mathematics

MDPI

Article

A Novel Effective Vehicle Detection Method Based on Swin Transformer in Hazy Scenes

Zaiming Sun [1,*], Chang'an Liu [2], Hongquan Qu [2] and Guangda Xie [3]

[1] School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China
[2] Information College, North China University of Technology, Beijing 100144, China; liuchangan@ncut.edu.cn (C.L.); qhqphd@ncut.edu.cn (H.Q.)
[3] School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China; 2020413010103@mail.ncut.edu.cn
* Correspondence: sunzm@ncepu.edu.cn

Abstract: Under bad weather, the ability of intelligent vehicles to perceive the environment accurately is an important research content in many practical applications such as smart cities and unmanned driving. In order to improve vehicle environment perception technology in real hazy scenes, we propose an effective detection algorithm based on Swin Transformer for hazy vehicle detection. This algorithm includes two aspects. First of all, for the aspect of the difficulty in extracting haze features with poor visibility, a dehazing network is designed to obtain high-quality haze-free output through encoding and decoding methods using Swin Transformer blocks. In addition, for the aspect of the difficulty of vehicle detection in hazy images, a new end-to-end vehicle detection model in hazy days is constructed by fusing the dehazing module and the Swin Transformer detection module. In the training stage, the self-made dataset Haze-Car is used, and the haze detection model parameters are initialized by using the dehazing model and Swin-T through transfer learning. Finally, the final haze detection model is obtained by fine tuning. Through the joint learning of dehazing and object detection and comparative experiments on the self-made real hazy image dataset, it can be seen that the detection performance of the model in real-world scenes is improved by 12.5%.

Keywords: Swin Transformer; image dehazing; vehicle detection; multi-scale feature

MSC: 68T01; 68T07

Citation: Sun, Z.; Liu, C.; Xie, G. A Novel Effective Vehicle Detection Method Based on Swin Transformer in Hazy Scenes. *Mathematics* **2022**, *10*, 2199. https://doi.org/10.3390/math10132199

Academic Editors: Fan Zhang, Songhe Feng, Yongsheng Zhou and Junlin Hu

Received: 27 May 2022
Accepted: 22 June 2022
Published: 23 June 2022

1. Introduction

With the maturity of advanced driver assistance systems (ADAS) and autonomous driving systems, the range of vehicle perception solutions is also diversified. The detection performance of unmanned driving is mainly responsible for the sensor, and the most commonly used is the camera, which collects visible light images for the perception of the environment. However, under bad weather conditions, such as hazy scenes, the outdoor images captured by the camera are usually affected by dynamic targets, small particles suspended in the atmosphere or water droplets, resulting in color distortion and texture blurring due to blur, occlusion, and abnormal illumination [1]. The degraded image makes the human naked eye and the traditional computer vision system unable to capture and perceive the features of the object more accurately, which makes it difficult to separate the area of interest from background clutter. It seriously affects the accurate perception of the vehicle to the traffic information and increases the rate of traffic accidents. Therefore, how to effectively extract vehicle features in hazy conditions to achieve more accurate vehicle perception is of great significance.

In recent years, with the development of deep learning, CNN-based object detection methods have made breakthrough progress in intelligent vehicle environment perception. By determining the category and location of the object, it assists the vehicle to achieve safe driving in a complex driving environment. However, most of the existing object detection algorithm research is aimed at objects in a simple and clean environment, and such a model has achieved good results on the existing object detection dataset [2–9]. For hazy scenes, the existing CNN-based detection framework has two problems. On the one hand, due to the inherent properties of convolution, the shared convolution kernel makes the properties of each region of the image easy to be ignored, and the principle of local inductive bias also invalidates the construction of convolutional dependence on remotes. On the other hand, in order to improve the detection performance, a two-stage solution is generally adopted; firstly, the image is dehazed to improve quality, and then it is detected. However, such a dehazing model cannot completely recover the potential clear image, and as a means of pre-processing, it cannot always improve the performance of object detection, and this two-stage model cannot meet the real-time requirements of intelligent vehicle driving.

