IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 July 2024, accepted 8 August 2024, date of publication 16 August 2024, date of current version 26 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3442484

RESEARCH ARTICLE

Super-Resolution GAN and Global Aware Object Detection System for Vehicle Detection in Complex Traffic Environments

HONGQING WANG[1], JUN KIT CHAW[1], SIM KUAN GOH[2], (Senior Member, IEEE), LIANTAO SHI[3], TING TIN TIN[4], NANNAN HUANG[1], AND HONG-SENG GAN[5]

[1]Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia
[2]School of Electrical Engineering and Artificial Intelligence, Xiamen University Malaysia, Sepang, Selangor 43900, Malaysia
[3]Institute for Carbon-Neutral Technology, Shenzhen Polytechnic University, Shenzhen 518055, China
[4]Faculty of Data Science and Information Technology, INTI International University, Nilai, Negeri Sembilan 71800, Malaysia
[5]School of Al and Advanced Computing, Xi'an Jiaotong–Liverpool University, Suzhou, Jiangsu 215400, China

Corresponding author: Jun Kit Chaw (chawjk@ukm.edu.my)

This work was supported in part by Geran Galakan Penyelidik Muda under Grant GGPM-2022-063, and in part by INTI International University.

ABSTRACT Intelligent vehicle detection systems have the potential to improve road safety and optimize traffic management. Despite the continuous advancements in AI technology, the detection of different types of vehicles in complex traffic environments remains a persistent challenge. In this paper, an end-to-end solution is proposed. The image enhancement part proposes a super-resolution synthetic image GAN (SSIGAN) to improve detection of small, distant objects in low-resolution (LR) images. An edge enhancer (EE) and a hierarchical self-attention module (HS) are applied to address the loss of high-frequency edge information and texture details in the super-resolved images. The output super-resolution (SR) image is fed into detection part. In the detection part, we introduce a global context-aware network (GCAFormer) for accurate vehicle detection. GCAFormer utilizes a cascade transformer backbone (CT) that enables internal information interaction and generates multi-scale feature maps. This approach effectively addresses the challenge of varying vehicle scales, ensuring robust detection performance. We also built in a cross-scale aggregation feature (CSAF) module inside GCAFormer, which fuses low- and high-dimensional semantic information and provides multi-resolution feature maps as input to the detection head, so as to make the network more adaptable to complex traffic environments and realize accurate detection. In addition, we validate the effectiveness of our proposed method on a large number of datasets, reaching 89.12% mAP on the KITTI dataset, 90.62% on the IITM-hetra, 86.83% on the Pascal VOC and 93.33% on the BDD-100k. The results were compared to SOTA and demonstrated the competitive advantages of our proposed method for Vehicle Detection in complex traffic environments.

INDEX TERMS Intelligent vehicle detection, self-attention, multi-scale semantic feature, generative adversarial network, feature aggregation, transportation.

I. INTRODUCTION

Nowadays, with the rising number of vehicles and an overburdened transportation system, we are facing problems such as longer traffic waiting times, heightened environmental pollution, and higher accident rate. Therefore, the development of intelligent vehicle detection has become a top priority for solving transportation problems [57], [58], [59]. Nonetheless, the utilization of existing vehicle detection techniques in intricate traffic conditions is constrained, primarily due to the low resolution of the current vehicle dataset and the absence of detailed feature information for smaller targets. In crowded urban environments, small vehicles and pedestrians are a crucial part of the traffic. However, due to their small size

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

and less distinctive features, these small objects may appear insignificant. They may be affected by factors such as masking, blurring, and light interference. For instance, small-target vehicles at a distance, as they approach each other, some of which blend in with the color of the highway, the surrounding green belt, or the color of pedestrians' clothing, can result in missed detections and false detections. In many practical applications such as UAV monitoring, video monitoring, and intelligent security, image resolution is often reduced to cut costs and enhance real-time detection performance. However, this can lead to a loss of detail in the image and reduce the feature information that the network can learn, thus affecting detection accuracy. Detection in complex environments is challenging, and timely and accurate detection of these objects is crucial to prevent accidents and improve road safety. In applications that require a real-time response, such as path planning for emergency vehicles, the ability to quickly and accurately detect objects of all sizes can provide vital information for decision support systems [1], [48].

The development of intelligent vehicle detection technology can effectively solve a series of problems due to the increase in the number of vehicles, such as longer waiting time, heightened environmental pollution and increased accident rate. Among them, the accuracy of target detection plays an important role in solving these problems. Longer waiting time in traffic congestion can be solved by accurately detecting the position and speed of vehicles and uploading the real-time data to the traffic management center, so as to optimize the timing of traffic signals and reduce the waiting time of vehicles at intersections. In addition, traffic flow can be predicted and traffic strategies can be established to avoid congestion. In terms of environmental pollution, vehicle target detection can be reduced by reducing ineffective waiting and idling time. Vehicle target detection can measure the distance between vehicles and the stopping time of vehicles through real-time data, which can be fed back to the traffic management center to issue timely warnings. In terms of traffic accidents, the vehicle detection system can also identify illegal behaviors, such as speeding and running red lights, and reduce the incidence of accidents through punitive measures. Therefore, an accurate vehicle detection model is essential for the development of intelligent vehicle detection technology [52]. However, in practice, image quality is a key factor affecting the accuracy of vehicle detection, so super-resolution technology plays a crucial role in traffic image processing, especially in small target detection. Traffic images usually have low imaging resolution due to the limitation of hardware cost or noise interference during transmission. Super-resolution techniques improve the resolution by increasing the pixel density of the image. This means that the image contains more pixel points within the same physical size, which makes the image more detailed [13], [53]. For small targets in vehicle images, the increase in resolution can make the features of the target more obvious for recognition and classification by detection model. In summary, this research is devoted to the development of a system capable of generating SR vehicle images and detecting vehicles in complex traffic environments.

LR images are typically 72 to 150 pixels per inch. Common LR images include 320 × 240 pixels and 640 × 480 pixels. At present, some industries that require real-time monitoring and transmission will choose relatively LR monitoring equipment due to cost considerations, which may lead to poor monitoring quality and is not conducive to the real-time transmission and processing of images. HR refers to the presence of a large number of pixels in an area of the image, usually, images with a resolution of 800 × 600 pixels or more are considered HR. HR images have rich detail information, high quality, and high clarity [56].

Most of the existing vehicle datasets consist of low resolution (LR) images which are often disturbed by noise. This leads to poor detection performance, especially when the objects are very small [2]. Even in high resolution (HR) images, small objects are not detected as well as large objects [3]. For improving the accuracy of small target detection at a distance, some researchers have used CNN-based techniques to generate super-resolution (SR) images, which are then used for target detection [4], [5]. GAN are currently the main method for SR generation, and models such as SRGNN and ESRGAN have shown excellent performance when applied to both enhanced noise and noise-free low-resolution (LR) images [6], [7]. These two models consist of two sub-modules: the generator (G) and the edge enhancer (EE). The generator (G) generates a high resolution (HR) image from the LR image, while the discriminator (D) determines whether the generated image is a true HR image or a G-processed LR image. While this approach improves the LR image and makes it look realistic, some of the high-frequency details and edge information may not correspond to the actual HR image labels. Some studies have also demonstrated that edge information is a key feature for object detection. Thus, enhancing the high-frequency and edge feature information is a key method to improve the detection accuracy [8].

The generation of SR images greatly contributes to enhanced object detection. In the initial stages of research, features were extracted from vehicle images using traditional image processing (TIP) methods, which are time consuming and their performance is adversely affected by background noise [9]. Currently, there are several methods that use CNNs as a backbone to extract features and utilize the detection head for regression prediction. These methods have superior performance over TIP methods [10]. The CNN-based methods are mainly of two types: single-stage and two-stage. Single-level methods segment the image into grid cells and then use regression to predict the location and class of vehicles. Although single-stage provides good operation speed, it faced with the variable scale of the vehicle image and small target features, there is the problem of detecting the vehicle inaccurately [38]. Although single-stage provides good operation speed, it suffers from inaccurate detection in the face of vehicle images and small target features with variable scales. The two-stage method determines the head region of interest (ROI) and performs detection within each ROI. This improves the accuracy of the detection but also increases the complicated of the network and decreases the prediction speed [41]. Some works proposed feature pyramid networks (FPNs), embedded between the backbone and the detection head, for extracting multi-scale feature information [11]. However, FPN performs feature fusion from top to bottom paths, and although it can provide multiscale features, this introduces background noise to the high-dimensional feature maps. Vision-Transformer has been proposed to realize the intention of local features to correlate with the global, but it also results in exponential growth in computation, which affects the training time of the model and increases the complexity of the model [12]. In this paper, we proposed two networks: SSIGAN that generates SR images and GCAFormer that performs the detection for a vehicle detection system. The proposed method takes LR image as input to SSIGAN to get SR image which is then fed into GCAFormer for detection. The main contributions of SSIGAN and GCAFormer are as follows:

- The SSIGAN model is proposed to reconstruct the SR image from LR image and enhance the traffic feature information in the image for down-streaming detection tasks.
- An HS module is embedded in generator (G) and edge enhancer (EE)is designed to effectively enhance the edge information in the image to generate sharper SR images with detailed textures.
- The GCAFormer model is proposed with a cascade self-attention backbone, which can efficiently realize the information interchange between global features and finally generate multi-resolution feature maps.
- The proposed CSAF module, which is capable of bi-directional path fusion of low-dimensional and high-dimensional semantic information and noise suppression by self-adaptive (SA) block, provides accurate multi-scale feature information for vehicle detection.

II. RELATED WORK

In this section, the previous research work on vehicle detection is discussed. The contents are divided into two parts, one focuses on introducing SR image generation techniques while other part discusses the recent advances in target detection modeling.

A. SUPER-RESOLUTION IMAGE

Many existing methods used convolutional neural networks for SR image generation, such as SRCNN to enhance LR images with end-to-end training [13]. Related researchers introduced densely connected networks and residual networks to improve SR generation. Liebel et al. proposed a deep CNN for SR network applied to remote sensing images [14]. Jiang et al. introduced an edge enhancement network based on the network architecture of GAN in order to facilitate the acquisition of smooth edge information [8]. Bai et al. used a CNN-based image enhancement approach along with a single-stage detection model for simultaneous processing [15]. Ji et al. also proposed a method for SR image generation and vehicle detection on remotely sensed images [5].

B. OBJECT DETECTION

Early vehicle detection methods relied on background subtraction and template matching of images, but the accuracy and speed of these methods were low due to background variations and noise [16]. Later, researchers started using manual features such as shape context, Haar-like features, HOG, and SIFT for vehicle detection [17], [18]. These methods perform well in dealing with illumination changes and vehicle occlusion but are complex and time-consuming as they require manual feature extraction.

Yin et al. proposed a domain-adaptive Faster R-CNN method for vehicle detection in various types of different weather such as sunny, cloudy and snowy days. This method improves the accuracy of Faster R-CNN detection in complex environments, but it is more suitable for orderly traffic scenarios rather than chaotic and unorganized traffic [42]. Instead, Liu and his colleagues designed a single-shot multibox detector (SSMD) to achieve a balance between accuracy and speed. However, its accuracy is significantly lower when dealing with vehicles of different sizes [31].

Mao et al. proposed a vehicle detection method based on the YOLOv3 algorithm. The method uses inverse residual blocks and non-maximum suppression (NMS) to solve the problem of complex and variable vehicle features in vehicle detection. Although it runs at a commendable speed, it is limited to vehicle detection. It has poor detection performance in complex traffic environment such as vehicle shading and variable weather conditions [32] and hence is not suitable for vehicle detection in irregular traffic conditions. Junayed et al. proposed a real-time vehicle detection technique for congested metropolitan cities to detect vehicles from the front view by YOLOv3 algorithm to provide valuable data for autonomous driving [44]. This method recognizes vehicles from the front but struggles to detect occluded vehicles. Despite the impressive speed, the accuracy is greatly reduced. Roy et al. proposed a vehicle detection and counting method [38], implemented by the YOLOv4 algorithm, capable of detecting five types of vehicles.

Compared to traditional Convolutional Neural Networks (CNNs), Transformer is not constrained by a fixed receptive field and is more flexible in capturing long-range dependencies [61]. In vehicle detection, this means that different parts of a vehicle (e.g., front, rear, and wheels) can be effectively associated even if they are spatially far apart. In addition, the self-attention mechanism assigns different weights to each feature region, which helps to recognize the most important parts of the image and improves the detection accuracy. Currently transformer-based detection methods are still less

TABLE 1. Summary of vehicle detection methods.

