DPPNet: A Deformable-Perspective-Perception Network for Safety Helmet Violation Detection

Yahya Alassaf
Department of Civil Engineering, College of Engineering, Northern Border University, Saudi Arabia

Yahia Said
Department of Electrical Engineering, College of Engineering, Northern Border University, Saudi Arabia
said.yahia1@gmail.com (corresponding author)

Received: 16 November 2023 | Revised: 5 December 2023 | Accepted: 7 December 2023

ABSTRACT

The issue of worker safety at construction sites has become increasingly prominent within the construction industry. Safety helmet usage has been shown to reduce accidents among construction workers. However, there are instances when safety helmets are not consistently worn, which may be attributed to a variety of factors. Therefore, an automated system based on computer vision needs to be established to track protective gear appropriate usage. While there have been studies on helmet detection systems, there is a limited amount of research specifically addressing helmet detection. Also, various challenges need to be addressed such as small object miss-detection and occluded helmet detection. To fix these issues, a Deformable Perspective Perception Network (DPPNet) is proposed in this paper. Two modules make up the proposed DPPNet: Background/Image Spatial Fusion (BISF) and Grayscale Background Subtraction (GBS). While the BISF module utilizes channel attention to blend feature maps from a current frame and the background, the GBS submodule in particular incorporates background spatial information into a current frame. Additionally, the DPPNet facilitates occluded and small helmet detection. Excessive training and testing experiments have been performed using the Safety Helmet Wearing Detection (SHWD) Dataset. Experimental results demonstrate the effectiveness of the proposed DPPNet network. The obtained findings exhibit that the suggested module significantly enhances the detection capabilities of small objects. Effective mean average precision results have been obtained on the SHWD dataset coming up to 97.4% of mAP.

Keywords-safety helmet; construction sites; traffic accidents; violation detection; deep learning; DPPNet

I. INTRODUCTION

Safety is an enduring and noteworthy consideration in several sectors, with special emphasis on hazardous construction sites like chemical plants and building sites. Protective equipment, including safety clothes and helmets, plays a critical role in ensuring employees safety at construction sites with elevated risk levels. Helmets are very efficient in mitigating head injuries resulting from falling or splashing items. Similarly, safety gear serves as a protective barrier, shielding the body and arms from potential harm caused by exposure to dangerous chemicals and liquids. The absence of appropriate safety attire and helmets often has negative consequences for both families and society. Hence, monitoring the usage of safety apparel and helmets at factories or construction sites bears great importance and has wide-ranging applicability. The existing legal framework and safety protocols mandate that both the individual in a position of authority and the contractor have the responsibility of ensuring the provision, oversight, and upkeep of personal protective equipment inside construction sites. Nevertheless, some employees may exhibit a decrease in their level of attentiveness as a result of insufficient information on safety measures or experiencing pain from prolonged use, thereby elevating the likelihood of safety incidents.

Surveillance cameras are extensively used in contemporary society to oversee construction sites, mostly to supervise employees and enforce compliance with safety protocols. Although the installation of security cameras does not provide a significant obstacle, the ongoing task of watching the live feed proves to be burdensome. Continuous monitoring is necessary for human people, who are susceptible to committing mistakes. Individuals often experience a decline in concentration and vigilance after a few hours, which may lead to a lack of awareness of significant safety breaches. Furthermore, due to the presence of several cameras recording various areas inside the building site, it becomes impractical to watch all places simultaneously. Automated surveillance application offers a convenient and effective solution to address this issue. This approach implementation may be achieved by using machine learning methods or utilizing deep learning models. This study objective is to address the issue of safety helmet recognition in real-time recordings, specifically in the context of industrial or construction sites or traffic management systems, to facilitate more efficient monitoring measures.

The increasing advancements in computational capabilities and technological progress have led to the use of Artificial Intelligence methodologies, such as machine learning and deep learning, in addressing diverse challenges. These approaches have shown their efficacy in situations where conventional algorithms lack the necessary resilience to be effectively deployed across many scenarios. Although there have been previous efforts in this domain, the current body of research has yielded inadequate outcomes that lack practical applicability in real-world contexts. The limitations of previous studies mostly stemmed from the restricted range of datasets and inadequate identification of safety headgear. To maintain worker safety in construction sites, there is a growing demand for the development of an effective and dependable intelligent system that does not rely on human observers to determine whether or not the worker wears a helmet. Creating such a system is extremely desirable yet difficult due to factors such as lighting, occlusion, or low-quality security cameras. Deep learning is a potential method for automating the identification of worker helmets.

In recent years, deep learning-based techniques gained undertakable success in solving various computer vision tasks including indoor object detection [3], wayfinding assistance [4], medical imaging [5], pedestrian detection [6], road sign detection [7], traffic light detection [8] and face recognition [9]. Furthermore, background subtraction technique has been successfully integrated into the object detection model. Background subtraction helps in isolating the foreground objects from the static or slowly changing background in a scene. By subtracting the background model from the current frame, the moving or dynamic elements (objects of interest) stand out. It is particularly effective for detecting moving objects in a video stream. By continuously updating the background model and identifying pixels that deviate significantly from this model, it is possible to detect and track objects based on their motion. Background subtraction is computationally efficient, and so suitable for real-time applications, while its straightforward nature, makes it easy to implement. It is a basic and effective method for scenarios where the assumption of a relatively static background holds. Advanced background subtraction techniques are designed to handle dynamic backgrounds where parts of the scene may change over time. These methods can adapt to gradual changes in the background, ensuring robust performance in a wide range of scenarios. Motivated by the great success of deep learning and background subtraction techniques, a combination of both was performed with many novel modifications to handle the studied task.

The main aim of this work is to build a smart system used to detect helmets and to reduce the number of violations to ensure safety. To build such a system, deformable convolution has been employed. Deformable convolutions add adaptability to the convolutional kernels, allowing them to be adjusted to capture more precise and contextually relevant information from objects with varying shapes, orientations, and sizes, like helmets. This is in contrast to traditional convolutional layers, which use fixed grids for feature extraction. Different helmet designs, positions, and head shapes are frequently present in helmet detection scenarios. By allowing the convolutional filters to deform and align with the specific properties of helmets, deformable convolutions improve the network's capacity to manage such variability, leading to better localization and recognition accuracy. Moreover, the Grayscale Background Subtraction (GBS) module was deployed. Numerous advantages result from the incorporation of GBS in helmet detection, including enhanced foreground segmentation, noise reduction, contrast enhancement, adaptability to changes in illumination, real-time processing capabilities, decreased computational requirements, and generalization across various environments. The Background/Image Spatial Fusion (BISF) is the second module included in the proposed network. This module focuses on efficiently merging details from the current frame with its background to enhance scene understanding. The originality of this work is based on the use of different modules alongside with the CNN. Also, it is the first that uses deformable convolution in a background subtraction module. The suggested work ensures various contributions which are the following:

- Proposing a neural network for safety helmet detection.
- Proposing a sub-module used to enhance context understanding.
- Evaluating the proposed model on real-world challenging conditions.
- Achieving new state-of-the-art performances for safety helmet detection.

