IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 May 2023, accepted 9 July 2023, date of publication 17 July 2023, date of current version 25 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3296309

RESEARCH ARTICLE

A Swin Transformer-Based Approach for Motorcycle Helmet Detection

AYYOUB BOUHAYANE[1,2], ZAKARIA CHAROUH[3], (Senior Member, IEEE), MOUNIR GHOGHO[2,4], (Fellow, IEEE), AND ZOUHAIR GUENNOUN[1], (Senior Member, IEEE)

[1]ERSC Team, Mohammadia Engineering School, Mohammed V University, Rabat 10090, Morocco
[2]TICLab, College of Engineering and Architecture, International University of Rabat, Rabat 11103, Morocco
[3]Majal Berkane, Berkane 63300, Morocco
[4]Faculty of Engineering, School of EEE, University of Leeds, LS2 9JT Leeds, U.K.

Corresponding author: Ayyoub Bouhayane (ayyoub.bouhayane@uir.ac.ma)

This work was supported by the National Road Safety Agency (NARSA) and the Moroccan Ministry of Equipment, Transport, Logistics, and Water, through the National Center for Scientific and Technical Research (CNRST).

ABSTRACT Video surveillance-based automated detection of helmet use among motorcyclists has the potential to improve road safety by aiding in the implementation of enforcement initiatives. Despite that, the current detection approaches have many limitations. For instance, they are unable to detect multiple passengers or to function effectively in complex conditions. In this paper, we address the challenging problem of automated monitoring of helmet use using computer vision and machine learning. We propose a method based on deep neural network models known as transformers. We apply the base version of the Swin transformer as a backbone for feature extraction, and then combine a Feature Pyramid Network (FPN) neck with the Cascade Region-based Convolutional Neural Networks (RCNN) framework for final detection. The effectiveness of our proposed method is demonstrated through extensive experiments and is compared to existing approaches. Our method achieves a mean Average Precision (mAP) of 30.4, thus outperforming state-of-the-art detection methods.

INDEX TERMS Deep learning, helmet detection, intelligent transport systems, motorcycle safety, road safety, transformers.

I. INTRODUCTION

According to the World Health Organization, more than 1.3 million people die in road traffic accidents each year [1]. The leading cause of road accidents is human error or bad driving behavior. A large number of these accidents involve motorcycles. These accidents are a leading cause of fatal injuries in underdeveloped countries [1]. Helmets can reduce the risk of head injuries and fatalities among motorcycle riders. Helmet enforcement is therefore an efficient strategy to decrease fatalities and mortality. Many countries rely on traffic police officers to enforce compliance with traffic laws by directly observing drivers on a daily basis. It is, however, generally expensive and logistically challenging to deploy many police officers across the country in order to ensure universal and strict enforcement.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Xiao.

Advances in computer vision (CV) have enabled the development of video-based efficient methods to monitor driving behavior and other application in intelligent transportion systems (ITS) [2], [3]. For example, a CV method for monitoring driving behavior at a road intersection was proposed in [4]. Monitoring driver distraction including the use of mobile phone while driving was investigated in [5]. Automating the monitoring of driving behavior through CV techniques can facilitate law enforcement, thereby improving driving conditions. Considering that motorcycles are the most common mode of motorized transportation in many developing nations [6], an automated CV system is necessary to analyze helmet wear behavior among motorcycle riders and passengers. The system can be utilized to target enforcement programs and implement effective education programs.

Recent advancements in deep learning have led to the development of powerful computer vision models, such as Convolutional Neural Networks (CNNs), that have

shown tremendous success in various computer vision tasks [7], [8], [9], [10], [11]. One area where CNNs have been extensively applied is road safety, particularly in detecting pedestrians and vehicles using video surveillance cameras installed on roads [12], [13]. Although CNNs have been the dominate approach in computer vision, Transformer architectures, inspired by Natural Language Processing (NLP) achievements [14], have demonstrated higher accuracy in many computer vision tasks [15], [16] including image classification, object detection, and semantic segmentation.

Over the past few years, transformers have been used for a range of vision-related tasks, but they have not yet drawn much attention in the context of ITS. The problem addressed here, which is helmet detection, has been investigated in the literature using only CNNs. To the best of our knowledge, no research has considered the transformers architecture for helmet detection.

One specific problem in road safety is the detection of helmet usage by motorcyclists. Previous research has addressed this problem using CNNs, primarily with two different approaches: binary classification and multi-class classification. In binary classification, the motorcycle and its occupants are detected as a single object, and the top half of the object is cropped to classify the head region as either wearing or not wearing a helmet [17]. Multi-class classification also involves detecting the motorcycle and its occupants as the entire object and classifying it into a specific class, such as the driver and passenger are both wearing a helmet which is one example of the classes described in [18]. While these methods have shown some success, no research has considered the use of Transformer-based models for helmet detection in ITS.

In this paper, we propose a new system for helmet use detection in ITS using a combination of Transformers for vision and the Cascade RCNN framework. Specifically, we use the Swin Transformer as a backbone for feature extraction, which has shown state-of-the-art performance in computer vision tasks [16]. We also incorporate a Feature Pyramid Network (FPN) neck with the Cascade RCNN framework for object detection to improve accuracy. To evaluate our proposed method, we use a publicly available dataset, the Helmet Dataset [19], and compare our results with several CNN-based methods.

Our proposed method outperforms the state-of-the-art CNN-based approaches with a mean average precision (mAP) of 30.4. Our contributions include using a state-of-the-art Transformer-based model for feature extraction, incorporating the Cascade RCNN framework with FPN neck for object detection, and conducting a comparative analysis with other existing models to determine the efficacy of our proposed approach. This study demonstrates the potential of Transformer-based models for solving computer vision problems in ITS, specifically in the context of helmet use detection, which has important implications for road safety.

