IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 September 2023, accepted 25 September 2023, date of publication 2 October 2023,
date of current version 13 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3321290

RESEARCH ARTICLE

Automatic Identifier of Socket for Electrical Vehicles Using SWIN-Transformer and SimAM Attention Mechanism-Based EVS YOLO

V. C. MAHAADEVAN[1], R. NARAYANAMOORTHI[1], (Member, IEEE), RADOMIR GONO[2], (Senior Member, IEEE), AND PETR MOLDRIK[2]

[1]Electric Vehicle Charging Research Centre, Department of Electrical and Electronics Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai 603203, India
[2]Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, 708 00 Ostrava, Czech Republic

Corresponding author: R. Narayanamoorthi (narayanamoorthi.r@gmail.com)

This work was supported in part by the Government of India, Department of Science and Technology (DST), Science and Engineering Research Board (SERB) Core Research under Grant CRG/2020/004073; and in part by SGS Grant from VSB-Technical University of Ostrava under Grant SP2023/005.

ABSTRACT Electric vehicle (EV) technology is emerging as one of the most promising solutions for green transportation. The same growth occurs in the charging infrastructure development and automating the EV charging process. Globally, EVs has different types of charging sockets and it's located at the various positions in the Vehicle. In simple, EV has a diversity in socket type and socket location. Hence, correctly identifying the socket type and location is mandatory to automate the charging process. The recent development in computer vision and robotic systems helps to automate EV charging without human intervention. Image processing and deep learning-based socket identification can help the EV charging infrastructure providers automate the process. Moreover, the deep learning techniques should be simple enough to implement in the real-time processing boards for experimental viability. Hence, this paper proposes a new You Only Look Once (YOLO) model called the Electric Vehicle Socket (EVS) YOLO that uses YOLOv5 as its base architecture with the addition of a vision-type transformer called the SWIN-Transformer and an attention mechanism called SimAM for better performance of the model in detecting the correct charging port. A dataset of 2700 images with six types of classes has been used to test the model, and the EVS -YOLO also evaluated with varying mechanisms of attention positioned at various places along the head. The paper contrasts the suggested model with alternative deep learning architectures and analyzes respective performances.

INDEX TERMS SWIN-transformer, attention mechanism, YOLOv5, electric vehicles, socket detection, SimAM.

I. INTRODUCTION

The current technological revolution, the electric vehicle, uses batteries instead of fuel-based technologies to help reduce emissions [1], [2]. Since then, EVs have advanced significantly, from efficient designs to many sorts of charging sockets [3], [4]. Even though there has been growth since different EV models employ various kinds of sockets, the techniques used to charge electric vehicles now rely significantly on manual operation, which can result in problems including parking the vehicle according to socket location, heavyweight handling approx. 1.5 kg and choose the correct plug and docking of charging ports, which needs some effort. The EV has a charging port at five locations: front right & left, rear right & left, and front middle [5], as shown in Figure 1. Most vehicles' charging ports are located at the right rear at 36%, followed by 28% at the left end and 22% at the left front. Only 10 % of the global vehicle has front-end charging, which enable ease of parking, and the rest need some effort to align the vehicle according to the charging

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

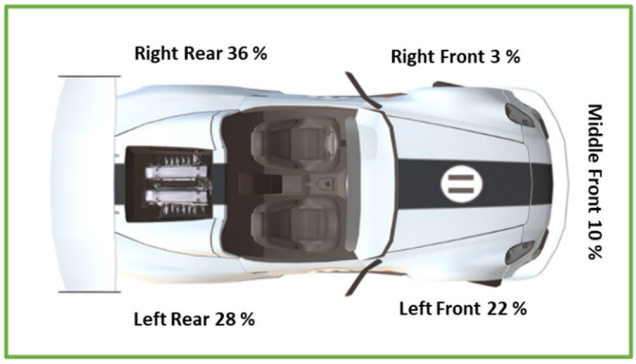

FIGURE 1. Global charging port position.

station gun. From this study, there is a need for automated charging robots enabled by machine vision. Electric shock is another possible safety risk brought on by old insulation and manual operation. The need for intelligent and humanized services [6] in the electric car charging sector is rising in the age of intelligence. To enhance the charging experience for owners of electric vehicles and to reduce the hazards [7], research on automated identification methods based on image recognition and the creation of automatic charging control systems. Robotic, automatic charging is one of the many methods investigated in the study of electric cars automated charging systems. For instance, TESLA [8] has created an autonomous charging robot as a snake that can stoop down to locate the charging port and attach the charging gun to the charging socket. The Volkswagen Automotive Company created the E-Smart Connect system, employing a camera to detect the vehicle's location and interface while also controlling a KUKA robot to charge [9]. Other studies have concentrated on very accurate charging port locations utilizing binocular vision or ultrasonic-based techniques [10], [11], [12], [13]. [14] Samsung's EVAR is the latest autonomous charging system for electric cars.

Hyundai Motors recently developed an Automatic charging robot (ACR) [15]. Most automated charging systems proposed, conceived & developed are limited to specific sockets & connectors and particular locations at the vehicle and more end-user involvement. Hence, a 6 DOF robot with a 3D vision sensor mount on AGV is proposed that moves in the parking line as guided to ensure the entire system moves around the vehicle. The robot's vision sensor with image processing capability will detect the charging port by moving the whole robot system on the parking line, get the correct charging gun from the station through a suitable control system, and achieve charging by plug-in with image processing support. The process will reverse to keep the charger gun back at the station. Through this proposed automated system, any electric car with any charging socket located in any location of a vehicle can be charged.

The image processing area alone is discussed in this paper. Safety is a top priority when designing automated charging systems for electric vehicles [16]. Hence, it is crucial to distinguish between purposeful encounters and unintentional accidents when robots and vehicles come into touch during the charging process. Despite research advances, little is known about these vehicle-robot interaction components. Given the significance of new energy-based vehicles and the rising demand for intelligent charging solutions, there is a need for more research and development in electric car charging facilities. The electric vehicle charging industry can enhance the effectiveness, security, and user experience of electric vehicle charging by utilizing cutting-edge technologies like image recognition, robotics, and automation. It will help to promote the overall growth of the electric vehicle ecosystem and the international efforts to conserve energy and reduce emissions.

To address such concerns, applying deep learning through object detection can play a role by automatically detecting the type of charging socket required by the electric vehicle, eliminating the need for the driver to select the appropriate one. s manually. Using deep learning algorithms, EV charging stations can detect a wide range of charging socket types, including older or less common types, making it easier for EV drivers to charge their vehicles. By automating the socket detection process, EV charging stations can reduce the need for additional personnel, resulting in operator cost savings. It is possible through object detection, which forms the basis of an automatic EV charging system. Similarly, charging sockets can also be detected. Every sort of socket will have a unique quality. For example, the type 1 plug, which can support a power injection of up to 7.4 kW (230 V, 32 A), is a single-phase connector. Type 2 plug is a triple-phase plug. Type 2 sockets at public charging stations provide power levels of up to 43 kW (400 V, 63 A, AC). The Combined Charging System plugs in short Combination plugs or CCS with power levels of up to 170 kW. However, in practical scenarios, the value frequently remains in the vicinity of 50 kW. A wide variety of electric vehicles can use the CHAdeMO plug, which is built for them and offers a charge of 50kW, designed specifically for Tesla electric vehicles. These chargers come with a charging capacity between 150kW and 250kW.

Identifying the existence and placement of objects in an image or video is object detection in computer vision. Object detection algorithms coupled with a label specifying the item type produce Bounding boxes that encircle the things in the picture. Some challenges associated with this task include occlusions, background clutter, and scale variations. Several approaches are developed, which typically revolve around learning complex representations of objects and their features to solve these issues. In deep learning, a feature refers to a pattern or characteristic detected or extracted from input data relevant to the problem being solved. Features are frequently utilized to express the input data more concisely and appropriately to make it simpler. For instance, in image recognition, a feature may be a specific edge or texture pattern typical of a particular class of objects. A distinctive word or phrase to a feeling or topic might be a feature in natural language

processing. Feature methods like Histogram of Oriented Gradients (HOG) [17] and Scale-Invariant Feature Transform (SIFT) [18] are categorized as handcrafted feature extraction methods.

A. MOTIVATION

Handmade feature extraction methods involve manually designing algorithms or heuristics to identify and extract features from input data. These methods typically use domain knowledge and human expertise to develop effective feature extractors. HOG and SIFT are examples of widely used handcrafted feature extraction methods in computer vision. HOG works by computing the distribution of gradient orientations in an image, while SIFT extracts features by detecting key points and describing their local appearance using scale-invariant descriptors. The capacity of more recent deep learning techniques to automatically learn the most pertinent and discriminative features from the data, without the need for human skill or domain knowledge, is one of their key benefits over more traditional feature extraction techniques like HOG and SIFT. Deep learning models can achieve cutting-edge performance on a broad range of tasks by learning to extract complicated and abstract characteristics from the data by training on vast volumes of labeled data.

B. CONTRIBUTIONS

The main contribution of this paper includes:

- It effectively detects charging ports embedded in electric vehicles using a novel network architecture EVS-YOLO that uses a Swin transformer to improve self-attention.
- A SimAM attention mechanism is added in 3 locations across the detection heads in the architecture, and a comparison of different attention mechanisms is to come to a consensus on using the SimAM attention module.
- A dataset [18] of 2700 images with six types of classes was used and the same was operated on different models and results are compared between proposed EVS-YOLO network architecture and existing object detectors. The proposed EVS-YOLO model has a Mean Average Precision (mAP) of 81.4, the highest compared to the other object detectors.

II. LITERATURE SURVEY

A. REAL-TIME OBJECT DETECTORS

Modern-day deep learning frameworks are aggregated into two types: single-stage and two-stage detectors. The bounding boxes and class probabilities of objects in an image are directly predicted by a single-stage detector, sometimes called a one-stage detector. These versions are perfect for real-time applications since they are often quicker and easier to use than two-stage detectors. Contrarily, two-stage detectors have two steps. The model creates region proposals in the first step, potential placements for objects in the picture. The model uses these suggestions in the second step to categorize items and improve the bounding box coordinates. An example of this type is R-CNN, proposed by Girshick et al. [19] This approach uses a method to extract features from an image, followed by a step of the regional proposal to identify potential object locations and a classifier to classify objects in each proposed region.

