands, and both foot. These images were collected from five individuals with different body types. Notably, due to the varying resolution requirements for each body part, the exact quantity and resolution for each part are provided in Table 2. In total, this substantial dataset, comprising 7,921 images, aids in training the neural network. An 80/20 split for the training and validation sets is applied to ensure the model learns effectively and has a sufficient amount of unseen data for testing to accurately assess its performance. An additional 74 images were collected as a test set, which are not involved in the training process. To provide a better understanding of the passing and failing criteria for each body part, Figure 4 displays examples of binary classification along with the respective descriptions of their pass and fail criteria.

4.2. Hardware Devices Setup

In this work, the experiments were trained and evaluated on a computer equipped with an Intel(R) Core(TM) i7-9750H 2.60GHz CPU and an NVIDIA GTX 1660Ti GPU. MATLAB R2023b was selected as the development platform due to its rich suite of image processing and machine learning tools. For camera utilization, three identical Logitech C270 webcams were employed. To enhance the functionality of the webcams, they were equipped with wide-angle lenses to provide a broader field of view. Table 3 provides the detailed specifications of the webcam and wide-angle lenses utilized.

4.3. Network Parameter Configuration

Given that there are six body parts as regions of interest, six distinct deep learning networks for binary classification were designed. The input layer of each network has a size of 200×200, except for images of foot, which were resized to 150×225. This adjustment was necessary because foot are elongated, and resizing them to 200×200 would cause excessive distortion, potentially affecting the accuracy of the recognition results. The optimizer used was Adam, with a learning rate of 0.0001 and a mini-batch size of 64, with shuffling every epoch. On the other hand, Table 4 provides the epoch values applied to each selected network, indicating that VGG-19 was trained for 40 epochs, GoogleNet for 50 epochs, ResNet-18 for 40 epochs, and Inception-V3 for 60 epochs. The variation in the number of epochs is due to the nature of the networks, which have distinct depths and varying numbers of learnable parameters.

The six body parts along with their respective distributions in the training and testing datasets

		Pass	Fail	Total
Head	Train	546	557	1103
	Validation	137	139	276
	Test	47	27	74
	Total	730	721	1453
Body	Train	601	1010	1611
	Validation	150	252	402
	Test	20	54	74
	Total	771	1316	2087
Left hand	Train	558	521	1079
	Validation	139	130	269
	Test	20	54	74
	Total	717	705	1422
Right hand	Train	716	459	1175
	Validation	179	115	294
	Test	20	54	74
	Total	915	628	1543
Left feet	Train	341	359	700
	Validation	85	90	175
	Test	26	48	74
	Total	452	497	949
Right feet	Train	297	373	670
	Validation	74	93	167
	Test	26	48	74
	Total	397	514	911

4.4. Performance Metrics

To evaluate the effectiveness and robustness of the proposed classification system, performance metrics such as accuracy and F1-score were utilized. These metrics are mathematically expressed as follows:

$$\text{Accuracy} = \frac{\square\square + \square\square}{\square\square + \square\square + \square\square + \square\square} \tag{1}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

where

$$\text{Precision} = \frac{\square\square}{\square\square + \square\square} \tag{3}$$

and

The specifications of the webcam and wide-angle lenses adopted

Feature	Description
Webcam	
Model	Logitech C720
Resolution	1280×720
Frame/ second	30fps
Camera mega pixel	0.9
Focus type	Fixed
Diagonal field of view	55°
Wide-angle lens	
Model	036
Optical format	0.36×wide-angle
Magnification	15×macro

Table 4
The training configuration specifically the epoch parameters applied to each selected network

	Epoch
VGG-19	40
GoogleNet	50
ResNet-18	40
Inception-V3	60

$$\text{Recall} = \frac{\square\square}{\square\square + \square\square} \tag{4}$$

where

- TP (True Positive) indicates that the model correctly identifies the presence of protective equipment.
- TN (True Negative) indicates that the model correctly predicts the absence of protective equipment.
- FN (False Negative) indicates that the model incorrectly classifies worn protective equipment as absent.
- FP (False Positive) indicates that the model incorrectly identifies the absence of protective equipment as present.

5. Results

The results of the protective equipment identification and cleanroom air showering process verification tasks are elucidated in Section 5.1 and Section 5.2, respectively. Additionally, the constraints and limitations of the proposed pipeline are thoroughly discussed in Section 5.3.