Method	Principle of work	Advantage	limitation
TIP [16]	Background subtraction	Simple, easy to deploy, and fast.	Excessive background noise interference.
TIP [18]	The main work consists of extracting histograms, scale invariance and accelerating robust features.	Less noise interference.	Complex modeling, poor results on different resolution images.
CNN-based [31]	Regression prediction using single-stage.	Fast detection, high accuracy, and no need for hand-made features.	Without multi-scale feature information, it is not applicable to complex traffic environments.
CNN-based [38]	The method employs regression to estimate the location of the vehicle and utilizes a single-stage vehicle detection network equipped with a Feature Pyramid Network (FPN).	The detection speed is fast, and the optimization strategy includes various aspects of data processing, backbone, and loss function with different degrees of optimization.	Performs poorly when dealing with small or dense targets. Also requires significant computing resources and storage space.
Transformer-based [45]	Self-attention processing CNN backbone network, simple feedforward network final prediction.	Simple training with high accuracy.	It is easy to lead to a decrease in the speed of model training, difficult to converge and effect of detection of targets with insufficient features check.
Vision Transformer-based [46]	Multi-scale feature extraction based on FPN benchmark network and window self-attention.	Accuracy and less training time.	Computationally intensive and insensitive to small target detection.

common compared to CNNs. Just in 2020, a Transformer-based target detection model, DTER, was proposed, which takes a fixed set of objects as query key inputs, models object association global dependencies, and ultimately outputs prediction results [45]. Li et al. in 2022 proposed a simple non-hierarchical approach backbone (ViTDeT) to perform the detection task, which acts as a simple FPN for windowed attention mechanisms [46]. Zheng et al. proposed SwinNet, a cross-modal fusion model based on Swin-Transformer, for object detection. The model consists of a Swin-Transformer as a baseline network to extract discriminative hierarchical features, augmented by an attentional mechanism to bridge the gap between the two modalities, and finally utilizes edge information to highlight the detected target. Although the aforementioned model is able to achieve more accurate detection results through its own algorithmic optimization, low-resolution images with small targets still have few pixels, which makes it difficult for their features to be accurately captured [47]. Li et al. proposed a YOLOSR-IST deep learning method for small target detection in images. Based on the improved deep learning network of YOLOv5, the feature extraction and target detection capabilities are enhanced by introducing coordinate attention, high-resolution feature map fusion and Swin Transformer model [50]. Zhao et al. proposed the SatDetX-YOLO model to improve the precision of small target detection in satellite remote sensing images. It significantly improves the detection performance of vehicle targets by adopting FasterNet as the backbone network, introducing the TBDetect decoupling header, integrating the DAM attention mechanism, and applying the MPDIoU loss function, where the MPDIoU loss function optimizes the similarity between the predicted frame and the real frame through the minimum point distance, which is effective even when the bounding boxes do not overlap, thus further enhancing the accuracy and regression capability of the model. [51]. Wu et al. proposed YOLO-SE, which is an improved version of YOLOv8, specifically designed for detecting and recognizing small objects in remote sensing images. It introduces a lightweight SEConv convolution and SEF module to reduce the number of parameters. YOLO-SE also integrates the SPPFE module with the EMA attention mechanism to improve the efficiency of feature extraction and includes a dedicated prediction head for small object detection [55]. Transformer-based object detection technique demonstrates superior performance in comparison to the CNN method. The SR generation process and the detection procedure presented in this paper draw inspiration from the Transformer. Table 1 summarizes some vehicle detection methods based on TIP, CNN and Transformer.

III. METHOD

The vehicle detection system proposed in this study consists of two parts, the first one is super-resolution image generation via SSIGAN. The main framework of SSIGAN is composed of generator (G), discriminator (D) and edge enhancer (EE). We design a hierarchical self-attention module (HS) that is

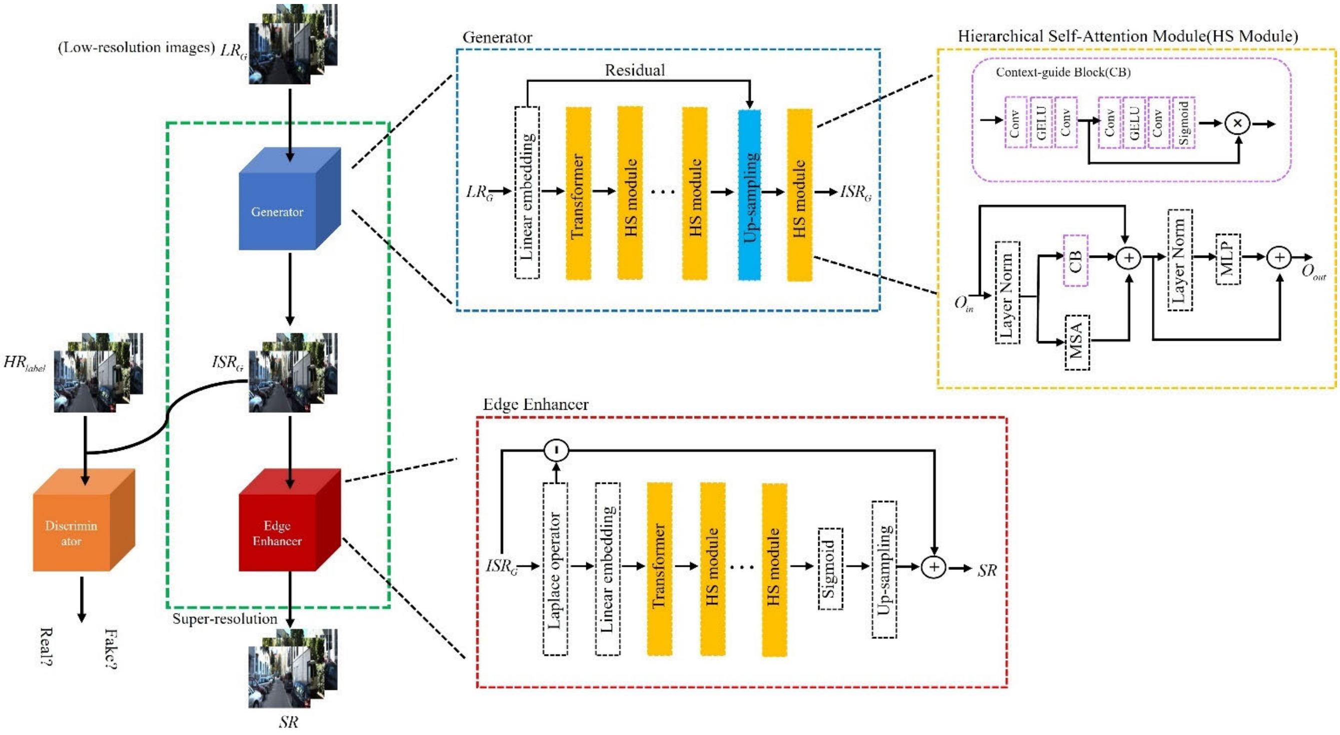

FIGURE 1. Super-resolution synthesis image with generative adversarial network (SSIGAN). SSIGAN is composed of a generator (G), an edge enhancer (EE) and a discriminator (D). The upper right part represents the flow of the generator processing the image. HS module denotes the hierarchical self-attention module, which internally includes context-guided blocks in combination with regular self-attention blocks.

embedded in both G and EE. The proposed framework for super-resolution vehicle image generation is shown in Fig. 1. The HS is designed to utilize the global context information of the Transformer, but considering the introduction of a large amount of computational and redundant information, a context correction block (CG) is designed inside the HS, which is able to model the global dependencies while reducing the redundancy information. The detection part is composed of GCAFomer and its built-in modules cascade transformer backbone (CT), Cross-Scale Aggregation Feature module (CSAF), and detection header, which are mainly designed to adapt to vehicle inspection and classification in complex environments. Fig. 2 illustrates our proposed vehicle detection model. As shown in Fig. 4, the main contribution of CT is the creation of a cascading pattern of attention modules that deeply interact with the information between patches. As inputs from stage1 to stage4 are received to obtain different resolution feature maps, resulting in CT being able to fully utilize the multi-scale information. These multi-scale feature maps are fed into the CSAF for further processing. As shown in Fig. 6, the CSAF module also employs a top-down approach to generate feature maps through a top-down approach. The module adopts a bidirectional feature fusion strategy to effectively combine high-dimensional and low-dimensional semantic information to correct each other. As shown in the Fig. 7, in the output layer of CSAF, we proposed a novel Self-Adaption Block (SA), which realizes the condensation of the outputs from different stages and effectively suppresses the background noise. The detection head receives multi-scale feature maps from the CSAF, adapts itself to the vehicle information of the complex environment, and provides accurate detection and classification results. The parts of the above two are described in detail as follows.

A. SUPER-RESOLUTION SYNTHESIS IMAGE WITH GENERATIVE ADVERSARIAL NETWORK (SSIGAN)

This study aims to improve the detection of low-resolution targets during vehicle detection. To this end, we designed a network architecture, Super Resolution Synthetic Imaging based on Generative Adversarial Networks (SSIGAN), in which the network consists of three parts: generator (G), discriminator (D) and edge enhancement (EE), and we embed a novel hierarchical self-attention module (HS) into G and EE. As shown in Fig. 1, for the original vehicle image with LR, the intermediate super-resolution image (ISR) is generated by G, and then the SR is output by EE. D receives the HR and ISR from G and the real labels, respectively, and updates the gradient backpropagation into G by calculating the computational discriminative loss, which is used to guide G training. The feature information obtained from the ISR is then edge enhanced by the EE module.

The D architecture adopts the idea of augmented super-resolution GAN [7], which removes all the Batch Normalization layers and reduced complexity of the network. The main framework of the D is VGG-19, which has significant advantages as a discriminator in Generative Adversarial

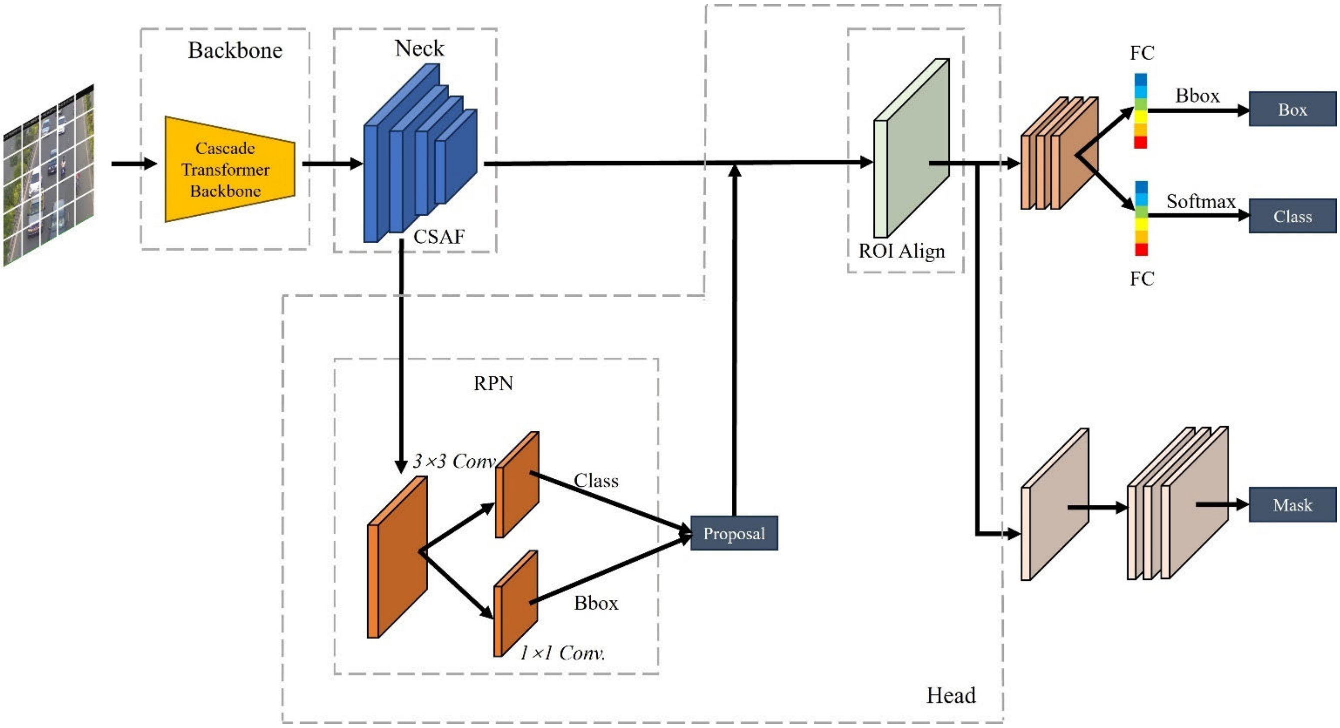

FIGURE 2. Global-Context Awareness Network for Vehicle Detection in Complex Traffic Environments (GCAFormer). GCAFormer consists of a backbone, a neck and a head. The backbone is improved the Cascade Transformer (CT) backbone. The neck features the Cross-scale Aggregate Feature (CSAF) module designed by us, and the head includes RPN and ROI components.

Networks (GANs). The VGG-19 deep architecture offers advanced feature extraction capabilities, crucial for differentiating between authentic and synthetic images. Its robustness ensures consistent and reliable discriminative outcomes across a diverse array of image processing tasks. Moreover, VGG-19 exhibits strong generalization abilities, enabling it to effectively distinguish images that were not part of the training dataset. The architecture is designed to be straightforward and easy to train and optimize, resulting in faster convergence. Extensive empirical validation further demonstrates its efficacy in similar tasks [19]. In this paper, a hierarchical self-attentive module (HS) is proposed to replace the CNN dense link block, and the HS can effectively extract more discriminative feature information. The HS module has been seamlessly integrated into both the generator and the discriminator components of the SSIGAN architecture, and we prioritize the description of the HS module in order to facilitate subsequent introductions between the different modules. Assume that the given input is $O_{in} \in \mathbb{R}$. The following is a detailed description of the HS module.

$$O_m = MSA(LN(O_{in}, \theta)) + CB(LN(O_{in}, \theta)) + O_{in} \quad (1)$$

$$O_{out} = MLP(LN(O_m, \varpi)) + O_m \quad (2)$$

where $MSA(\cdot)$ is the multi-head self-attention function, which is computed by dividing the channel of input features equally into d parts, i.e., the number of heads is d, and then executing d times $SA(\cdot)\,(SA(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{D}})V)$. $CB(\cdot)$ is composed of the convolution, the activation function GELU, and the channel attention together. $LN(\cdot)$ is the layer normalization, $MLP(\cdot)$ is the multilayer perceptron, and ϖ is the correlation coefficient with respect to O_m.We use $HSM(\cdot)$ to represent the HS module.

The main process of the generator is to process the input of the generator by linear embedding, then use Transformer block and HS modules to obtain high-level semantic information features, and finally output ISR feature maps by establishing residual connections. Assume that the low-resolution given input is $I_{LR} \in \mathbb{R}^{H\times W\times C}$, where H, W and C denote the height width and number of channels, respectively. A detailed description of the generator is given below.

$$I'_{LR} = UP(HSM_{\times 3}(TF(le(I_{LR}), \rho))) \oplus le(I_{LR}) \quad (3)$$

$$I_{ISR} = HSM(I'_{LR}) \quad (4)$$

where $le(\cdot)$ denotes the linear embedding that transforms patches into fixed dimensional vectors. ρ denotes the positional encoding that provides the network with spatial location information for each patch. $TF(\cdot)$ denotes the regular self-attention module without context-guide blocks, which includes layer normalization, $MSA(\cdot)$ and $MLP(\cdot)$. $UP(\cdot)$ denotes the up-sampling. $\oplus$ denotes the pixel-wise addition operation. The final output feature map I_{ISR} will be fed into the D and EE for the following operations.