II. RELATED WORKS

An alarming rise in accidents has been largely attributed to the failure to wear helmets, particularly in situations involving constructions or activities that carry inherent risks. This disrespect for helmet safety precautions has augmented the number of fatalities, severe brain injuries, and head injuries among accident victims. For example, failure to wear a helmet by motorcyclists and cyclists has significantly increased the severity of brain injuries in traffic accidents. Detecting helmet-use violators is a primary need to be resolved by the newest smart traffic management systems. Various works have been elaborated to tackle this issue. In the following, we will review the state-of-the-art works concerning the detection of helmet use violations.

Various studies on biker and motorcycle helmet detection have been conducted. Two types of studies have been proposed: classical solutions and deep learning-based solutions. Hand-designed feature descriptors including Local Binary Pattern (LBP), Histogram of Oriented Gradient (HOG), and Scale-Invariant Feature Transform (SIFT) are used to extract the features of motorcycles and other vehicles. Last but not least, motorcycles are categorized using binary classifiers like Support Vector Machines (SVMs) and K-Nearest Neighbor (KNN). The challenge of determining motorcycle helmet wear was divided into two components in [10]. The first stage is to divide and classify the vehicle images. This phase objective is to find every moving object in the environment. In the second stage, called helmet detection, an SVM classifier is involved to distinguish between images with and without helmets while a hybrid descriptor is used to extract image attributes. Authors in [11] employ the background subtraction method. In recent years, academics have suggested solutions based on deep learning. Motorcycle riders can significantly lower their risk of fatalities and head injuries in an accident by wearing helmets. The challenge of policing the helmet law and ensuring rider compliance continues to exist in many nations. Using computer vision and deep learning techniques, authors in [12] offer a unique framework that can distinguish between the driver and passengers and identify any riders who are not wearing helmets. The authors use a second head detection module and a unique tracking algorithm that takes advantage of auxiliary data like moving direction to enhance small objects detection. Experiment results prove the efficiency of the proposed work.

The success of attempts to increase road safety through teaching and enforcement can be improved by video surveillance-based automatic detection of motorcycle helmet wear. However, there is still potential for improvement in current detection techniques, as seen by the inability to identify specific motorcycles or distinguish between drivers and passengers based on the use of helmets. Authors in [13] present a system for detecting and identifying individual motorcycles and tracking riders' specific helmet usage. Experimental results demonstrate the effectiveness of their work. They achieved a score of 0.7754 on the AI City 2023 Challenge Track 5 public leaderboard. Applications of deep learning-based image processing frequently use the recognition of safety helmets worn by construction workers as a target detection issue [34]. In [14], the authors present a study of an improved YOLOv5-based approach that takes into account the difficulties posed by dense targets, intricate backgrounds in construction environments, and the irregular shapes of safety helmets. Experiments results showed that the enhanced model's detection accuracy is 91.6%, up 2.3% from the original network model, and its detection speed is 29 frames per second. To keep motorcycle riders safe on the road, it is essential to identify both helmeted and un-helmeted riders. In this context, Authors in [15] proposed a helmet detection system based on the YOLO v4 model. The achieved results demonstrate that this work presents good performances on traffic videos.

The majority of traffic collisions involve motorcycles, and they can cause extensive damage [33]. The majority of places require motorcycle riders to wear helmets, yet for a variety of reasons, most people choose not to follow the law. In [16], authors proposed to build a helmet detection system based on YOLO v2 model. The authors demonstrate that when compared with traditional techniques this work provided better results. In [17], a real-time YOLOv5 Deep Learning (DL) model for detecting motorcycle riders and passengers and determining if the discovered person is wearing a helmet was developed and evaluated. During the experiments, the authors fed the model 100 videos, each of them lasting 20 seconds and shot at 10 frames per second. To improve the results of their work, the authors employed the data augmentation technique. This model achieved a mean average precision of 0.5267, ranking 11th on the AI City Track 5 public leaderboard. In [18], a real-time helmet violation detection system is proposed. The suggested system makes use of a novel data processing technique called few-shot data sampling to create a robust model with fewer annotations and a single-stage object detection model called YOLOv8 to detect helmet violations in real time from video frames. This system placed seventh in the 2023 AI City Challenge, Track 5, with an experimental validation score of mAP of 0.5861.

To reduce traffic accidents and increase public safety, it is essential to identify and punish such riders. Authors in [19] introduced a method for identifying, following, and tracking motorcycle riding infractions in dashboard camera footage. To effectively handle complex situations like occlusions, they use an object detector based on curriculum learning. To improve robustness and address the rider-motorcycle relationship, they provide a brand-new trapezium-shaped object boundary representation. Also, they present a regressor that produces bounding boxes for the riders who are obscured. Experimental achievements demonstrate this work efficacy evaluated on the SHWD dataset [20]. Wearing safety helmet is mandatory on construction sites to ensure safety. To detect safety helmet-wearing [35], the YOLO v7 was deployed with many modifications. First, the input images with 3 channels in RGB space color was replaced by 16 channel input. Second, SIoU was deployed as a loss function. Finally, structured pruning was applied on the network head to reduce model size. The proposed method assessment compared to state-of-the-art object detection models proved its efficiency. Authors in [36] proposed a safety helmet detection approach utilizing a finetuned YOLO model. It was first trained on a large-scale dataset then the transfer learning technique was applied and the model was fine-tuned on safety helmet detection dataset. Good results were achieved but the model still struggles in detecting small objects.

Numerous works have been proposed in the literature to lower the frequency of helmet use violations, but few of them achieve a better balance between processing time and detection accuracy. In this study, we propose a brand-new helmet and non-helmet use detection system using the benefits of deep learning, which can function in real-time and produce superior accuracy than the findings of the state-of-the-art systems. The proposed system can be incorporated into an intelligent traffic management system.

III. THE PROPOSED APPROACH

The development of a reliable safety helmet detection system is essential for upholding security rules and successfully reducing accident impact. This technological advancement not only ensures adherence to safety standards but also significantly increases safety levels. For this purpose, a DPPNet model was developed explicitly for detecting helmets. Basically, the proposed model is composed of two main modules which are BISF and GBS. While the BISF module deploys channel attention to blend frame and its background feature maps, spatial features of the background were incorporated into a current frame using the GBS module. In the following, we will detail the proposed architecture used to build the helmet violation detection system. Figure 1 illustrates the overall architecture of the proposed GBS and BISF modules.