The rest of the article is structured as follows: First, we provide a literature review of existing approaches in automated helmet use detection. Next, we detail our proposed method. We then present the experimental section, where we describe the dataset used, experimental settings, evaluation metrics, and evaluate the performance of our method compared to other models.

II. RELATED WORK

The scientific community has grown more interested in employing computer vision techniques for automating tasks associated with motorcycle helmet detection. The majority of these studies use one of two methods: classical or statistical methods, or deep learning methods.

A. CLASSICAL METHODS

Classical approaches are based on hand-crafted features that are extracted manually from the image using a feature descriptor like Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and many others. These features are then fed into a classifier that classifies them into one of the categories studied. These methods are performed in three different stages. The first stage starts by applying background subtraction for detecting moving vehicles. The second stage consists of training a classifier on features extracted from the foreground image to classify motorcycles from other moving objects. Finally, the third stage is for the helmet detection part, where the rider's head is defined as the Region of Interest (ROI) to classify the helmet use.

The authors in [20] used an improved adaptive Gaussian Mixture Model (GMM) for the background subtraction. They employ the wavelet transform (WT) to extract vehicles' features which are fed to a Random Forest algorithm for motorcycle classification. Next, they apply Circular Hough Transform (CHT) and HOG descriptors for extracting helmet features from the head region, where the ROI is determined by cropping the top one-fifth of the image where the head region is situated. Finally, they use a Multi-Layer Perceptron (MLP) for helmet classification. In addition, they compared other descriptors for feature extraction like Local binary patterns (LBP), Speeded Up Robust Features (SURF) and different classifiers such as Support Vector Machines (SVM), Naive Bayes, and MLP. They adopted a private dataset that is collected from recorded videos on public roads recording the motorcycle's rear view. In this method, images are captured primarily from a top view angle that restricts the field of view of a typical surveillance camera.

In [21], the authors also make use of the HOG descriptor for feature extraction after applying a background subtraction using GMM. The resulting foreground image is used to train a SVM classifier for classifying both motorcycles and helmet features. However, the methods for extracting the head region features and classification were not determined in the paper. They constructed a private dataset based on video clips extracted from a CCTV, with simple scenarios and non-overlapping vehicles.

In [22], the authors performed the task first by applying GMM for the background subtraction, then extracting vehicle features in three different ways, by using HOG, Local binary

patterns (LBP), and by combining both features extracted using these descriptors. These features are then classified into motorcycle and non-motorcycle using a SVM classifier. For the helmet detection part, they adopted two approaches and compared them. The first one is by building a classifier based on combined features extracted by HOG, LBP, and Harlick features. They compared several classifiers which are Naive Bayes, SVM, Random Forest, and Logistic Regression. And the second one is by training a custom CNN architecture.

Although these methods produced satisfactory results, they have many drawbacks: 1) they require calibration for each new scene, as most of the datasets were captured at a restricted distance and view angle; 2) they are prone to failing in many different real-world scenarios since the features are designed and extracted manually; they cannot handle the changes in illumination, and heavy traffic congestion hinders their ability to cope with occlusions and overlapping objects; 3) they do not take into account multiple motorcycle passengers. They are limited to detecting the rider's helmet only.

B. DEEP LEARNING METHODS

Computer vision has undergone a revolution due to deep learning (DL) techniques. In pattern recognition, the techniques have proven robust in classifying images despite various levels of distortion and transformation (noise, scale, rotation, displacement, illumination variance). When it comes to object recognition, feature representations derived from DL often outperform popular features such as LBP, SURF, and HOG.

The studies in [23], [24], and [25] adopted GMM for background subtraction to get moving vehicles and a CNN to classify motorcycles from other vehicles. They apply a second CNN for the helmet classification. In [25], the authors used the whole blob extracted from the previous step and consider that helmets must be worn by all motorcycle passengers to be classified as wearing a helmet; otherwise, if only one of the passengers is not wearing it, it is classified as not wearing a helmet. Nonetheless, [23] and [24] employed a technique of cropping the top one-fourth of the blob detected for the classification, as they suppose it is the ROI where the head of the rider is located. Their method takes the motorcycle's rider only into account for the helmet classification and does not consider multiple pillions.

In [26] and [27], the authors used an object detector to detect motorcycles and persons separately. To identify motorcycle riders and passengers from pedestrians, the method used in [26] combine the overlapped motorcycle and person bounding boxes. The resulting blob is then used to classify the helmet-wearing using an InceptionV3 model. Reference [27] calculates the euclidean distance between the centers of the overlapping bounding boxes to count the number of passengers using a single motorcycle. Then, they crop the top one-fourth of the image for the helmet classification. However, it is not possible to handle other overlapping motorcycles and persons using these methods.

In [17], the authors built a two-modules helmet violation detection system. The first module is for object detection and the second module is used to classify the detected object. They apply YOLOv3 to detect a motorcycle and its users jointly. The YOLOv3 algorithm was fine-tuned to specifically detect instances of the motorcycle-person class, which involves identifying both the motorcycle and its riders as a single object with a shared bounding box. Then, they crop the top one-half of the resulting detection to classify the helmet usage using GoogLeNet model. In their work, multiple passengers riding a motorcycle with only one of them not wearing a helmet is considered a violation. Furthermore, the authors did not provide any information regarding the configuration of both models, including details about the hyperparameters used during training.

The work in [28] implemented a YOLOv5 model for the detection part, where they also combine motorcycles and their users as a single object. They consider the helmet use problem, unlike other works, as a detection task. As in [17], they cut the top one-half of the motorcycle patches detected from the previous stage and use it as an input to train another YOLOv5 for helmet detection.