The D uses the network framework of VGG-19, with a dependency between the discriminator and the generator. The role of the D is to predict probability values that the real image I_{HR} is relatively more realistic than the generated intermediate image I_{ISR}. Equations (5) and (6), which formulate the relativistic average discriminator for our architecture.

$$\mathrm{D}(I_{HR}, I_{ISR}) = \sigma\{C(I_{HR}) - E_{I_{ISR}}[C(I_{ISR})]\} \Rightarrow 1 \quad (5)$$

$$\mathrm{D}(I_{ISR}, I_{HR}) = \sigma\{C(I_{ISR}) - E_{I_{HR}}[C(I_{HR})]\} \Rightarrow 0 \quad (6)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, since the final output of the D is a probability value. $C(\cdot)$ denotes the output feature maps that have been processed by the VGG-19 architecture. These feature maps are usually high-dimensional and capture important visual features of the input image. $E_{I_{ISR}}[\cdot]$ denotes the operation of calculating the average of all generated intermediate images in a small batch.

$E_{I_{HR}}[\cdot]$ denotes the operation of calculating the average of all real images in a small batch.

The procedure of EE involves processing I_{ISR} with the Laplacian operator, which is an edge detection operator effective at identifying edges in an image. By applying the Laplacian operator, the high-frequency parts of the image, i.e., the edge regions, are highlighted [60]. The feature map then through subsequent processing, which include linear embedding, combinations of HS modules, a sigmoid activation function, and up-sampling. The final SR image output is obtained by capturing long-range dependencies through residual connections, a technique that enhances feature learning and image quality.

$$I'_{SR} = UP(\sigma(HSM_{\times 3}(TF(le(lo(I_{ISR}), \kappa))) \quad (7)$$

$$I_{SR} = lo(I_{ISR}) - I_{ISR} \oplus I'_{SR} \quad (8)$$

where $lo(\cdot)$ is the Laplace operator, κ is positional encoding, and - and $\oplus$ denote pixel-wise addition and subtraction operations

The loss functions for G and D are as follows, respectively.

$$L_G = -E_{I_{HR}}[\log(1 - D(I_{HR}, I_{ISR}))] - E_{I_{ISR}}[\log(D(I_{ISR}, I_{HR}))] \quad (9)$$

$$L_D = -E_{I_{HR}}[\log(D(I_{HR}, I_{ISR}))] - E_{I_{ISR}}[\log(1 - D(I_{ISR}, I_{HR}))] \quad (10)$$

As shown in equations (9), (10), the loss functions of G and D is symmetric and contain both the intermediate super-resolution image I_{ISR} and the high-resolution image I_{HR}. The generated intermediate image is created by the generator where $I_{ISR} = G(I_{LR})$. Therefore, the generator benefits from the gradient of the generated data and the real data in adversarial training, and this design approach can help the generator to learn more details and texture information.

The structure of the edge enhancer (EE) is shown in the lower part of Fig. 1, the HI generated by the generator is used as the input, the edge information is firstly extracted by the Laplace operator, and then the features are extracted by a number of HS modules after the linear embedding operation, and then the weights are reshaped for the distribution by using the Sigmoid activation function, and finally, the enhanced edge information is fused with the ISR for the feature fusion at the same time of subtracting the edge information extracted by the Laplace operator. Thus, the SR image is obtained.

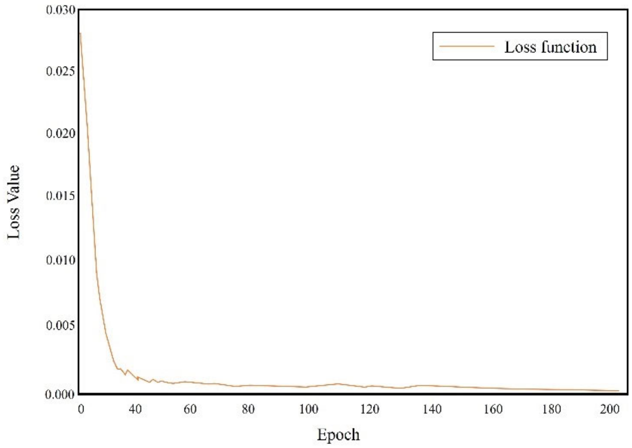

FIGURE 3. Convergence of the loss function.

B. GLOBAL-CONTEXT AWARENESS NETWORK FOR VEHICLE DETECTION (GCAFORMER)

GCAFormer transforms the input images of different sizes into 512 × 512 for better training. The input RGB three-channel image is divided into patches and these patches are transformed into a one-dimensional sequence of vectors, which are then fed into the cascade transformer block (CT) encoder. First, the input image is assumed to be $I \in \mathbb{R}^{H \times W \times 3}$.

In this context, H, W, and 3 represent the spatial height, width, and the number of channels of the image, respectively. Subsequently, I is divided into $(H \times W)/K^2$ grids, each of size $K \times K \times 3$. Every patch is regarded as a token, with its attributes defined as a segment of the original image's RGB values. All the patches in the grid are flattened to form a sequence $x \in \mathbb{R}^{N \times P}$, where $N = HW/K^2$, $P = K \times K \times 3$. In the sequence x, we apply a learnable projection $l : x_i \Rightarrow e_i \in \mathbb{R}^P (i \in 1, \ldots .N)$ to obtain the sequence $e_i \in \mathbb{R}^{N \times P} (i \in 1, \ldots .N)$. Finally, the sequence is fed into the CT for encoding.

CT is used as the backbone of GCAFormer's hierarchical feature extraction, which not only employs a window self-attention design, but also introduces the position encoding of CNN based on the Transformer. CT restricts the self-attention computation to each window, which guarantees that the information within the window can be fully interacted with each other, and greatly reduces the computation amount. CT also introduces the shifted window so that the information between different windows can be exchanged. CT also introduces the shifted window so that the information between different windows can be exchanged. Therefore, CT can effectively introduce multi-scale feature information to solve multi-task vehicle detection in complex traffic environments. The framework of CT is described in Fig. 3, where W-MSA is to divide the image into non-overlapping windows, and each window carries out self-attention computation, and SW-MSA is a kind of shifted window special multi-head self-attention module. An LN layer is used before MSA and MLP computation, and there is a residual connection after both MSA and MLP computation. Multiple CT blocks receive inputs from

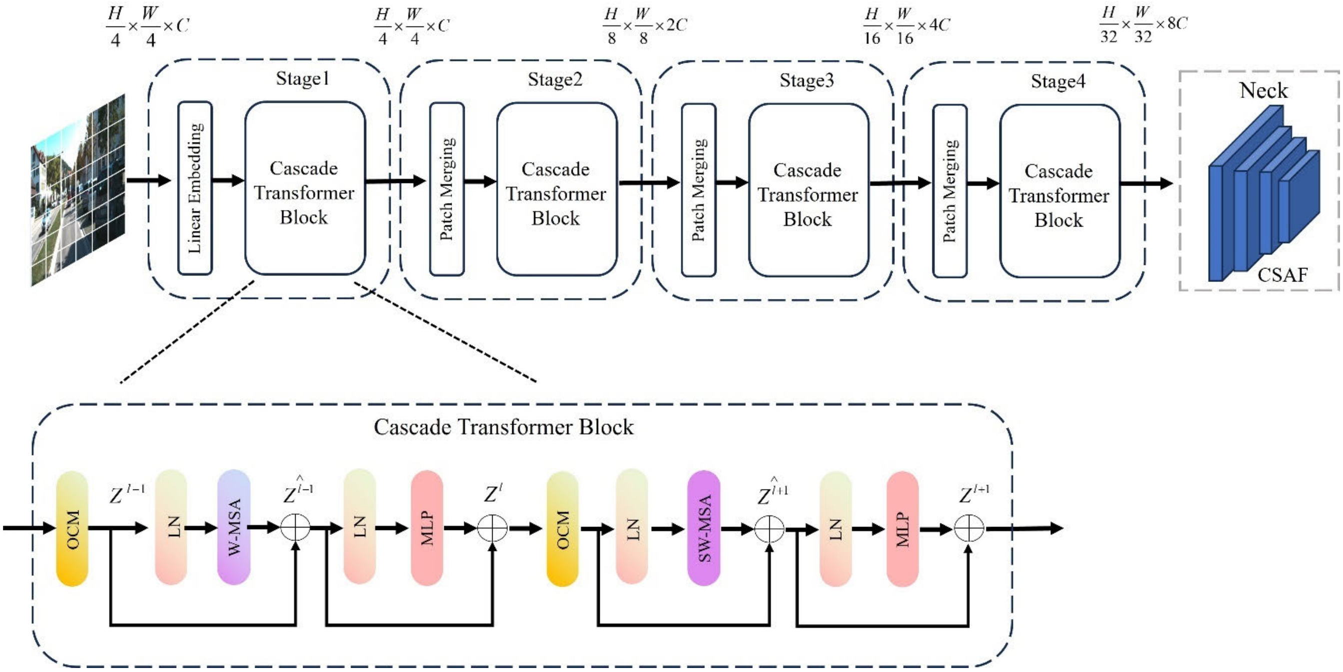
$\frac{H}{4}\times\frac{W}{4}\times C$
$\frac{H}{4}\times\frac{W}{4}\times C$
$\frac{H}{8}\times\frac{W}{8}\times 2C$
$\frac{H}{16}\times\frac{W}{16}\times 4C$
$\frac{H}{32}\times\frac{W}{32}\times 8C$
Stage1
Linear Embedding
Cascade Transformer Block
Stage2
Patch Merging
Cascade Transformer Block
Stage3
Patch Merging
Cascade Transformer Block
Stage4
Patch Merging
Cascade Transformer Block
Neck
CSAF
Cascade Transformer Block
OCM
Z^{l-1}
LN
W-MSA
$\hat{Z}^{l-1}$
LN
MLP
Z^{l}
OCM
LN
SW-MSA
$\hat{Z}^{l+1}$
LN
MLP
Z^{l+1}

FIGURE 4. Cascade transformer (CT) backbone.

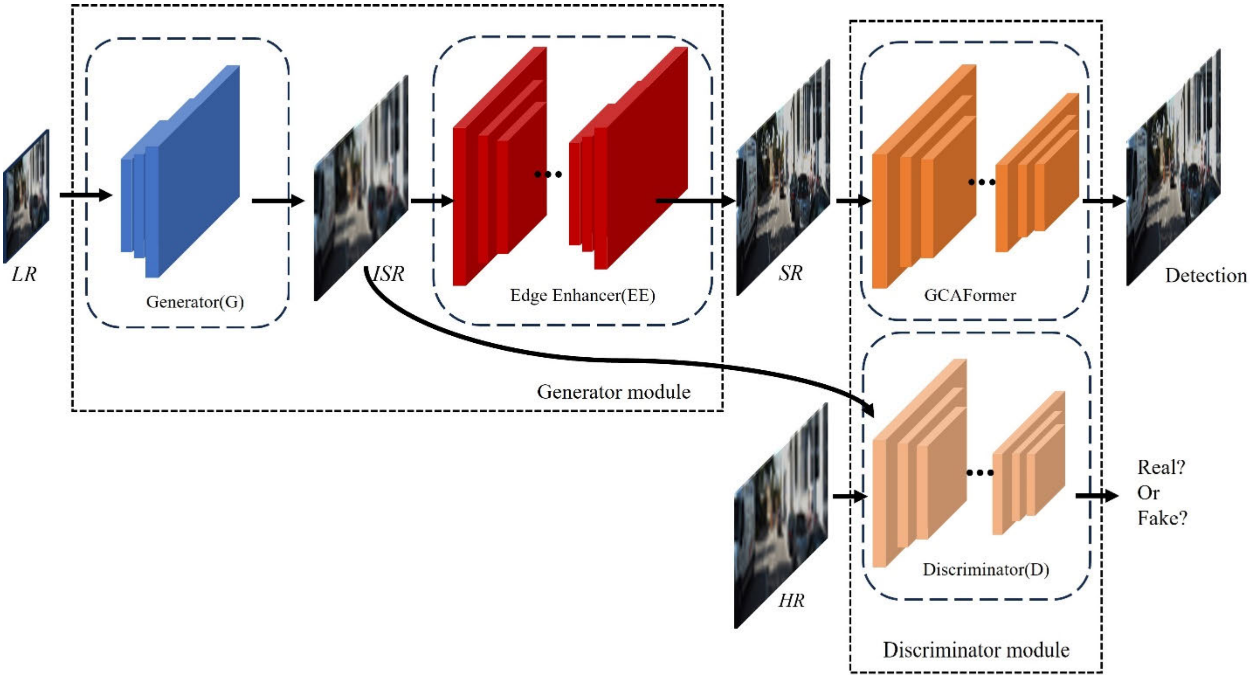

FIGURE 5. Super-Resolution GAN and Global aware object detection system.

the previous step. The backbone of GCAFormer consists of a total of 4 stages. Stage1 consists of linear embedding and CT blocks, which maintains the dimensions of the input token $H/4 \times W/4 \times C$. In Stage2, patch merging is used to merge the attributes of each set of neighboring 2×2 patches. A linear layer is then applied to the merged patches to change their dimensions from 4c to 2c for output. The CT block is then used for feature extraction and output, while the resolution is changed to $H/8 \times W/8$. The above process of stage2 is repeated twice, resulting in stage3 and stage4. The output resolution is $H/16 \times W/16$ and $H/32 \times W/32$. The above feature maps are generated in a hierarchical manner similar to traditional CNNs with the same resolution feature maps.