A. Grayscale Background Subtraction

It was suggested that the RGB properties of the backdrop and the instance's context should be combined. Initially, the Laplacian Pyramid Blending technique [21] was used to construct a background image. Subsequently, an additional grayscale channel was acquired through the process of background subtraction for a given frame. The GBS submodule is centered around a background image that lacks any discernible items. Nevertheless, there are two inherent issues with utilizing raw background images extracted from a dataset collected along roadways. The main issue is the persistence of diminutive items. Due to the substantial influx of individuals and automobiles, locating a time in metropolitan crossroads where the background is devoid of any objects proves to be a challenging task. This phenomenon greatly hinders the ability to effectively exclude tiny things, such as individuals and distant cars, from the background, hence introducing confusion to the overall background structure. Based on the findings of previous studies [22, 23], the accurate detection of small objects within a restricted receptive field is highly dependent on the use of global context regional, and spatial information. Convolutional neural networks use a shallow layer with a great resolution to effectively capture the spatial color information and limited receptive field. In accordance with the aforementioned approach, the GBS submodule was developed to acquire authentic local background information and a global context.

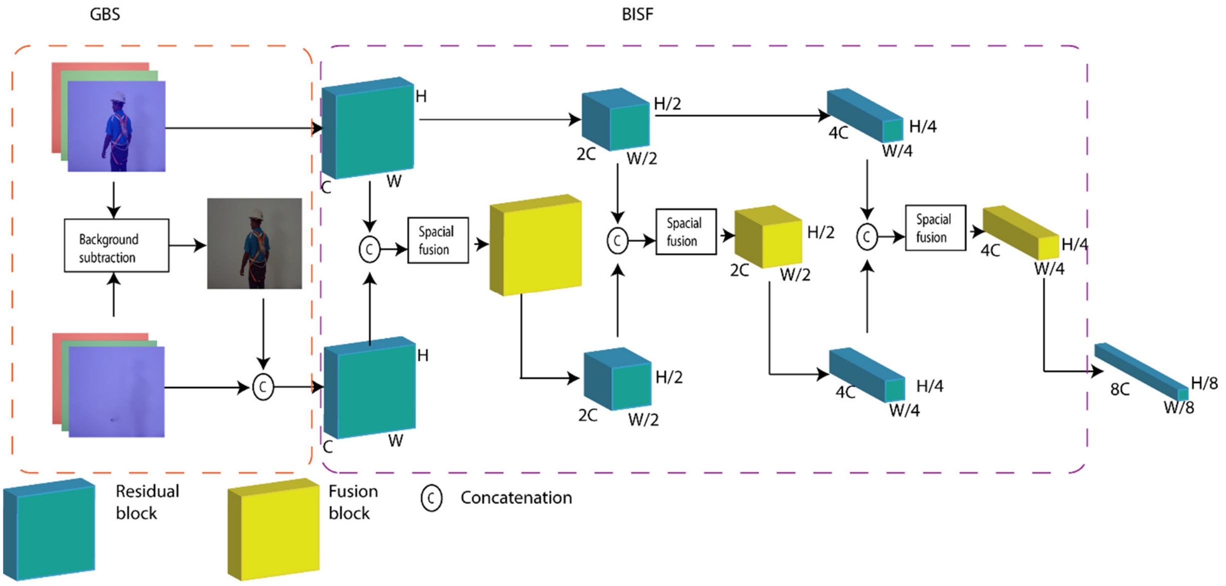

Fig. 1. Proposed GBS and BISF modules.

Initially, the background of the grayscale images was used to create a GBS channel. To preserve both global information and regional spatial information to the greatest extent feasible, the GBS channel will be included in the present frame as an extra channel. The proposed modification involves the addition of a GBS channel to the existing R, G, and B channels to alter the present frames. To achieve channel unification for the GBS channel, the initial step was performed to normalize the average illumination to a predetermined constant value. In addition, a spatial attention module was employed to direct the attention of the networks towards the regions corresponding to smaller objects. A GBS channel was obtained by the utilization of a grayscale image since the GBS submodule primary objective is to facilitate the acquisition of contextual information and spatial features. The inclusion of RGB information was deemed unneeded for the purpose of accurately identifying and finding tiny objects. Nevertheless, when the input is high-dimensional, it might result in feature maps that are not well defined, especially if the backbone layers are not expanded. Indeed, the presence of a GBS channel introduces challenges in the process of feature extraction and diminishes the overall network resilience. To address this issue, a batch normalization layer [24] was suggested as a technique for normalizing illumination. This technique aims to enhance the effectiveness of extracting features by standardizing the mean luminance of the GBS channel, which is adjusted to a pre-established value.

f_{gray}uses a standard formulation [25], presented in (1), to convert the input $I_{background}$ and $I_{current}$ to grayscale.

$$I_{gray} = V_R \times 0.299 + V_G \times 0.587 + V_B \times 0.11 \quad (1)$$

The red, green, and blue channels are represented by V_R, V_G, and V_B, respectively, I_{gray} is the combination of the current frame and the background grayscale. Generating a grayscale image by subtracting the L1 norm from the original is computed as:

$$I_{subtract}(i,j) = \left|I_{gray1}(i,j) - I_{gray2}(i,j)\right| \qquad (2)$$

where $I_{gray1}(i,j)$ is equal to $I_{gray\ current}$, and $I_{gray2}(i,j)$ is equal to $I_{gray\ background}$ background. $I_{subtract}(i,j)$ is the pixel value of (i, j) in I_{gray} background subtract.

A method for normalizing light levels was created to standardize brightness when a gray background subtraction image was obtained:

$$I_{normalized}(i,j) = I_{subtract}(i,j) + V_b \qquad (3)$$

$$V_b = V_f - V_i$$

$$V_i = \frac{1}{w\,.\,h}\sum_{k=1}^{w}\sum_{r=1}^{h} I_{subtract}(k,r)$$

The variable V_i is the GBS channel average illumination, whereas V_f represents the *fi* bias that is employed to modify the pixel value. The (i, j) pixel in the normalized gray background subtraction can be denoted as $I_{normalized}(i,j)$. Therefore, the image of background subtraction is adjusted to have a standardized illumination. In this work, the value of V_f is held constant at 20. Authors in [26], suggest that to achieve enlightenment, it is necessary to assign varying weights to distinct pixel places. Based on the aforementioned, the spatial attention concept was devised to acquire spatial knowledge pertaining to diminutive entities. As depicted in Figure 2, the function f_{max} is devised to consolidate the pixel's max value within the input feature maps f_{input}, while f_{avg} is devised to consolidate the pixel's average value within the same feature maps.