Other works like [29], [30], and [31] built a system to extract the license plate of motocycles whose riders do not wear a helmet. In [31], they adopted a hybrid approach by using a SVM to classify features extracted by HOG descriptor and a CNN for the helmet classification based on the head region. Despite that, they did not mention how to detect multiple pillions of a single motorcycle. A YOLOv3 was fine-tuned in [29] on 5 classes separately: Person, motorcycle, helmet, no helmet and license plate. Nevertheless, it is not clear how their method combines those detected objects when there are occlusions and overlapping between pedestrians and motorcyclists. The study in [30] also used a YOLOv3 algorithm for detecting a motorcycle and its passengers as a single object, the head region, and the license plate. The head region is then classified using a ResNet model [11]. However, the authors did not describe how head regions of motorcyclists only can be detected and processed, while ignoring any other head regions like those associated with pedestrians. the training process based on the head region was not described, which poses a problem if there are pedestrians.

In [32], the authors built a system to track motorcycles and detect their users' helmet use. They followed previous approaches to detecting motorcycles along with their users as a first step, then, unlike other methods, for the helmet classification part, these motorcycle patches are classified as a whole instead of considering only the top of the image. They apply a multi-task learning approach using a siamese network, where the same CNN model is duplicated with shared weights, for the helmet use classification. The process is done by training the siamese network on two patches, either similar or different ones, extracted from the previous step.

In contrast to other techniques, the authors in [33] proposes a different approach for preprocessing their data to detect and track motorcycle riders on unconstrained roads.

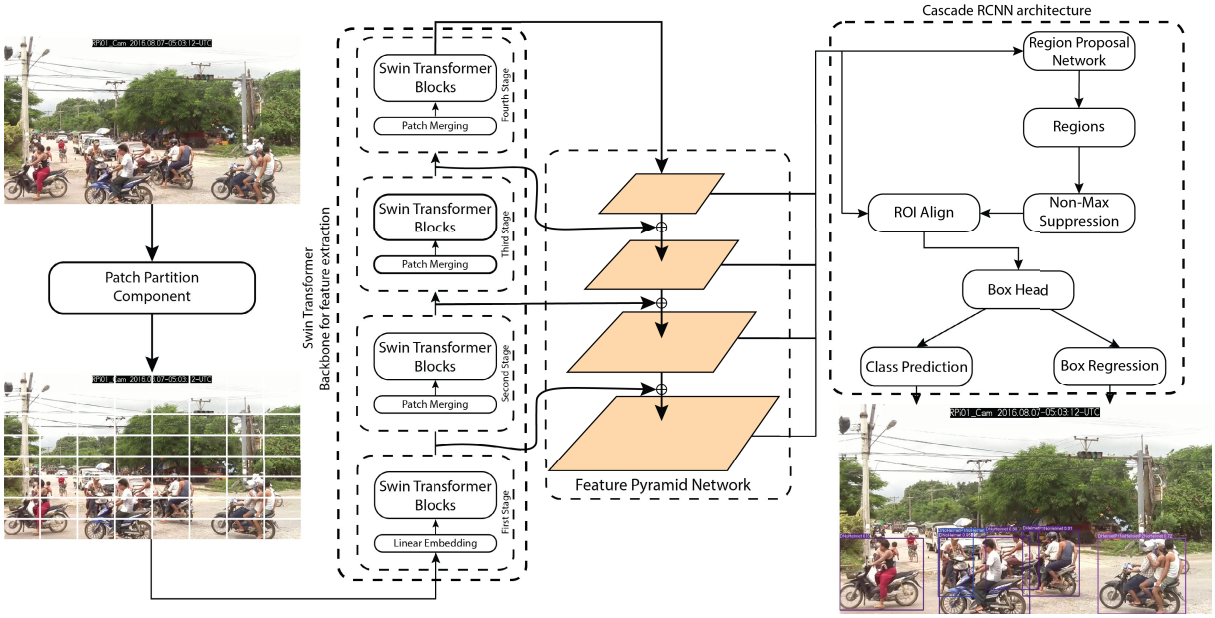

FIGURE 1. Overview of the proposed Helmet use detection method. We first split the input image into non-overlapping patches, which are treated as tokens. These patches are processed using the Swin Transformer backbone to extract features across multiple stages. At each stage, a hierarchical representation is generated, and these representations are then used in a Feature Pyramid Network and a Cascade RCNN framework for detection.

They use a trapezium bounding box instead of a rectangular bounding box to better detect and associate multiple riders on a motorcycle. For this purpose, they train a regressor to transform the rectangular bounding box into a trapezium one. They also employ a curriculum learning-based object detector to improve detection in challenging scenarios such as occlusions. First, they fine-tune a YOLOv4 [34] model on detecting motorcycles, and then they further fine-tune the resulting model to detect both motorcycles and riders. Additionally, they train another YOLOv4 model to detect helmet use by motorcycle riders, and use DeepSort [35] algorithm for object tracking.

Unlike other methods, which treat the problem in multiple stages, [18] adopted one single object detector for both motorcycle and helmet use detection. They fine-tuned RetinaNet [36], a one-stage object detection model, on 36 classes based on the Helmet dataset [19]. Their approach gave better results compared to a human observer and the riders' positions and number on the motorcycle could be accurately determined. However, for motorcycles with a high number of riders or motorcycles with unusual rider compositions, the algorithm proved less accurate. Since each unique composition of riders and their helmet use is considered as a class, the dataset has a much higher number of examples of motorcycles with one or two riders than most other classes, which leads to poor results for detecting the under-represented classes.

Although the majority of these methods showed promising results, they still need lots of improvements. Classical methods have many drawbacks as they require different stages for feature extraction and helmet use detection. Additionally, they are not accurate enough when used with complex scenarios where many occlusions or illumination changes can occur. Moreover, the existing deep learning methods used for helmet use detection are all based on CNNs. To the best of our knowledge, no work has been proposed for helmet use detection based on transformers models for vision. These models showed superior performance over CNNs in many CV tasks. This motivates our work to employ new methods based on vision transformers for helmet use detection.