The CSAF module receives low-dimensional feature maps with rich spatial location information and high-dimensional low-resolution feature maps from stage1 to stage4,

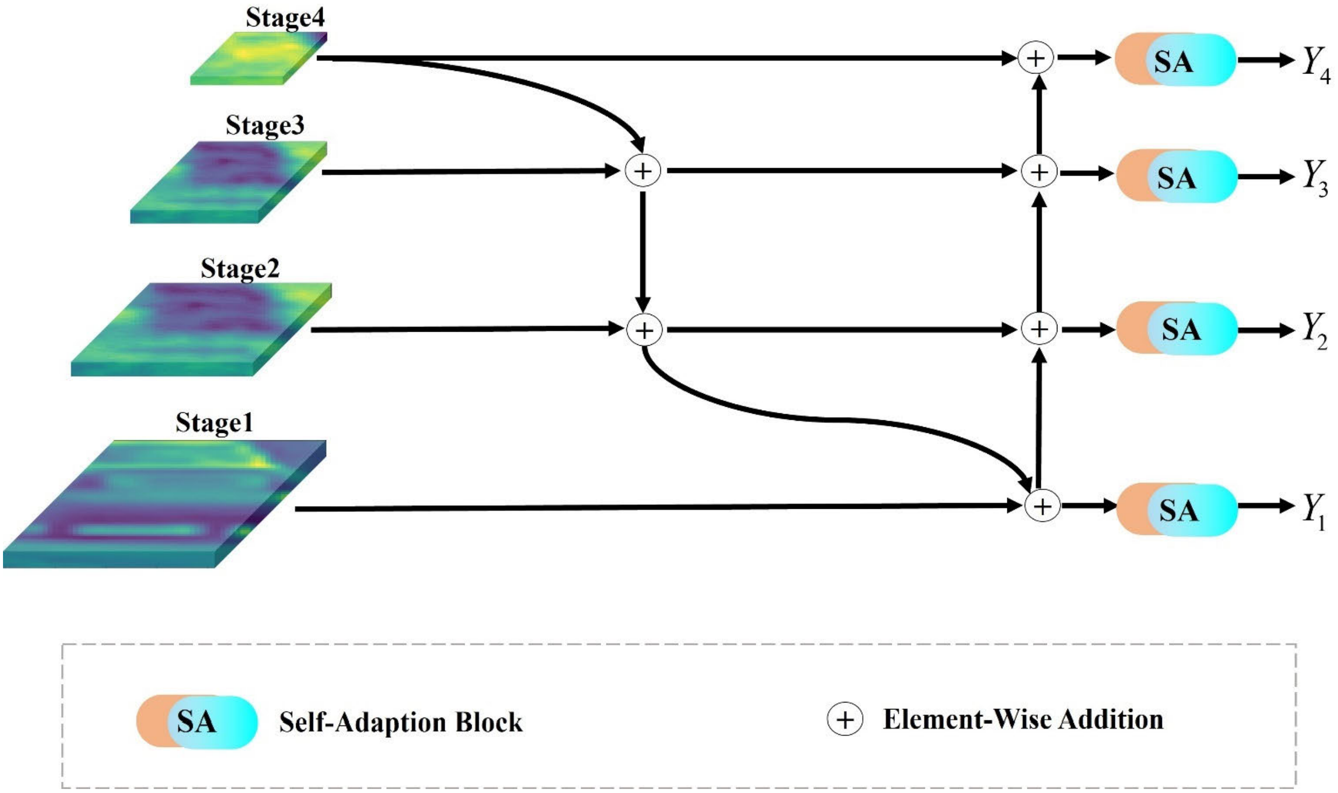

FIGURE 6. Cross-scale aggregation feature module.

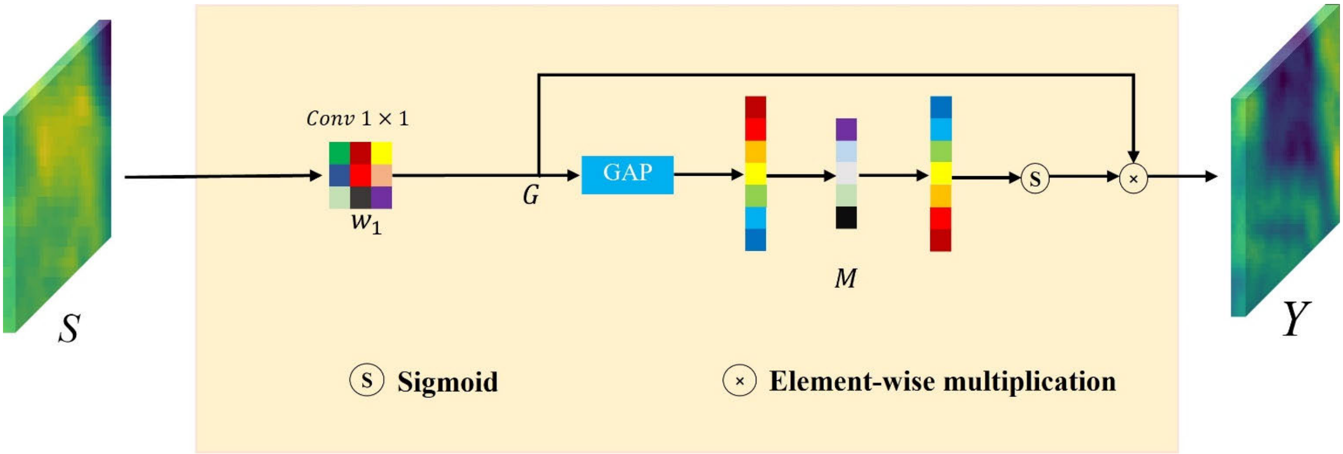

FIGURE 7. Self-adaption (SA) block.

respectively. Since deep feature maps and shallow feature maps involve different semantic information, the combination of the two types of information can be adapted to vehicles of different sizes and shapes in the traffic environment. Therefore, in this study, we design the CSAF module for receiving the multi-scale feature maps of the CT backbone network and then processing them to facilitate the fusion of multi-scale information. The process is illustrated in Fig. 6, where the CSAF module employs a bi-directional path fusion strategy (top-down and bottom-up) and an adaptive module (SA), where Fig. 7 demonstrates the SA process. CSAF cross-fuses the multi-scale feature maps from the first to the fourth stage, and the SA module performs background noise suppression on the fused feature maps. Since the conventional way of adjusting the resolution and then performing pixel-level addition operations is not sufficient to cope with vehicle detection in complex traffic environments. Some previous studies have pointed out that feature maps with different resolutions have different weight distributions [20], so in the CSAF module, we add additional weights on each input branch, which enables the network to learn the important feature information on each input. For additional weights, we use the fast normalized fusion strategy $O = \sum_i \left(\theta / \left(\sigma + \sum_j \theta_j\right)\right) \cdot I_i$ to adjust the weights of the CSAF module, where θ_i is the learnable weights and I_i is the input feature map from stages 1 to 4, which is then processed by θ_i using the ReLU activation function to make the weights of F greater than 0 [20]. The value of E is ascertained through a process of trial-and-error and is set to 1e-3. The output feature map of the CSAF module can be depicted as follows.

$$Y_1 = SA(\frac{\theta_1' \cdot S_1^{in} + \theta_2' \cdot re\left(S_2^{mid}\right)}{\theta_1' + \theta_2' + \sigma}) \tag{11}$$

$$Y_2 = SA(\frac{\theta_1' \cdot S_2^{mid} + \theta_2' \cdot re\left(Y_1\right)}{\theta_1' + \theta_2' + \sigma}) \tag{12}$$

$$Y_3 = SA(\frac{\theta_1' \cdot S_3^{mid} + \theta_2' \cdot re\left(Y_2\right)}{\theta_1' + \theta_2' + \sigma}) \tag{13}$$

$$Y_4 = SA(\frac{\theta_1' \cdot S_4^{in} + \theta_2' \cdot re\left(Y_3\right)}{\theta_1' + \theta_2' + \sigma}) \tag{14}$$

The input feature maps for Stage1, Stage2, Stage3, and Stage4 are denoted by S_1^{in}, S_2^{in}, S_3^{in} and S_4^{in} respectively, and the output features are denoted by Y_1, Y_2, Y_3, and Y_4. θ_1' and θ_2' denote the weights for each level, respectively, and $re\left(\cdot\right)$ denotes the feature maps are resized to match the inputs. Since the CSAF module as a whole is cross-fertilization of different Stages among the feature maps, it is easy to introduce background noise to affect the detection effect, so we proposed self-adaption block (SA), the purpose is to inhibit the shallow stage semantic information among the interfering information as shown in Fig. 7, about the specific description of the SA is as follows.

$$G = w_1(S, \lambda) \tag{15}$$

$$Y = (G \otimes \delta\left(M\left(GAP(G, \gamma)\right)\right)) \tag{16}$$

where $w_1\left(\cdot\right)$ is a 1×1 convolution operation, λ is the correlation coefficient with respect to the input S, $\otimes$ denotes the pixel-level multiplication operation, $\delta\left(\cdot\right)$ denotes the sigmoid activation function mapping the input to a 0 to 1 distribution, $M\left(\cdot\right)$ is the information interaction between the channels, $GAP\left(\cdot\right)$ is the global average pooling operation generating the statistics of the channel dimensions, and γ is the correlation coefficient with respect to G. S_2^{mid} and S_3^{mid} denote the intermediate feature layers, as follows.

$$S_2^{mid} = Conv(\frac{\theta_1 \cdot S_2^{in} + \theta_2 \cdot re\left(S_3^{mid}\right)}{\theta_1 + \theta_2 + \sigma}) \tag{17}$$

$$S_2^{mid} = Conv(\frac{\theta_1 \cdot S_2^{in} + \theta_2 \cdot re\left(S_4^{in}\right)}{\theta_1 + \theta_2 + \sigma}) \tag{18}$$

where $Conv(\cdot)$ denotes the convolution operation and θ_1 and θ_2 denote the corresponding weights. The detection header consists of two stages designed to predict the object class and output the detection box. Through the experiments we conducted, we discovered that the performance of Mask-RCNN aligns with our requirements for the detection head component [21]. Therefore, we have chosen to utilize Mask-RCNN as the detection head in this study. As depicted in the head section of Fig. 2, it comprises a fully connected layer, a Region Proposal Network (RPN), bounding box prediction, category prediction, a Region of Interest (ROI), and a loss function. It aligns between pixels by ROI align and inputs vehicle detection box (bbox), vehicle class (class) and vehicle mask (mask).

The detection head acquires multi-scale features from the CSAF module and uses RPN to generate candidate regions containing information about the approximate location of the target. These suggestions are further optimized by feeding them into two parallel fully connected layers, which are used as bounding box regression and bounding box classification, respectively. In the second stage, ROIs are used to classify and positionally refine the candidate regions, aligning the outputs of the PRN and CSAF modules.

GCAFormer utilizes global context awareness through the CT and CSAF modules. The CT module employs self-attention to capture global features, considering the relevance of all image regions, which aids in identifying different parts of the vehicle and their spatial relationships. The CSAF module further enhances this capability by integrating multi-scale feature maps to provide richer semantic information, assisting in the understanding of vehicle features and their distribution.

C. LOSS FUNCTION

1) LOSS FUNCTION OF SSIGAN

We employed two loss functions in the generator (G) part: the perceptual loss function (L_{aware}) and ($L_{content}$) the contextual loss function B [7]. The perceptual loss function A is calculated using the VGG feature mapping ($vgg(\cdot)$). The content loss function calculates the 1-norm distance between I_{ISR} and I_{HR}. Here, I_{ISR} and I_{HR} denote the low-resolution image J that has been processed by the generator (G) and the high-resolution image labeled by us, respectively. The detailed equations are as follows.

$$L_{aware} = E_{I_{LR}}||vgg(G(I_{LR}) - vgg(I_{HR}))||_1 \tag{19}$$

$$L_{content} = E_{I_{LR}}\|G((I_{LR}) - I_{HR})\|_1 \tag{20}$$

Inspired by previous edge enhancement networks Jiang et al. [8] proposed the use of a consistency loss function, applied between I_{ISR} and I_{HR}, which can effectively preserve the detail information of the edges. Therefore, we invoked the consistency loss function for the computation of edges (L_{edge}), and we also used the Charbonnier loss to evaluate the edge information($I_{SR_{edge}}$) extracted from the super-resolution image I_{SR} generated by Edge Enhancer(EE) and the edge information($I_{HR_{edge}}$) extracted from I_{HR} [49]. We used two consistency loss functions to compute the image and edges separately, and then summed the two losses. This is shown in equations (21) and (22), where $\eta\left(\cdot\right)$ represents the penalty term of the Charbonnier loss.

$$L_{image} = E_{I_{SR}}[\eta(I_{HR} - I_{SR})] \tag{21}$$

$$L_{edge} = E_{I_{SR_{edge}}}[\eta(I_{HR_{edge}} - I_{SR_{edge}})] \tag{22}$$

$$L_{ee} = L_{edge} + L_{image} \tag{23}$$

We sum up the above loss functions of the generator (G) and edge enhancer (EE) to finally get the total loss function regarding our generator (G) module. As shown in equation (24), where λ denotes the weights accounted for different loss functions. We set λ_1, λ_2, λ_3 and λ_4 as 1, 0.001, 0.03 and 3 respectively according to empirical experience.

$$L^{G}_{final} = \lambda_1 L_{aware} + \lambda_2 L_G + \lambda_3 L_{content} + \lambda_4 L_{ee} \tag{24}$$

2) LOSS FUNCTION OF GCAFORMER

The selection and design of the loss function plays a very important role in the training of the model. It provides insight into the performance of a model on training data by evaluating the difference between the model's predicted and true results. In the head part of the network, the loss function (L) is composed of three parts: classification loss (L_{class}), position loss (L_{loc}) and mask loss (L_{mask}) [21]. It is represented as follows.

