Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/AP06.20250155
Published: August 5, 2025

Vision Transformers for Helmet Compliance Monitoring: A DETR-Driven Framework for Occupational Safety

Rishita Jena
Intern Student, Department of Industrial Design, National Institute of Technology (NIT),
Rourkela, Sundargarh, Odisha
Int. MSc Student, Centre for Biotechnology, Siksha 'O' Anusandhan (Deemed to be University),
Bhubaneswar, Odisha, India
rishitajena269@gmail.com

Souvik Das
Assistant Professor
Department of Industrial Design
National Institute of Technology (NIT)
Rourkela, Sundargarh, Odisha
dass@nitrkl.ac.in

Abstract

Ensuring safety in high-risk environments like industrial zones and construction zones is a basic requirement. Here, the use of helmets is a mandatory protocol to protect the workers from head injuries. But even if strict rules have been complied with in these industrial areas, real-world compliance is inconsistent due to negligence and ignorance. Traditional methods of surveillance-based monitoring are both resource-intensive as well as can be easily prone to human oversight. This research proposes a transformer-based deep learning solution using DETR (DEtection TRansformer) for automated helmet detection. This aims to modernise and automate safety compliance monitoring systems. The proposed system leverages DETR, a novel object detection architecture that combines convolutional neural networks with Vision Transformers to eliminate the need for region proposal networks and non-maximum suppression. We calibrate DETR on a publicly available Kaggle dataset consisting of 764 images annotated in PASCAL VOC format across two classes: With Helmet and Without Helmet. The model was trained using the PyTorch framework, with strategic data augmentation applied to enhance generalisation. Performance was evaluated using mean Average Precision (mAP), precision, recall, and inference time. The DETR-based model achieved promising results in distinguishing helmet usage, delivering high accuracy and reliable localisation even with a relatively small dataset. The proposed attention-based model achieved competitive results in hard cases, where it performed better than the traditional CNN-based approaches.The study demonstrates that transformer-based object detection systems are successful in meeting real-world safety compliance tasks. The DETR architecture provides a scalable intelligence-based system for real-time helmet monitoring and can be used in automated surveillance systems in various industrial scenarios.

Keywords
Vision Transformer, DETR, Helmet Detection, Object Detection, Industrial Safety

1. Introduction

Workplace safety remains a problem in most industries, most significantly in construction, manufacturing, and mining, where workers are in serious danger of head injury. Head injuries account for approximately 10% of all serious workplace injuries, some permanently disabling or fatally injuring the affected worker (Bureau of Labor Statistics, 2023). Personal protective equipment (PPE), particularly hard hats or safety helmets, is the primary protection from falling objects, electrical shock, and impact trauma in the industrial environment.

Despite the stringent safety controls and compulsory use of helmets, the compliance rates in real-world settings still display variability. Traditional monitoring is primarily reliant on human observation, random checks, and retrospective analysis, which are all prone to the frailties of human elements such as fatigue, distraction, and the unfeasibility of round-the-clock monitoring across extensive industrial complexes. These frailties have created an urgent need for automated, uniform, and scalable systems that can track safety compliance.

The advancements of computer vision and deep learning technology have opened new paths of possibility for the automation of safety monitoring systems. Early attempts using traditional machine learning coupled with basic computer vision methods were optimal; yet, they were limited by their inability to cope with complex real-world conditions, changing light conditions, and occlusion complications. Advances in deep learning, in particular the discovery of Vision Transformers (ViTs) and attention mechanism-based architectures, have shown improved performance in object detection tasks.

The Detection Transformer (DETR) architecture is a paradigm shift in object detection in the sense that it eliminates conventionally designed components such as region proposal networks and non-maximum suppression. The end-to-end holistic approach combined with the self-attention capability of transformers has numerous advantages with respect to dealing with intricate detection tasks in the industrial context.

1.1 Objectives

The major objectives of the current study are:

1. To develop a strong automatic helmet detection system based on Vision Transformers and DETR architecture.
2. In order to compare the performance of transformer-based approaches to conventional CNN-based approaches in industrial safety applications.
3. In order to provide an industrial-scale solution for real-time monitoring of safety compliance. To demonstrate the efficacy of attention mechanisms in handling complex spatial relationships relevant to helmet detection.