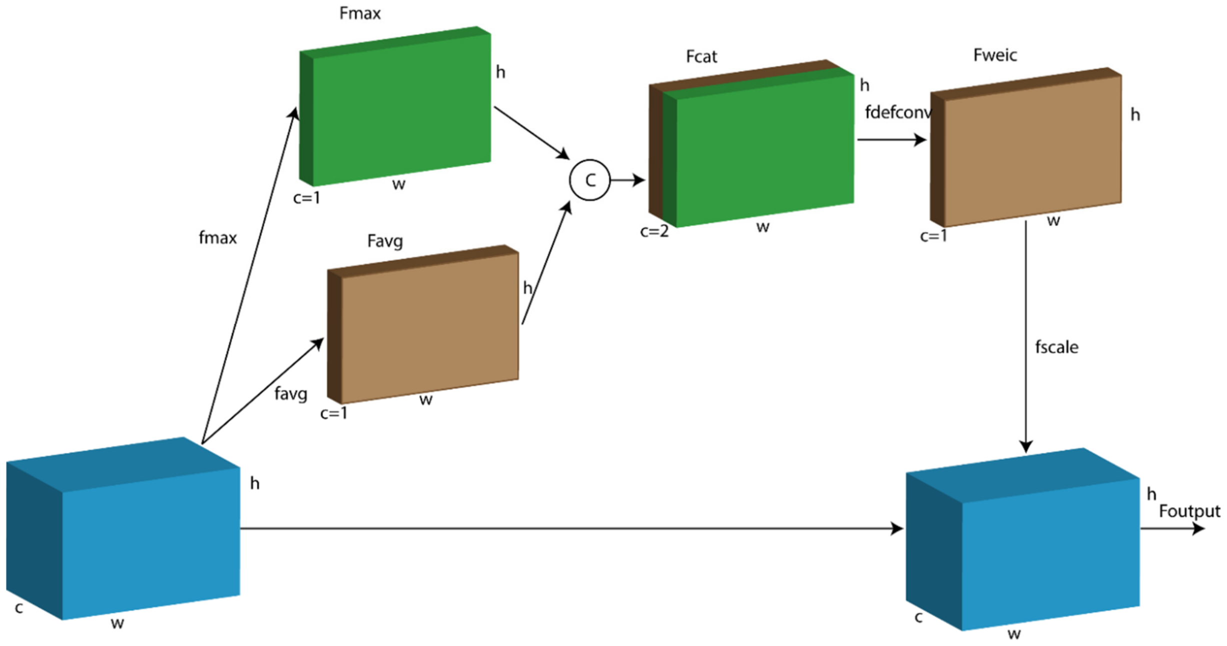

Fig. 2. Proposed spatial attention module.

Ultimately, two feature maps, namely f_{max} and f_{avg}, are produced, whereby the value of each pixel (i, j) may be computed as:

$$f_{max}(i,j) = \max\{u_1(i,j), u_2(i,j), \ldots\ldots, u_{Nc}(i,j)\} \quad (4)$$

$$f_{avg}(i,j) = \frac{\sum_{k=0}^{Nc} u_n(i,j)}{Nc}$$

The variable $u_n(i,j)$ represents the value at position (i, j) within the nth channel. Nc represents the aggregate quantity of channels. The application of a deformable convolution layer to the feature maps f_{cat} results in the determination of the weights assigned to each pixel.

B. Background/Image Spatial Fusion

This module was designed to extract features from the current frame and its background to generate feature maps through a fusion technique. First, two backbones were charged to extract semantic features from the current frame and its background. Second, a special fusion was applied to aggregate features based on channel attention on residual convolution layers.

The BISF module was partitioned into two distinct components: a cascade fusion module and the extraction of semantic features from the current frame background to combine feature maps obtained from two frames. Initially, two distinct backbones were devised to extract semantic information from both the background and the present image in a different manner. The applied methodology to integrate semantic features included the utilization of cascade fusion module maps using channel attention and residual convolution.

The BISF architecture utilizes a common backbone for extracting spatial and semantic characteristics from both background and current images. This process is performed separately for each frame. The first box seen in Figure 1 serves as the background, while the second box, colored blue, represents the current frame. The fusion module requires the presence of relevant feature maps in each backbone, as anticipated. For instance, semantic feature map for the ith backbone layer's jth channel is denoted by F_{ij}. The current image's lighting is described using this map. The inclusion of the background backbone equivalent channel 〖F'〗$_{ij}$ is necessary in order to accurately depict the total illumination in a comparable manner, utilizing the same backbone. Using consistent spatial and semantic expressions helps improve input correlation in the fusion module. In addition, the BISF backbone module assures that the input of fusion layers possesses feature maps with the same dimensions.

Features were extracted from the background and the current image using a postulated backbone structure. However, it is not certain that the best results would be produced by simply joining the two feature maps. To provide accurate background context, a module that combines background and present information was suggested. This module incorporates residual convolution and channel attention methods. As suggested by previous studies [27, 28], it is advisable to provide distinct weights to various channels within the feature maps. Regarding the BISF module, it incorporates channels that consist of semantic feature mappings derived from both current frames and the background. Therefore, a channel attention mechanism was devised to acquire distinct channel weights. Nevertheless, it is important to note that channel attention mechanisms are limited in their ability to just determine the varying weights of distinct channels. Consequently, it was not possible to execute a comprehensive semantic fusion. The combined feature map was made from the channel attention resultsusing a residual convolution method. As seen in Figure 3, the process involves merging two feature maps, namely $F_{background}$ and $F_{current}$, into a single feature map denoted as F_{input}. To achieve distinct channel weights, $F_{channel}$, which utilizes a global average descriptor, is employed. Each channel's feature map value in F_{aggc} can be calculated as:

$$F_{channel}\left(u_c\right) = V_c = \frac{1}{w\,.\,h}\sum_{i=1}^{w}\sum_{j=1}^{h} u_c(i,j) \qquad (5)$$

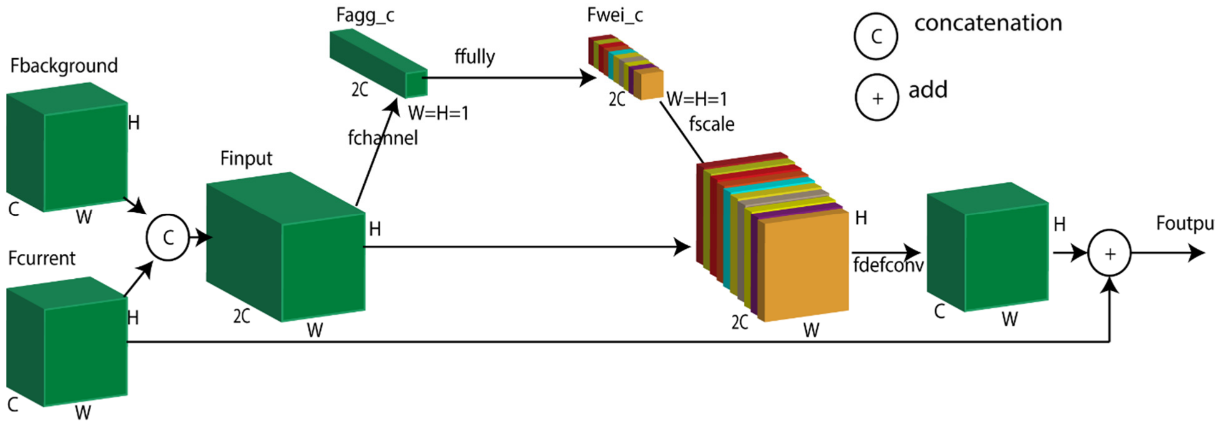

Fig. 3. Proposed fusion module

The variable V_c represents channel c value within F_{aggc}. The variable u_c represents channel c value in F_{input}. The variables W and H represent channel c width and height, respectively. In order to maximize the utilization of the collected data from F_{aggc}, F_{fully} has been developed to incorporate channel-wise weights. Producing a one-hot tensor involves the use of a sigmoid activation function and using a fully connected (FC) layer. This can be computed as:

$$F_{fully}\left(V_c\right) = w_c = \sigma\left(f_1(w_1)\right) = \sigma\left(f_1\left(\delta\left(f_2(V_c)\right)\right)\right) \quad (6)$$