Fast RCNN [20], [21] improved upon R-CNN by using a single CNN to extract features for both the region proposal and object classification stages, resulting in faster training and inference. One another example that portrays the two-stage detector method is Faster RCNN [22], which introduced a Region Proposal Network (RPN) that shares convolutional features with the object detection network, allowing for even faster training and inference. The RPN generates region proposals more efficiently than the external region proposal algorithms used in R-CNN. Even though R-CNN was a significant advance in object detection, it had several limitations, including slow training and inference and limited flexibility. Its successors, such as Fast R-CNN Faster R-CNN, addressed these limitations and significantly improved the accuracy and efficiency of object detection models. On the other hand, in SSD [23] and YOLO [24], [25], [26] the input images are fragmented into a grid of cells, and for each cell, bounding boxes and probabilities of class are predicted. On the other hand, SSD uses multiple layers with different aspect ratios to manage objects of different shapes and scales. RetinaNet [27] uses an FPN 0662to identify objects at various dimensions and resolutions that includes a novel focal loss function that addresses the class imbalance problem in object detection.

B. ATTENTION MODULES

Attention modules are utilized in object identification to assist the model in concentrating on the areas of an image that are most important for detecting objects. According to their importance to the task, distinct parts of a picture are given varied weights by attention modules. The model's attention is directed to the areas of a concept that are the most informative using weights learned during training. It enables the model to concentrate on an image's most crucial elements while avoiding unimportant or distracting aspects. To upgrade the model's capacity to focus on critical properties of the input data, deep learning models employ two different types of attention methodologies: channel and spatial.

A CNN's channel attention mechanism may be trained to highlight or suppress particular feature map channels selectively. Contrarily, a spatial attention technique may be instructed to selectively emphasize or hide certain spatial positions in a CNN's feature maps. SE attention [28] module consists of the squeeze and excites operations. The squeezing process aggregates the spatial dimensions of each feature map into a single scalar value by a global average pooling technique. It decreases the dimensions of the feature maps while retaining the channel-wise information. The final step is to apply the generated vector to a two-layer perceptron

TABLE 1. Summary of related works.

References	References	Models	Techniques	Reference
Zhang *et al.* 2016	Images	Based on HSI Color Model	Image filtering, threshold segmentation, morphology processing, edge detection, feature extraction	[47]
Miseikis *et al.* 2017	Images	CATIA v5	Region-based Convolutional Neural Networks, Deep learning	[48]
Jiang *et al.* 2019	Image , Video Frames	YOLO v3	Multi-agent deep reinforcement learning	[59]
Bochkovskiy *et al.* 2020	Video Frames	YOLO v4	Bag of Freebies (BoF) and Bag of Specials (BoS), Data Augmentation, CmBN	[26]
Mingqiang et al. 2020	Images	Based on CNN	Convolution neural network, Deep learning	[46]
Dirir *et al.* 2021	Video Frames , Images	YOLO v2	Region-based Convolutional Neural Networks, Single Shot Detector	[55]
Zhou *et al.* 2021	3D point cloud data	Based on PV-RCNN model	3D point cloud technique	[60]
Shibl *et al.* 2021	Images	Based on LSTM	K-Nearest Neighbors, Deep learning	[61]
Park *et al.* 2022	Video Frames	DQN,VCE	Deep Reinforcement Learning, convolution neural network,RNN	[62]
Guney *et al.* 2022	Images	Based on YOLO v4, CNN	HSV Color Space	[50]
Lin *et al.* 2022	Images	Based on CNN, LSTM	Deep Convolutional Neural Network, Support Vector Machine, KNA	[63]
Li *et al.* 2022	Images	DUAL-200M-030T160	Semi-global block matching, Fast Library for Approximate Nearest Neighbors	[49]
Guney el al.2022	Images	YOLO v5	Portable and image-based ADAS system for real-time detection of trafc signs, vehicles, and pedestrians.	[42]
ElKashlan *et al.* 2023	Network traffic data(IoT)	Based on CNN	Convolution neural network, Deep learning	[64]
Karanam *et al.* 2023	Time series data	LSTM, HMM	Recurrent neural network, SVM	[65]

that learns to represent the channel-wise correlations. The excitation operation, a sigmoid activation function, scales the learned attention map produced by a two-layer perceptron. CBAM [29] attention module consists of channel and spatial attention modules, wherein the SAM records the spatial dependencies, and the CAM learns the channel-wise feature maps. Therefore, the CBAM attention module can adaptively recalibrate the feature maps across both spatial and channel dimensions by integrating the CAM and SAM processes.

The coordinate attention [30] module aims to improve by ingraining positional information with channel attention with the help of two blocks: coordinate information embedding and coordinate attention generation blocks. The first block replaces the typically used global pooling operation with a two 1- D encoding structure to better preserve the positional information. The second block lets the model concentrate on an image's most crucial elements while avoiding unimportant or distracting aspects. Deep learning models employ two types of attention methodologies, channel and spatial, to upgrade the model's capacity to focus on key properties of the input data. A CNN's channel attention mechanism may be trained to highlight or suppress particular feature map channels selectively. Contrarily, a technique called spatial attention the captured positional data to identify the regions of interest precisely, and it is also capable of effectively capturing inter-channel relationships.

C. DETECTION METHODS BASED ON THE YOLO SERIES

The popularity of the YOLO series due to its speed, its real-time inference of images and videos as it is a one-stage detector, its accuracy, its versatility to be trained on different datasets, and its ability to be able to detect a wide variety of objects it is has been employed in many fields [31], [32], [33], [34], [35], [36], [37], [38] and performs very well concerning other methods. Tian et al. [39] Proposed a method for detecting apple lesions using a YOLOv3-dense network; it also included the usage of cycle GAN as an image augmentation tool to improve the results as traditional YOLOv3 networks don't have the inbuilt ability for data augmentation.

The consequent model in the YOLO series is YOLOv4, which introduced a CSP-based structure into the backbone and added SPP and PAN to improve the multi-scale fusion of features. It also added an improved head and data augmentation technique for better performance. Cai et al. [40] introduced YOLOv4-5D for object detection in autonomous driving in which the existing CSPdarknet53 backbone was combined with deformable convolutional layers, and the network was modified by adding two large-scale layers to be effective in detecting smaller objects. The proposed model was also pruned in terms of parameters. Since the proposed model had five different detection scales, the feature fusion network was modified to accompany that change. One another example that shows the versatility of the YOLO series is [41] in the field of marine target detection, wherein the

CBAM attention module was added to the three branches of the feature fusion network to improve the accuracy of the result. Guney et al. [42] used YOLO v5 for the ADAS system in cars to recognize sign boards on roads to assist drivers in real-time. YOLOv5 was improved upon the predecessor by introducing changes to its backbone; it is also a shallower model, thus making it faster to train on. In 2021, Zhu et al. [43] introduced a modified YOLOv5 network called transformer prediction heads YOLOv5 to aid object detection in drone scenarios. It was improved by the addition of transformer heads for prediction and the addition of the CBAM attention module. Therefore, when tested on the VisDrone dataset, it outperformed the base model by 7%. Due to the ability of real-time and fast inference, the YOLO models are helpful in applications involving monitoring and sorting. Wang et al. [44] portray the application of a modified YOLOv5 network called YOLO-BS which includes a SimAM attention module by detecting more giant coal blocks to aid congestion in underground mine scraper conveyors.

D. RELATED WORKS ON EV CHARGING SYSTEMS

The capacity to identify and locate the charging outlet is essential for autonomous charging as it also affects the system's dependability. It is critical to attenuate the negative impacts of a complicated environment by using the necessary algorithms. For an autonomous vehicle charging system to work efficiently, detecting the charging port used in an electric vehicle is essential; therefore, the algorithm used for the detection must aid the robotic system accurately to get the best results. Zhao et al. [45] proposed a method for fast identification and localized detection of the socket with a combination of a modified YOLOv4 network for quick recognition and mean shift clustering to improve the success rate by removing noise and an affine transformation method for the correction of coordinates. Mingqiang et al. [46] use a modified Lenet-5 model in which the ReLU function is used, the dimensionality of the Final convolution is changed, and the learning rate is optimized for the recognition of the socket. The socket was located by using a feature circle method. Zhang et al. [47] describe a process for using the vision software HALCON to automatically extract the characteristic parameters of an electric vehicle charging hole. The method involves filtering the original image using various image processing techniques.

Miseikis et al. [48] presented an approach that combines shape-based template matching, stereo cameras, and a robot with a connector plug to localize and approach the charging socket of an EV or PHEV. The method uses markerless eye-to-hand calibration to estimate the location and orientation of the charging socket and observes the forces exerted on the robot's end-effector to prevent misalignment. The approach has succeeded in lab conditions using a custom-made charging port holder and indoor illumination. Li et al. [49] propose a method for accurately identifying and positioning charging ports, even under varying light intensities and backgrounds. The technique uses the SIFT feature extraction algorithm and FLANN matching algorithm to attain a high-precision mapping of points. It then employs the SGBM algorithm for binocular ranging for calculating the depth of the socket. The proposed method was validated through binocular range and image identification experiments, demonstrating high-precision results. Behl et al. [50], This technology employs HSV color space to recognize and monitor the location of the female socket. It combines a mobile male socket on the shore charging station with a stationary female socket aboard the ship. The system was trained using the YOLO model for quicker and more precise identification, and an application interface was created for real-time monitoring. Table 1 depicts the summary of related works that clearly illuminates data such as method, technology, and description of various studies.

III. METHODOLOGY

A. YOLOV5 STRUCTURE

YOLOv5, being one of the widely used object detection algorithms due to its speed and effectiveness, has intrigued researchers to explore its possibilities in bringing change in its overall architecture, which comprises the head, neck, and backbone. Feature information gets extracted in the backbone and gets gathered in the neck. Based on the feature maps created, the head detects the predictions. CSPDarket53 framework with Spatial Pyramid Pooling-Fast (SPPF) layer makes up the backbone gets employed, PANet as neck and detection head completes the YOLOv5 basic architecture. The feature pyramids are obtained using the PANet. With the accuracy and speed of a pyramid in mind, the Feature Pyramid Network (FPN) [51] feature extractor was developed. As compared to older models like the quicker RCNN, it creates numerous layers of feature maps with greater quality information than the usual feature pyramid. The FPN is made up of top-down feature pyramids and a bottom-up path. The bottom-up method extracts features using a typical convolutional network.