III. METHOD

In this section, we provide a detailed description of our proposed method. It consists of three main modules, including preprocessing, feature extraction, and helmet use detection. Figure 1 depicts the flowchart of the proposed method

A. PREPROCESSING MODULE

Throughout the training phase, we utilized common preprocessing and data augmentation techniques, including random resizing, random resizing combined cropping, and random flipping horizontally and vertically. In the testing stage, we only apply a fixed resizing to images. In both stages, we applied pixel normalization by dividing each pixel intensity by 255.

B. FEATURE EXTRACTION MODULE

Feature extraction is an essential step that serves to generate a representation of the input image. This representation is then used by the next module to generate objects' localizations and classify them. Hence, this module represents the core component of the framework.

1) SWIN TRANSFORMER

The Swin transformer architecture takes a sequence of tokens as input. These tokens are generated by applying a patch partition layer on the input image to split it into N patches. The hidden layers are composed of several blocks, each of which

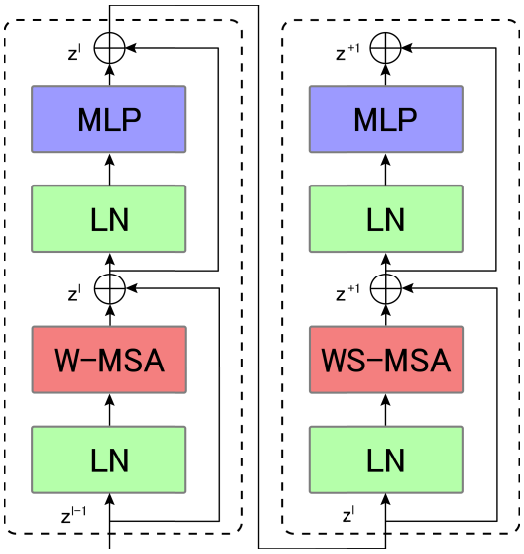

FIGURE 2. Two consecutive Swin Transformer blocks. The first consists of a regular window multi-head self-attention module (W-MSA), while the latter uses a shifted window configuration (WS-MSA).

is consisted of a multi-head self-attention module (MSA) as illustrated in figure 2. This module applies an attention function over a set of query Q, key K, and value V vectors where it maps the query to a set of key-value pairs, to an output. The process is done by performing a dot product of the query vector with all the key vectors. A softmax function is used then to scale the inner products and normalize them into k weights as given in Eq.1

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^T/\sqrt{d_k})V \quad (1)$$

where d_k is the key dimension and normalization is performed by dividing by $\sqrt{d_k}$. Swin Transformer applies instead, a window-based Multi-head self-attention module (W-MSA) and a shifted window multi-head self-attention module (SW-MSA). The W-MSA module calculates attention locally, where it applies self-attention on non-overlapping windows. To perform the cross-window self-attention calculations, the SW-MSA module performs the same computations as in the W-MSA module after shifting the windows.

The following equation Eq.2 is used to compute two consecutive blocks:

$$\begin{aligned} \hat{z}^l &= \text{W-MSA}(\text{LN}(z^{l-1})) + z^{l-1}, \\ z^l &= \text{MLP}(\text{LN}(\hat{z}^l)) + \hat{z}^l, \\ \hat{z}^{l+1} &= \text{SW-MSA}(LN(z^l)) + z^l, \\ z^{l+1} &= \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1} \end{aligned} \quad (2)$$

where $\hat{z}^l, z^l, \hat{z}^{l+1}$ and z^{l-1} denote the output features of the W-MSA, MLP, SW-MSA and MLP modules respectively, and LN denotes Layer Normalization.

2) FEATURE EXTRACTION PROCESS

In our work, we use the Swin Transformer for the feature extraction process. First, the input image is split into patches of size 4×4, and a linear projection layer is applied to flatten the patches and to transform each patch to a random dimension C (in this study we set $C = 128$). This procedure is required as transformers demand a sequence of tokens as input. We employed the base version of the Swin Transformer architecture where the overall architecture comprises 4 stages. Every stage includes two consecutive blocks, except for the 3rd stage which contains 18 consecutive blocks.

We followed the bottom-up pathway approach to extract the features. As opposed to common works, the proposed framework generates feature maps for every stage instead of just the last one. This gives us the ability to obtain a representation on multiple scales. Therefore, the framework can detect objects at different scales, from small to large objects. For instance, in our study, objects look smaller when they are far away from the optical sensor because of perspective transformation. The output from each stage is used for the purpose of enhancing the top-down pathway through a lateral connection in the next module.

C. HELMET USE DETECTION

This module is responsible for the final predictions of the bounding boxes and their classes. It is based on the Cascade RCNN framework [37] for object detection. The original architecture utilizes a ResNet50 [11] backbone.

In our implementation, we added a Feature Pyramid Network (FPN) neck [38] on the bottom of the Cascade RCNN module. The FPN neck allows higher-resolution layers to be constructed from semantically rich layers by following a top-down path. It acts as a feature extractor for the RPN head by taking the output of each stage in the Swin transformer backbone and then generating a new pyramid of feature maps. These feature maps are then provided to the RPN head for the anchor generation process, which produces region proposals that may contain objects. Finally, classification and bounding box regression networks are used to process the features of those region proposals. Upon detection, a bounding box along with a confidence score are produced for each object.

IV. EXPERIMENTAL RESULTS

In this section, we start by describing the dataset used in this work. Next, we present the models' architectures implemented for our experiments. Then, we give a detailed explanation of our experimental settings for both approaches. Finally, we report the results obtained and evaluate the performance of each model and approach.