Let $w_c \in \mathbb{R}^C$ represent channel c value in F_{weic}. The parameters f_1, δ, and f_2 pertain to a constriction in a neural network architecture consisting of two fully connected layers and a Rectified Linear Unit (ReLU) activation function. Let $w_1 \in \mathbb{R}^{C/r}$ be a feature map for dimensionality reduction, where r is the reduction ratio. Next, a sigmoid activation function σ is utilized to calculate the weight of each channel.

Following the process of channel attention, a deformable convolutional layer will be utilized on F_{output}, using a kernel size of 3. This operation aims to extract semantic information from the feature maps generated by channel attention. In conclusion, the residual link integrates textures with the current feature, hence maximizing the utilization of the present feature. The output values may be determined using (7):

$$F_{output} = F_{defconv} + F_{current} \qquad (7)$$

The current frame value is $F_{current}$, used in the computation. F_{output} is a tensor with the same dimension as $F_{current}$, ensuring compatibility with residual block dimension. However, varied layers encompass a multitude of distinct pieces of information. Specifically, shallow layers primarily capture local information, whereas contextual and semantic information from the surrounding environment are captured with deep layers, which are essential for detecting tiny objects. In order to aggregate multi-level correlation data, a cascade fusion module with three BISF fusion layers was designed. Traditionally, the first fusion module was used to join spatial information. Contextual features were combined utilizing the third fusion module.

To enhance computing efficiency during the inference phase, it is possible to pre-obtain all background feature maps without the need for repeated inference. This is feasible due to the fixed nature of the background image and the backbone parameters. Given this information, it can be concluded that the inclusion of a cascade module is the sole supplementary computational expense during the inference phase. This particular module offers a favorable equilibrium between cost and accuracy.

In order to accomplish the task of detecting small helmets on the roadside, we developed DPPNet by integrating an FPP module with the YOLOv5 network, which is a commonly employed framework for object detection. As seen in Figure 4, the inclusion of the GBS submodule inside the image preprocessing stage of the present image was undertaken to ameliorate the availability of superficial context and spatial information. The BISF submodule was included in the current frame underlying structure to enhance tiny objects detection accuracy by acquiring background contextual information. To upgrade the conciseness of the network design, a spatial attention layer from the GBS was integrated with the BISF fusion module. A cascade fusion architecture was employed to strike a balance between model complexity and detection accuracy. This architecture consisted of three fusion modules, utilized to capture more detailed regional textures from the current frame and broader global context from the background. Simultaneously, the PANet methodology was employed to facilitate tiny object detection using detectors that provide a wide receptive field. At the end of the network, three distinct detection heads were exploited to identify varying sizes objects. The ultimate output size for each detector head is determined as:

$$S_{output} = 3 \times w \times h \times (n_c + 5) \tag{8}$$

where S_{output} represents the ultimate outcome of the forecast. The numerical value 3 portrays the quantity of 3 anchors allocated for each grid. The width and height of the final feature maps denoted as w and h respectively, are specified in this study. Specifically, w_1 and h_1 are both equal to 52, w_2 and h_2 are both equal to 26, and w_3 and h_3 are both equal to 13.

The variable n_c responds to the total number of classes. The height, width, and the expected bounding box position, as well as the forecast confidence are represented by the number 5. The loss function is composed of three distinct components, namely the location loss, confidence loss, and category loss. The location loss is computed using the Generalized Intersection Over Union (GIOU) loss [29], whereas the confidence loss and category loss are calculated using the cross-entropy loss. The loss function is computed by (9):

$$Loss = L_{iou} + L_c - L_{cls} \tag{9}$$

$$L_{iou} = \lambda_{iou} \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij}^{obj} L_{giou}$$

$$L_c = \lambda_{cls} \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij}^{obj} \lambda_c \left(C_i - \widehat{C_\iota}\right)^2 + \lambda_{cls} \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij}^{noobj} \lambda_c (C_i - \widehat{C_\iota})^2$$

$$L_{cls} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij}^{obj} \sum_{c \epsilon classes} \lambda_c \times (\widehat{p_\iota}(c) \log\left(p_i(c)\right) + \left(1 - \widehat{p_\iota}(c)\right) \log\left(1 - p_i(c)\right))$$