5.1. Protective equipment verification

To establish a fair and reliable evaluation methodology, the training, validation, and test datasets are distributed to prevent the model from overfitting. Specifically, the test images, totaling 74, are acquired in real-time and include individuals with varying body types. Table 5 provides a summary of the performance metrics for the four network models considered in this work: VGG-19, ResNet-18, GoogleNet, and Inception-v3. The respective results for different body parts (i.e., head, body, left hand, right hand, left feet, and right feet) are tabulated. VGG-19 achieves perfect scores (100%) in all metrics for four body parts (i.e., head, left hand, left feet, and right feet). However, for the body and right

The performance results for the binary classification task for six distinct body parts using different pre-trained networks

		Head	Body	Left hand	Right hand	Left feet	Right feet	Overall
GoogleNet	Accuracy	100	90.5	100	100	98.6	93.2	97.07
	Precision	100	74.0	100	100	96.3	83.9	92.44
	Recall	100	100	100	100	100	100	100
	F1-score	100	85.1	100	100	98.1	91.2	96.07
ResNet-18	Accuracy	100	93.2	100	94.6	100	97.3	97.52
	Precision	100	80.0	100	100	100	90.9	95.68
	Recall	100	100	100	80.0	100	100	97.48
	F1-score	100	88.9	100	88.9	100	95.2	96.57
Inception-V3	Accuracy	100	94.6	100	94.6	100	95.9	97.52
	Precision	100	83.3	100	100	100	89.7	93.53
	Recall	100	100	100	80.0	100	100	100
	F1-score	100	91.0	100	88.9	100	94.6	96.66
VGG-19	Accuracy	100	94.6	100	95.9	100	100	98.42
	Precision	100	83.3	100	100	100	100	97.50
	Recall	100	100	100	85.0	100	100	98.11
	F1-score	100	91.0	100	91.9	100	100	97.80

Table 6
The confusion matrix of recognition result (%) for the binary classification task for six distinct body parts using employing VGG-19 network

(a) Head

		Desired	
		Pass	Fail
Predicted	Pass	**100%**	0%
	Fail	0%	**100%**

(b) Body

		Desired	
		Pass	Fail
Predicted	Pass	**100%**	7.4%
	Fail	0%	**92.6%**

(c) Left hand

		Desired	
		Pass	Fail
Predicted	Pass	**100%**	0%
	Fail	0%	**100%**

(d) Right hand

		Desired	
		Pass	Fail
Predicted	Pass	**85%**	0%
	Fail	15%	**100%**

(e) Left feet

		Desired	
		Pass	Fail
Predicted	Pass	**100%**	0%
	Fail	0%	**100%**

(f) Right feet

		Desired	
		Pass	Fail
Predicted	Pass	**100%**	0%
	Fail	0%	**100%**

hand, VGG-19 achieves lower F1 scores of 91.0% and 91.9%, respectively. Conversely, GoogleNet, while performing perfectly for the head, left hand, and right hand, exhibits the lowest performance for the body, left feet, and right feet, with F1 scores of 85.1%, 98.1 and 91.1%, respectively.

The results indicate that the head is the most accurately predicted body part across all models. This could be due to the distinct nature of the binary classification associated with the head (e.g., black hair versus white protective gear), making the classification more straightforward. On the other hand, the body shows the lowest precision, which may be due to significant variations in this body part, such as differences in zipper closures or misidentification of targets as hat edges, making classification more challenging.

For a closer inspection into the performance of every individual class, the confusion matrices for two selected networks that performed the best and the poorest are tabulated in Table 6 and Table 7. These are VGG-19, with an average accuracy of 98.42%, and GoogLeNet, with an average accuracy of 97.07%, respectively. Specifically, VGG-19 shows relatively stronger performance overall, achieving 100% accuracy in detecting the head, left hand, left feet, and right feet, with only minor errors observed in the body and right hand classifications. Additionally, VGG-19 achieves 92.6% accuracy for the body and 85% for the right hand, compared to GoogLeNet, which exhibits lower accuracy in

Table 7
The confusion matrix of recognition result (%) for the binary classification task for six distinct body parts using employing GoogLeNet network