A. DATASET

Datasets for motorcycle helmet detection are limited since most do not have helmets annotated as an object. As a result, the vast majority of datasets have motorcycles and their rides all annotated as a single object as shown in figure 4. The dataset [19] we used in this work is the same one used by [18] and [32]. It comprises 91,000 images extracted from 242 hours of traffic video, captured in Myanmar from 12 different observation sites over the course of two months period in 2016.

TABLE 1. Average precision on test set for each class by model.

Class	nb of examples	Position					Average Precision AP (%) of Helmet Use Detection for Each Class by Each Model							
		D	P1	P2	P3	P0	Swin-C	Swin-F	DETR	D-DETR	RetinaNet	YOLOv7	PP-YOLOE	PP-YOLOv2
1	22159	✓	-	-	-	-	76.2	76.6	77.4	71.2	77.1	79.4	73.5	76.1
2	11331	✓	✓	-	-	-	76.2	75.0	75.7	69.6	76.3	78.8	71.8	74.1
3	9971	✗	-	-	-	-	71.2	67.8	64.7	61.6	67.4	70.8	64.5	67.6
4	4924	✗	✗	-	-	-	68.4	67.9	64.8	58.4	67.4	70.1	57.9	63.3
5	1970	✓	✗	-	-	-	44.2	34.0	34.8	25.1	39.6	40.8	31.5	31.4
6	936	✗	✗	✗	-	-	32.4	37.3	31.6	24.5	31.8	37.3	17.6	23.6
7	446	✗	✓	-	-	-	32.1	24.2	25.0	18.5	23.4	29.6	17.4	4.6
8	445	✓	✗	✓	-	-	9.0	11.5	8.1	7.7	13.2	14.3	5.6	6.1
9	390	✓	✗	✗	-	-	36.6	31.5	24.7	21.5	30.1	21.6	25.4	14.9
10	466	✗	✗	-	-	✗	3.7	5.5	6.4	9.8	2.4	6.7	4.9	5.3
11	393	✓	✓	-	-	✗	29.1	25.3	17.9	15.3	24.1	28.3	11.7	2.4
12	230	✗	✗	✗	-	✗	42.1	37.7	44.7	31.1	34.9	43.5	28.6	11.5
13	279	✓	-	-	-	✗	43.0	45.6	32.0	21.4	38.1	43.7	1.8	0.0
14	174	✗	-	-	-	✗	4.2	3.6	0.6	33.7	3.5	37.4	1.5	0.2
15	22	✓	✗	-	-	✗	44.2	8.7	12.0	0.7	2.4	4.8	2.2	0.6
16	69	✓	✓	✓	-	-	53.6	29.0	27.8	3.2	22.2	15.5	4.4	2.7
17	137	✓	✓	-	-	✓	13.4	10.7	11.4	4.7	13.0	15.7	3.3	0.3
18	22	✓	✗	✓	-	✗	11.5	8.3	15.0	6.2	6.1	34.4	1.5	1.2
19	13	✓	✗	✗	✓	-	38.0	61.9	78.8	61.1	44.1	8.6	5.4	9.3
20	42	✓	✗	✓	-	✓	0.0	1.1	0.3	3.3	3.7	3.3	3.5	1.4
21	0	✓	-	-	-	✓	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
22	0	✓	✗	✗	-	✗	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
23	0	✗	✗	✓	-	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
24	0	✗	✗	✗	✗	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
25	0	✓	✓	✗	-	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
26	0	✓	✗	✗	✗	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
27	0	✓	✓	✓	-	✗	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
28	0	✗	✓	✓	-	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
29	0	✗	✓	-	-	✗	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
30	0	✓	✗	✗	-	✓	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
31	0	✗	✗	-	-	✓	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
32	0	✗	✗	✗	✗	✗	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
33	18	✓	✓	✓	-	✓	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
34	14	✗	✗	✓	-	✗	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
35	49	✓	✗	✗	✓	✗	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
36	29	✓	✗	✗	✗	✗	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total	54529	**mAP**					**30.4**	27.6	27.2	22.9	25.9	28.6	18.3	16.6
		Weighted mAP					69.54	68.37	67.77	62.23	68.74	**71.44**	63.93	66.28

✓ Helmet worn by motorcycle rider/pillion in the corresponding position
✗ Helmet is not worn by motorcycle rider/pillion in the corresponding position
- The corresponding position does not have a motorcyclist
N/A No results are available for this class, as there are no test data for this class
Swin-C stands for the combination of the Swin transformer architecture and the Cascade R-CNN framework
Swin-F stands for the combination of the Swin transformer architecture and the Faster R-CNN framework

The dataset contains 36 classes, annotated by drawing a bounding box around each motorcycle with its rider and passengers. Each class describes the number of riders using the motorcycle, their positions, and their helmet use. These motorcycle users are identified as a rider, several passengers from a single one to three passengers, and in front of the driver, a child passenger standing on the motorcycle's floorboard as shown in figure 3.

B. EXPERIMENTAL SETTINGS

Our methodology is based on using transformers for vision. We used Swin Transformer [16] as a backbone for feature extraction and Cascade RCNN [37] as object detector model. Specifically, we employed the base version of Swin Transformer pre-trained on the ImageNet-1K dataset with an embedding dimension of $C = 128$. Our choice of the base version was due to its ability to achieve a good balance between model size and computational complexity, while still providing high accuracy.

We fine-tuned the model for 12 epochs, using a batch size of 2. We apply the AdamW optimizer with an initial learning rate of 10^{-4} using a piecewise decay learning rate scheduler for the 7^{th} and the 10^{th} with a weight decay of 0.05 and 1000 steps of linear warm-up. We applied different data augmentations as described in section III-A.

For the other models used for comparison, table 2 presents details about each hyperparameter used for each model. We fine-tuned each model with different hyperparameters based on their nature, as CNNs are trained differently from transformer models.