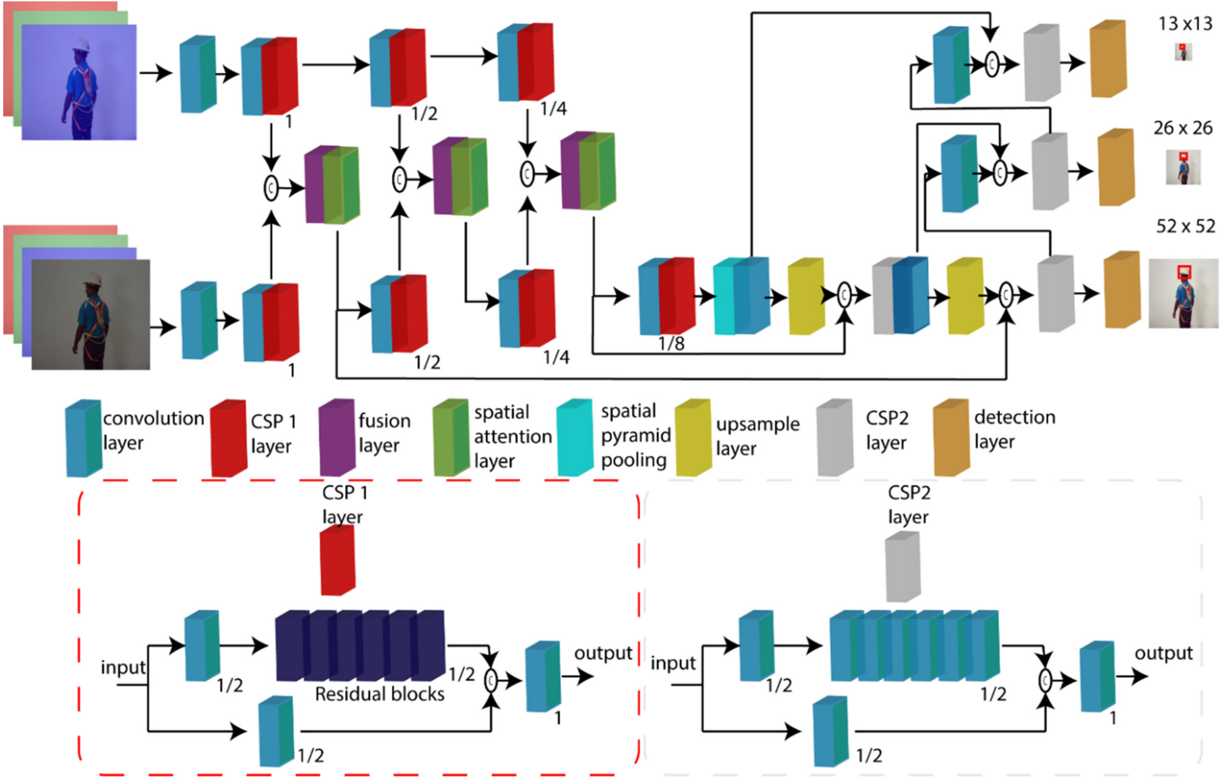

Fig. 4. Proposed DPPNet for helmet detection.

The abbreviations L_c, L_{cls}, and L_{iou}, characterize the concepts of category loss, confidence loss, and location loss. The variable S describes the final feature map dimensions, specifically 13, 26, and 52. The variable B represents each grid's allotted tally of anchor boxes. In the context of this article, B is defined as 3. The variable L_{iou} denotes the location loss, specifically for a single item. The indices i and j are used to indicate that the comparison is made between the predicted bounding box and the Ground Truth (GT) object. If the overlapping area between the predicted bounding box and one GT object is greater than the overlapping area with another GT object, the former is included into the location loss function. In a similar vein, the notation 1noobj reflects the scenario when the overlap is below a certain threshold and is hence excluded from the loss function. L_c encompasses both true prediction and false prediction. The variable C_i portrays the confidence score for the predicted box, whereas (C_i)ˆ symbolizes the confidence of the ground truth. The value of (C_i)ˆ is equivalent to 1. The acronym L_{cls} refers to the concept of category loss. In this context, (p_i)ˆ(c) responds to the true category of the GT item, whereas p_i (c) represents the category assigned to the predicted object. The normalizer parameters in YOLOV5, denoted as λ_{iou}, λ_{cls}, and λ_c, are provided.

IV. EXPERIMENT AND RESULT

A. Data and Environment Setup

Experiments including training, validation, and testing have been carried out using the SHWD dataset [20]. It comprises a collection of 7581 images. Among those images, there were a total of 9,044 instances of human safety helmet-wearing objects and 111,514 instances of regular head objects. The aforementioned images were obtained from authentic building sites and included a diverse range of perspectives. The provided dataset is appropriate for conducting both single-class (Helmet alone) and multi-class (Helmet and No Helmet) detections.

One of the most prevalent challenges encountered in deep learning-based networks is the presence of class imbalance, which gives rise to many issues, such as suboptimal detection outcomes. In order to mitigate this issue, a data augmentation strategy is used. Multiple data augmentation techniques were implemented, including horizontal flipping, vertical flipping, picture translation, random cropping, brightness adjustment, and random translation. Pytorch was employed for network development with the support of CUDA acceleration. The NVIDIA GTX 960 GPU was engaged for both training and inferencing the networks.

Three widely used indices, mAP, $AP_{50,}$ and $AP_{50\text{-}95}$, were derived from the MS COCO dataset and utilized to assess the proposed DPPNet performance. When the IOU threshold between the GT bounding box and the prediction bounding box is set to 0.5, as it is in AP_{50}, the average n is represented. The AP is calculated as an average of the mAPs obtained using the following IOU cutoffs: 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95. To optimize the parameters of our DPPNet, we employed Stochastic Gradient Descent (SGD) with a momentum of 0.9, a weight decay of 2e – 4, and a batch size of 16. The models underwent training until the training loss reached convergence during a span of 200 epochs. This was achieved by utilizing an initial learning rate of 0.01, which was then reduced to 1e–4 via cosine annealing.

B. Evaluation and Discussion

According to the data shown in Table I, the proposed DPPNet demonstrated a significant improvement in the identification of tiny helmets, achieving a performance of 93.7 for mAP. These results surpassed the performance of the original YOLOv5s model [31] by 9.6%. The performance of DPPNet surpasses that of the YOLOX [32] in terms of speed and detection accuracy. In comparison to the state-of-the-art object detection network known as DINO [30], DPPNet demonstrates a small reduction in AP (0.9%) specifically for tiny items. However, it presents a noteworthy improvement regarding speed, being around 10 times quicker.

TABLE I. OBTAINED RESULTS OF THE DPPNET EVALUATION ON THE SHWD DATASET IN COMPARISON TO STATE-OF-THE-ART OBJECT DETECTION NETWORKS

Model	Backbone	Parameters	FPS	$AP_{50\text{-}95}$(%)	mAP(%)	AP_{50}(%)
DINO [30]	ResNet 50	46606102	3	71.3	94.6	98.2
Yolo v5s [31]	CSPNets	7035811	38	61.5	84.1	95.6
Yolo Xs [32]	CSPDarkNet	8939617	26	58.3	72.5	89.6
DPPNet	CSPNets	8229449	29	69.8	93.7	97.4

DPPNet has the capability to enhance detection outcomes despite its uncomplicated structure and low computational requirements. The experimental results shown in Figure 5 provide evidence that FPPNet is capable of detecting tiny helmets in various lighting settings, including both daytime and hard evening circumstances. Additionally, it demonstrates applicability not just in long-range detection scenarios but also in the identification of badly obscured objects. This finding suggests that the proposed model can be easily implemented in challenging situations. Also, the inference was accelerated by precompiling the feature mappings of the background frame so that it does not have to be inferred from scratch every time, resulting in greater computing efficiency during the inference. The BISF module consists of the attention mechanism and three convolutional layers and performs extra work during the inference. The GBS module is part of the data preprocessing step and so has little impact on processing speed. Increases in computational complexity and inference time are offset by the proposed modules while offering superior accuracy.

In conclusion, the findings shown in Table I illustrate that DPPNet has strong generalization capabilities across different object detectors. Moreover, it yields significant advancements in tiny object detection and Average Precision (AP). Specifically, it has shown greater improvements for lightweight networks by offering a substantially higher amount of spatial and contextual information that is essential for such networks.

C. Ablation Study

In this part, we will detail the ablation experiments that we carried out in order to test the proposed modules' effectiveness.

The DPPNet parameters and settings were used for the hyper-parameters. The efficacy of the fusion module, background subtraction procedure, and the GBS module was tested via a series of experiments.