$$\begin{aligned} &L_{sum}(GCAFomer_{p_i}\left(G(I_{LR})\right), GCAFomer_{t_i}\left(G(I_{LR})\right)) \\ &\quad = L_{class} + L_{loc} + L_{mask} \qquad (25) \\ &L_{class} = \frac{1}{N_{class}} \sum_i L_{class}(GCAFomer_{p_i}\left(G(I_{LR})\right), p_i^*) \\ &L_{loc} = \frac{\beta}{N_{loc}} \sum_i p_i^* \cdot L_1^{smooth}(GCAFormer_{t_i}(G(I_{LR})) - t_i^*) \end{aligned} \tag{26}$$

where $GCAFomer_{p_i}\left(G(I_{LR})\right)$ and p_i^* refer to the confidence level that the prediction frame is a vehicle and the true value of the label, respectively. $GCAFormer_{t_i}(G(I_{LR}))$ represents four coordinates of the prediction, and t_i^* is the true coordinate. N_{class} is the regularization term with respect to L_{class}, N_{loc} is the regularization term with respect to L_{loc}, and β balance the weights between L_{class} and L_{loc}. L_{class} and L_{loc} are specified as follows.

$$\begin{aligned} &L_{class}(GCAFomer_{p_i}\left(G(I_{LR})\right), p_i^*) \\ &\quad = -p_i^* ln(GCAFomer_{p_i}\left(G(I_{LR})\right)) \\ &\qquad - (1 - p_i^*) ln(1 - GCAFomer_{p_i}\left(G(I_{LR})\right)) \end{aligned} \tag{27}$$

$$L_1^{smooth} = \begin{cases} |z| & if\ |z| > b; \\ \frac{1}{|b|} z^2 & if\ |z| \leq b \end{cases} \tag{28}$$

where $|z|$ and b are the absolute error values and hyperparameters, respectively, and L_{mask} is denoted as follows.

$$L_{mask} = -\frac{1}{m^2} \sum_{1 \leq i,j \leq m} \begin{bmatrix} ln\left(GCAFomer_{y_{ij}}\left(G(I_{LR})\right)\right)^{y_{ij}} \\ + \\ ln(1 - GCAFomer_{y_{ij}}\left(G(I_{LR})\right))^{(1-y_{ij})} \end{bmatrix} \tag{29}$$

where y_{ij} is the label of cell (i, j) in the true mask and $GCAFomer_{y_{ij}}\left(G(I_{LR})\right)$ is the predicted value of the true label.

3) TRAINING

Based on the previous description of equation (5) and (6), we define the discriminator (D) for training our generator (G). The structure of the discriminator (D) adopts VGG-19. In the methods section, we have defined GCAFormer as our detection network. The discriminator (D) and GCAFormer will jointly act as a discriminator (D) for the generator module.

For the training part, we provide two approaches. The first approach, separate training, involves training SSIGAN and GCAFormer separately. The second approach involves end-to-end training. We will focus on both approaches in the following sections.

In the separate training approach, we train the super-resolution image generation network SSIGAN (generator module and discriminator (D)) and the detection model GCAFormer separately. This means that the loss function of GCAFormer is not backpropagated into the generator module (generator (G) and edge enhancer (EE)). The generator module only receives feedback from the discriminator (D).

In the end-to-end training approach as shown in Fig. 5, the system (SSIGAN and GCAFormer) is trained end-to-end. This means that the loss function of the detection network can be back-propagated to the generator module. Since we treat GCAFormer and the discriminator (D) as a whole, the generator module receives gradients from both GCAFormer and the discriminator (D). As shown in equations (30) and (31), we obtain the final discriminator (D) loss and the total loss function of our system.

$$L^{D}_{final} = L_D + L_{sum} \tag{30}$$

$$L_{system} = L^{G}_{final} + L^{D}_{final} \tag{31}$$

In this paper, the network is trained for 200 epochs, and the total loss of the network shows good convergence at the 40th epoch, as shown in Fig. 3.

IV. EXPERIMENTS

A. DATASET AND EVALUATION

In image enhancement part of the system, an adversarial network (SSIGAN) for generating super-resolution vehicle images is proposed, which requires LR as inputs and high-resolution images as labels during the training process of SR generation, and generates SR images through the joint action of G, D, and edge EE, which helps in the subsequent target recognition and analysis tasks in complex traffic environments. The labeling in this study was done by capturing high-definition traffic images of no less than 30 cm from Bing maps and cutting the images into 512 × 512 pixels image blocks, where it was ensured that each block contained at least one vehicle target [22]. For the production of low-resolution images, double cubic interpolation down-sampling was used to reduce the size of the high-resolution image (HR) by a factor of four, resulting in LR image blocks of 128 × 128 pixels [23]. We used the production dataset to train SSIGAN, making the model sensitive to traffic environment

TABLE 2. Detection of LR (low resolution) and SR (super resolution) images using the same dataset. Calculate the mAP values under different detection categories.

Model	Traning Image Resolution-Test Image Resolution	IITM-hetra Dataset (mean average precision test results)	KITTI Dataset (mean average precision test results)
GCAFormer without CSAF	LR	79.44%	77.20%
	SR	82.42%	79.59%
GCAFormer with CSAF	LR	86.50%	84.32%
	SR	90.62%	**89.11%**

TABLE 3. The proposed system (SSIGAN and GCAFormer) is analyzed in comparison with the current SOTA model on the dataset KITTI (E, M, and H for easy, medium, and hard, respectively).

Method	Average precision (%)									mAP (%)	Runtime (ms)
	Car			Person			Cyclist				
	E	M	H	E	M	H	E	M	H		
SSMD [31]	87.61	87.93	79.26	51.02	48.27	43.70	48.56	53.16	52.92	61.38	2000
YOLOv3 [32]	88.65	78.94	77.93	85.36	80.24	77.08	89.16	83.82	84.29	82.72	28
YOLOv3-MT [33]	88.61	81.43	80.57	85.86	82.54	79.71	88.96	84.33	84.29	84.03	w/o
SINet [34]	w/o	w/o	w/o	w/o	w/o	w/o	w/o	w/o	w/o	85.98	200
Finding every car	w/o	w/o	w/o	w/o	w/o	w/o	w/o	w/o	w/o	82.32	**25**
YOLOv5-NAM [35]	88.87	89.66	81.55	86.25	81.56	78.55	89.50	87.66	83.25	85.54	w/o
Swin-Transformer [27]	86.40	78.66	80.33	79.32	77.64	69.53	81.32	79.58	80.24	79.23	65
YOLOSR-IST [50]	92.23	88.65	**83.92**	89.34	87.32	79.66	88.52	**83.65**	80.23	87.54	120
SatDetX-YOLO [51]	94.99	87.63	81.32	90.45	85.66	**80.58**	88.74	86.42	81.59	89.09	94
Ours	**96.11**	**90.32**	83.88	**93.33**	**88.45**	80.23	**91.42**	**88.65**	**84.62**	**90.33**	422

TABLE 4. The proposed system is analyzed in comparison with the current SOTA model on the dataset IITM-hetra.

Method	Backbone	mAP (%)	FPS
Faster-RCNN [37]	ResNet101	84.2	3
Faster-RCNN-UTC [36]	VGG16	80.2	6
YOLOv3	Darknet53	84.32	24
YOLOv4 [38]	Darknet53	85.5	19
EfficientDet [20]	EfficientB3	85.5	19
Swin-Transformer	Swin-T	80.4	10
YOLOSR-IST	YOLOv5+Swin-T	87.37	12
SatDetX-YOLO	YOLOv8	91.44	**48**
Ours	SSIGAN+GCAFormer	**92.31**	24

features for subsequent application to other traffic datasets for SR image generation.

Due to the currently available datasets, there are small targets with far away vehicles and pedestrians with fewer pixel points, which do not provide sufficient contextual information. Without the use of image related techniques, this would result in the possible loss of texture and detail feature information, which will affect the effectiveness of the detection. To ensure a comprehensive analysis of the experiments, the proposed GCAFormer model was employed for object detection on the IITM-hetra, KITTI, and Pascal VOC datasets, which was enhanced with the SSIGAN model. The

TABLE 5. The proposed system is analyzed in comparison with the current SOTA model on the dataset Pascal VOC.

Method	Backbone	Input size	mAP (%)
Single-stage Detectors:			
SSMD	VGG16	512×512	73.52
IMFRE [39]	VGG16	512×512	79.66
MREFP-Net [40]	VGG16	512×512	75.20
YOLOv3	Darknet53	512×512	81.32
YOLOv5-NAM	CSP-NAM	512×512	84.77
SatDetX-YOLO	YOLOv8	512×512	87.02
Two-stage Detectors:			
Faster R-CNN [37]	ResNet101	512×512	72.31
Faster-RCNN	VGG16	512×512	76.69
Domain adaptive [42]	ResNet101	512×512	86.66
Ours	SSIGAN+GCAFormer	512×512	**88.41**

TABLE 6. The proposed system is analyzed in comparison with the current SOTA model on the dataset BDD-100K.

Method	Resolution	Backbone	mAP (%)	FPS
YOLOv3	HR	Darknet53	77.68	23
YOLOv4	HR	Darknet53	83.00	15
EfficientDet	HR	EfficientB3	79.88	22
Swin-Transformer	HR	Swin-T	85.12	17
YOLOSR-IST	HR	YOLOv5+Swin-T	86.32	18
SatDetX-YOLO	HR	YOLOv8	91.33	34
YOLO-SE [55]	HR	YOLOv8	90.45	**49**
Ours (GCAFormer)	LR	GCAFormr	78.43	32
Ours (SSIGAN+**GCAFormer)**	SR	SSIGAN+GCAFormer	**93.33**	21

IITM-hetra dataset encompasses complex traffic conditions, while both the KITTI and Pascal VOC datasets serve as standard benchmarks for vehicle detection [24], [25], [26]. Fig. 8 shows the image enhancement and detection results using the IITM-hetra dataset under proposed super-resolution GAN with global-aware target detection system. SSIGAN assists GCAFormer to achieve higher average accuracy. From the figure, it can be clearly seen that, comparing the generated SR image with the input LR image, the SR image restores the feature information of the vehicle targets in the LR image and enhances the detail information of the proximal targets, which enables GCAFormer to correctly detect most of the objects.

GCAFormer uses the pre-trained weights of Swin-Transformer on ImageNet for feature extraction [27], [28]. GCAFormer uses the same hyper-parameter settings on the batch-size of the KITTI, Pascal VOC and IITM-hetra datasets was set to 20 and trained for a total of 40 cycles using the AdamW [29] optimizer. The initial learning rate is 0.0001 and the weight decay is 0.05. The number of epochs for system is set to 200. Using data augmentation techniques, the input training set is subjected to random transformations such as horizontal rotation, scaling and brightness difference to prevent model overfitting. The input image size was resized to at least 512. Our experimental environment is under the Ubuntu 20.04 operating system, using the PyTorch framework and an NVIDIA 4090 graphics.

We evaluate the model using the following metrics: mean average precision (mAP), frames per second (FPS), and

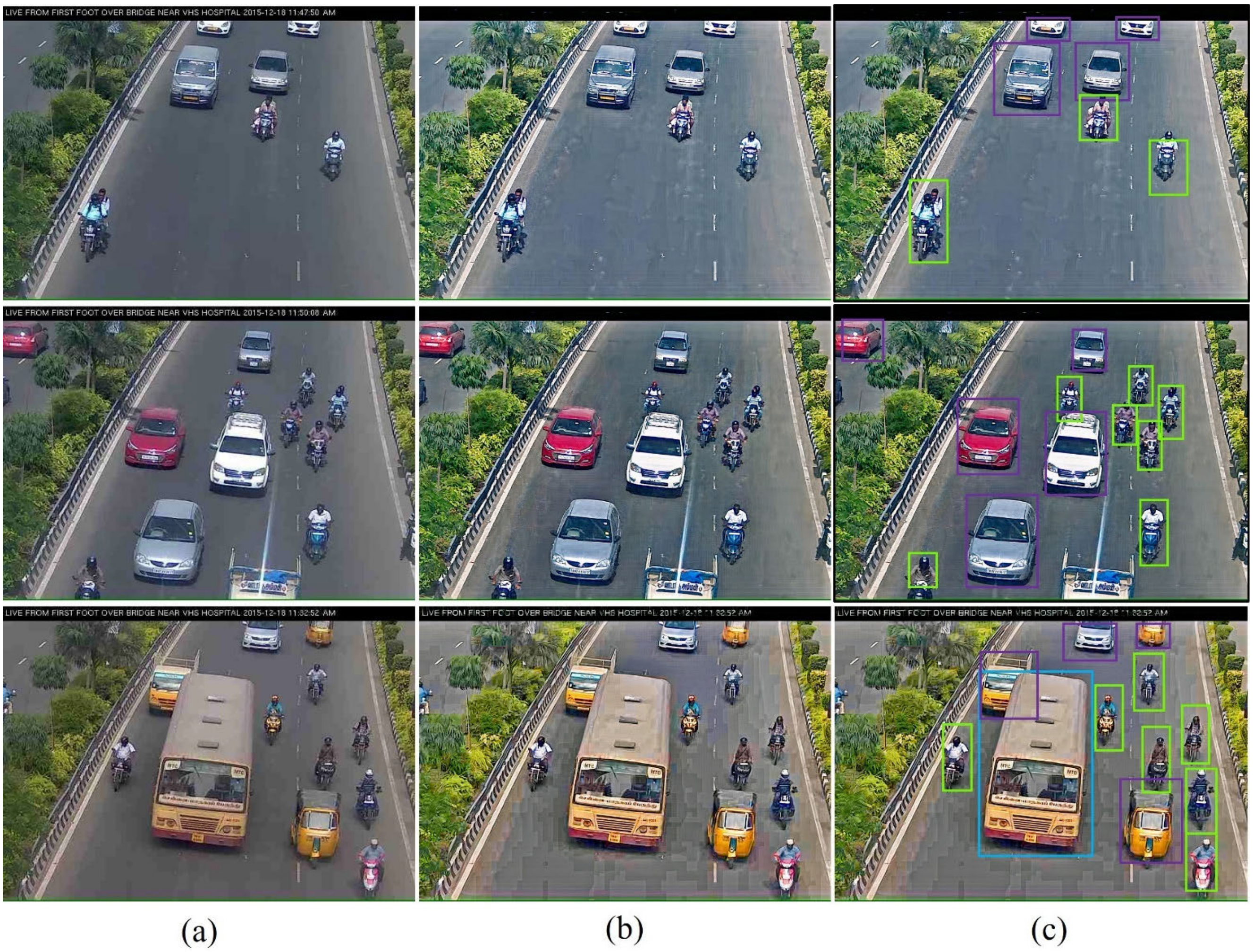
LIVE FROM FIRST FOOT OVER BRIDGE NEAR VHS HOSPITAL 2015-12-18 11:47:50 AM
LIVE FROM FIRST FOOT OVER BRIDGE NEAR VHS HOSPITAL 2015-12-18 11:50:08 AM
LIVE FROM FIRST FOOT OVER BRIDGE NEAR VHS HOSPITAL 2015-12-18 11:32:52 AM
(a)
(b)
(c)

FIGURE 8. **The image enhancement and detection results using the dataset IITM-hetra dataset. (a) The input LR image. (b) The SR image generated by SSIGAN. (c) The detection results from our proposed model.**

runtime (ms). mAP metric is the most critical metric in this study, and is used to evaluate the mean accuracy of all classes. The mAP metric is the most critical metric in this study, which was used to calculate the average accuracy across all categories. It is affected by average precision (AP) and intersection over union (IoU), where AP needs to be evaluated in combination with precision and recall, and Precision is a measure of the percentage of correct predictions made by the model in question, as follows.

$$Pre = \frac{TP}{TP + FP} \tag{32}$$

where TP denotes a positive positive and FP a false positive. Recall is defined as follows.

$$Re = \frac{TP}{TP + FN} \tag{33}$$

where FN is denoted as false negative. In summary AP can be expressed as.

$$AP = \frac{1}{N} \sum_{Re_i} Pre(Re_i) = 1 \tag{34}$$

where N is the number of samples for accuracy and recall.