We conduct our experiments based on the PaddleDetection[1] toolbox. We performed these tests on an Intel Core i9

[1] https://github.com/PaddlePaddle/PaddleDetection

TABLE 2. Models hyperparameters.

Model	Learning Rate	Regularizer		Optimizer	Batch Size	Epochs	Training Time (Hours)
		Type	Factor				
Swin Transformer (Cascade RCNN)	0.0001	Weight decay	0.05	AdamW	2	12	58
Swin Transformer (Faster RCNN)	0.0001	Weight decay	0.05	AdamW	2	15	67
Deformable DETR	10^{-5}	Weight decay	0.0001	AdamW	2	15	50
DETR	0.0001	Weight decay	0.0001	AdamW	8	20	31
RetinaNet	0.01	L2	–	SGD*	12	15	14
YOLOv7-L	0.001	L2	–	SGD*	32	50	47
PP-YOLOE	0.01	L2	–	SGD*	20	15	21
PP-YOLOv2	0.001	L2	–	SGD*	10	15	22

* Refers to SGD with Momentum

TABLE 3. Results of the different models applied to the helmet dataset.

Method	Model	Backbone	mAP (%)	mAP@50(%)	mAP@75(%)	Avg FPS
Transformers	Cascade R-CNN	Swin Transformer	**30.4**	**33.4**	**32.7**	20
	Faster R-CNN	Swin Transformer	27.6	31.1	30.2	20
	DETR	ResNet-50	27.2	30.6	29.6	50
	Deformable DETR	ResNet-50	22.9	26.9	25.9	20
CNN	RetinaNet	ResNet-50	25.9	29.0	28.0	20
	Yolov7-L	ELANNet	28.6	31.9	31.1	39
	PP-YOLOE	CSPResNet	18.3	21.0	20.0	18
	PP-YOLOv2	ResNet-50	16.6	19.9	19.0	21

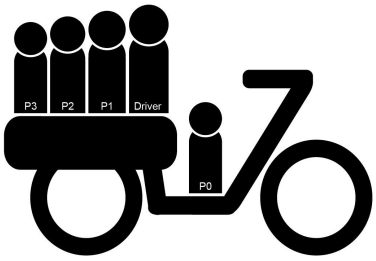

FIGURE 3. Description of positions of riders. The dataset is annotated with a maximum of 5 riders in a single motorcycle. P1, P2, and P3 are the first, second, and third pillions respectively, while P0 is determined as a child sitting before the rider.

CPU at 4.00 GHz clock frequency, 64 GB of RAM, and a single 24 GB Nvidia RTX3090 GPU.

C. EVALUATION METRIC

In this study, we utilized the mean average precision (mAP) metric to evaluate the performance of our object detection models. The mAP metric provides a comprehensive measure of the models' effectiveness in detecting objects across different classes.

The mAP is a metric that combines precision and recall, considering the trade-off between them at various confidence thresholds. It allows us to assess both the accuracy and completeness of the models in object detection tasks.

To calculate the mAP, we first compute the average precision (AP) for each individual class. The AP represents how well the model detects objects of a specific class by considering different confidence thresholds.

To get the AP, we first determine the precision-recall curve for a certain class by varying the detection algorithm's confidence threshold as presented in the COCO challenge [39]. Recall is the percentage of true positive detections over the entire number of ground truth objects in that class, whereas precision is the percentage of true positive detections over the total number of true positive detections (detections that match a ground truth object). The AP for that class is then calculated by integrating the precision-recall curve. Finally, we get the mAP score for the model by averaging the APs across all classes.

D. PERFORMANCE EVALUATION

In this work, we tested several object detectors based on CNNs and transformers for vision models. Tables 3 and 1 summarize the comparison results for each model.

Table 3 presents the evaluation results of all the models on the test set using the mean average precision (mAP) metric, which is calculated by averaging the precision at different IoU thresholds ranging from 0.5 to 0.95 with a step of 0.05. Additionally, the mAP@50 and mAP@75 metrics are also reported, which measure the performance of the models at an IoU threshold of 0.5 and 0.75, respectively. Our proposed model, which employs the Swin transformer as a backbone and the Cascade R-CNN with an FPN neck as a detection module, achieves the highest scores in all three metrics compared to other models. These results demonstrate the effectiveness of our approach in accurately detecting helmet use. The detection results on several example images are also shown in Figure 4.

In our study, we compare the performance of transformers and conventional convolutional neural networks (CNNs) in the task of helmet detection for motorcycle riders and passengers. Our results demonstrate that the transformer-based

FIGURE 4. **Results of helmet use detection utilizing Swin Transformers and Cascade RCNN on some sampled images from the test set. Each class detected is represented by a different colored bounding box.**

models, particularly the Swin Transformer architecture, outperformed the CNN-based models in terms of accuracy. This highlights the effectiveness of the attention mechanism, which is an integral part of the transformers' architecture.

We further analyzed the results by category and found that classes with up to two occupants per motorcycle yield better results, as presented in table 1. This is likely due to the fact that in real-world scenarios, most motorcycles are driven by just one or two riders most of the time, resulting in a larger number of samples for these classes. However, our analysis also revealed some limitations of the dataset, particularly imbalanced classes. Some of the classes have missing results due to the insufficient number of samples in the training and test sets. In particular, classes 21 to 32 lacked any examples, whereas classes 33 to 36 had fewer samples in the training set and no samples in the test set, limiting the ability of the models to detect these categories accurately.