The GBS component entails both background elimination and spatial awareness. We included the GBS module into the base YOLOv5s to see how well it handles background subtraction. This was achieved by involving a gray background subtraction channel in the input frames and using a three-spatial attention module in the neural network's central processing unit. In Table II, Row 1 depticts the original YOLOv5s, Row 2 displays the input data with an extra background image channel but no spatial attention, and Row 3 illustrates the full GBS module application. As seen by the data presented in Table II, the incorporation of the GBS submodule resulted in a notable improvement in the performance of tiny object identification. The gray background subtraction channel demonstrated a 7.5% increase in mAP while maintaining a modest level of model complexity and computational expense.

TABLE II. IMPACT OF THE GBS MODULE ON THE PERFORMANCE

Model	Parameters	AP_{50-95}(%)	mAP(%)	AP_{50}(%)
Yolo v5s	7045823	53.7	70.9	89.1
+ background image	7046974	55.8	76.2	91.5
+ GBS module	7047238	59.3	78.4	93.2

Additional spatial information, such as shadows and regional textures, may be included through the background subtraction channel. Simultaneously, the inclusion of spatial attention yielded a performance improvement of more than 2.1% when compared to the model that did not include spatial attention. This suggests that spatial attention has the ability to efficiently retrieve feature information within the spatial domain of tiny objects.

The importance of the created background image was assessed by contrasting the outcomes achieved when using a basic background image against a normalized one. Considering the same network configuration, the findings are shown in Table III. It is worth noting that the performance of a normalized background image demonstrates a considerable level of robustness when compared to a simple background image. The results consistently demonstrated a 0.5% improvement when comparing raw backgrounds across all studies. This indicates that using a background image generation approach enhances the availability of semantic information and spatial features spatially when using normalized images.

The fusion module was designed to combine feature maps from two different sources (the background and another backbone). We decided to replace the fusion module with two feature maps to see what effect it would have on the final product. Additionally, an experiment was conducted to assess how the deformable convolution layer influences the fusion component. According to the data shown in Table 4, it is evident that the fusion module exhibited superior performance compared to YOLOv5s, with a margin of 3.4% in terms of mAP. Additionally, the inclusion of the deformable convolution layer resulted in a 2% improvement in performance compared to the module without it. The findings show that the fusion module successfully acquired contextual information from the background while also preserving regional features from the present image.

TABLE III. IMPACT OF BACKGROUND IMAGE NORMALIZATION ON THE PERFORMANCE OF THE GBS MODULE

Model	Image	AP_{50-95}(%)	mAP(%)	AP_{50}(%)
Yolo v5s	N/A	53.7	70.9	89.1
+ GBS module	Simple	58.9	77.9	92.8
+ GBS module	Normalized	59.3	78.4	93.2

TABLE IV. IMPACT OF THE DEFORMABLE CONVOLUTION ON THE PERFORMANCE OF THE FUSION MODULE

Model	Fusion	AP_{50-95}(%)	mAP(%)	AP_{50}(%)
Yolo v5s	N/A	53.7	70.9	89.1
+ BISF module	normal	55.3	73.8	91.6
+ BISF module	Deformable convolution	56.2	74.3	92.7

V. CONCLUSIONS

This study examines the major importance of ensuring safety in engineering practices within real-world contexts. Workers may guarantee engineering safety by wearing safety helmets in accordance with the prescribed rules. In order to achieve this objective, we have put forward a DPPNet model that consists of two modules: a GBS submodule and a BISF submodule. By subtracting the previous frame from the present one, we may extract the GBS module. The resulting one-dimensional grayscale pictures are integrated into the existing scene. The networks attention was narrowed down on certain locations using spatial attention. The BISF component separates the current and background frames and creates feature maps for each. The maps are then fused together utilizing a cascade attention module.

In accordance with the proposed modules, a novel architectural framework known as DPPNet was introduced with the aim of detecting safety helmets. The experimental findings indicate that DPPNet demonstrated favorable outcomes of 69.8%, 93.7%, and 97.4% in relation to $AP_{0.5}$, mAP, and $AP_{0.5:0.95}$, respectively. Additionally, the proposed model achieved a processing speed of 29 frames per second while getting a low computation complexity, which makes it suitable for real-world applications based on embedded systems. However, the proposed method presents a limitation regarding the generation of the background image that is made manually. This process limits the deployment of the suggested method in a large-scale manner. In future works, the point cloud technique may be integrated to make denser expressions and enhance the detection performance of small objects at complex backgrounds.

ACKNOWLEDGMENT

The authors gratefully acknowledge the approval and the support of this research study by the grant no. ENGA-2023-12-

2229 from the Deanship of Scientific Research at Northern Border University, Arar, KSA.

REFERENCES

[1] K. Dahiya, D. Singh, and C. K. Mohan, "Automatic detection of bike-riders without helmet using surveillance videos in real-time," in *International Joint Conference on Neural Networks*, Vancouver, BC, Canada, Jul. 2016, pp. 3046–3051, https://doi.org/10.1109/IJCNN.2016.7727586.

[2] "CDC Works 24/7," *Centers for Disease Control and Prevention*, Dec. 06, 2023. https://www.cdc.gov/index.htm.

[3] M. Aftf, R. Ayachi, Y. Said, E. Pissaloux, and M. Atri, "Indoor Object C1assification for Autonomous Navigation Assistance Based on Deep CNN Model," in *International Symposium on Measurements & Networking*, Catania, Italy, Jul. 2019, pp. 1–4, https://doi.org/10.1109/IWMN.2019.8805042.

[4] M. Afif, R. ayachi, Y. Said, E. Pissaloux, and M. Atri, "Recognizing signs and doors for Indoor Wayfinding for Blind and Visually Impaired Persons," in *5th International Conference on Advanced Technologies for Signal and Image Processing*, Sousse, Tunisia, Sep. 2020, pp. 1–4, https://doi.org/10.1109/ATSIP49331.2020.9231933.

[5] M. Afif, R. Ayachi, S. Yahia, and M. Atri, "COVID-19 Disease Detection Using Deep Learning Techniques in CT Scan Images," in *Advanced AI and Internet of Health Things for Combating Pandemics*, M. Lahby, V. Pilloni, J. S. Banerjee, and M. Mahmud, Eds. New York, NY, USA: Springer, 2023, pp. 177–191.

[6] R. Ayachi, M. Afif, Y. Said, and A. B. Abdelaali, "pedestrian detection for advanced driving assisting system: a transfer learning approach," in *5th International Conference on Advanced Technologies for Signal and Image Processing*, Sousse, Tunisia, Sep. 2020, pp. 1–5, https://doi.org/10.1109/ATSIP49331.2020.9231559.

[7] R. Ayachi, M. Afif, Y. Said, and A. B. Abdelali, "Real-Time Implementation of Traffic Signs Detection and Identification Application on Graphics Processing Units," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 35, no. 7, Jun. 2021, Art. no. 2150024, https://doi.org/10.1142/S0218001421500245.