(a) Head

		Desired	
		Pass	Fail
Predicted	Pass	**100%**	0%
	Fail	0%	**100%**

(b) Body

		Desired	
		Pass	Fail
Predicted	Pass	**100%**	12.9%
	Fail	0%	**87.1%**

(c) Left hand

		Desired	
		Pass	Fail
Predicted	Pass	**100%**	0%
	Fail	0%	**100%**

(d) Right hand

		Desired	
		Pass	Fail
Predicted	Pass	**100%**	0%
	Fail	0%	**100%**

(e) Left feet

		Desired	
		Pass	Fail
Predicted	Pass	**100%**	2%
	Fail	0%	**98%**

(f) Right feet

		Desired	
		Pass	Fail
Predicted	Pass	**100%**	10.4%
	Fail	0%	**89.6%**

Table 8
The confusion matrix of the dust-cleaning inspection system, where the participants were asked to execute 20 self-rotations with deliberately designed combinations of false and standard postures.

(a) Participant 1

		Desired	
		Pass	Fail
Predicted	Pass	**90%**	10%
	Fail	0%	**100%**

(b) Participant 2

		Desired	
		Pass	Fail
Predicted	Pass	**100%**	0%
	Fail	0%	**100%**

several categories, such as 87.1% for the body, 89.6% for the right feet, and 98% for the left feet. These results indicate that VGG-19 has higher precision and reliability, particularly in classifying smaller or more distinct body parts, making it a stronger model for this specific binary classification task.

To visualize the network performance, Grad-CAM is employed for both correctly and incorrectly classified classes. Specifically, Figure 7 displays some classification results for the body parts. It is observed that a common error occurs when the model misinterprets headgear edges as the target for judgment, resulting in the misclassification of originally inappropriate attire as meeting the passing criteria. Our analysis suggests that these misclassifications mainly stem from the model's challenge in handling edge cases, particularly in features with high variability. Hence, these findings hold significance for enhancing future model designs, particularly in addressing high-variability features and extreme binary classification problems.

5.2. Air-shower detection

To assess the reliability of the proposed dust-clean inspection system in the air shower, two individuals with heights of 165 cm and 170 cm participated in the validation process. The primary objective was to evaluate the system's feasibility across different body sizes.

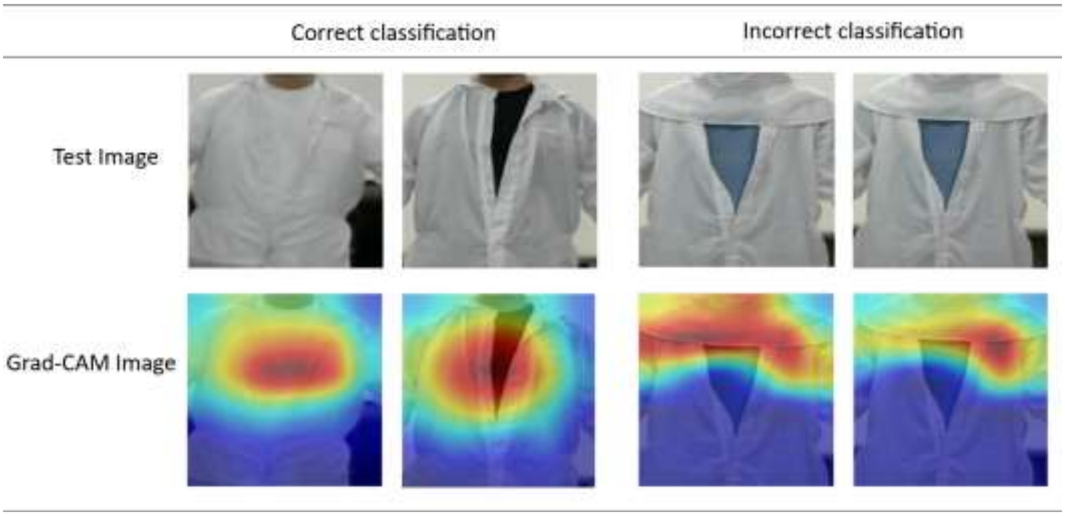


Figure 7: Sample visualization of the Grad-CAM activation for correct and incorrect classifications for the body parts

The experiment is divided into two parts. The first part focuses on verifying the stability of joint point recognition across individuals of varying heights. Participants were instructed to raise their hands in front of the camera to simulate cleaning movements in the air shower. Recordings were made at 30 fps, totaling 90 frames over three seconds. The skeleton detection results, illustrated in Figure 9, show occasional missed joints and limbs, potentially due to errors in the dust-cleaning inspection system. Specifically, Figure 8 provides detailed body landmark detected results, indicating an average error rate of approximately 5%, with 4 to 5 frames out of 90 being incorrect. Nonetheless, this error rate does not significantly impact the overall system performance.