Our proposed model, which uses Swin Transformer as a backbone combined with a Cascade RCNN framework for object detection, performs better than all other models in terms of mean average precision (mAP) over 36 classes, especially the YOLOv7 [40] model which achieved the second highest mAP score. This can be attributed to the more effective feature extraction capabilities of the Swin Transformer, which leverages the attention mechanism to extract relevant features from the input image. In contrast, YOLOv7 utilizes a different approach that relies on predefined anchor boxes to detect objects in the image and a CNN-based backbone to extract features. This approach may not be as effective as the attention mechanism in identifying and extracting relevant features, leading to lower performance in terms of mAP. Although the YOLOv7 model performs slightly better in terms of weighted mean average precision, it fails to outperform the Swin Transformer and Cascade RCNN model in the overall mAP. Furthermore, our model demonstrates its efficacy not only for classes with ample examples but also for those with limited samples, as demonstrated by its better performance for classes 15 and 16, which contain a small

number of examples. This is particularly important in real-life scenarios, where rare classes can pose a significant challenge for object detection. Therefore, our results highlight the effectiveness of our approach for monitoring helmet usage among motorcyclists and support the use of the Swin Transformer and Cascade RCNN model.

V. CONCLUSION

In this paper, we propose the use of transformer models, specifically the Swin transformer, for helmet use detection in images. The Swin transformer is used as the backbone for feature extraction, which extracts features from the input image that can be used for object detection.

In order to handle scale variation, we combine a Feature Pyramid Network (FPN) with the Cascade RCNN framework for final detection. The FPN neck takes the feature maps produced by the backbone network and fuses them with up-sampled feature maps from a top-down pathway. This results in a set of multi-scale feature maps that can be used for object detection, enabling the detector to better handle objects of different sizes.

The Cascade RCNN framework is used as the object detection head, which takes the multi-scale feature maps produced by the FPN and uses them to detect objects. The framework is trained and evaluated using a public dataset.

The proposed method is compared to other state-of-the-art detection methods, and it achieves promising results in terms of accuracy. Specifically, the framework achieves a mean average precision (mAP) of 30.4 and a weighted mAP of 69.54, which outperforms other state-of-the-art detection methods.

Overall, the proposed method shows the potential of using transformer models for detecting helmet use in images. The combination of the Swin transformer, FPN neck, and Cascade RCNN framework achieves promising results and outperforms other state-of-the-art detection methods.

REFERENCES

[1] (Jun. 2022). *Road Traffic Injuries*. WHO. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[2] H. A. Abdelali, O. Bourja, R. Haouari, H. Derrouz, Y. Zennayi, F. Bourzex, and R. O. H. Thami, "Visual vehicle tracking via deep learning and particle filter," in *Advances on Smart and Soft Computing*, F. Saeed, T. Al-Hadhrami, F. Mohammed, and E. Mohammed, Eds. Singapore: Springer, 2021, pp. 517–526.

[3] H. Derrouz, A. Elbouziady, H. A. Abdelali, R. O. H. Thami, S. El Fkihi, and F. Bourzeix, "Moroccan video intelligent transport system: Vehicle type classification based on three-dimensional and two-dimensional features," *IEEE Access*, vol. 7, pp. 72528–72537, 2019.

[4] Z. Charouh, A. Ezzouhri, M. Ghogho, and Z. Guennoun, "Video analysis and rule-based reasoning for driving maneuver classification at intersections," *IEEE Access*, vol. 10, pp. 45102–45111, 2022.

[5] A. Ezzouhri, Z. Charouh, M. Ghogho, and Z. Guennoun, "Robust deep learning-based driver distraction detection and classification," *IEEE Access*, vol. 9, pp. 168080–168092, 2021.

[6] J. Misachi. (Aug. 2019). *Countries With the Highest Motorbike Usage*. [Online]. Available: https://www.worldatlas.com/articles/countries-that-ride-motorbikes.html

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[12] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao, "From handcrafted to deep features for pedestrian detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4913–4934, Sep. 2022.

[13] S. Srivastava, S. Narayan, and S. Mittal, "A survey of deep learning techniques for vehicle detection from UAV images," *J. Syst. Archit.*, vol. 117, Aug. 2021, Art. no. 102152.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[17] A. Chairat, M. N. Dailey, S. Limsoonthrakul, M. Ekpanyapong, and D. Raj K. C., "Low cost, high performance automatic motorcycle helmet violation detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3549–3557.

[18] F. W. Siebert and H. Lin, "Detecting motorcycle helmet use with deep learning," *Accident Anal. Prevention*, vol. 134, Jan. 2020, Art. no. 105319.

[19] H. Lin and F. W. Siebert. (2020). *Helmet Dataset, the Largest Annotated Motorcycle Helmet Use Dataset*. [Online]. Available: https://osf.io/4pwj8/

[20] R. R. V. E. Silva, K. R. T. Aires, and R. D. M. S. Veras, "Detection of helmets on motorcyclists," *Multimedia Tools Appl.*, vol. 77, no. 5, pp. 5659–5683, Mar. 2018.

[21] V. L. Padmini, G. K. Kishore, P. Durgamalleswarao, and P. T. Sree, "Real time automatic detection of motorcyclists with and without a safety helmet," in *Proc. Int. Conf. Smart Electron. Commun. (ICOSEC)*, Sep. 2020, pp. 1251–1256.

[22] L. Shine and J. CV, "Automated detection of helmet on motorcyclists from traffic surveillance videos: A comparative analysis using hand-crafted features and CNN," *Multimedia Tools Appl.*, vol. 79, pp. 14179–14199, May 2020.

[23] C. Vishnu, D. Singh, C. K. Mohan, and S. Babu, "Detection of motorcyclists without helmet in videos using convolutional neural network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 3036–3041.

[24] B. Yogameena, K. Menaka, and S. S. Perumaal, "Deep learning-based helmet wear analysis of a motorcycle rider for intelligent surveillance system," *IET Intell. Transp. Syst.*, vol. 13, no. 7, pp. 1190–1198, Jul. 2019.