The IoU's are used to calculate the difference between the predicted and true boxes. It is specified as follows.

$$IoU = \frac{A_p \cap A_g}{A_p \cup A_g} \tag{35}$$

where A_p is the predicted bounding box and A_g is the true bounding box. mAP is calculated by considering the mean of the estimated IoU thresholds for all classes in the AP dataset, and FPS and runtime provide model speed on the runtime platform.

B. ABLATION EXPERIMENTS ON SSIGAN

The SSIGAN model generates super-resolution images, and in order to verify whether it can help the detector improve its detection accuracy, we used LR and SR as inputs for the target detection task based on GCAFormer, respectively.

In Table 2, the performance is examined using detectors with different training/testing combinations. When we train and test on LR images without the CSAF module (using a conventional FPN instead of the CSAF module [30]),

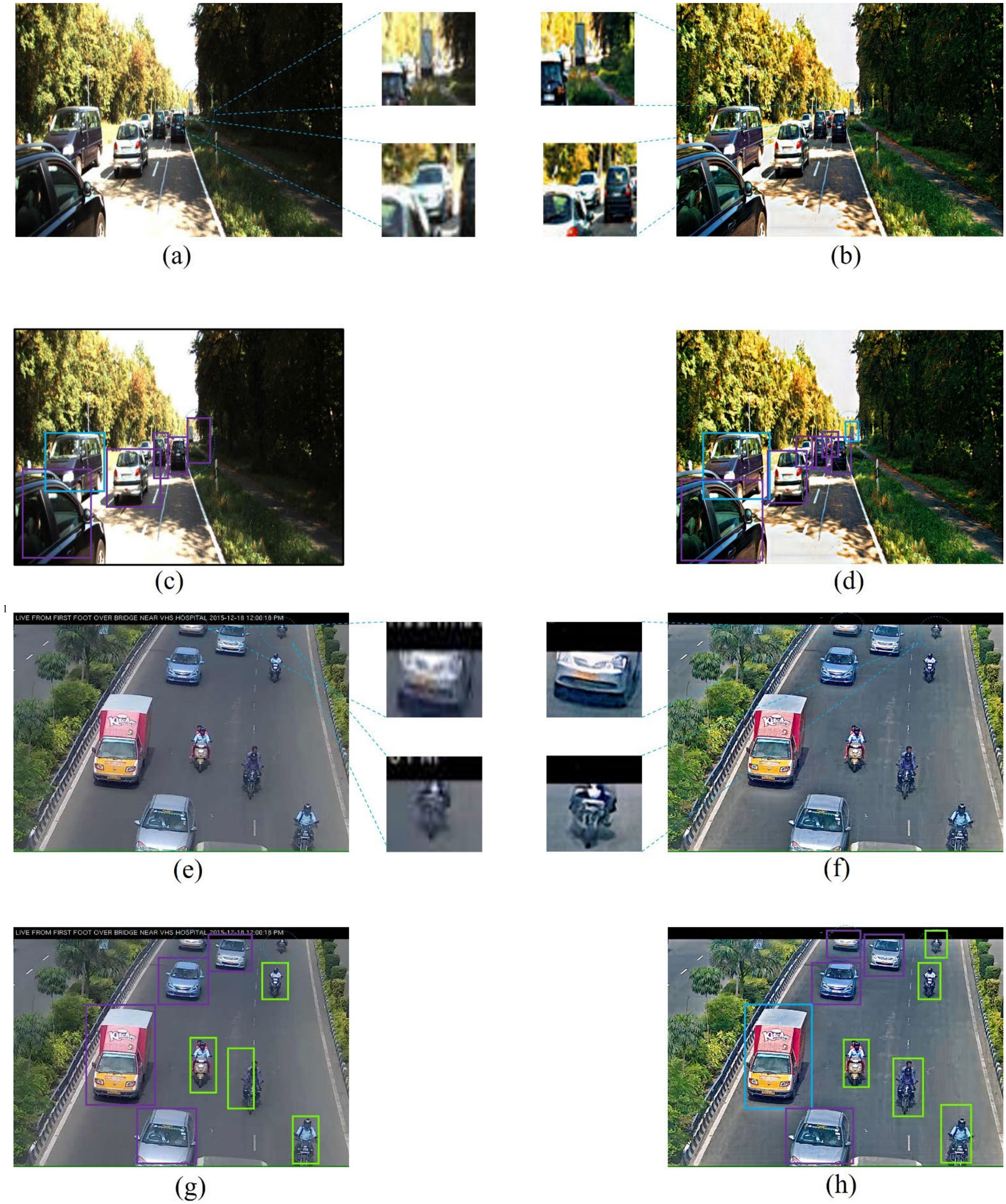
(a)
(b)
(c)
(d)
1
LIVE FROM FIRST FOOT OVER BRIDGE NEAR VHS HOSPITAL 2015-12-18 12:00:18 PM
(e)
(f)
LIVE FROM FIRST FOOT OVER BRIDGE NEAR VHS HOSPITAL 2015-12-18 12:00:18 PM
(g)
(h)

FIGURE 9. (a) and (e) show the LR image without enhancement using SSIGAN model. (b) and (f) are the SR image enhanced with SSIGAN model. (c) (g) and (d) (h) are the detection result images using GCAFormer under LR and SR, respectively.

we observe that the mAP of the GCAFormer is only 79.44% and 77.20%, and its average accuracy improves by 2.98% and 2.39% for training and testing on SR. Therefore, providing the model with high-quality images as input for detection can

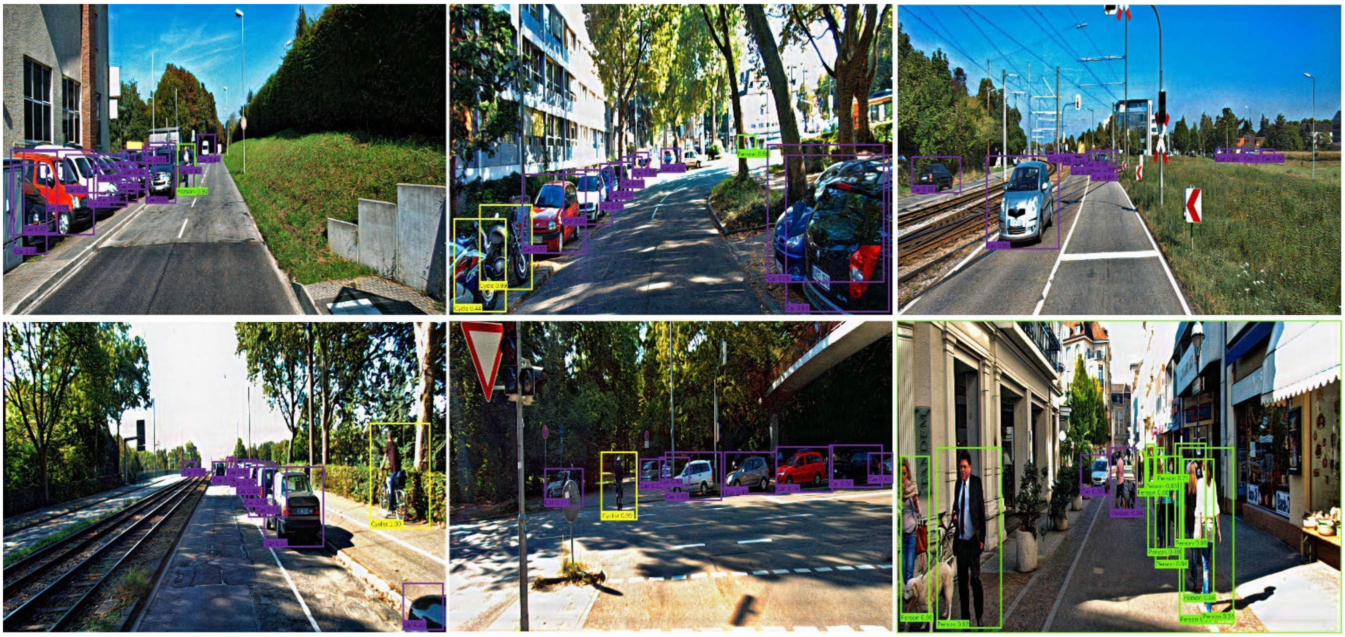

FIGURE 10. Some result images on the KITTI dataset are shown in figure. Augmented by our proposed SSIGAN model and then detected by GCAFormer.

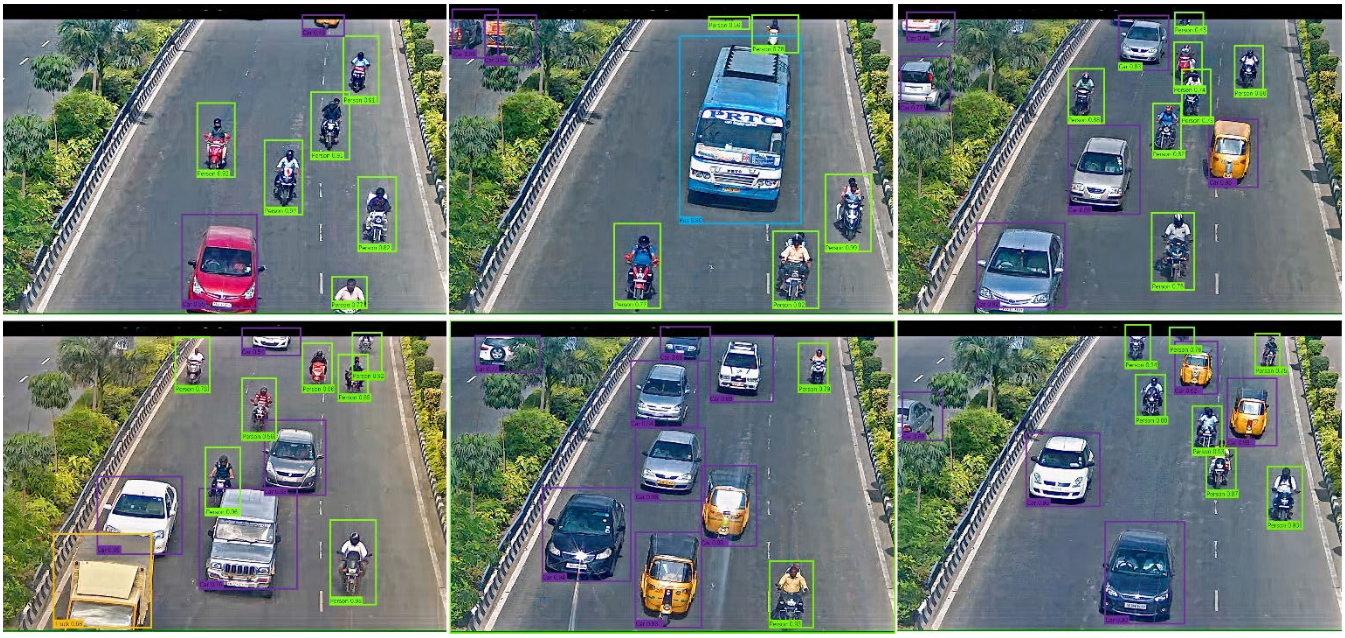

FIGURE 11. Some result images on the IITM-hetra dataset are shown in figure. Augmented by our proposed SSIGAN model and then detected by GCAFormer.

recover target details and provide rich feature information to the detection model.

In the last two lines of Tabel 2, we add the CSAF module to GCAFormer, which achieves a mAP accuracy of 90.62% on the SR image, which is 11.18% ahead of that of the first line of the test on LR. 11.18%. This not only shows the significant impact of resolution on the quality of target detection, but also demonstrates that our proposed bidirectional path fusion feature strategy of CSAF module provides more efficient multi-scale feature information to the model, which can be adapted to the task of vehicle detection in complex environments.