The subsequent stage of the dust-cleaning process involves a body turning experiment. Users are required to perform a 180° turn after cleaning the front side to ensure thorough cleaning of both sides. Both testers were asked to execute 20 self-rotations with their hands raised. Specifically, 10 of these rotations were deliberately designed to include instances where users either turned back to the front or disappeared from the camera's view post-turn, with the expected outcome being a failure to pass the dust-cleaning system identification. In contrast, the remaining 10 rotations were performed at a constant speed, ensuring that the front and rear of each individual were captured to complete the air showering process. As a result, the actions of one participant were fully identified, while the other participant had a false detection. The experimental results of these two participants are summarized in Table 8.

Nevertheless, this dust-cleaning process demonstrates the system's robust capability in detecting proper cleaning procedures during the air shower and effectively alerting users to errors. Moreover, the proposed system is valuable for personnel training, highlighting correct postures and those that may lead to errors. Furthermore, the insights gained from this dust-clean inspection system underscore its potential to not only enhance operational efficiency but also to contribute significantly to maintaining stringent cleanliness standards essential for various industrial applications.

5.3. Limitation

The equipment recognition system has limitations rooted in its dependency on fixed locations and predefined actions for accurate recognition. These constraints restrict its adaptability to dynamic environments where equipment and actions may vary beyond the predefined scenarios. Similarly, the dust cleanroom inspection system implemented in the air shower room faces limitations due to the absence of real-world filming in actual air shower rooms during experimentation. Relying on online references for the experimental environment lacks the fidelity of actual camera setup conditions and the complexities posed by real-time interactions. Furthermore, detecting skeletons becomes challenging amidst the constraints imposed by wearing protective clothing.

Future work could focus on enhancing the adaptability of the equipment recognition system by integrating more robust and flexible recognition algorithms capable of autonomously learning from dynamic environments. Additionally, the algorithms should prioritize low computational cost to enable deployment on edge devices. For the air shower room recognition system, future research could involve conducting experiments in real air shower