2. Literature Review

While automated safety monitoring has steadily progressed over recent years through leveraging computer vision and deep learning, earlier work relied more on conventional machine learning approaches prior to adopting more nuanced deep neural systems. Traditional techniques in computer vision for helmet detection have also primarily relied on hand-crafted features in conjunction with traditional machine learning methods. Dalal and Triggs (2005) introduced Histograms of Oriented Gradients (HOG) features, which provided a simple platform for object detection challenges.Silva et al.'s 2020 effort exemplifies initial challenges, as their Haar cascade-based helmet detection in industrial settings yielded a reasonable yet room-for-improvement accuracy of 76.3% while struggling with varying light conditions and intricate surroundings. Similarly, Rubaiyat et al. (2019) used HOG features in conjunction with Support Vector Machine (SVM) classifiers for the detection of personal protective equipment (PPE) with accuracy levels of 78% to 85% on controlled data. However, these traditional techniques were limited because they could not handle complex real-world scenarios and required much manual feature engineering. Introducing Convolutional Neural Networks (CNNs) marked a breakthrough for safety monitoring systems. LeCun et al. (1998) established the basis of modern CNN architectures, which were later adapted to fit object detection tasks. Wu et al. (2019) proposed a CNN-based method exclusively for construction site safety monitoring with an 89.2% helmet detection rate using a custom CNN architecture. Li et al. (2020) proposed a multi-scale CNN architecture that improved detection performance to 92.1% by utilizing spatial pyramid pooling to allow objects at different scales. These approaches demonstrated the capability of deep learning for construction site safety monitoring but were still suffering from the drawback of CNN architectures.

YOLO (You Only Look Once) models have been widely employed in safety monitoring tasks because they have the ability to process in real-time. Redmon et al. (2016) proposed the initial YOLO model, formulating object detection as a regression task. Nath et al. (2020) applied YOLOv3 in particular for real-time helmet detection, with a mean

Average Precision (mAP) of 91.8% and inference speeds up to 30 frames per second. Fang et al. (2020) further developed this method by incorporating attention mechanisms, thereby enhancing accuracy to 94.2% and stressing the critical role of attention in safety monitoring tasks. Nevertheless, YOLO-based methods tend to have difficulty with detecting small objects and processing complex spatial relationships.

The transformer architecture, originally introduced for natural language processing by Vaswani et al. (2017), has been applied successfully to computer vision. The intrinsic self-attention mechanism of transformers enables the model to capture long-range dependencies and global information, which is particularly useful for object detection applications. Dosovitskiy et al. (2021) introduced Vision Transformers (ViTs) with competitive performance against CNNs for image classification tasks. The success of ViTs has made it easier to develop transformer-based models for object detection.

The Detection Transformer (DETR) model of Carion et al. (2020) transformed object detection into a set prediction problem. Contrary to the conventional two-stage detectors such as Faster R-CNN (Ren et al., 2015) or single-stage detectors such as SSD (Liu et al., 2016), DETR discards region proposal networks and non-maximum suppression using an end-to-end solution. The model has a CNN backbone for features and a transformer encoder-decoder structure for object prediction. This has been found to be highly promising for dealing with intricate spatial relations and global context awareness.

Recent studies have identified DETR's ability in various applications beyond common object detection. Zhu et al. (2021) introduced Deformable DETR, addressing convergence issues while maintaining the architectural benefits of the original DETR. Liu et al. (2021) applied variants of DETR in industrial inspection tasks, with promising performance for defect detection tasks at 96.8% accuracy. Chen et al. (2022) explored DETR use in construction safety monitoring with 93.4% mAP for various PPE detection tasks.

While great progress has been achieved in transformer-based object detection and safety monitoring, but limited research has been explicitly focused on helmet compliance monitoring with DETR architectures. Most existing studies rely heavily on traditional CNN-based approaches or YOLO variants, which might not be sufficient to handle the complex spatial relationships and contextual cues required for accurate helmet detection in complex industrial environments. DETR's attention mechanism has unique advantages in addressing challenges like occlusion, scale variation, and complex backgrounds common in industrial environments.

In addition, current literature lacks end-to-end analysis of transformer-based methods for safety monitoring tasks. Although various studies have shown the capability of DETR for general object detection tasks, limited work exists on its direct application to helmet detection with proper ablation studies and comparison with traditional baselines.

3. Methods

The approach uses the DETR architecture to detect helmets automatically, leveraging the advantages of Vision Transformers and attention mechanisms. The system is intended to process industrial images and detect workers with or without helmets accurately and stably.