As we climb, the spatial resolution decreases. Each layer becomes more important semantically when more high-level structures are found. The YOLO deep network uses residual and dense blocks to overcome the vanishing gradient problem, enabling information to go to the deepest layers. However, one advantage of having thick and residual blocks is the problem of recurring gradients. CSPNet [52] solves this problem by discretizing the gradient flow. Convolutional neural networks are designed to perform better, and one sort of feature aggregation module, CSPNet, seeks to do just that (CNNs) [53]. The CSPNet module serves as the backbone of the network architecture in YOLOv5. The object detection head uses feature maps produced by the spine to forecast the predictions. The three phases of the CSPNet module in YOLOv5 [54], [55] each feature a set of convolutional layers followed by a cross-pathway link. It achieves this by dividing the network into central and cross pathways.The

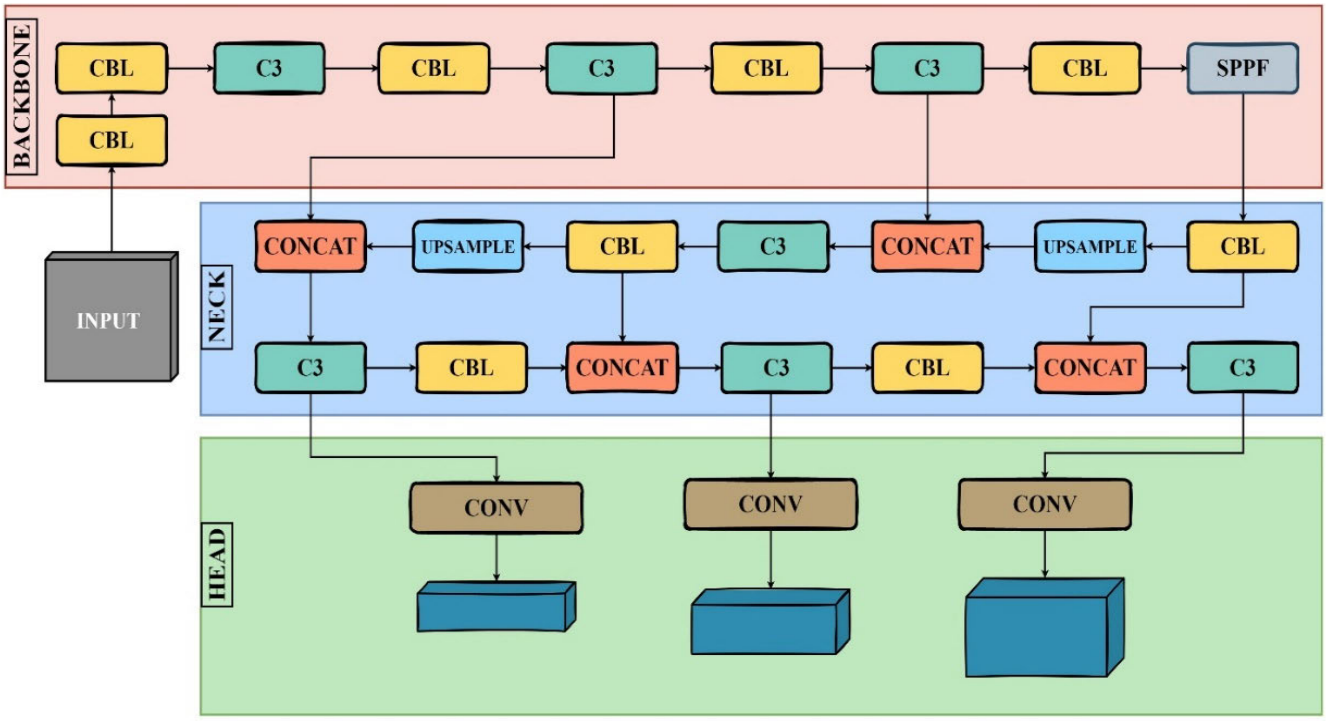

FIGURE 2. Base architecture of yolov5 showing three regions backbone, neck and head.

main pathway analyses the input data and creates feature maps. More information flow is made possible by doing this, which also lessens the chance that important data may be lost while being processed. Incorporating CSPNet in YOLOv5 enables maintaining a smaller size and a higher inference speed. This is because CSPNet enhances the network's information flow and enables more effective feature aggregation. The final product is an application-friendly object detection model that is very effective and accurate. CSPNet decreases the model's parameters and Flops, which not only enhances inference speed and accuracy but also addresses the problems with recurrent gradient information in large-scale backbones and trims down the model's size.Fast and accurate data detection is essential, and the model's size also determines the effectiveness of its inference on devices with minimal computational resources. The SPPF block produces a fixed-length result after it has combined the data it received from the inputs. As a result, without degrading the network's performance, it has the advantage of greatly increasing the receptive field using this block in earlier iterations of YOLO; however, to increase network speed in YOLOv5, SPPF just another variation of the SPP block was utilized. The classes of the discovered objects, their bounding boxes, and the abjectness scores are the three outputs. YOLOv5 produces. In calculating the location loss, CIoU [56] loss is used. The following equation provides the ultimate loss formula.

$$\text{Loss} = \lambda 1\text{Lcls} + \lambda 2\text{Lobj} + \lambda 3\text{Lloc} \quad (1)$$

Different activation functions, attention mechanisms, and modifying the backbone have been done previously to improve the performance of the YOLOv5 detection algorithm. Sigmoid linear units are the standard activation function used in YOLOv5. YOLOv5 comes in five different sizes, namely. YOLOv5x, YOLOv5N, YOLOv5M, YOLOv5L, and YOLOv5s. The model architecture in all five models remains the same, but changes in the width and depths can be noticed. YOLOv5X being the largest and YOLOv5n being the smallest. The larger models tend to perform better, but they are computationally intensive. We use the YOLOv5s architecture a base for our proposed EVS-YOLO architecture. The architecture of the YOLOv5 model is depicted in Figure 2. Base architecture of yolov5, which is predominantly made of three regions backbone, neck and head, wherein CBL consists of convolution, batch normalization and activation layer. C3 refers to a cross stage partial network with 3 convolutions and concat refers to the concatenation operation. SPPF refers to a faster version of spatial pyramid pooling which reduces the dimensionality and improves the network speed.

B. SWIN TRANSFORMER

In addition to effectively modeling global contextual information, the transformer also exhibits great transferability to downstream tasks when pretraining on a large scale. It offers new opportunities for visual feature learning and has observed the performance of the transformer in various deep-learning avenues. The transformer creates a method for global

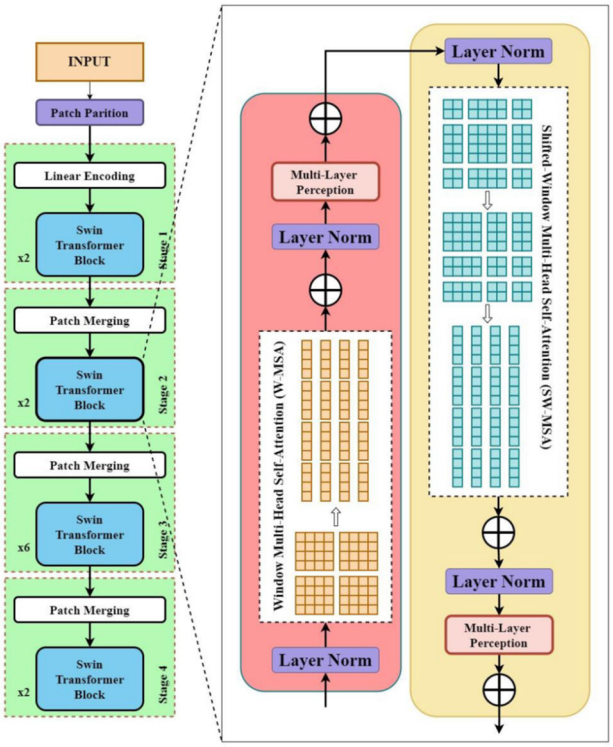

FIGURE 3. The Swin transformer with W-MSA and SW-MSA.

information exchange that aids in creating a suitable feature representation. Nevertheless, using the transformer for visual tasks has two big drawbacks. On the one hand, the transformer's use is severely constrained by the high computing cost of its transformation, which employs sequences as input.

On the other hand, Transformer mines correlations from global linkages instead of local inductive bias in convolution and needs training with a lot of data to provide good results. The introduction of the Swin transformer expands the potential uses of transforms in visual activities. The computational overhead of the swing transformer is minimal. Hierarchical structures are built as it analyses pictures, enabling the swing-based model to tackle multiscale heavy jobs. The proposed work modifies the YOLOv5s network topology to incorporate the Swin module, allowing the network to do global modeling while using fewer computing resources. The window based multi-head self-attention (W-MSA) module was proposed by Swin Transformer. There are several windows split up into the image. Swin Transformer reduces the computing complexity to a linear relationship by performing attention computations exclusively on the window pixel areas. Importantly, the Swin Transformer interacts with information across non-overlapping windows utilizing a multi-head self-attention module for shifted windows (SW-MSA). A shifted window partitioning strategy in the Swin transformer is incorporated such that the switches between two partitioning configurations in subsequent Swin Transformer blocks allow for cross-window connections to ensure the effectiveness of non-overlapping

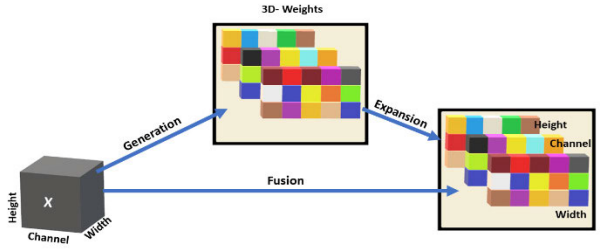

FIGURE 4. SimAM module.

window computation. It is difficult to gather global contextual information due to CNN's constrained perceptual field. The Swin Transformer, in comparison, makes use of more adaptable self-attention information transmission and performs very well in obtaining global semantic information and effectiveness.