In this article, an effective end-to-end detection model for haze vehicles is proposed, which solves the problem of vehicle detection in haze weather with poor visibility. To achieve this goal, Swin Transformer [10], one of the best object detectors at present, has been adopted as the backbone network. On this basis, a hazy image feature recovery module is proposed, which extracts multi-scale features through Transformer hierarchical construction and multi-stage processing. A recovery subnet that can enhance image sharpness is constructed and trained in an end-to-end manner to simultaneously learn visibility enhancement, object classification, and localization. Through this scheme, clean features can be recovered from the input hazy blurred image, so as to achieve more accurate vehicle detection in severe weather.

The main features of our proposed method are summarized as follows:

(1) The low quality of hazy images makes feature extraction difficult. To solve this problem, a model of dehazing based on an attention mechanism is proposed in this paper. Firstly, the global semantic features of the image are extracted by the encoding–decoding module, and then the high-quality haze-free image is generated by the image reconstruction module. The purpose of generating haze-free images is not to serve as input for detecting subnets but to generate clean features by learning visibility enhancement tasks.

(2) To solve the problem of too few hazy datasets for vehicle detection, this paper collects and labels the dataset Haze-Car for model training. The Real Haze-100 dataset of real hazy scenes is used to test the model.

(3) A new end-to-end haze detection model is formed by fusing the dehazing module and Swin Transformer detection module. The dehazing module is responsible for extracting clean features from blurred images, and the detection module is responsible for object classification and localization. In the training stage, the hazing model and Swin-T are used to initialize the hazy detection model parameters by means of transfer learning. Finally, the final hazy detection model is obtained by fine tuning.

(4) Comparing the algorithm proposed in this paper with the frontier object detection algorithms YOlO, SSD, Faster-RCNN, EfficientDet, Swin Transformer, etc. The experiments show that the model proposed in this paper has a certain real-time performance and achieves higher detection accuracy.

2. Related Work

2.1. Hazy Object Detection

In recent years, the deep learning object detection algorithm based on 2D images has become a powerful tool for automatic driving road object detection. In fact, deep

convolutional networks have achieved amazing success in the field of vehicle object detection [11]. CNN has a strong image feature learning ability and can perform multiple related tasks, such as classification and bounding box regression [12]. The existing methods are divided into two categories: two-stage and one-stage. The one-stage method does not generate candidate boxes but directly transforms the localization problem of object bounding boxes into a regression problem for processing. Typical algorithms include You Only Look Once (YOLO) [13] and Single-Shot Multibox Detector (SSD) [14]. The two-stage method generates the candidate boxes of the object through various algorithms and then classifies the object through the CNN. The typical algorithm is the region-CNN (R-CNN) [4] algorithm based on the candidate boxes, R-CNN, Fast R-CNN, and Faster R-CNN. Although these models achieved satisfactory performance in clear weather conditions, none of them worked efficiently in hazy scenarios without some kind of adjustment.

The general Idea of object detection In blurred scenes Is to adopt a two-stage method, that is, to dehaze the image first and then perform object detection. Early single image dehazing methods are generally based on handcrafted priors, such as dark channel prior (DCP) [15], color attenuation prior (CAP) [16], and haze-line prior (HLP) [17]. However, these methods can only achieve good results if a prior is valid; otherwise, they may generate unnatural artifacts to degrade the image quality. As a result, pre-processing images as the input of the object detector is not always guaranteed to improve the performance of detection [18]. In recent years, with the rapid development of deep learning, many CNN-based image dehazing methods have been proposed. These methods generally outperform prior-based methods because deep networks can implicitly learn the relevant features of haze in images and overcome the limitations of a single specific prior [19]. The existing dehazing models based on deep learning include DehazeNet [20] proposed by Cai et al., which used convolutional neural networks to learn the characteristics of hazy images. Through end-to-end learning and estimation of the mapping between fuzzy images and their transmission images. Li et al. [21] constructed an AOD-Net neural network dehazing model, jointly optimized dehazing and detection, absorbed the characteristics of DenseNet, and directly generated a clear image model by using lightweight CNN, achieving better results than the traditional two-stage method. Li et al. [18] studied the effect of dehazing on various detectors and found that image dehazing as a pre-treatment is not very helpful and sometimes even damages the Image features. The main reason Is that the existing dehazing methods cannot reconstruct high-quality and clear images well for subsequent high-level vision tasks. To solve this problem, Zhang et al. [22] proposed an end-to-end optimized dehazing network embedded in the atmospheric scattering model. Two generation networks are used to estimate the transmission map and atmospheric light intensity, and the two networks are fused together for reverse propagation using a boundary-aware loss function.