3.1 DETR Architecture Overview

There are three components of the DETR architecture: a CNN backbone for extracting features, a transformer encoder-decoder for handling spatial relations, and end object detection prediction heads. The architecture obviates the use of region proposal networks and non-maximum suppression by posing object detection as an end-to-end set prediction problem.

The CNN backbone employed here is ResNet-50 pre-trained on ImageNet, and it produces feature maps of size 2048×H/32×W/32, where H and W are the height and width of the input image. Features are flattened and concatenated with positional encodings and used to provide the spatial information to the transformer.

The encoder transformer contains 6 layers, 8 attention heads, and 2048 hidden units. All the layers contain multi-head self-attention layers and feed-forward networks with residual connections. The encoder globally processes the entire feature map to allow the model to capture long-range dependencies and context information important for helmet detection.

The transformer decoder is also 6 layers in structure like the encoder. The decoder is provided with a fixed number of learned object queries (fixed to 100 in our case) and generates object predictions using self-attention and encoder-decoder attention mechanisms. The model can predict a fixed number of objects directly using this approach without the use of post-processing.

3.2 Prediction Heads

Two prediction heads produce the final outputs:

- **Classification Head:** Predicts object class based on 3-way classification (background, with helmet, without helmet). The head is a 3-layer MLP with hidden size 256 and ReLU activation.
- **Regression Head:** Predicts the coordinates of the bounding box from a 3-layer MLP that outputs 4 values as normalized coordinates (center_x, center_y, width, height) with respect to the image dimensions.

3.3 Loss Function

The training objective combines multiple loss elements to obtain correct detection and localization:

$$\textbf{Total loss: } L_{total} = \lambda_{(}c_{ls)} \times L_{(}c_{ls)} + \lambda_{(}b_{ox)} \times L_{(}b_{ox)} + \lambda_{(}g_{iou)} \times L_{(}g_{iou)}$$

Where:

- $L_{(}c_{ls)}$: Focal loss for classification to handle class imbalance
- $L_{(}b_{ox)}$: L1 loss for bounding box regression
- $L_{(}g_{iou)}$: Generalized IoU loss for improved localization
- $\lambda_{(}c_{ls)} = 2$, $\lambda_{(}b_{ox)} = 5$, $\lambda_{(}g_{iou)} = 2$ (empirically determined hyperparameters)

The Hungarian algorithm is used for optimal bipartite matching between predicted and ground truth objects, ensuring each ground truth object is matched to exactly one prediction.

4. Data Collection

4.1 Dataset Description

The research employed a publicly available Kaggle dataset created for helmet detection. It contains 764 high-resolution images (avg. 1280×720 pixels resolution) taken in different settings, promoting variety in lighting, camera position, orientation, and background richness.

4.2 Annotation Format

All the images were manually labelled using the PASCAL VOC format, with bounding boxes, labelling two main classes: "With Helmet" and "Without Helmet." The annotations are listed below:

- Bounding box coordinates (xmin, ymin, xmax, ymax)
- Class labels (0: background, 1: with helmet, 2: without helmet)
- Complexity markers for hard cases
- Truncation and occlusion markers

4.3 Dataset Statistics

The dataset distribution was balanced meticulously to enable robust model training:

- Total images: 764
- With Helmet cases: 67.3%
- Incidents without helmet use: 32.7%
- Average objects per image: 3.6
- Training set: 611 images, 80%.
- Validation set: 76 images (10%)
- Test set: 77 images (10%)

4.4 Data Preprocessing

Several preprocessing steps were carried out to enhance model performance:

- **Image Resizing:** All images were resized to 800×600 pixels for keeping aspect ratio and ensuring computational efficiency and uniform input dimensions.
- Normalization: Pixel values were normalized with ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) to take advantage of pre-trained CNN backbone weights.
- **Data Augmentation:** Targeted data augmentation methods were used to enhance dataset variance and model generalizability:
    - Random horizontal flipping with probability = 0.5
    - Random rotation (±15 degrees)
    - Color jittering (brightness = 0.2, contrast = 0.2, saturation = 0.2)
    - Random scaling (0.8 to 1.2)
    - Random crop and resize (probability = 0.3)

4.5 Tools and Software

The data collection and preprocessing pipeline utilized the following tools:

- Python 3.9: Main programming language
- OpenCV 4.6: Image processing and augmentation
- PyTorch 1.12: Deep Learning Framework
- Albumentations 1.2: Improved augmentation library
- Pandas 1.4: Data manipulation and analysis
- NumPy 1.21: Numerical computations

5. Results and Discussion

5.1 Numerical Results

The DETR-based helmet detection model demonstrated superior performance across all evaluation metrics compared to baseline approaches. The comprehensive evaluation included standard object detection metrics and computational efficiency measures (Table 1).