The architecture of the Swin Transformer is shown in Figure 3. The Swin transformer predominantly contains two blocks namely Windows Multi-Head Self-Attention module (W-MSA) and Shifted Windows Multi-Head Self-Attention module (SW-MSA). The SW-MSA is crucial to know the functions of Swin transformer as it allows for information exchange across non overlapping windows. The computational complexity of a global MSA module and a window based are given as

$$\Omega(\mathrm{MSA}) = 4\mathrm{h}w\mathrm{C}2 + 2(\mathrm{h}w)2\mathrm{C} \quad (2)$$

$$\Omega(\mathrm{W}-\mathrm{MSA}) = 4\mathrm{h}w\mathrm{C}2 + 2\mathrm{M}2\mathrm{h}w\mathrm{C} \quad (3)$$

The shifted window partitioning approach is used to compute the successively aligned Swin transformer blocks.

$$\hat{\mathrm{z}}\mathrm{l} = \mathrm{W}-\mathrm{MSA}(\mathrm{LN}(\mathrm{zl}-1)) + \mathrm{zl}-1 \quad (4)$$

$$\mathrm{zl} = \mathrm{MLP}(\mathrm{LN}(\hat{\mathrm{z}}\mathrm{l})) + \hat{\mathrm{z}}\mathrm{l} \quad (5)$$

$$\hat{\mathrm{z}}\mathrm{l}{+}1 = \mathrm{SW}-\mathrm{MSA}(\mathrm{LN}(\mathrm{zl})) + \mathrm{zl} \quad (6)$$

$$\mathrm{zl}+1 = \mathrm{MLP}(\mathrm{LN}(\hat{\mathrm{z}}\mathrm{l}+1)) + \hat{\mathrm{z}}\mathrm{l}{+}1 \quad (7)$$

where, $\hat{z}l$ and zl stands for the respective block l output characteristics of the SW-MSA module and MLP module.

C. SIMAM ATTENTION MECHANISM

Attention mechanisms are a great way to utilize the most significant features in an input sequence by using the weighted combination of the input vectors. For our proposed model, we have incorporated the SimAM attention mechanism in the YOLOv5 structure to improve its overall performance. The attention module in the system uses a complex set of filters to focus on a single object while having a variety of different things in our field of view. To filter out feature combinations that are helpful to the recognition of the feature, we include a 3D attention module called SimAM here. Moreover, the issue of feature misalignment brought on by the direct stacking of components with various scales may be resolved. To capitalize on the value of neurons, the SimAM module suggests an improved energy function based on neuroscience theory.

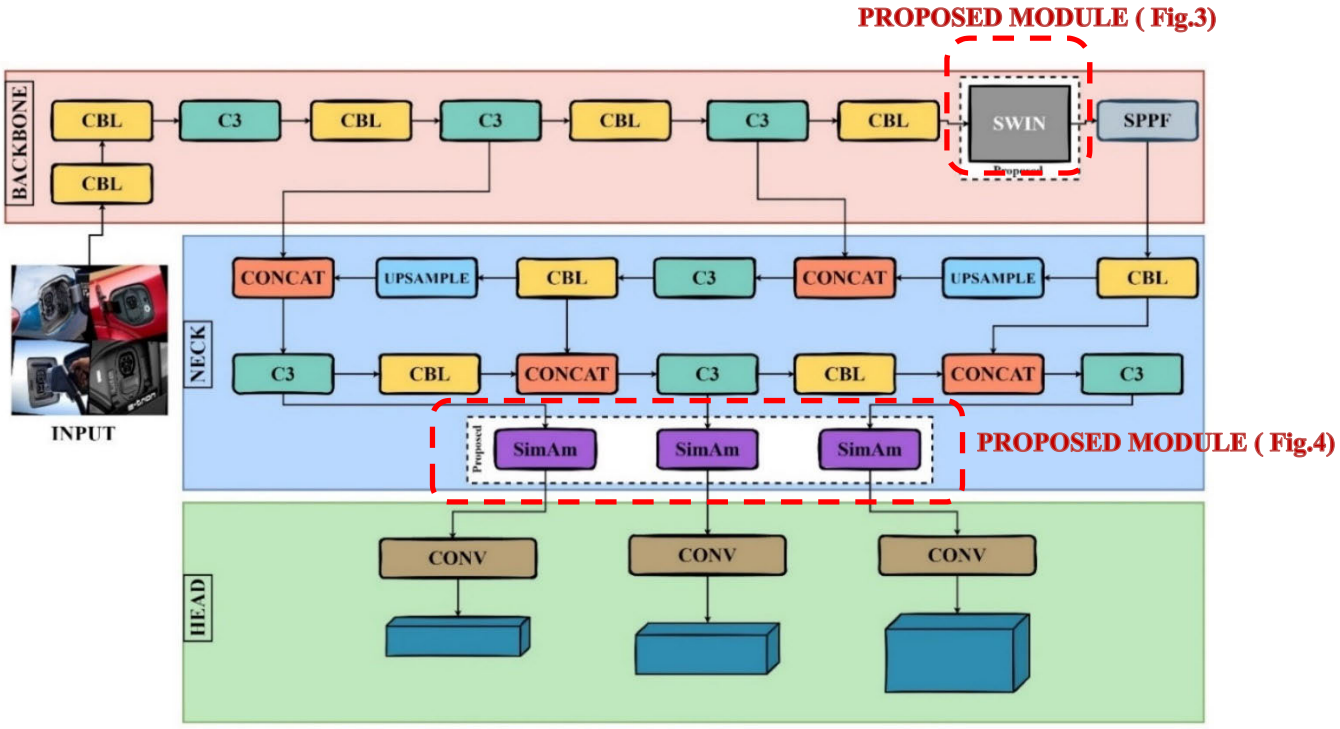

FIGURE 5. Proposed EVS-YOLO architecture Model.

It then generates an analytical solution to the energy function to expedite the calculation of attention weights. SimAM is a unique attention mechanism because of its complete usage of 3D weights and the energy function, which accelerates the computation of weights. The entire SimAM is a lightweight module as it is non-parametric; the number of parameters accounts for zero compared to the other attention mechanisms. The SimAM attention module will be placed in different places along the head, and the performance will be tabulated for each positioning. The architecture of the SimAM attention module can be seen in Figure 4. SimAM completely uses the 3d weights and the energy function to accelerate the weights calculation along with an added advantage of it being a non-parametric module Energy function of each neuron will be:

$$e_{t(w_t,b_t,y,x_i)} = (y_t - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}\left(y_o - \hat{x}_i\right)^2 \tag{8}$$

Here, $\hat{t} = w_t t + b_t$ and $\hat{x}_i = w_t x_i + b_t$ are linear transformations of t and x_i, where t is the target neuron and x_i is the other neurons in a single channel of the input feature. Final energy function will be:

$$e_{t(w_t,b_t,y,x_i)} = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(w_t x_i + b_t))^2 + (1-(w_t t + b_t))^2 + \lambda w_t^2 \tag{9}$$

where, w_t and b_t can be easily obtained by:

$$w_t = -\frac{2\,(t-\mu_t)}{(t-\mu_t)^2 + 2\sigma_t^2 + 2\lambda} \tag{10}$$

$$b_t = -\frac{1}{2}\,(t+\mu_t)\,w_t \tag{11}$$

where,

$$\mu_t = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i \tag{12}$$

$$\sigma_t^2 = \frac{1}{M-1}\sum_{i}^{M-1}(x_t - \mu_t)^2 \tag{13}$$

are mean and variance calculated for all neurons except t.

Minimal energy can be computed with:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t-\hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{14}$$

where,

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i \tag{15}$$

$$\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{\mu})^2. \tag{16}$$

The above equation (14) depicts the lower energy e_t^*, the neuron t is more unique from neighbouring neurons.

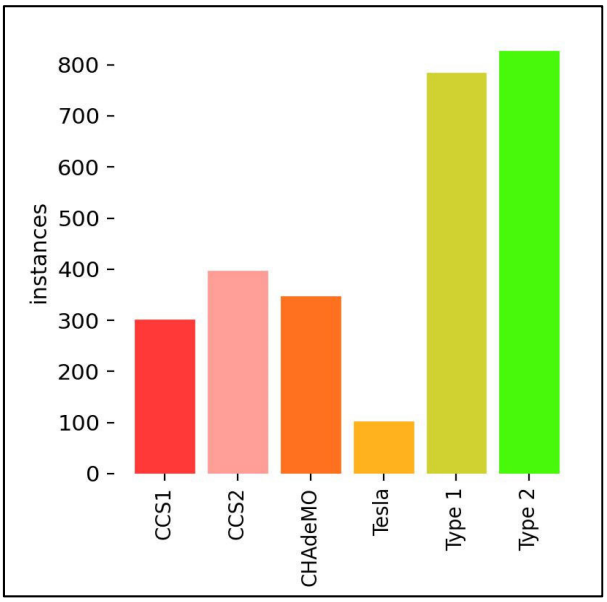

FIGURE 6. Instances of each class of the dataset.

D. PROPOSED EVS-YOLO ARCHITECTURE

In the proposed EVS-YOLO model, we use the Swin module in the backbone of the YOLOv5S architecture. Adding a transformer in the backbone improves the model's overall performance by focusing on the necessary information rather than focusing on every aspect of the image that includes the unwanted part, thus decreasing the overall accuracy and performance. Researchers have created a better network by fusing CNN with the transformer inspired by the visual transformer. The transformer works better for dense and obstructed images and scenarios. It has a more significant capacity to collect global information than CNN, thus being an essential addition to the overall network in enhancing its performance.

Further, to push the model's performance, the SimAM module is embedded before the three detection heads in the architecture. On using an attention mechanism, the EVS-YOLO model tends to focus better on the target rather than concentrating more on the unimportant features, thus enhancing the model's overall performance. Attention mechanisms can selectively attend to informative regions of the input image, which can improve the discrimination ability of the model. By listening to relevant regions, the attention module can help the YOLO model better distinguish between objects with similar features, reducing false positives or false negatives. The SimAM attention [57] is a lightweight and non-parametric algorithm. It directly uses 3D- weights, thus making it robust attention to be embedded along with the Swin Transformer [58] in the YOLO architecture. The proposed EVS-YOLO architecture is depicted in Figure 5 with SWIN transformer module embedded in the backbone of the network before the pooling operation done by the SPPF layer. The SimAM attention modules are added in the neck before passing to the head through the different detection heads.

IV. EXPERIMENT

A. DATASET

To evaluate the performance and robustness of the EVS-YOLO algorithm for detecting the EV charging socket, we used a dataset derived from two publicly available datasets consisting of over 2,500 images. The dataset consists of different EV charging sockets, and it was categorized into six different types of classes manually.

The dataset consists of 6 classes, namely CCS1, CCS2, Type1, Type2, Tesla, and CHAdeMO. All the images in the combined dataset have been manually labeled and then categorized into their respective type of sockets. The dataset is partitioned in a ratio of 8:1:1 for the training, validation, and testing sets, respectively.

Figure 6 represents the instances of different types of classes in the dataset used for the analysis of EVS-YOLO model. An overview of the dataset showing different sockets namely CCS1, CCS2, CHAdeMO, Type1, Type2 and Tesla is illustrated in Figure 7. Dataset preparation includes images with different conditions and scenarios as shown in Figure 8 for calculation the accuracy of the proposed model so as to attain the real-time implementation success containing all the different cases such as image samples with different camera angles, different brightness, weather conditions, different colors, environments and images with socket and camera - dirty, clean conditions.