2.2. Vision Transformer

Transformer [23] was first applied in the field of natural language processing (NLP) and has been widely used in computer vision in recent years. Usually in visual problems, CNN is considered to be the most basic component [24], but now, Dosovitskiy [25] et al. directly divided images into block sequences and used the visual Transformer (ViT) to perform image classification tasks. Unlike CNN-based algorithms, Transformer was able to obtain semantic information between each image block through the attention mechanism, which enabled it to gain a global perception field from the beginning, making full use of the contextual semantic information. It had better recognition ability for small targets, and the computing resources required were also greatly reduced, showing strong performance beyond CNN in image and video visual tasks, such as image classification [26–29], object detection [10,30–32], semantic segmentation [33–35], and crowd counting [36,37].

Specifically, in 2018, Parmar [38] first applied Transformer to image generation and proposed the image Transformer model. In 2020, Carion et al. [30] combined CNN with Transformer to propose a complete end-to-end DETR object detection framework, which applied Transformer to object detection for the first time and obtained parallel computing capability comparable to CNN. Zhu et al. [39] proposed a deformable DETR model based on a variable convolutional neural network, which had an excellent detection effect on small objects. Zheng et al. [40] proposed an ACT algorithm to reduce the computational complexity of the self-attention module. Pyramid Vision Transformer (PVT) [41] applied Transformers to lower-resolution features, which greatly reduced computational costs. Local grouping self-attention [42] was proposed in Swin Transformer [10], where the input features were separated into a grid of non-overlapping windows, and the visual Transformer operates only in each window. Many methods have been proposed to introduce inductive bias into ViT. LocalViT [26] brought a locality mechanism to ViT by using deep convolution in the feedforward network.

These studies divided the image into multiple image patches and used the linear embedding sequence of these image patches as the input of Transformer. Then, the image patches are processed in the way of processing tokens in the NLP field, and the image classification model is trained in a supervised manner. By exploring the global interaction between different regions, it learns to pay attention to important regions in the image.

3. Proposed Method

There are two problems with outdoor traffic images. Firstly, the image quality is easily affected by bad weather. Secondly, when the image is taken, the distance between the vehicle and the camera varies greatly, resulting in large changes in the size of the vehicle. In response to the above problems, a new hazy vehicle detection network will be introduced in this section to detect all vehicle targets in the image. This network combines the two modules of dehazing and object detection to perform end-to-end multi-task learning. The method of the CNN model is to first extract the high-level features of the image and then obtain the bounding box of the specific object through classification regression. However, unlike training CNN, in order to make full use of image context information and improve the detection effect of vehicles in hazy scenes, this paper explores a robust vehicle detection method with a hierarchical visual Transformer architecture with shifted windows. Their overall architecture is based on the structure of encoder and decoder. First of all, the framework of Swin Transformer will be introduced in Section 3.1 as the backbone of the dehazing module and hazy image vehicle detection. Then, the image dehazing module will be introduced in Section 3.2 to enhance the visibility of hazy images. Finally, Section 3.3 will introduce how the feature extraction module based on hierarchical vision Transformer replaces CNN for object detection and propose an end-to-end effective vehicle detection overall framework.

3.1. Swin Transformer

Swin Transformer works by first deeply merging image patches and then replacing the standard multi-head self-attention (MSA) module in the Transformer block with a shifted window-based module, which has lower computational complexity than ViT. Figure 1a shows a schematic diagram of two successive Swin Transformer blocks connected in series. Different from the conventional MSA module, as shown in Figure 1b, W-MSA and SW-MSA are multi-head self-attention modules with regular and shifted windowing configurations, respectively. In layer n, a regular window partitioning scheme is adopted, and self-attention is computed within each window. In the next layer n + 1, the window partitioning Is shifted, resulting In new windows. In addition, the 2-layer MLP is between non-linear layers, a LayerNorm (LN) layer is applied between each MSA module and MLP, and a residual connection is applied after each module. The purpose of this window partition mechanism is to provide the connection of information between adjacent Windows. By introducing local ideas, each layer only models the local

relationship, while continuously reducing the width and height of the feature map, so as to expand the receptive field and maintain the efficient calculation of non-overlapping windows.