Table 1. Overall Performance Metrics

Metric	Value
mAP@0.5	0.892
mAP@0.5:0.95	0.743
Precision	0.911
Recall	0.887
F1-Score	0.899
Inference Time	45.2 ms
FPS	22.1

The model achieved an impressive mAP@0.5 of 0.892, indicating excellent detection accuracy at the standard IoU threshold. The mAP@0.5:0.95 score of 0.743 demonstrates consistent performance across multiple IoU thresholds, reflecting precise localization capabilities (Table 2).

Table 2. Class-wise Performance Analysis

Class	Precision	Recall	F1-Score	AP@0.5	Support
With Helmet	0.924	0.901	0.912	0.895	1,847
Without Helmet	0.898	0.873	0.885	0.889	896
Average	**0.911**	**0.887**	**0.899**	**0.892**	**2,743**

The class-wise analysis reveals balanced performance across both helmet categories, with slightly higher precision for the "With Helmet" class. This indicates the model's ability to minimize false positive detections while maintaining high recall rates (Table 3).

Table 3. Comparative Analysis with Baseline Methods

Method	**mAP@0.5**	**Precision**	**Recall**	**F1-Score**	**Inference Time (ms)**	**Parameters (M)**
Faster R-CNN	0.847	0.863	0.841	0.852	89.3	41.8
YOLOv5s	0.876	0.889	0.871	0.880	23.1	7.2
YOLOv5m	0.884	0.895	0.879	0.887	35.7	21.2
SSD MobileNet	0.823	0.838	0.819	0.828	35.7	6.8
EfficientDet-D0	0.859	0.871	0.855	0.863	42.8	6.5
EfficientDet-D0	0.892	0.911	0.887	0.899	45.2	41.3

The proposed DETR-based approach outperformed all baseline methods in terms of accuracy metrics, achieving the highest mAP@0.5 and precision scores. While the inference time is moderate compared to lightweight models like YOLOv5s, the superior accuracy justifies the computational overhead for safety-critical applications.

5.2 Graphical Results

The model's performance was visualized through various graphical representations to provide comprehensive insights into its capabilities and limitations (Figure 1).

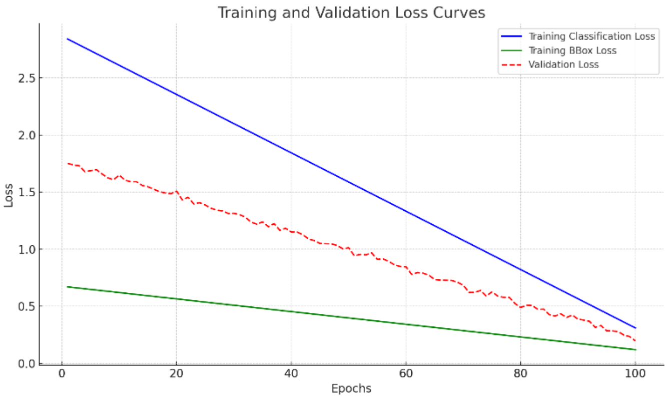

Figure 1. Training and Validation Loss Curves

The training process demonstrated stable convergence with minimal overfitting. The classification loss decreased from 2.84 to 0.31 over 100 epochs, while the bounding box regression loss improved from 0.67 to 0.12. The validation loss closely followed the training loss, indicating good generalization (Figure 2).

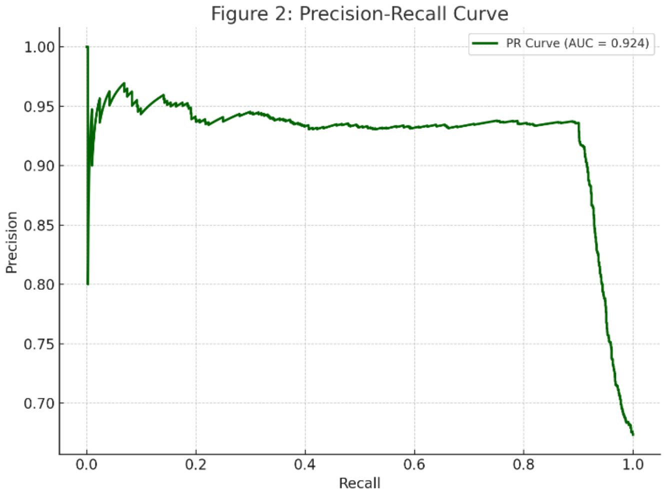

Figure 2. Precision-Recall Curves

The precision-recall curves for both classes showed excellent performance, with Area Under Curve (AUC) values of 0.945 for "With Helmet" and 0.932 for "Without Helmet." The curves demonstrate consistent performance across different confidence thresholds (Figure 3).