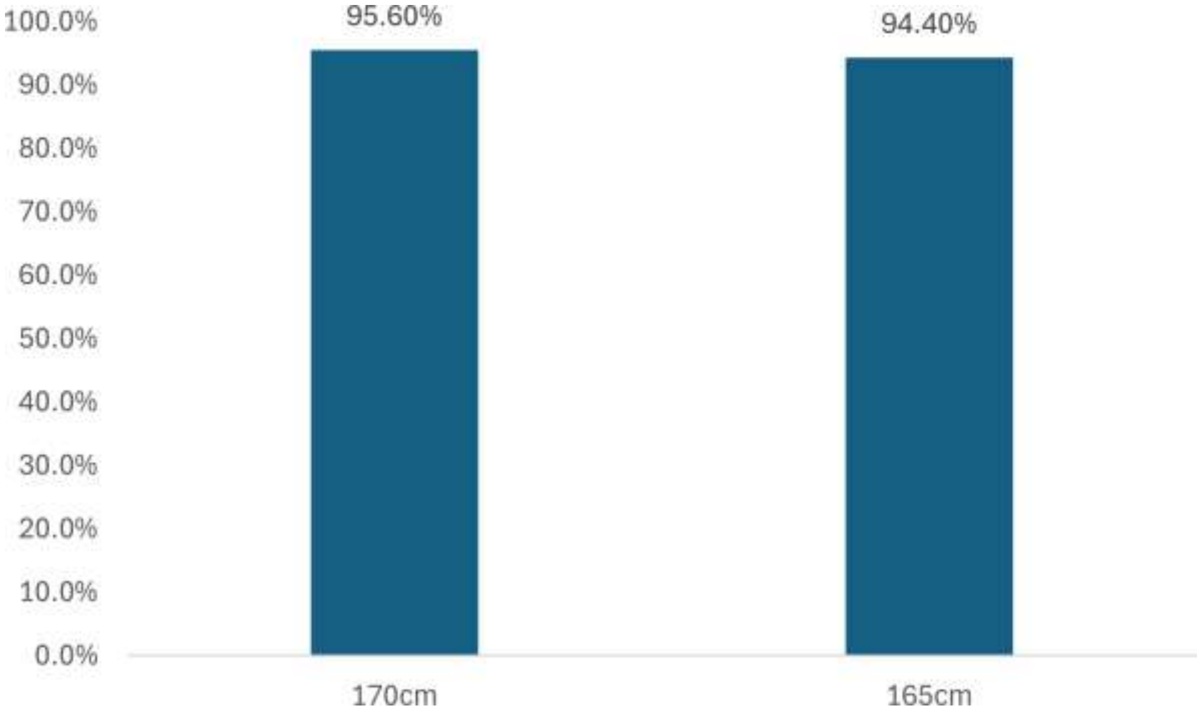


Figure 8: The body landmark detection results when the two participants of different heights were instructed to raise their hands in front of the camera to simulate cleaning movements in the air shower room.

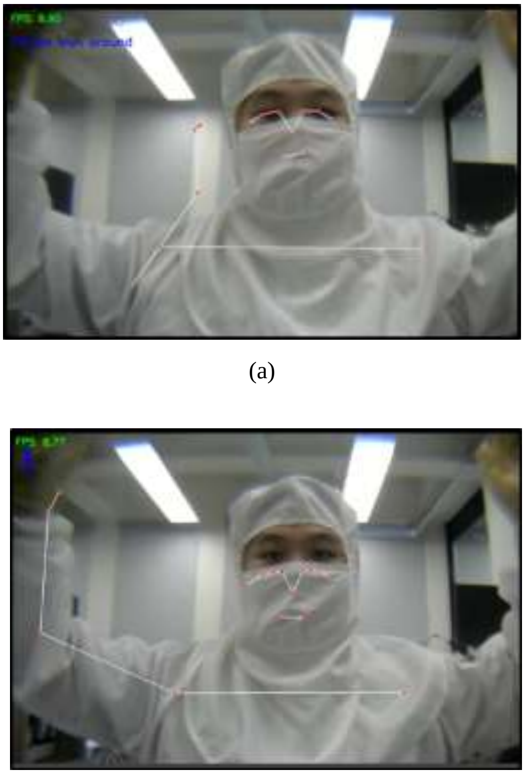
environments to validate and refine the system under realistic conditions. Additionally, exploring advanced body skeleton detection techniques or sensor fusion methods could mitigate the challenges posed by protective clothing and fisheye lens distortions, thereby improving recognition accuracy in confined spaces.

6. Conclusion

This study introduces two inspection systems suitable for deployment in dust-free cleanroom environments: the protective equipment verification system and the body pose verification system in cleanroom inspections. The former aims to develop a robust binary classification system designed to automate the recognition of proper wearing of protective clothing. Through extensive experimentation and analysis, the effectiveness of our algorithm has been rigorously validated. The adoption of VGG-19 as the foundational model for training yielded an impressive average accuracy rate of 98.42%, confirming its capability to reliably identify adherence to protective clothing protocols.

Additionally, the second system designed to detect human body poses in cleanroom inspections demonstrated not able accuracy in capturing and recording actions within the dynamic environment ofthe air shower room. This dual capability not only ensures workplace safety within semiconductor facilities but also extends its applicability to diverse industries such as healthcare and pharmaceuticals. By leveraging advanced machine learning techniques, our system enhances operational efficiency and safety protocols, paving the way for broader implementation across various sectors.

Looking forward, future research could explore enhancing the system's adaptability to diverse operational environments and expanding its capabilities beyond binary classification. Integrating real-time feedback mechanisms and leveraging edge computing technologies could further optimize performance and scalability. Moreover, conducting field trials in real-world air shower rooms and refining the system's algorithms to address challenges like protective clothing and environmental variations would validate its robustness and reliability in practical settings. These advancements not only bolster workplace safety standards but also drive forward automation technologies in critical industrial processes, fostering safer and more efficient workplaces globally.