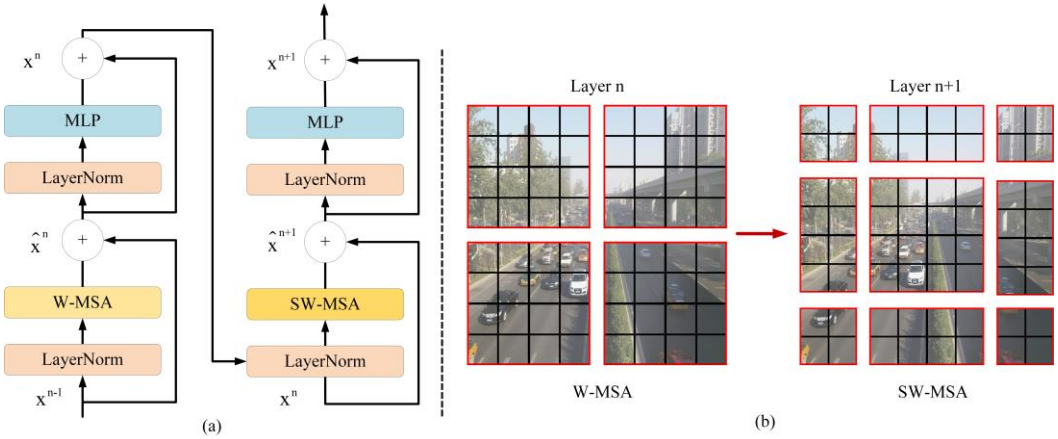

Figure 1. (**a**) Two successive Swin Transformer blocks; (**b**) W-MSA and SW-MSA.

In this work, the tiny version of Swin Transformer (Swin-T) [10] is used as the default backbone, and the shifted window partition method is used. The successive Transformer blocks are calculated as:

$$\hat{x}^n = \text{W-MSA}\left(\text{LN}\left(x^{n-1}\right)\right) + x^{n-1} \tag{1}$$

$$x^n = \text{MLP}\left(\text{LN}\left(\hat{x}^n\right)\right) + x^n \tag{2}$$

$$\hat{x}^{n+1} = \text{SW-MSA}\left(\text{LN}\left(x^n\right)\right) + x^n \tag{3}$$

$$x^{n+1} = \text{MLP}\left(\text{LN}\left(\hat{x}^{n+1}\right)\right) + \hat{x}^{n+1} \tag{4}$$

Among them, $\hat{x}^n$ and x^n are the output feature results of the multi-head attention mechanism W-MSA (SW-MSA) and MLP, respectively; W-MSA and SW-MSA represent window-based multi-head self-attention using regular and shifted window partitioning configurations, respectively.

3.2. Dehazing Network

As shown in Figure 2, the input of the dehazing network is a hazy image, which is first extracted shallow features through dense blocks and then sent to the Swin Transformer block architecture. In particular, each Swin Transformer block is followed by a convolution to reduce spatial resolution and double the number of channels. This operation causes the inductive bias to be introduced into the Transformer encoder. By connecting with the features of the same stage encoder, the decoder can effectively alleviate the loss of spatial information caused by downscaling. Finally, the features are transferred to the reconstruction section to obtain high-quality haze-free output.

The dehazing network includes encoder E for extracting multi-scale features and decoder D for generating haze-free images with enhanced visibility. As shown in Figure 2, the encoder and decoder are described as follows:

$$e^0 = F_E^0\left(I_H\right), e^n = F_E^n\left(e^{n-1}\right) \tag{5}$$

$$I_D = F_D^0\left(d_D^1, e^0\right) \tag{6}$$

where F_E^0 represents the feature extraction layer used to extract shallow features e^0, I_H represents the input hazy image, $n \in [1,\cdots,N]$, represents the different stages of the encoder, F_E^n represents the n-th stage of the encoder E, and e^n represents the deep feature of the stage n. Decoder D predicts the multi-scale features of haze-free images and finally generates high-quality haze-free images, where F_D^n represents the n-th stage of decoder D, $n \in [1,\cdots,N]$, d^n represents the multi-scale characteristics of the decoder D at stage n, and F_D^0 is the image reconstruction layer [43]. The shallow feature e^0 of the image and the feature d_D^1 recovered by the decoder are concatenated as its input. The final haze-free images I_D are generated through image reconstruction layers F_D^0.