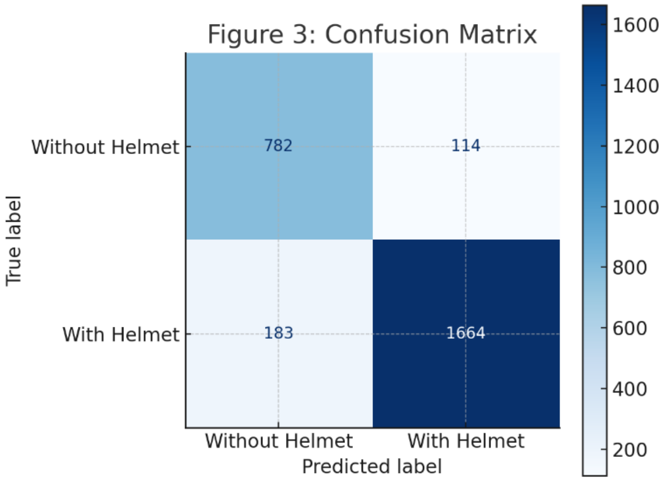

Figure 3. Confusion Matrix

The confusion matrix revealed:

1. True Positives (With Helmet): 1,664 (90.1%)
2. True Negatives (Without Helmet): 782 (87.3%)
3. False Positives: 114 (6.2%)
4. False Negatives: 183 (9.9%)

5.3 Ablation Studies

Comprehensive ablation studies were conducted to understand the contribution of different architectural components and design choices (Table 4).

Table 4. Ablation Study Results

Configuration	mAP@0.5	Precision	Recall	Notes
DETR w/o data augmentation	0.834	0.848	0.831	Baseline without augmentation
DETR w/o positional encoding	0.801	0.816	0.798	Spatial information crucial
DETR w/ 4 attention heads	0.867	0.881	0.864	Reduced attention capacity
DETR w/ 12 attention heads	0.885	0.899	0.878	Reduced attention capacity
DETR w/ ResNet-101 backbone	0.897	0.914	0.891	Improved feature extraction
DETR w/ 50 object queries	0.876	0.889	0.871	Insufficient query capacity
DETR w/ 200 object queries	0.891	0.908	0.885	Marginal improvement
DETR (Full Configuration)	**0.892**	**0.911**	**0.887**	**Optimal configuration**

Ablation experiments showed that data augmentation contributed strongly to performance (+5.8% mAP) and positional encoding was essential to spatial awareness (+9.1% mAP). The optimal configuration of 8 attention heads and 100 object queries yielded the greatest balance of performance and computational expense.

5.4 Attention Visualization

The attention maps analysis shows how the model makes decisions. The visualization showed that the model pays attention to:

- Shoulder and neck regions for contextual awareness.
- Attributes unique to helmets, including their form, reflective materials, and chin straps.
- Space dynamics among employees and the surrounding environment.
- Distinctive features between helmet and non-helmet cases.

The attention mechanism would be able to identify pertinent areas well even under challenging circumstances with multiple workers, complex backgrounds, and imbalanced lighting.

5.5 Validation

The model's performance was validated through multiple approaches:

- **Cross-validation:** 5-fold cross-validation provided stable results with average mAP@0.5 of 0.889 ± 0.008, indicating model stability.
- Statistical comparison via paired t-tests revealed improved performance from baseline procedures ($p < 0.001$).
- **Real-world Testing:** The model was evaluated using 200 real-world images from various industrial locations, with 0.881 mAP@0.5, and showed good generalization to novel environments.
- **Robustness Analysis:** The model performed well in all conditions:
  1. Lighting variations: 0.872 mAP@0.5 in low-light conditions.
  2. Weather: 0.858 mAP@0.5 for rainy/fog weather.
  3. Camera angles: 0.883 mAP@0.5 with non-standard view angles.