(a)

(b)

Figure 9: The errors occurred when adopting the keypoint landmark detector: (a) joints and limbs were not detected, and (b) the location of face landmarks was shifted

References

[1] F. Gong, X. Ji, W. Gong, X. Yuan, C. Gong, Deep learning based protective equipment detection on offshore drilling platform, Symmetry 13 (6) (2021) 954.

[2] Z. Liu, T. Weise, Z. Wu, Detection of personal protective equipment in factories: A survey and benchmark dataset, in: International Conference on Intelligent Computing, Springer, 2022, pp. 448–459.

[3] S. Chen, K. Demachi, Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph, Automation in construction 125 (2021) 103619.

[4] N.-T. Nguyen, Q. Tran, C.-H. Dao, D. A. Nguyen, D.-H. Tran, Automatic detection of personal protective equipment in construction sites using metaheuristic optimized yolov5, Arabian Journal for Science and Engineering (2024) 1–19.

[5] B. Wu, C. Pang, X. Zeng, X. Hu, Me-yolo: Improved yolov5 for detecting medical personal protective equipment, Applied Sciences 12 (23) (2022) 11978.

[6] Q. Zhang, Z. Pei, R. Guo, H. Zhang, W. Kong, J. Lu, X. Liu, An automated detection approach of protective equipment donning for medical staff under covid-19 using deep learning, Cmes-Computer Modeling in Engineering & Sciences (2022) 845–863.

[7] N. D. Nath, A. H. Behzadan, S. G. Paal, Deep learning for site safety: Real-time detection of personal protective equipment, Automation in Construction 112 (2020) 103085.

[8] M. Ferdous, S. M. M. Ahsan, Ppe detector: a yolo-based architecture to detect personal protective equipment (ppe) for construction sites, Peer Computer Science 8 (2022) e999.

[9] G. Gallo, F. Di Rienzo, F. Garzelli, P. Ducange, C. Vallati, A smart system for personal protective equipment detection in industrial environments based on deep learning at the edge, IEEE Access 10 (2022) 110862–110878.

[10] X. Ji, F. Gong, X. Yuan, N. Wang, A high-performance framework for personal protective equipment detection on the offshore drilling platform, Complex & Intelligent Systems (2023) 1–16.

[11] J. P. Cheng, P. K.-Y. Wong, H. Luo, M. Wang, P. H. Leung, Vision-based monitoring of site safety compliance based on worker re-identification and personal protective equipment classification, Automation in Construction 139 (2022) 104312.

[12] M. I. B. Ahmed, L. Saraireh, A. Rahman, S. Al-Qarawi, A. Mhran, J. Al-Jalaoud, D. Al-Mudaifer, F. Al-Haidar, D. AlKhulaifi, M. Youldash, et al., Personal protective equipment detection: A deep-learning-based sustainable approach, Sustainability 15 (18) (2023) 13990.

[13]

J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767 (2018).

[14] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, Yolox: Exceeding yolo series in 2021, arXiv preprint arXiv:2107.08430 (2021).

[15] M. Ferdous, S. M. M. Ahsan, Chvg dataset, https://figshare.com/articles/dataset/CHVG_Dataset/19625166 (2022).

[16] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, Yolov4: Optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934 (2020).

[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[18] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, Centernet: Keypoint triplets for object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6569–6578.

[19] M. Tan, R. Pang, Q. V. Le, Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10781–10790.

[20] J. Glenn, Yolov5 by ultralytics, https://github.com/ultralytics/yolov5 (2020-5-29).

[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[22] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Learning generalisable omni-scale representations for person re-identification, IEEE transactions on pattern analysis and machine intelligence 44 (9) (2021) 5056–5069.

[23] P. K.-Y. Wong, H. Luo, M. Wang, J. C. Cheng, Enriched and discriminative convolutional neural network features for pedestrian re-identification and trajectory modeling, Computer-Aided Civil and Infrastructure Engineering 37 (5) (2022) 573–592.

[24] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE transactions on pattern analysis and machine intelligence 39 (6) (2016) 1137–1149.

[25] G. Jocher, ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, https://github.com/ultralytics/yolov5 (Oct. 2020). doi:10.5281/zenodo.4154370.
URL https://doi.org/10.5281/zenodo.4154370

[26] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, M. Grundmann, Mediapipe: A framework for perceiving and processing reality, in: Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019, 2019.
URL https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf

[27] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.