The SR image generated after SSIGAN is shown in Fig. 9. Fig. 9(a) shows the LR image in the original dataset, and Fig. 9(b) represents the SR reconstructed image, and the visual effect of the reconstructed image is significantly enhanced as seen from the left side of Fig. 9(b). The centered signboard in Fig. 9(c) is recognized as a vehicle, and the white vehicle near the back is not detected by GCAFormer due to less feature information. Fig. 9(d) shows a plot of the

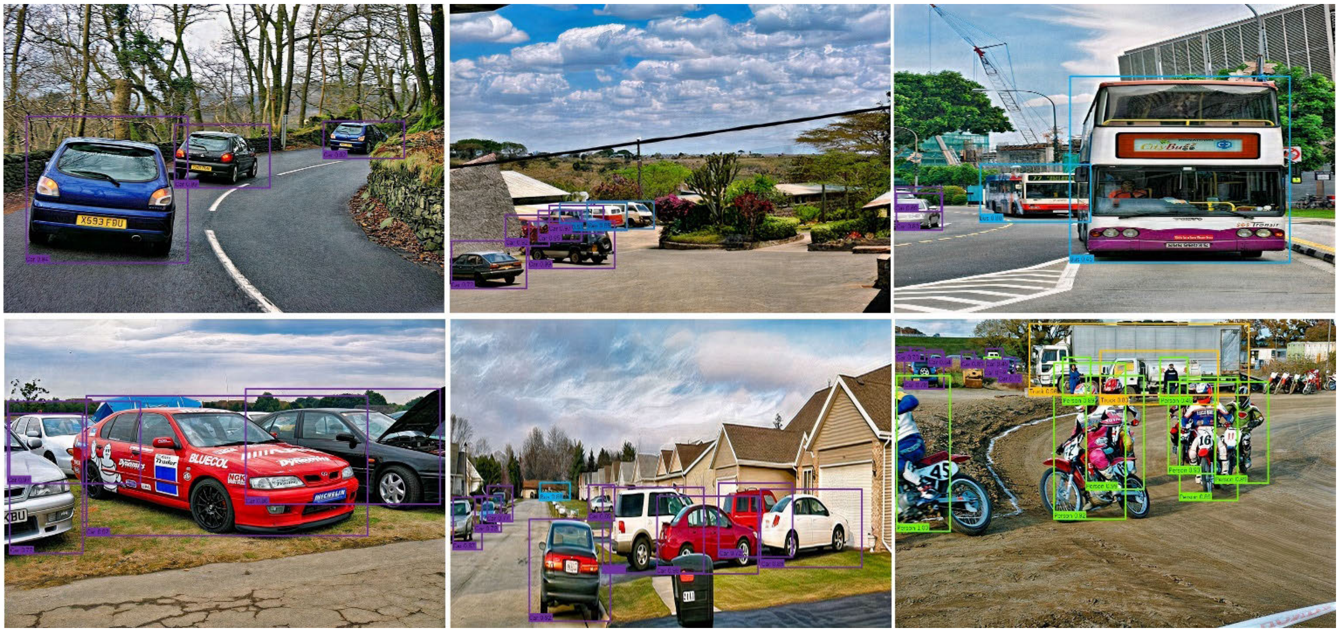

FIGURE 12. Some result images on the pascal VOC dataset are shown in figure. Augmented by our proposed SSIGAN model and then detected by GCAFormer.

detection results using the GCAFormer model after SSIGAN reconstruction, in which the false detection of the centered signage is corrected and the white vehicle near the rear is correctly detected. It can also be seen from Fig. 9(e) and Fig. 9(f) that the feature information at the far end of the image after SR reconstruction is enhanced. Comparing the detection results after LR and SR reconstruction (Fig. 9(g) and (h)), it can be seen that the blurred motorcycles and cars at the distance are also accurately detected.

C. EXPERIMENTAL RESULTS ON KITTI DATASET

The KITTI dataset is currently a popular dataset for vehicle detection with 7,481 images, including cars, people, bicycles, vans, and trucks, but according to the official evaluation, only three categories are considered (car, person, and cyclist). In order to verify the validity and reliability of our proposed models, all comparative models were evaluated for performance at KITTI. Since the dataset has no corresponding labels, this study divides the dataset according to the method proposed by Xiang, Choi et al. The training and test sets are 3712 and 3769, respectively [43]. This study presents a system that includes two models (SSIGAN and GCAFormer). We will use an end-to-end training approach where the edge enhancer (EE) and the generator (G) are grouped into the generation module as a whole. The discriminator (D) and the detection model GCAFormer will be grouped into the discriminator module as a whole. By integrating the above loss functions into a unified system loss function, the generator module can receive the gradient from the discriminator module, as shown in Fig. 5. The evaluation of KITTI is categorized into three modes, i.e., easy, moderate, and difficult, depending on the height of the bounding box and the level of occlusion. The results of comparison with the SOTA model are shown in Table 3, where our method achieves 90.33% mAP and 422 ms Runtime (ms). It is 7.61% and 4.79% ahead of YOLOv3 [32] and YOLOv5-NAM [33], respectively, but since our backbone is based on the self-attention module, it results in a slower runtime than the former two. SINet is the stronger competitor of our method, which also uses SR images as input [34]. YOLOSR-IST and SatDetX-YOLO are current state-of-the-art algorithmic models, and although they are leading in different categories of target recognition tasks, mAP still behind our system. Table 3 demonstrates that our model performs well in all three difficulty levels, proving that our strategy of SSIGAN generating SR images as input and GCAFormer sensing global information for detection is accurate and efficient. Fig. 10 shows the specific detection results of our proposed model on the KITTI dataset.

D. EXPERIMENTAL RESULTS ON IITM-HETRA DATASET

The IITM-hetra dataset contains 1417 images and we have divided the dataset into two parts, training and testing, where training contains 1200 images and testing contains 215 images. The detection targets are made up of a total of four: cars, people, autos, and buses. People riding motorcycles are labeled as people, and large vehicles are labeled as bus. In the IITM-hetra dataset, the input images are cropped to 512×512 and then passed into system. Hetra dataset before evaluation, we have cropped the image to 512 × 512 size and fed it as input to our system for training and inference. For a fair comparison with the SOTA model, we trained it on IITM-hetra again using transfer learning as well. As shown in Table 4, we introduce two experimental metrics are backbone network and FPS. The backbone network includes the

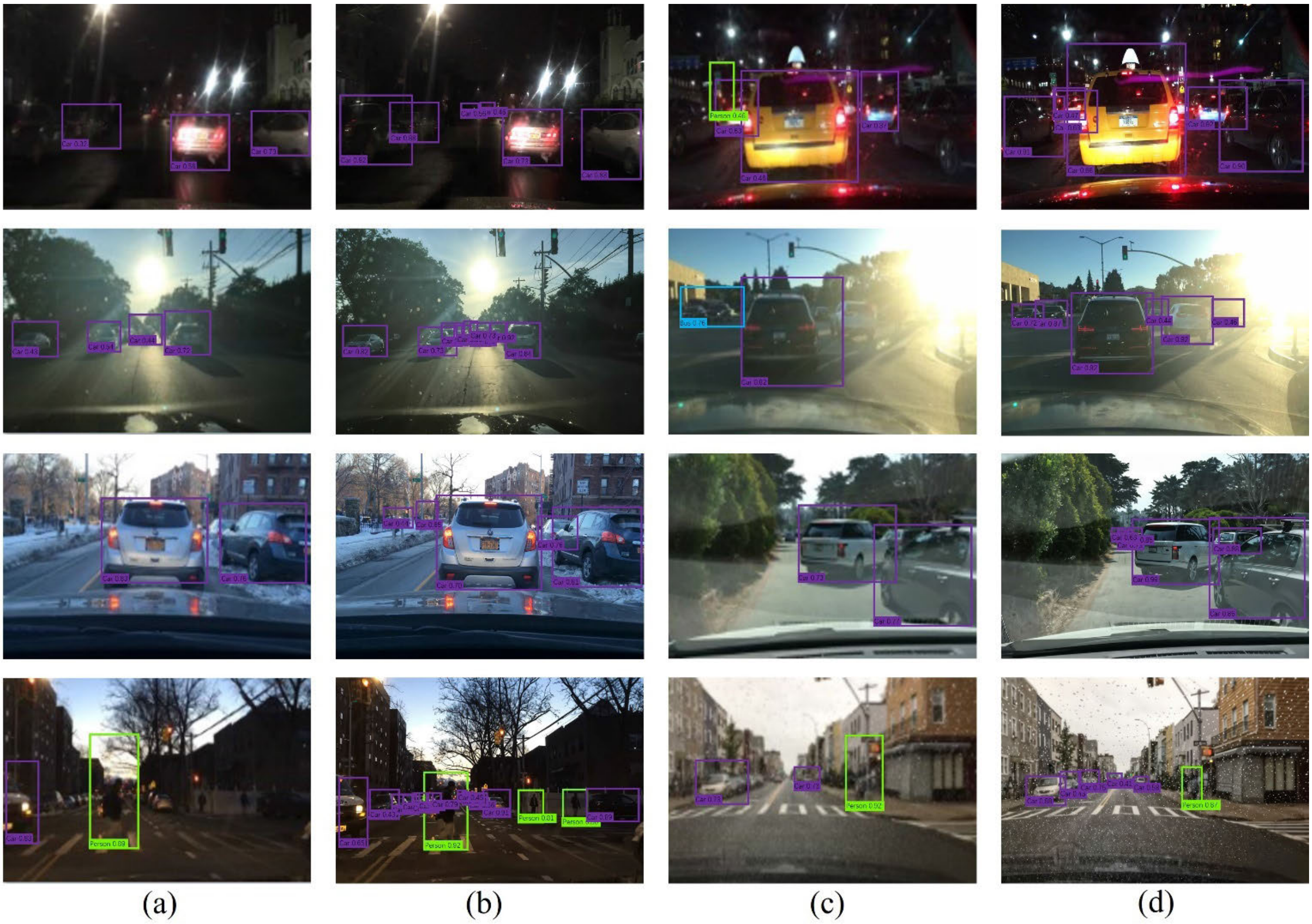

FIGURE 13. (a) and (c) show the resultant images of GCAFormer model detection at low resolution. (b) and (d) show the SR images enhanced by the SSIGAN model and then detected by the GCAFormer model.

swin-transformer and YOLOv8 commonly used in current SOTA models. Our method leads YOLOv4, EfficientDet and Swin-Transformer in mAP by close to ranging from 6% to 12%. Although YOLOv4 and EfficientDet perform well for cars, bicycles, and buses, they are less effective in small target detection, and Swin-Transformer, although sensitive to small targets at long distances, is also ineffective due to the absence of SR images and multi-scale information to guide detection. As our strongest competitor, SatDetX-YOLO reaches 48 in FPS, but still lags behind our proposed system in accuracy. Our proposed method provides SR images for training and more powerful multi-scale features in the detection session, which can achieve finer-grained detection results. The detection results are shown in Fig. 11, from which the accuracy of our proposed work can be proved.

E. EXPERIMENTAL RESULTS ON PASCAL VOC DATASET

The Pascal VOC dataset has 9963 pictures in 20 groups. It has 2501 pictures for training, 2510 for checking, and 4952 for testing. The primary task of our system is to locate the vehicle position in a complex traffic environment. Because the Pascal VOC dataset's pictures have many classes and scenarios we need to spot cars are complex, we tested our system on this dataset to make sure it could work well in different situations. As can be seen in Table 5, our model achieves 88.41% mAP on the test set, outperforming the single and two-stage SOTA models in the table. Fig. 12 illustrates some of the model detection results.

F. EXPERIMENTAL RESULTS ON BDD-100K DATASET

The BDD-100k dataset is a large-scale dataset created by the Berkeley Deep Driving Project team, containing high-resolution images and detailed annotations of more than 100,000 driving scenarios, with a large number of driving scenarios with day/night and bad weather images [54]. We selected 2000 images of complex driving environments as our training sets and 200 images as test sets. Since the BDD-100k dataset itself is a high-resolution dataset, we used bicubic down-sampling on its original images, capturing the low-resolution images for training and keeping the high-resolution original images for labeling. Fig. 13 shows the driving scene images in various types of complex traffic environments, such as alternating day and night, light interference, alternating seasons, and interference from surrounding obstacles, as shown in column (a). Column (c) also shows disturbances such as bad weather. Column

(a) and (c) show the detection effect in the low-resolution image, and columns (b) and (d) show the detection effect in the high-resolution image. It can be clearly observed that there are fewer pixels in the low-resolution image for the long- distance target, and the detail information is not rich enough. The presence of a large amount of noise in the low-resolution image suppresses the feature information of the target region and reduces the detection accuracy. When using the SR image generated by our proposed system, the detection algorithm is able to extract more detailed feature information at the distance, which is more sensitive to the edge information and background noise. SR image allows the same size image to contain more pixel points while suppressing the noise. As shown in Table 6, we divided the resolution, the down-sampled images of the BDD-100K dataset as LR images, the original HD images as HR images, and the ones that have undergone SSIGAN to generate the images as SR images. In Table 6, we first used the untrained and untested LR image dataset as input and used our proposed detection model GCAFormer for prediction, and the resulted mAP is only 78.43%, which is not enough to compare with the current SOTA model. However, when we integrated the SR image generation model SSIGAN and the GCAFormer detection model into one system. When we use this system for prediction, the mAP reached 93.33%, surpassing the strongest competitor, YOLO-SE, by 2.88%.

V. CONCLUSION

In this study, we proposed a framework that combines a super-resolution generative adversarial network (GAN) and a global-aware vehicle detection system. The framework consists of two main parts: a super-resolution image generation model (SSIGAN) and a global context-aware model for vehicle detection (GCAFormer).

SSIGAN is a model consisting of a generator (G), a discriminator (D) and an edge enhancer (EE). We embed a hierarchical self-attention module (HS) in G and EE, which effectively reduces the problems of matching errors and loss of texture details at the later stage of super-resolution image reconstruction. The input of SSIGAN is a low-resolution (LR) traffic image, and the output is a super-resolution (SR) image with clear edges and textures. This provided detailed feature information for subsequent detection tasks.