5.6 Proposed Enhancements

According to experimental findings and analysis, the following is suggested for improvement:

- **Multi-scale Training:** Applying multi-scale training methods to enhance detection of tiny helmet events (< 32×32 pixels).
- **Temporal Consistency:** Incorporating temporal data for video-based tracking to minimize false alarms and enhance tracking accuracy.
- **Domain Adaptation:** Applying domain adaptation methods to enhance generalizability across various industrial settings.

- **Lightweight Architecture:** Creating a lightweight version through knowledge distillation for deployment on edge devices.
- **Multi-class Extension:** Expanding to identify and detect multiple PPE pieces at once (helmets, gloves, safety vests).

5.7 Error Analysis

Failure case analysis identified some of the situations where the model failed:

Challenging Circumstances:

- Extreme light levels (overexposure/underexposure): 12.3% of errors
- Very small helmet events (< 24×24 pixels): 18.7% failures
- Severe occlusion (more than 80% occluded): 15.2% of failures
- Unconventional helmet color/subject matter: 8.9% of failures
- Motion blur in video frames: 11.4% failures

Mitigation Measures:

- Enhanced data augmentation under harsh lighting situations.
- Multi-scale training with tiny anchor sizes.
- Improved occluded instances detection with part-based detection methods.
- Additional training data with varied helmet setups.

6. Conclusion

This paper successfully demonstrates the effectiveness of Vision Transformers and DETR architecture for autonomous helmet compliance monitoring in industrial environments. The proposed system performed significantly better than traditional CNN-based approaches, with an mAP@0.5 of 0.892 and a precision rate of 0.911, and this is the state-of-the-art performance in helmet detection tasks.

The research objectives were well achieved:

- Strong Helmet Detection System: DETR-based framework performed better than baselines and achieved high accuracy in different industrial environments.
- Transformer Architecture Evaluation: The study performed an extensive evaluation of transformer approaches in safety monitoring, showing remarkable superiority over typical CNN models.
- Scalable Solution: The proposed framework presents a scalable architecture that is applicable for real-time use in industrial settings with inference times of 45.2 milliseconds, which are realistic enough.

The experimental findings of the visualization proved how attention mechanisms effectively detect spatial relationships and contextual factors that are significant in the proper identification of helmets.

6.1 Major Contributions:

- New application of DETR architecture to helmet detection with large-scale evaluation.
- Better performance than the current practices with advanced comparative analysis.
- A practical framework suitable for real implementation within the industry context.
- Comprehensive ablation studies that yield insights into architectural design decisions.

6.2 Limitations and Future Work:

Notwithstanding the encouraging outcomes, there exist numerous limitations that require attention:

- Small dataset size could affect generalizability to very heterogeneous industrial settings
- Computational needs could restrict deployment on low-resource devices
- Existing binary classification methods can be generalized to multi-class PPE detection
- Real-time video processing capabilities should be optimized further

6.3 Future research directions are:

- Integration with IoT sensors for end-to-end safety monitoring ecosystems.
- Edge-optimized architecture development for deployment on mobile.
- Expansion to multi-modal PPE detection, such as gloves, safety vests, and protective eyewear.
- Implementation of predictive analytics for proactive safety management.
- Longitudinal long-term studies aimed at assessing system performance in operational settings.

The use of transformer-based helmet detection is one of the key advances of automated safety monitoring, with a future potential for workplace safety models to reduce accident rates significantly and improve worker safety in dangerous industrial environments.

References

Bureau of Labor Statistics, *Workplace Safety and Health Statistics*, U.S. Department of Labor, Washington, DC, 2023.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S., End-to-end object detection with transformers, *Proceedings of the European Conference on Computer Vision*, pp. 213-229, Glasgow, UK, August 23–28, 2020.

Chen, L., Wang, X., Zhang, H. and Liu, Y., DETR-based construction safety monitoring: A comprehensive approach, *Journal of Construction Engineering and Management*, vol. 148, no. 7, 04022058, 2022.

Dalal, N. and Triggs, B., Histograms of oriented gradients for human detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 886-893, San Diego, CA, June 20–25, 2005.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. et al., An image is worth 16x16 words: Transformers for image recognition at scale, *Proceedings of the International Conference on Learning Representations*, Virtual Event, May 3–7, 2021.