B. EXPERIMENTAL AND ENVIRONMENTAL SETTINGS

To confirm the efficacy and reliability of the EVS-YOLO algorithm, we performed ablation and comparison tests on the dataset. Windows 11 is the operating system employed for the trials. The processor is an Intel 10th Gen H type. Python 3.8.13 is the programming language used. Pytorch 1.10.0 is the deep learning framework, and the acceleration environment is CUDA 11.4. For the analysis, the epochs were set at 125, the weights used were YOLOv5.pt, and the batch size was set as 16. Figure 9, the overall representation and distribution of the objects in the dataset are uniform.

V. RESULTS AND DICUSSION

To analyze the improvement that happens with adding different modules in the ES-YOLO algorithm, we performed ablation experiments that would better help us understand how adding each module improves the base model. Table 2 shows the experimental results of proposed EVS-YOLO ablation on the respective datasets. The general trend is that the base model's performance is enhanced by including a process in terms of precision, recall, mean average accuracy, and F1 score. It lags in inference speed as there are many more mathematical operations to be performed, as indicated by the Gflops column. In scheme 2, the SWIN transformer module is added to the backbone of the base network to improve the model's ability to capture long-range dependencies and contextual information. This enhanced precision by 1.6%, recall by 6.2%, mAP by 1.0%, and f1 score by 4.79%, but conversely, the inference speed increased to 7.9ms from

FIGURE 7. A Representation of different classes of EV sockets present in the dataset [18].

7.7ms in scheme 1. Gigaflops (GFlops) refer to the number of mathematical operations required for the model to have an entire pass. These metric increases, so there is an increase in inference speed. The lower the inference speed, the faster we have the result.

In scheme three, along with the Swin transformer in the backbone, SimAM attention module is added to three detection heads, allowing the neural network to selectively focus on different parts of the input data, assigning varying levels of importance to other regions or features, helping the network capture relevant information while filtering out irrelevant or redundant information. Therefore, this improves precision by 5.6%, recall by 0.5%, mAP by 1.6%, and F1 score by 2.59%, and is slightly slower as there is an increase in inference speed. On comparison of the results of scheme one and scheme 3, we can infer that the EVS-YOLO model outperforms the YOLOv5s model in all evaluation indicators. The precision reached 95.2%, the recall reached 78.7%, and the mAP and F1 scores reached 81.4% and 86.16%, respectively, with a slight increase in inference speed but still fast enough for real-time applications. Figure 10 represents the class-wise mAP comparison between the EVSYOLO and base YOLOv5s. Drastic changes is observed in the tesla and CHAdeMO classes where the base model significantly underperforms.

For other classes, the proposed model either slightly improves or has similar accuracy values. The confusion matrix of the YOLOv5s model and the proposed EVS-YOLO model is shown in Figure 11. It can be seen that the EVS-YOLO model outperforms the YOLOv5s model in every class. Table 3 represents the performance of different attention modules with a Swin transformer backbone. The other attention models are placed before the three detection heads of the structure. The proposed model comprises a SimAM attention module, a lightweight and non-parametric module, reflected in the results as the proposed model has the lowest number of parameters at 7,163,019.

TABLE 2. Experimental results of the proposed EVS-YOLO ablation on the respective dataset.

Scheme	YOLOv5s	Swin	SimAM	Precision	Recall	mAP (0.50)	F1 Score (%)	gflop	Speed (ms)
1	✓	-	-	87	72	80.8	78.79	15.8	7.7
2	✓	✓	-	89.6	78.2	77.8	83.51	55.7	7.9
3	✓	✓	✓	95.2	78.7	81.4	86.16	55.7	8.3

Compared to the next best-performing model in terms of accuracy with CBAM, which has 7207276 parameters, it is 44,257 less. The proposed model also performs the best in precision, recall, mAP, and f1 scores. It outperforms the next best model, which includes CBAM as its attention module in terms of accuracy by 16.8%, mAP by 0.3%, and f1 score by 6.3%.

It significantly outperforms when compared to the model which includes the SK Attention module, as it has five times less the number of parameters and twice as few Gflops, and it improves on precision by 5.8%, recall by5.0%, mAP by 5.7%, f1 score by 5.3%. Typically, the lower the parameters, the faster the model is in inference. The addition of SimAM improves the accuracy compared with another attention module called the Coordinate attention module by 0.8% while having 35,680 lesser parameters. It also outperforms

FIGURE 8. Sample images taken under different environmental conditions Source: ©Roboflow.com [18].

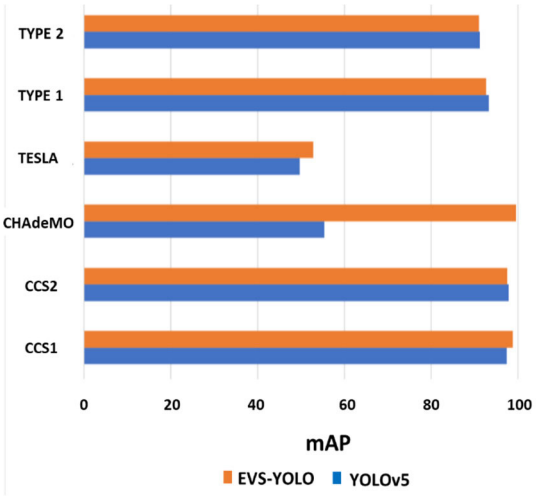

FIGURE 9. Location size and distribution of objects.

attention modules like NAM and SE attention modules in all metrics while still having fewer parameters. Figure 12 shows the class wise comparison of Mean Average Precision for

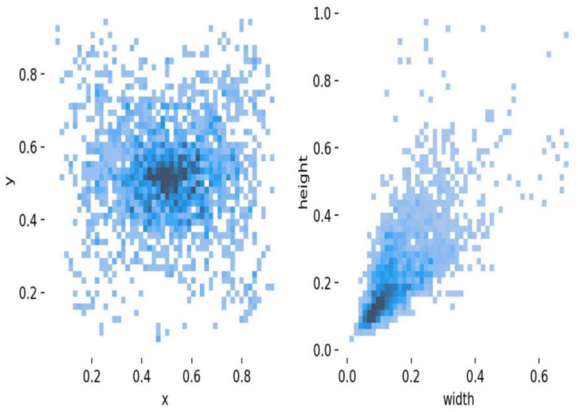

FIGURE 10. Class wise performance comparison between EVS-YOLO and YOLOv5s model.

TABLE 3. Comparison of different positioning of SimAM attention module on different detection heads on the EVS-YOLO model.

Backbone	Attentions Model	Precision	Recall	mAP (0.50)	F1 Score (%)	Param	gflop
YOLOv5 + Swin	NAM	91.7	76.8	78.2	83.59	7164811	55.7
	SE	91.6	98.1	99.4	94.73	7206027	55.8
	CA	95.8	79.1	80.6	86.65	7198699	55.8
	CBAM	78.4	80.7	81.1	79.53	7207276	55.8
	SKATTNTION	89.4	73.7	75.7	80.79	3.6E+07	108.7
	SIMAM	95.2	78.7	81.4	86.16	7163019	55.7

different attention modules with Swin Transformer embedded in the backbone. It justifies that adding the SimAM attention module improves precision for certain classes such as CCS1, CHAdeMO and Tesla. YOLOv5 uses three different detection heads, each responsible for predicting objects at different scales or resolutions in the input image. The large Detection Head predicts things at the highest resolution in the input image. It has a larger receptive field, which allows it to detect smaller objects with fine details. It outputs a tensor with higher spatial resolution and smaller object anchor boxes for detecting smaller objects. Medium Detection Head operates at an intermediate resolution in the input image and is responsible for detecting objects of medium size. It has a medium-sized receptive field, allowing it to detect moderate-sized objects. The Small Detection Head operates at the lowest resolution in the input image and detects larger objects. Its larger receptive field will enable it to detect objects of larger size with coarser details. Combining these three detection heads at different resolutions helps YOLOv5 to accurately detect objects of varying sizes and scales. Table 4 represents the results of the influence of the positioning of the SimAM attention module in these three detection heads. Since there are three detection heads, seven possible combinations of positions are likely. The best-performing model is scheme seven, where an attention module is present in all three places with a precision, recall, mAP, and f1 score

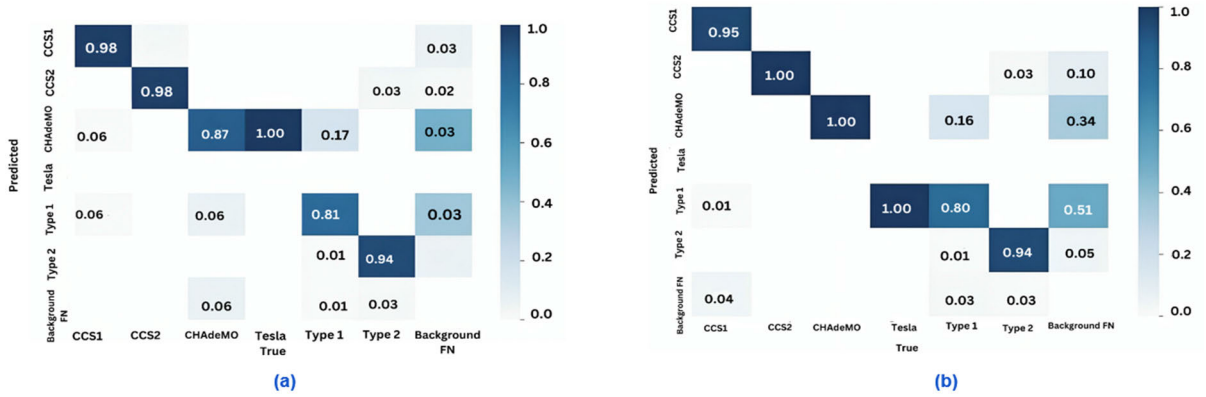

FIGURE 11. Confusion matrix obtained from (a) . YOLOv5s base model and (b). proposed EVS-YOLO model.