[8] R. Ayachi, M. Afif, Y. Said, and A. B. Abdelali, "An Embedded Implementation of a Traffic Light Detection System for Advanced Driver Assistance Systems," in *Industrial Transformation*, Boca Raton, FL, USA: CRC Press, 2022, pp. 237–250.

[9] Y. Said, M. Barr, and H. E. Ahmed, "Design of a Face Recognition System based on Convolutional Neural Network (CNN)," *Engineering, Technology & Applied Science Research*, vol. 10, no. 3, pp. 5608–5612, Jun. 2020, https://doi.org/10.48084/etasr.3490.

[10] D. Contractorr, K. Pathak, S. Sharma, S. Bhagat, and T. Sharma, "Cascade Classifier based Helmet Detection using OpenCV in Image Processing," in *National Conference on Recent Trends in Computer and Communication Technology*, May. 2016, pp. 195–200.

[11] L. Shine and C. V. Jiji, "Automated detection of helmet on motorcyclists from traffic surveillance videos: a comparative analysis using hand-crafted features and CNN," *Multimedia Tools and Applications*, vol. 79, no. 19, pp. 14179–14199, May 2020, https://doi.org/10.1007/s11042-020-08627-w.

[12] V. H. Duong, Q. H. Tran, H. S. P. Nguyen, D. Q. Nguyen, and T. C. Nguyen, "Helmet Rule Violation Detection for Motorcyclists Using a Custom Tracking Framework and Advanced Object Detection Techniques," in *Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, Jun. 2023, pp. 5381–5390.

[13] D. N.-N. Tran *et al.*, "Robust Automatic Motorcycle Helmet Violation Detection for an Intelligent Transportation System," in *Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, Jun. 2023, pp. 5341–5349.

[14] L. Wang *et al.*, "Investigation Into Recognition Algorithm of Helmet Violation Based on YOLOv5-CBAM-DCN," *IEEE Access*, vol. 10, pp. 60622–60632, 2022, https://doi.org/10.1109/ACCESS.2022.3180796.

[15] N. Kharade, S. Mane, J. Raghav, N. Alle, A. Khatavkar, and G. Navale, "Deep-learning based helmet violation detection system," in *International Conference on Artificial Intelligence and Machine Vision*, Gandhinagar, India, Sep. 2021, pp. 1–4, https://doi.org/10.1109/AIMV53313.2021.9670937.

[16] P. Sridhar, M. Jagadeeswari, S. H. Sri, N. Akshaya, and J. Haritha, "Helmet Violation Detection using YOLO v2 Deep Learning Framework," in *6th International Conference on Trends in Electronics and Informatics*, Tirunelveli, India, Apr. 2022, pp. 1207–1212, https://doi.org/10.1109/ICOEI53556.2022.9776661.

[17] G. Agorku *et al.*, "Real-Time Helmet Violation Detection Using YOLOv5 and Ensemble Learning." arXiv, Apr. 14, 2023, https://doi.org/10.48550/arXiv.2304.09246.

[18] A. Aboah, B. Wang, U. Bagci, and Y. Adu-Gyamfi, "Real-Time Multi-Class Helmet Violation Detection Using Few-Shot Data Sampling Technique and YOLOv8," in *Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, Jun. 2023, pp. 5350–5358.

[19] A. Goyal, D. Agarwal, A. Subramanian, C. V. Jawahar, R. K. Sarvadevabhatla, and R. Saluja, "Detecting, Tracking and Counting Motorcycle Rider Traffic Violations on Unconstrained Roads," in *Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp. 4303–4312.

[20] njvisionpower, "Safety-Helmet-Wearing-Dataset." github.com, Dec. 08, 2023, [Online]. Available: https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset.

[21] S. T. Mpinda Ataky, J. de Matos, A. de S. Britto, L. E. S. Oliveira, and A. L. Koerich, "Data Augmentation for Histopathological Images Based on Gaussian-Laplacian Pyramid Blending," in *International Joint Conference on Neural Networks*, Glasgow, UK, Jul. 2020, pp. 1–8, https://doi.org/10.1109/IJCNN48605.2020.9206855.

[22] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-Outside Net: Detecting Objects in Context With Skip Pooling and Recurrent Neural Networks," in *Conference on Computer Vision and Pattern Recognition*, Jun. 2016, pp. 2874–2883.

[23] X. Tang, D. K. Du, Z. He, and J. Liu, "PyramidBox: A Context-assisted Single Shot Face Detector," in *European Conference on Computer Vision*, Munich, Germany, Sep. 2018, pp. 797–813.

[24] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *32nd International Conference on Machine Learning*, Lile, France, Jul. 2015, pp. 448–456.

[25] *BT.601-5 - Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios*. ITU, 2011.

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *15th European Conference on Computer Vision*, Munich, Germany, Sep. 2018, pp. 3–19.

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[28] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in *Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, Jun. 2020, pp. 11534–11542.

[29] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression," in *Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, Jun. 2019, pp. 658–666.

[30] H. Zhang *et al.*, "DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection." arXiv, Jul. 11, 2022, https://doi.org/10.48550/arXiv.2203.03605.

[31] M. Karthi, V. Muthulakshmi, R. Priscilla, P. Praveen, and K. Vanisri, "Evolution of YOLO-V5 Algorithm for Object Detection: Automated Detection of Library Books and Performace validation of Dataset," in *International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems*, Chennai, India, Sep. 2021, pp. 1–6, https://doi.org/10.1109/ICSES52305.2021.9633834.

[32] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021." arXiv, Aug. 05, 2021, https://doi.org/10.48550/arXiv.2107.08430.

[33] M. Touahmia, "Identification of Risk Factors Influencing Road Traffic Accidents," *Engineering, Technology & Applied Science Research*, vol. 8, no. 1, pp. 2417–2421, Feb. 2018, https://doi.org/10.48084/etasr.1615.

[34] F. Siddiqui, M. A. Akhund, A. H. Memon, A. R. Khoso, and H. U. Imad, "Health and Safety Issues of Industry Workmen," *Engineering, Technology & Applied Science Research*, vol. 8, no. 4, pp. 3184–3188, Aug. 2018, https://doi.org/10.48084/etasr.2138.

[35] X. Chen and Q. Xie, "Safety Helmet-Wearing Detection System for Manufacturing Workshop Based on Improved YOLOv7," *Journal of Sensors*, vol. 2023, May 2023, Art. no. e7230463, https://doi.org/10.1155/2023/7230463.

[36] J. Li, Y. Li, J. F. Villaverde, X. Chen, and X. Zhang, "A safety wearing helmet detection method using deep leaning approach," *Journal of Optics*, Jul. 2023, https://doi.org/10.1007/s12596-023-01282-y.

[37] Y. Qian and B. Wang, "A new method for safety helmet detection based on convolutional neural network," *PLOS ONE*, vol. 18, no. 10, Sep. 2023, Art. no. e0292970, https://doi.org/10.1371/journal.pone.0292970.