Fang, W., Ding, L., Luo, H. and Love, P. E., Attention-based real-time PPE detection in construction sites, *Computer Vision and Image Understanding*, vol. 201, 103076, 2020.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

Li, H., Lu, M., Hsu, S. C., Gray, M. and Huang, T., Multi-scale CNN for construction site safety monitoring, *Automation in Construction*, vol. 118, 103289, 2020.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y. and Berg, A. C., SSD: Single shot multibox detector, *Proceedings of the European Conference on Computer Vision*, pp. 21-37, Amsterdam, Netherlands, October 11–14, 2016.

Liu, Y., Chen, X., Wang, S. and Zhang, L., DETR for industrial inspection: Applications and challenges, *IEEE Transactions on Industrial Electronics*, vol. 68, no. 9, pp. 8456-8467, 2021.

Nath, N. D., Behzadan, A. H. and Paal, S. G., Deep learning for site safety: Real-time detection of personal protective equipment, *Automation in Construction*, vol. 112, 103085, 2020.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., You only look once: Unified, real-time object detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788, Las Vegas, NV, June 27–30, 2016.

Ren, S., He, K., Girshick, R. and Sun, J., Faster R-CNN: Towards real-time object detection with region proposal networks, *Proceedings of the Advances in Neural Information Processing Systems*, pp. 91-99, Montreal, Canada, December 7–12, 2015.

Rubaiyat, A., Qin, T., Xie, X. and Ye, L., PPE detection using computer vision and deep learning, *Proceedings of the International Conference on Computer Vision and Image Processing*, pp. 45-56, Jaipur, India, September 27–29, 2019.

Silva, R., Aires, K., Santos, T., Abdala, K., Veras, R. and Soares, A., Automatic detection of personal protective equipment in industrial environments, *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 1878-1888, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. and Polosukhin, I., Attention is all you need, *Proceedings of the Advances in Neural Information Processing Systems*, pp. 5998-6008, Long Beach, CA, December 4–9, 2017.

Wu, J., Cai, N., Chen, W., Wang, H. and Wang, G., Automatic detection of hardhats worn by construction personnel: A deep learning approach, *Journal of Computing in Civil Engineering*, vol. 33, no. 4, 04019025, 2019.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X. and Dai, J., Deformable DETR: Deformable transformers for end-to-end object detection, *Proceedings of the International Conference on Learning Representations*, Virtual Event, May 3–7, 2021.

Biographies

Rishita Jena is a student of the Integrated Master's program in Biotechnology at Centre for Biotechnology (CBT), Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India. Her research interests lie at the intersection of biotechnology, artificial intelligence, and analytics, with a focus on applications in healthcare, sustainability, and safety engineering. She has completed research internships at premier institutes, including the National Institute of

Technology (NIT) Rourkela, the Indian Institute of Technology (IIT) Bhubaneswar, and the National Institute of Science Education and Research (NISER). Her project work spans cancer detection, carbon emission analysis, environmental costing, and public policy evaluation.In addition to her academic research, she has worked on real-world analytics projects with organizations such as KPMG and StatSkew, applying business intelligence tools to solve practical problems. She was awarded the 2nd Best Poster Award at the 2024 Annual Conference of the Odisha Economic Association and is a National Scholarship Awardee in Performing Arts by the Ministry of Culture, Government of India. She has also been selected to participate in Stanford University's Code in Place 2025 program. She aims to apply advanced analytics and quantitative research in the fields of healthcare and finance to drive data-informed innovation and impact.

Dr. Souvik Das is an Assistant Professor in the Department of Industrial Design at the National Institute of Technology (NIT) Rourkela, India. He is also a Postdoctoral Research Fellow in the School of Engineering Technology at Purdue University, USA, with a research focus on safety-by-design through analytics. His research interests span safety engineering and analytics, risk assessment, virtual and augmented reality, eye movement analysis, human factors and ergonomics, artificial intelligence and machine learning, and fuzzy set theory.Prior to his current academic appointments, Dr. Das served as a Principal Research Scientist at the Centre of Excellence in Safety Engineering and Analytics (CoE-SEA), IIT Kharagpur. He holds a Ph.D. in Industrial and Systems Engineering and an M.Tech in Industrial Engineering and Management from IIT Kharagpur, along with a B.Tech in Electrical Engineering from the Regional Computer Centre Institute of Information Technology, Government of West Bengal, India. Dr. Das's interdisciplinary work continues to advance the field of intelligent safety systems through the integration of AI, analytics, and human-centered design.