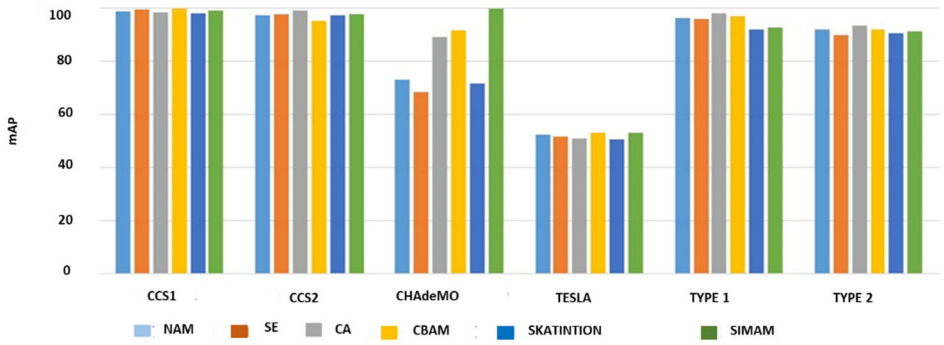

FIGURE 12. Class wise comparison of Mean Average Precision for different attention modules with Swin Transformer embedded in the backbone.

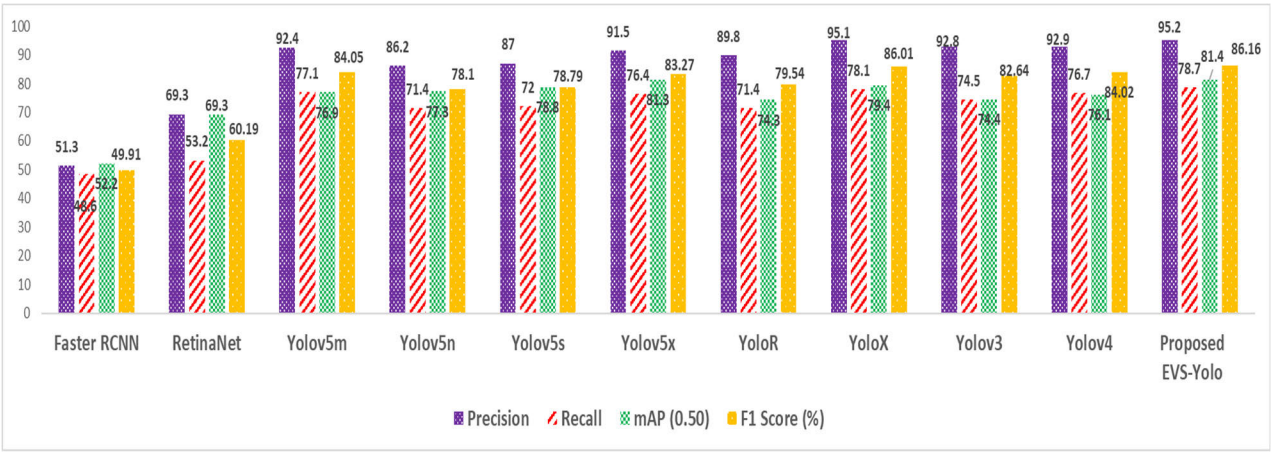

FIGURE 13. Graphical representation of various evaluation metrics of proposed EVS-YOLO model with existing object detection models.

value of 95.2%, 78.7%, 81.4%, and 86.16% as there is a combination of all three detection scales. The addition of attention modules in the small and large detection heads has similar results in terms of mAP at 80.6% in both cases. Still, including an attention method in the more giant head outperforms precision, recall, and f1 scores by 2.5%, 4.2%, and 3.56%. In terms of positioning in two detection heads, adding attention modules in the small and large detection heads gives the best outperforming the other schemes with similar positioning but still less than other positioning methods.

Table 5 compares the proposed model with other detection methods. It outperforms traditional methods like Faster RCNN and Retinanet regarding precision, recall, mAP, and F1 score by a considerable margin, and typically, stage

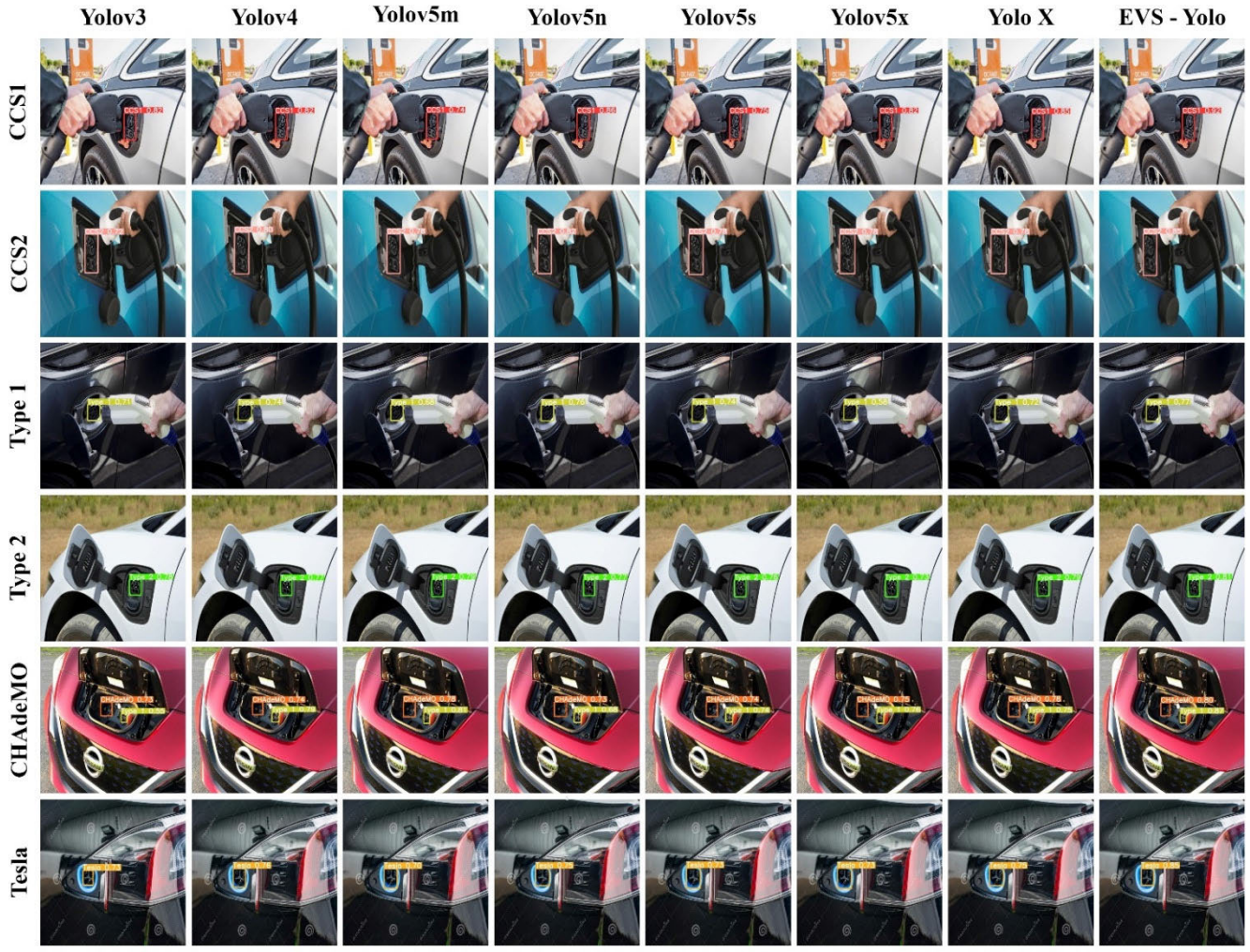

FIGURE 14. Class wise detection results of the proposed EVS-YOLO model with other object detection models.

TABLE 4. Performance comparison for different attention modules with Swin transformer embedded in the backbone.

Scheme	Attention Module Positioning			Precision	Recall	mAP (0.50)	F1 Score (%)
	Small	Medium	Large				
1	✓	-	-	90.1	72.5	80.6	80.34
2	-	✓	-	91.8	78.1	77.8	84.39
3	-	-	✓	92.6	76.7	80.6	83.9
4	✓	✓	-	92.6	75.9	79.2	83.42
5	✓	-	✓	91.3	79.4	80.1	84.93
6	-	✓	✓	76.7	79.3	79.4	77.97
7	✓	✓	✓	95.2	78.7	81.4	86.16

TABLE 5. Performance comparison of different detection models with EVS-YOLO.

Detection Models	Precision	Recall	mAP (0.50)	F1 Score (%)
Faster RCNN	51.3	48.6	52.2	49.91
RetinaNet	69.3	53.2	69.3	60.19
Yolov5m	92.4	77.1	76.9	84.05
Yolov5n	86.2	71.4	77.3	78.1
Yolov5s	87	72	78.8	78.79
Yolov5x	91.5	76.4	81.3	83.27
YoloR	89.8	71.4	74.3	79.54
YoloX	95.1	78.1	79.4	86.01
Yolov3	92.8	74.5	74.4	82.64
Yolov4	92.9	76.7	76.1	84.02
Proposed EVS-Yolo	95.2	78.7	81.4	86.16

methods are faster than two-stage methods, so this is an added advantage. Compared with other YOLO models, it outperforms YOLOv3 by 7% in terms of mAP and has a better F1 score, which is 3.7% better than YOLOv3. Its successor, YOLOv4, is slightly improved over YOLOv3 but still needs to catch up by 5.3% in average precision. Even though the YOLOX model has a better F1 score, it still needs to improve in accuracy by 2.0%.

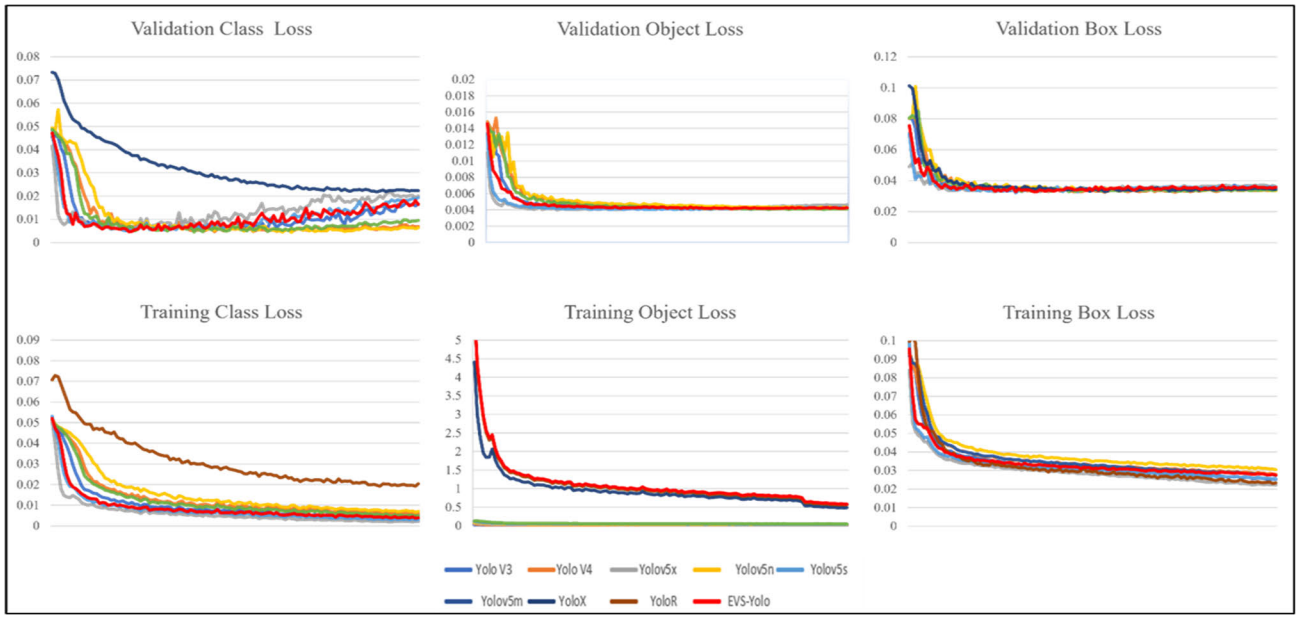

FIGURE 15. Performance comparison of three kinds of loss for the EVS-YOLO on the dataset used.