The global context-aware network for vehicle detection consists of GCAFormer and its built-in modules cascade transformer (CT), cross-scale aggregation feature (CSAF), and the detection head. The CT module provides multiscale information, and the CSAF module employs bi-directional feature fusion, which combines to allow the detection head to receive more detailed information about the vehicle. The CT module provides multi-scale information and the CSAF module adopts bi-directional feature fusion technology, both of which enable the detection head to receive more semantic feature information for accurate detection and classification on SR traffic images.

TABLE 7. List of abbreviations.

Abbreviation	Short terms
GAN	Generative Adversarial Network
HS	Hierarchical Self-Attention
EE	Edge Enhancer
SSIGAN	Super-Resolution Synthetic Image GAN
GCAFormer	Global Context-Aware Network for Object Detection in Complex Environments
CT	Cascade Transformer
CSAF	Cross-Scale Aggregation Feature
SR	Super Resolution
LR	Low Resolution
ISR	Intermediate Super-Resolution
G	Generator
D	Discriminator

We conducted extensive tests on the KITTI, IITM-hetra, Pascal VOC and DBB-100k datasets. The results show that our proposed system (SSIGAN with GCAFormer) can be adapted to complex traffic situations and is more accurate when compared to other car localization methods in most situations. Future work will focus on improving the robustness of the model and creating more diverse and realistic SR images in order to enhance the generalization capabilities of the model. We will also explore how to utilize more data and advanced computer vision techniques to improve our model performance further.

APPENDIX

Table 7 includes abbreviations of important short terms that appear in the paper.

ACKNOWLEDGMENT

We gratefully acknowledge the financial support from Geran Galakan Penyelidik Muda (Grant GGPM-2022-063) and INTI International University. Their contributions were instrumental to the success of our research.

REFERENCES

[1] M. A. S. Kamal, T. Hayakawa, and J.-I. Imura, "Development and evaluation of an adaptive traffic signal control scheme under a mixed-automated traffic scenario," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 2, pp. 590–602, Feb. 2020.

[2] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "$\mathcal{R}^2$-CNN: Fast tiny object detection in large-scale remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5512–5524, Aug. 2019.

[3] J. Shermeyer and A. Van Etten, "The effects of super-resolution on object detection performance in satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1432–1441.

[4] L. Shi, Z. Li, J. Li, Y. Wang, H. Wang, and Y. Guo, "AGCNet: A precise adaptive global context network for real-time colonoscopy," *IEEE Access*, vol. 11, pp. 59002–59015, 2023.

[5] H. Ji, Z. Gao, T. Mei, and B. Ramesh, "Vehicle detection in remote sensing images leveraging on simultaneous super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 676–680, Apr. 2020.

[6] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 105–114.

[7] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Comput. Vis.-ECCV Workshops*, 2018, pp. 63–79.

[8] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.

[9] N. C. Mithun, N. U. Rashid, and S. M. M. Rahman, "Detection and classification of vehicles from video using multiple time-spatial images," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1215–1225, Sep. 2012.

[10] W. Chen, D. Sharifrazi, G. Liang, S. S. Band, K. W. Chau, and A. Mosavi, "Accurate discharge coefficient prediction of streamlined weirs by coupling linear regression and deep convolutional gated recurrent unit," *Eng. Appl. Comput. Fluid Mech.*, vol. 16, no. 1, pp. 965–976, Dec. 2022.

[11] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7029–7038.

[12] P. Wang, Z. Cai, H. Yang, G. Swaminathan, N. Vasconcelos, B. Schiele, and S. Soatto, "Omni-DETR: Omni-supervised object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9357–9366.

[13] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[14] L. Liebel and M. Körner, "Single-image super resolution for multispectral remote sensing data using convolutional neural networks," *ISPRS-Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 41, pp. 883–890, Jun. 2016.

[15] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small object detection via multi-task generative adversarial network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 206–221.

[16] H. Sajid, S.-C.-S. Cheung, and N. Jacobs, "Motion and appearance based background subtraction for freely moving cameras," *Signal Process., Image Commun.*, vol. 75, pp. 11–21, Jul. 2019.

[17] M. Cheon, W. Lee, C. Yoon, and M. Park, "Vision-based vehicle detection system with consideration of the detecting location," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1243–1252, Sep. 2012.

[18] A. Kembhavi, D. Harwood, and L. S. Davis, "Vehicle detection using partial least squares," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1250–1265, Jun. 2011.

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

[20] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[22] (2023). *Bing Map*. Accessed: Dec. 5, 2023. [Online]. Available: https://www.bing.com/maps

[23] M. Saeed Rad, T. Yu, C. Musat, H. Kemal Ekenel, B. Bozorgtabar, and J.-P. Thiran, "Benefiting from bicubically down-sampled images for learning real-world image super-resolution," 2020, *arXiv:2007.03053*.

[24] D. Mittal, A. Reddy, G. Ramadurai, K. Mitra, and B. Ravindran, "Training a deep learning architecture for vehicle detection using limited heterogeneous traffic data," in *Proc. 10th Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2018, pp. 294–589.

[25] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[26] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[27] 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.

[28] S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, and D. J. Fleet, "Synthetic data from diffusion models improves ImageNet classification," 2023, *arXiv:2304.08466*.

[29] R. Llugsi, S. E. Yacoubi, A. Fontaine, and P. Lupera, "Comparison between adam, AdaMax and Adam w optimizers to implement a weather forecast based on neural networks for the Andean city of quito," in *Proc. IEEE 5th Ecuador Tech. Chapters Meeting (ETCM)*, Oct. 2021, pp. 1–6.

[30] 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.

[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 21–37.

[32] C. Kumar and R. Punitha, "YOLOv3 and YOLOv4: Multiple object detection for surveillance applications," in *Proc. 3rd Int. Conf. Smart Syst. Inventive*, 2020, pp. 1316–1321.

[33] K. Wang and M. Liu, "YOLOv3-MT: A YOLOv3 using multi-target tracking for vehicle visual detection," *Int. J. Speech Technol.*, vol. 52, no. 2, pp. 2070–2091, Jan. 2022.

[34] X. Hu, X. Xu, Y. Xiao, H. Chen, S. He, J. Qin, and P.-A. Heng, "SINet: A scale-insensitive convolutional neural network for fast vehicle detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 1010–1019, Mar. 2019.

[35] J. Wang, Y. Dong, S. Zhao, and Z. Zhang, "A high-precision vehicle detection and tracking method based on the attention mechanism," *Sensors*, vol. 23, no. 2, p. 724, Jan. 2023.

[36] S. H. Ahmed, M. Raza, S. S. Mehdi, I. Rehman, M. Kazmi, and S. A. Qazi, "Faster RCNN based vehicle detection and counting framework for undisciplined traffic conditions," in *Proc. IEEE 18th Int. Conf. Smart Communities, Improving Quality Life Using ICT, IoT AI (HONET)*, Oct. 2021, pp. 173–178.

[37] 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.

[38] A. M. Roy, R. Bose, and J. Bhaduri, "A fast accurate fine-grain object detection model based on YOLOv4 deep neural network," *Neural Comput. Appl.*, vol. 34, no. 5, pp. 3895–3921, Mar. 2022.

[39] Q. Zheng and Y. Chen, "Interactive multi-scale feature representation enhancement for small object detection," *Image Vis. Comput.*, vol. 108, Apr. 2021, Art. no. 104128.

[40] L. Aziz, M. S. B. H. S. Fc, and S. Ayub, "Multi-level refinement enriched feature pyramid network for object detection," *Image Vis. Comput.*, vol. 115, Nov. 2021, Art. no. 104287.

[41] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[42] G. Yin, M. Yu, M. Wang, Y. Hu, and Y. Zhang, "Research on highway vehicle detection based on faster R-CNN and domain adaptation," *Int. J. Speech Technol.*, vol. 52, no. 4, pp. 3483–3498, Mar. 2022.

[43] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Data-driven 3D voxel patterns for object category recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1903–1911.

[44] M. Shah Junayed, M. Baharul Islam, A. Sadeghzadeh, and T. Aydin, "Real-time YOLO-based heterogeneous front vehicles detection," in *Proc. Int. Conf. Innov. Intell. Syst. Appl. (INISTA)*, Aug. 2021, pp. 1–7.

[45] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[46] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 280–296.

[47] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4486–4497, Jul. 2022.

[48] R. Jing, W. Zhang, Y. Liu, W. Li, Y. Li, and C. Liu, "An effective method for small object detection in low-resolution images," *Eng. Appl. Artif. Intell.*, vol. 127, Jan. 2024, Art. no. 107206.

[49] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proc. 1st Int. Conf. Image Process.*, vol. 2, 1994, pp. 168–172.

[50] R. Li and Y. Shen, "YOLOSR-IST: A deep learning method for small target detection in infrared remote sensing images based on super-resolution and YOLO," *Signal Process.*, vol. 208, Jul. 2023, Art. no. 108962.

[51] C. Zhao, D. Guo, C. Shao, K. Zhao, M. Sun, and H. Shuai, "SatDetX-YOLO: A more accurate method for vehicle target detection in satellite remote sensing imagery," *IEEE Access*, vol. 12, pp. 46024–46041, 2024.

[52] Y. Hao, H. Pei, Y. Lyu, Z. Yuan, J.-R. Rizzo, Y. Wang, and Y. Fang, "Understanding the impact of image quality and distance of objects to object detection performance," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2023, pp. 11436–11442.

[53] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1646–1654.
[54] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," 2018, *arXiv:1805.04687*.
[55] T. Wu and Y. Dong, "YOLO-SE: Improved YOLOv8 for remote sensing object detection and recognition," *Appl. Sci.*, vol. 13, no. 24, p. 12977, Dec. 2023.
[56] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 251–260.
[57] J. R. J, B. L. R, and A. T. Al-Heety, "Moving vehicle detection from video sequences for traffic surveillance system," *ITEGAM- J. Eng. Technol. Ind. Appl. (ITEGAM-JETIA)*, vol. 27, no. 1, pp. 41–48, 2021.
[58] M. I. Pavel, S. Y. Tan, and A. Abdullah, "Vision-based autonomous vehicle systems based on deep learning: A systematic literature review," *Appl. Sci.*, vol. 12, no. 14, p. 6831, Jul. 2022.
[59] X. Liao, S. Sahran, and S. Abdul Shukor, "An experimental study of vehicle detection on aerial imagery using deep learning-based detection approaches," *J. Phys. Conf. Ser.*, vol. 1550, no. 3, May 2020, Art. no. 032005.
[60] B. Kamgar-Parsi, B. Kamgar-Parsi, and A. Rosenfeld, "Optimally isotropic Laplacian operator," *IEEE Trans. Image Process.*, vol. 8, no. 10, pp. 1467–1472, Oct. 1999.
[61] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 15908–15919.

HONGQING WANG received the B.Sc. degree in measurement and control technology from Shanghai Polytechnic University, and the M.Sc. degree in computer science from the University of Birmingham, U.K. He is currently pursuing the Ph.D. degree with the Institute of Visual Informatics, Universiti Kebangsaan Malaysia (UKM), Bangi. His academic journey includes significant research in deep learning, focusing on object detection and image segmentation. He has dedicated to advancing the field of computer vision and actively contributes to projects exploring innovative applications of AI technology. His work reflects a commitment to bridging the gap between theoretical research and practical implementation in the rapidly evolving world of artificial intelligence.

JUN KIT CHAW received the B.Eng. degree in computer engineering and the Ph.D. degree in electrical engineering from Universiti Teknologi Malaysia (UTM), Bangi, Malaysia. He is currently a Research Fellow with the Institute of Visual Informatics (IVI), UKM. Previously, he was a Senior Lecturer with the Faculty of Computing and Information Technology (FOCS), Tunku Abdul Rahman University of Management and Technology (TAR UMT). He has actively collaborated with industry partners, such as Advantech and HILTI Asia IT Services. His industry projects include optimization of manufacturing process and computer vision applications. His primary research interests include computer vision and deep learning.

SIM KUAN GOH (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees in electrical engineering and computational intelligence from the National University of Singapore, Singapore, in 2013 and 2019, respectively. He is currently an Assistant Professor with the School of Electrical Engineering and Artificial Intelligence, Xiamen University Malaysia, Sepang, Malaysia. Prior to joining Xiamen University Malaysia, he was a Research Fellow with the Air Traffic Management Research Institute, Nanyang Technological University, Singapore. His research interests include computational intelligence and its applications.

LIANTAO SHI received the B.S. degree from Xiamen University, in 2018, and the M.S. degree from Liaoning University of Science and Technology, in 2022. His main research interests include embedded systems and computer vision semantic segmentation.

TING TIN TIN received the B.S. and Ph.D. degrees in computer sciences from the University of Science, Penang, Malaysia, in 2003 and 2012, respectively.

In 2005, she joined Gemplus Technologies Asia Pte Ltd, Singapore, as Telecommunication Software Engineer. After the Ph.D. graduation, she started her career as an Educator and a Researcher. She has more than 12 years of lecturing, supervising projects, and research. She is currently a research track Educator mainly supervising postgraduate's projects with INTI International University, Malaysia. Simultaneously, she is a freelance Lecturer with Monash University, Tunku Abdul Rahman University of Management and Technology; and a Methodist with the College Kuala Lumpur. She has received her professional certification in project management from PMI and data analytics from SAS. Her research interests include big data analytics, information systems engineering, educational data mining, psycho-academic research, and software engineering.

NANNAN HUANG received the master's degree from the University of Birmingham, U.K., in 2021. He is currently pursuing the Ph.D. degree with the Faculty of Technology and Science, Universiti Kebangsaan Malaysia (UKM), Bangi, Malaysia. His research interests include deep learning, computer vision, and medical image analysis.

HONG-SENG GAN received the B.Eng. and Ph.D. degrees in biomedical engineering from the Universiti Teknologi Malaysia, in 2012 and 2016, respectively. Currently, he is a Faculty Member with the School of AI and Advanced Computing, Xi'an Jiaotong–Liverpool University, China. His research interests include medical image computing, deep learning, and computer vision.