[25] K. Jearanaitanakij, K. Iamthammarak, and N. Wangcharoen, "Fast classifying non-helmeted motorcyclists by using convolutional neural networks," *SNRU J. Sci. Technol.*, vol. 13, no. 1, pp. 1–10, 2021.

[26] C. A. Rohith, S. A. Nair, P. S. Nair, S. Alphonsa, and N. P. John, "An efficient helmet detection for MVD using deep learning," in *Proc. 3rd Int. Conf. Trends Electron. Informat. (ICOEI)*, Apr. 2019, pp. 282–286.

[27] M. Dasgupta, O. Bandyopadhyay, and S. Chatterji, "Automated helmet detection for multiple motorcycle riders using CNN," in *Proc. IEEE Conf. Inf. Commun. Technol.*, Dec. 2019, pp. 1–4.

[28] W. Jia, S. Xu, Z. Liang, Y. Zhao, H. Min, S. Li, and Y. Yu, "Real-time automatic helmet detection of motorcyclists in urban traffic using improved YOLOv5 detector," *IET Image Process.*, vol. 15, no. 14, pp. 3623–3637, Dec. 2021.

[29] L. Allamki, M. Panchakshari, A. Sateesha, and K. S. Pratheek, "Helmet detection using machine learning and automatic license plate recognition," *Int. Res. J. Eng. Technol.*, vol. 6, pp. 80–84, 2019.

[30] A. M. Vakani, A. K. Singh, S. Saksena, and H. R. Vanamala, "Automatic license plate recognition of bikers with no helmets," in *Proc. IEEE 17th India Council Int. Conf. (INDICON)*, Dec. 2020, pp. 1–5.

[31] K. C. D. Raj, A. Chairat, V. Timtong, M. N. Dailey, and M. Ekpanyapong, "Helmet violation processing using deep learning," in *Proc. Int. Workshop Adv. Image Technol. (IWAIT)*, Jan. 2018, pp. 1–4.
[32] H. Lin, J. D. Deng, D. Albers, and F. W. Siebert, "Helmet use detection of tracked motorcycles using CNN-based multi-task learning," *IEEE Access*, vol. 8, pp. 162073–162084, 2020.
[33] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 4302–4311.
[34] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[35] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.
[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[37] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, May 2021, doi: 10.1109/TPAMI.2019.2956516.
[38] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[39] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. Lawrence Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," 2014, *arXiv:1405.0312*.
[40] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

AYYOUB BOUHAYANE received the M.Sc. degree in intelligent and decision-making systems from the Faculty of Science Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco, in 2019. He is currently pursuing the Ph.D. degree with the Mohammadia School of Engineers and the International University of Rabat. In 2020, he was a Research Engineer on a project funded by NARSA that focused on monitoring driver behavior. His research interests include computer vision, deep learning, and intelligent transportation systems.

ZAKARIA CHAROUH (Senior Member, IEEE) received the master's degree from ENSA Kenitra and the Ph.D. degree from EMI, Mohamed V University, Rabat. It was his privilege to teach computer science at several Moroccan universities, such as the International University of Rabat and the Moroccan School of Engineering Sciences, as well as at several international institutes, such as the French M2I-Formation Institute and the Greek ZPeople Institute. As a Research Scientist, he has been involved in several research projects funded by NARSA, the Moroccan Ministry of Transportation, and the Moroccan region of RSK. These projects revolve around the use of AI, computer vision, and electric vehicles (EVs), on which he has published several research papers and secured patents. Currently, as an R&D Manager at Majal Berkane SA, he is spearheading multiple projects aimed at implementing a Smart-City strategy in the Province of Berkane. These initiatives are being carried out in collaboration with the Ministry of Energetic Transition, the National Water and Forest Agency of Morocco, and other Moroccan institutes.

MOUNIR GHOGHO (Fellow, IEEE) received the M.Sc. and Ph.D. degrees from the National Polytechnic Institute of Toulouse, France, in 1993 and 1997, respectively. He was an EPSRC Research Fellow with the University of Strathclyde, Scotland, from September 1997 to November 2001. In December 2001, he joined the School of Electronic and Electrical Engineering, University of Leeds, England, where he was promoted to a Full Professor, in 2008. While still affiliated with the University of Leeds, in 2010, he joined the International University of Rabat, where he is currently the Dean of the College of Doctoral Studies and the Director of ICT Research Laboratory (TICLab). He has coordinated around 20 research projects and supervised over 30 Ph.D. students in the U.K. and Morocco. His research interests include machine learning, signal processing, and wireless communication. He is a fellow of AAIA (Asia–Pacific AI Association), a recipient of the 2013 IBM Faculty Award, and a recipient of the 2000 U.K. Royal Academy of Engineering Research Fellowship. He is the Co-Founder and the Co-Director of the CNRS-Associated International Research Laboratory DataNet, in the field of big data and artificial intelligence. He served as an Associate Editor of many journals, including the *IEEE Signal Processing Magazine* and the IEEE TRANSACTIONS ON SIGNAL PROCESSING.

ZOUHAIR GUENNOUN (Senior Member, IEEE) received the engineering degree in electronics and telecommunications from the Electronics and Electrical Montefiore Institute, ULG Liege, Belgium, in 1987, and the M.Sc. degree in communication systems and the Ph.D. degree from the EMI School of Engineering, Rabat, Morocco, in 1993 and 1996, respectively. He visited the Centre for Communication Research (CCR), Bristol University, U.K., from 1990 to 1994, to prepare a split Ph.D. degree. From 1988 to 1996, he was an Assistant Lecturer with the EMI School of Engineering, where he has been a Professor/Lecturer, since 1996. His research interests include digital signal processing, error control coding, speech, and image processing. He is currently the ERSC Research Team Leader [previously known as the Laboratory in Electronics and Telecommunications (LEC)] of EMI. He is an ex-member of the Moroccan IEEE Section Executive Committee.