TABLE 6. Inference speed and parameters comparison of EVS-YOLO model with other detection models.

Detection Models	Params	gflops	speed (ms)
Yolov5m	20873139	47.9	13.9
Yolov5n	1772035	4.2	6
Yolov5s	7026307	15.8	7.7
Yolov5x	86207059	203.9	24.1
YoloR	36864968	80.6	16.2
YoloX	8042561	21.6	9.8
Yolov3	61524355	154.6	20.5
Yolov4	9124115	20.6	9.6
EVS-Yolo	7163019	55.7	8.3

When compared to other YOLOv5 models, the proposed model outperformsv5n, v5s, and v5m with regards to mean average precision by 4.1%, 2.6 %,4.5%, and f1 scores by 8.06%, 7.37%,2.11%. The YOLOv5x model outperforms the proposed model in terms of mAP by 1.9%. Inference speed and parameters comparison of proposed EVS-YOLO model and other detection models are tabulated in Table 6.

Even though the proposed model has significantly fewer parameters, it performed better compared to most one-stage methods, and it also performs well with inference speed, with only YOLOv5s and YOLOv5n having lesser inference speed owing to their more secondary parameters. Figure 13 depicts the graphical representation of various evaluation metrics such as precision, recall, mAP, and F1 score of the proposed EVS-YOLO model with existing object detection models. The proposed EVS Yolo model outperforms other models, as the graph shows. The subjective detection results of the proposed EVS-YOLO model with other object detection models for all six classes shown in Figure 14.

Figure 15 shows the performance comparison of all the losses in the training and validation set, such as bounding box, class, and object loss for the proposed EVS-YOLO on the dataset used. The first row is the loss of the validation set. The three figures in the first row from left to right are class loss, object loss, and box loss. The second row is the loss of the training set for the EVS-YOLO model on the dataset used, and the three figures in the second row from the left are the same as hose mentioned in the first row. Be it the validation set or the training set, the loss tends to decrease and eventually stabilize.

Despite the curve being close, when we compare the bounding box set of the validation loss, EVS-YOLO is comparatively low compared to the YOLOv5s, i.e., the base model. YOLOv5 loss stands at 0.0048 for 125 epochs, whereas EVS-YOLO is at 0.0042 for 125 periods, which means the proposed model accelerates the speed and tends to converge a lower loss value.

The terms ''bounding box loss,'' ''class loss,'' and ''object loss'' are used in object detection with deep learning to represent different types of losses that are computed during the model's training and validation stages. The bounding box loss is a specific type of loss that assesses the error in anticipating the bounding box coordinates of an object. The model is penalized when the expected coordinates differ from the actual coordinates, and the loss is frequently calculated utilizing a regression loss function like mean squared error or smooth L1 loss.

On the other hand, class loss is a type of loss that measures the model's error in predicting the object's class inside the bounding box. The model is punished when the predicted label varies from the true label. The loss is typically calculated using a classification function such as cross-entropy loss. Lastly, object loss is a type of loss that quantifies the model's error in detecting the object inside a boundary. The model is punished when it fails to detect an object's presence inside a bounding box or mistakenly detects a false positive. The loss is generally calculated using a binary classification function such as binary cross-entropy loss. During the training phase, the model tries to minimize the overall loss, an aggregated sum of the bounding box, class, and object loss. The weights are commonly chosen based on the relative importance of each loss term. During the validation phase, the loss terms are computed on a separate set of validation data to determine the model's performance. The end goal is to minimize the overall training loss and validation sets to achieve a high level of generalization performance.

VI. CONCLUSION AND FUTURE WORKS

An object detection algorithm called EVS-Yolo has been proposed to address the problems of identification of Electric Vehicle charging sockets, which hurdles the experience of the charging port efficiently for the users. We introduced a SimAM attention mechanism in the backbone section to enhance the network's ability to aggregate features and focus more on the object than the background. SimAM attention was applied in three positions across the proposed architecture to find the best possible result. We also embedded the Swin transformer module into the backbone part of the network, which can extract contextual features. To verify the satisfactory work of the algorithm, ablation experiments were performed. The experiment results show that the EVS-YOLO algorithm achieves an average detection accuracy of 81.4% on the test set, an improvement of 2.6%. The EVS-YOLO model achieves an inference speed of 8.3ms, attaining the accuracy and requirements for detecting Electric vehicle sockets. The EVSYOLO method is more suitable for this application than other object detection algorithms due to higher accuracy and real-time inference. It represents a significant advancement in the field of electric vehicle infrastructure. This innovative solution offers precise and efficient socket detection and showcases the potential for cutting-edge deep-learning techniques to revolutionize how we interact with and utilize electric vehicles. As we move towards a more sustainable future, this technology paves the way for enhanced user experience and increased accessibility to electric vehicle charging, ultimately accelerating the adoption of clean energy transportation solutions.

Further, the EVS-YOLO algorithm can apply to real-time samples with an automated application, which might act as a catalyst for improving the overall user Experience. Several avenues for future work can be explored, such as Integrating real-time data sources, availability, and charging rates to provide users with up-to-date information on the status of charging sockets, enhancing the user experience. Integrate energy management algorithms to optimize charging based on user preferences, electricity prices, and grid conditions, promoting efficient and sustainable charging practices. Implement security protocols to protect against potential cyber threats, ensuring the safety and integrity of electric vehicle charging transactions.

REFERENCES

[1] Z. Liu, H. Hao, X. Cheng, and F. Zhao, "Critical issues of energy efficient and new energy vehicles development in China," *Energy Policy*, vol. 115, pp. 92–97, Apr. 2018.
[2] D. Guo, W. Yan, X. Gao, Y. Hao, Y. Xu, X. Tan, and T. Zhang, "Forecast of passenger car market structure and environmental impact analysis in China," *Sci. Total Environ.*, vol. 772, Jun. 2021, Art. no. 144950.
[3] P. Yu, J. Zhang, D. Yang, X. Lin, and T. Xu, "The evolution of China's new energy vehicle industry from the perspective of a technology–market–policy framework," *Sustainability*, vol. 11, no. 6, p. 1711, Mar. 2019.
[4] Z. Chen and X. Huang, "Challenges and opportunities for the development of EV in large scale," *J. Electr. Eng.*, vol. 10, no. 4, pp. 35–44, 2015.
[5] J. Finnerty, "Where is the charge port on my electric car," Grid serve Sustain. Energy Ltd, U.K., Tech. Rep., Mar. 2023.
[6] J. Dixon, P. B. Andersen, K. Bell, and C. Træholt, "On the ease of being green: An investigation of the inconvenience of electric vehicle charging," *Appl. Energy*, vol. 258, Jan. 2020, Art. no. 114090.
[7] C. He, J. Chen, Q. Feng, X. Yin, and X. Li, "Safety analysis and solution of electric vehicle charging," *Distrib. Utility*, vol. 34, pp. 12–18, 2017.
[8] Tesla Develops. *Snake-Shaped, Metal Charging Robot That can Automatically Charge*. Accessed: Jan. 15, 2022. [Online]. Available: http://www.xinhuanet.com/world/2015-8/10/c_128111025.htm
[9] Coming! The Latest Development of the KUKA and Volkswagen Group. *E-Smart Connect*. Accessed: Jan. 16, 2022. [Online]. Available: https://www.imrobotic.com/news/detail/5755
[10] Y. Shi, "Research on robot-based electric vehicle charging system and its automatic plugin," Dept. Phys., Harbin Inst. Technol., Harbin, China, Tech. Rep., 2016.
[11] L. Ma, "Ultrasonic-based electric vehicle charging port positioning technology," Dept Inf. Sci. Eng., Harbin Inst. Technol., Harbin, China, Tech. Rep., 2018.
[12] L. Y. Zhao, "Research on the classification, identification and deblurring algorithm of electric vehicle charging ports based on deep learning," M.S. thesis, Harbin Inst. Technol., Dept. Inf. Sci. Eng., Harbin, China, 2021.
[13] S. Haddadin, A. De Luca, and A. Albu-Schäffer, "Robot collisions: A survey on detection, isolation, and identification," *IEEE Trans. Robot.*, vol. 33, no. 6, pp. 1292–1312, Dec. 2017.
[14] KoreaTechDesk. (Nov. 22, 2018). *EVAR: Samsung Electronics' Spinoff Brings an Autonomous Robotic Charger for Electric Vehicles*. Accessed: Jul. 29, 2019. [Online]. Available: https://koreatechdesk.com/evar-samsungelectronics-spinoff-brings-an-autonomous-robotic-charger-forelectric-vehicles/
[15] *Hyundai Introduces Robots to Charge Your EV*, Carpro, Amsterdam, The Netherlands, Mar. 2023.
[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, 2005, pp. 886–893, doi: 10.1109/CVPR.2005.177.
[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Jan. 2004, doi: 10.1023/B:VISI.0000029664.99615.94.
[18] B. Dwyer, J. Nelson, and J. Solawetz. (2022). *Roboflow (Version 1.0)*. [Online]. Available: https://roboflow.com.computer vision.
[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016, doi: 10.1109/TPAMI.2015.2437384.
[20] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, CL, USA, Jun. 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.

[21] E. Güney and C. Bayilmiş, "An implementation of traffic signs and road objects detection using faster R-CNN," *Sakarya Univ. J. Comput. Inf. Sci.*, vol. 5, no. 2, pp. 216–224, Aug. 2022.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[23] W. Liu, "SSD: Single shot MultiBox detector," in *Computer Vision—ECCV 2016* (Lecture Notes in Computer Science), vol. 9905, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, doi: 10.1007/978-3-319-46448-0_2S.

[24] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, 2017, pp. 6517–6525.

[25] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[26] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[28] Z. Zhong, Z. Q. Lin, R. Bidart, X. Hu, I. B. Daya, Z. Li, W.-S. Zheng, J. Li, and A. Wong, "Squeeze-and-attention networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13065–13074.

[29] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[30] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13713–13722.

[31] Z. Cao, T. Liao, W. Song, Z. Chen, and C. Li, "Detecting the shuttlecock for a badminton robot: A YOLO based approach," *Exp. Syst. Appl.*, vol. 164, Feb. 2020, Art. no. 113833.

[32] R. M. Fikri, B. Kim, and M. Hwang, "Waiting time estimation of hydrogen-fuel vehicles with YOLO real-time object detection," in *Information Science and Applications*. Singapore: Springer, 2020, pp. 229–237.

[33] Y. Jamtsho, P. Riyamongkol, and R. Waranusast, "Real-time bhutanese license plate localization using Yolo," *ICT Exp.*, vol. 6, no. 2, pp. 121–124, Jun. 2020.

[34] E. S. Kalhagen and L. O. Ørjan, "Hierarchical FSH species detection in real-time video using YOLO," M.S. thesis, Faculty Comput. Math. Sci., Univ. Agder-Kristiansand Campus, Kristiansand, Norway, 2020.

[35] P. Mohd and P. A. Nurul, "A real-time trafc sign recognition system for autonomous vehicle using YOLO," Universiti Teknologi MARA, Shah Alam, Selangor, Cawangan Melaka, Tech. Rep. 35625, 2020.

[36] P. Ren, "A novel squeeze YOLO-based real-time people counting approach," *Int. J. Bio-Inspired Comput.*, vol. 16, no. 2, pp. 94–101, 2020.

[37] R. Shi, L. Tianxing, and Y. Yasushi, "An attribution-based pruning method for real-time mango detection with YOLO network," *Comput. Electron. Agricult.*, vol. 169, Feb. 2020, Art. no. 105214.

[38] J. Wang, N. Wang, L. Li, and Z. Ren, "Real-time behavior detection and judgment of egg breeders based on YOLO v3," *Neural Comput. Appl.*, vol. 32, no. 10, pp. 5471–5481, May 2020.

[39] Y. Tian, G. Yang, Z. Wang, E. Li, and Z. Liang, "Detection of apple lesions in orchards based on deep learning methods of CycleGAN and YOLOV3-dense," *J. Sensors*, vol. 2019, pp. 1–13, Apr. 2019, doi: 10.1155/2019/7630926.

[40] Y. Cai, T. Luan, H. Gao, H. Wang, L. Chen, Y. Li, M. A. Sotelo, and Z. Li, "YOLOv4-5D: An effective and efficient object detector for autonomous driving," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021, doi: 10.1109/TIM.2021.3065438.

[41] H. Fu, G. Song, and Y. Wang, "Improved YOLOv4 marine target detection combined with CBAM," *Symmetry*, vol. 13, no. 4, p. 623, Apr. 2021, doi: 10.3390/sym13040623.

[42] E. Güney, C. Bayilmis, and B. Çakan, "An implementation of real-time traffic signs and road objects detection based on mobile GPU platforms," *IEEE Access*, vol. 10, pp. 86191–86203, 2022.

[43] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Montreal, BC, Canada, Oct. 2021, pp. 2778–2788, doi: 10.1109/ICCVW54120.2021.00312.

[44] Y. Wang, W. Guo, S. Zhao, B. Xue, W. Zhang, and Z. Xing, "A big coal block alarm detection method for scraper conveyor based on YOLOBS," *Sensors*, vol. 22, no. 23, p. 9052, 2022, doi: 10.3390/s22239052.

[45] P. Zhao, X. Chen, S. Tang, Y. Xu, M. Yu, and P. Xu, "Fast recognition and localization of electric vehicle charging socket based on deep learning and affine correction," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Jinghong, China, Dec. 2022, pp. 2140–2145, doi: 10.1109/ROBIO55434.2022.10011985.

[46] M. Pan, C. Sun, J. Liu, and Y. Wang, "Automatic recognition and location system for electric vehicle charging port in complex environment," *IET Image Process.*, vol. 14, no. 10, pp. 2263–2272, Aug. 2020, doi: 10.1049/iet-ipr.2019.1138.

[47] H. Zhang and X. Jin, "A method for new energy electric vehicle charging hole detection and location based on machine vision," in *Proc. 5th Int. Conf. Environ., Mater., Chem. Power Electron.*, 2016, p. 84, 10.2991/emcpe-16.2016.84.

[48] J. Miseikis, M. Ruther, B. Walzel, M. Hirz, and H. Brunner, "3D vision guided robotic charging station for electric and plug-in hybrid vehicles," 2017, *arXiv:1703.05381*.

[49] T. Li, C. Xia, M. Yu, P. Tang, W. Wei, and D. Zhang, "Scale-invariant localization of electric vehicle charging port via semi-global matching of binocular images," *Appl. Sci.*, vol. 12, no. 10, p. 5247, 2022.

[50] E. Guney, I. H. Sahin, S. Cakar, O. Atmaca, E. Erol, M. Doganli, and C. Bayilmis, "Electric shore-to-ship charging socket detection using image processing and YOLO," in *Proc. Int. Symp. Multidisciplinary Stud. Innov. Technol. (ISMSIT)*, Oct. 2022, pp. 1069–1073.

[51] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[52] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 390–391.

[53] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018, doi: 10.1016/j.patcog.2017.10.013.

[54] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: Challenges, architectural successors, datasets and applications," *Multimedia Tools Appl.*, vol. 82, no. 6, pp. 9243–9275, Mar. 2023, doi: 10.1007/s11042-022-13644-y.

[55] A. Dirir, H. Ignatious, H. Elsayed, M. Khan, M. Adib, A. Mahmoud, and M. Al-Gunaid, "An advanced deep learning approach for multi-object counting in urban vehicular environments," *Future Internet*, vol. 13, no. 12, p. 306, Nov. 2021, doi: 10.3390/fi13120306.

[56] D. Zhou, J. Fang, X. Song, C. Guan, J. Yin, Y. Dai, and R. Yang, "IoU loss for 2D/3D object detection," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2019, pp. 85–94.

[57] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter free attention module for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11863–11874.

[58] 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. 10012–10022.

[59] M. Jiang, T. Hai, Z. Pan, H. Wang, Y. Jia, and C. Deng, "Multi-agent deep reinforcement learning for multi-object tracker," *IEEE Access*, vol. 7, pp. 32400–32407, 2019.

[60] Z. Zhou, L. Li, R. Wang, and X. Zhang, "Deep learning on 3D object detection for automatic plug-in charging using a mobile manipulator," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 4148–4154.

[61] M. Shibl, L. Ismail, and A. Massoud, "Electric vehicles charging management using machine learning considering fast charging and vehicle-to-grid operation," *Energies*, vol. 14, no. 19, p. 6199, Sep. 2021, doi: 10.3390/en14196199.

[62] J.-H. Park, K. Farkhodov, S.-H. Lee, and K.-R. Kwon, "Deep reinforcement learning-based DQN agent algorithm for visual object tracking in a virtual environmental simulation," *Appl. Sci.*, vol. 12, no. 7, p. 3220, Mar. 2022, doi: 10.3390/app12073220.

[63] H. Lin, P. Quan, Z. Liang, Y. Lou, D. Wei, and S. Di, "Collision localization and classification on the end-effector of a cable-driven manipulator applied to EV auto-charging based on DCNN–SVM," *Sensors*, vol. 22, no. 9, p. 3439, Apr. 2022, doi: 10.3390/s22093439.

[64] M. ElKashlan, M. S. Elsayed, A. D. Jurcut, and M. Azer, "A machine learning-based intrusion detection system for IoT electric vehicle charging stations (EVCSs)," *Electronics*, vol. 12, no. 4, p. 1044, Feb. 2023, doi: 10.3390/electronics12041044.
[65] V. Karanam and G. Tal, "Developing a deep learning tool to detect electric vehicle supply equipment failures," in *Proc. EVS36 Int. Electr. Vehicle Symp. Exhib.*, Sacramento, CA, USA, Jun. 2023.

V. C. MAHAADEVAN received the B.E. degree in mechanical engineering from the University of Madras, the M.B.A. degree in marketing from Pondicherry University, and the M.Tech. degree in CAD from SRM University. He is currently pursuing the Ph.D. degree with the Department of Electrical and Electronics Engineering, SRM Institute of Science and Technology, Chennai. He is having more than 23 years of experience in automotive industry with quality management system, product development, and supplier chain management. He stared carrier in Tier 2 industry and moved to Hyundai ancillary for interior and seating system. He is also with Renault Nissan as a Global Benchmark Pilot and a Team Leader for body in white costing team.

R. NARAYANAMOORTHI (Member, IEEE) received the bachelor's degree in electrical engineering and the master's degree in control and instrumentation from Anna University, India, in 2009 and 2011, respectively, and the Ph.D. degree from the SRM Institute of Science and Technology, India, in 2019. He is currently an Associate Professor with the Department of Electrical and Electronics Engineering, SRM Institute of Science and Technology. His research interests include wireless power transfer, electric vehicles, power electronics, artificial intelligence and machine learning in renewable energy systems, and embedded system for smart sensors.

RADOMIR GONO (Senior Member, IEEE) received the M.Sc., Ph.D., Ph.D. Habilitate, and Professor degrees in electrical power engineering, in 1995, 2000, 2008, and 2019, respectively. He has been with the Department of Electrical Power Engineering, VSB-Technical University of Ostrava, Czech Republic, since 1999, where he is currently a Professor and the Vice-Head of the Department. His current research interests include electric power systems reliability, the optimization of maintenance, and renewable energy sources.

PETR MOLDRIK received the Graduate degree in electrical power engineering from the Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, Czech Republic, in 2003, and the Ph.D. degree in electrotechnics, communication and computer engineering, in 2008. Since 2003, he has been with the Department of Electrical Power Engineering. His research interests include the issue of the application of hydrogen fuel cells and related technologies, especially in the energy sector, in the field of hybrid electric vehicles.