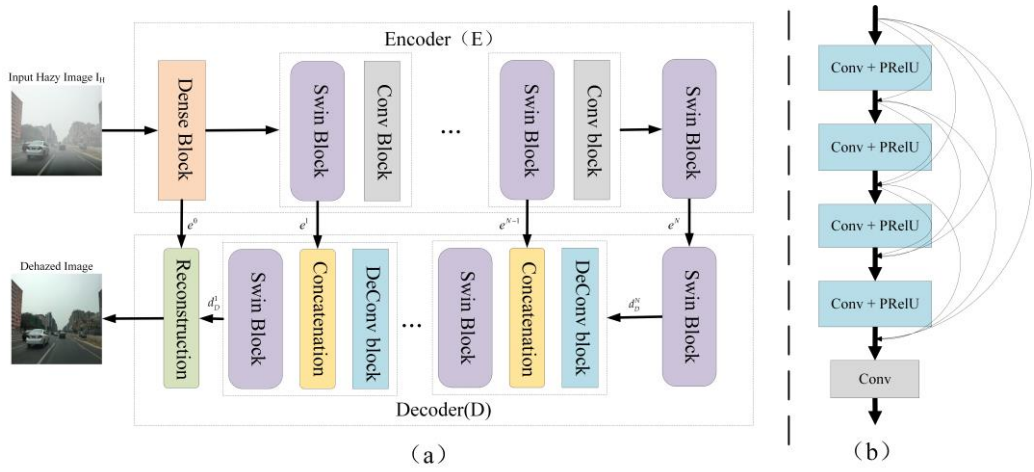

Figure 2. The architecture of dehazing network. (**a**) The overall dehazing network, which consists of an encoder and decoder. (**b**) The architecture of dense block.

Mean squared error (MSE), L2 loss, and smooth L1 loss are the most widely used loss functions for single image dehazing. However, they are based on pixel differences and do not take human visual perception into account, so even in the late stage, there is still a lot of noise on the image. Therefore, for haze-free image estimation, structure similarity index measure (SSIM) loss with faster convergence is used in the training. SSIM is an index that measures image similarity from three aspects: luminance, contrast, and structure. The mean is used as an estimate of luminance, the standard deviation is used as an estimate of contrast, and the covariance is used as a measure of structural similarity. From the perspective of image composition, SSIM defines structure information as properties that reflect the structure of objects in the scene independently of luminance and contrast, and models distortion as a combination of three different factors of luminance, contrast, and structure. The value range of SSIM is [0, 1]. The larger the value, the smaller the image distortion and the more similar. So, using the SSIM loss function is defined as:

$$\mathcal{L}_{\text{de}}\left(p^c, \hat{p}^c\right) = 1 - \text{SSIM}\left(p^c, \hat{p}^c\right) \tag{7}$$

where p^c denotes the ground truth clear image, and $\hat{p}^c$ denotes the dehazed image. The constant 1 here is added to ensure the loss value is non-negative.

3.3. Architecture Overview

The overall architecture of the proposed network is presented in Figure 3. This method is achieved by jointly learning two tasks: visibility enhancement and object localization, corresponding to two subnets: (1) detection subnet and (2) dehazing subnet. The dehazing subnet adopts the method of the encoder and decoder. The encoder is responsible for extracting deep features, while the decoder is responsible for generating clear features, and then the reconstruction module is used to obtain clean haze-free images. The detection subnet is based on the Swin Transformer block, which can model local and global dependencies, and the computational cost is lower than the ordinary Transformer block (ViT). It shares a common block (CB) module with the dehazing subnet and is responsible for object classification. The model structure is shown in Figure 3. The two subnets share the CB module to ensure that the features generated by the module can be used in the two subnets during joint learning. The detection subnet can be used to train the whole network end to end and predict the object. Through the joint optimization scheme, the clear features generated by the dehazing sub-network from the blurred image can be shared, so as to better learn the vehicle detection in the detection sub-network and improve the vehicle detection performance in the hazy scene.

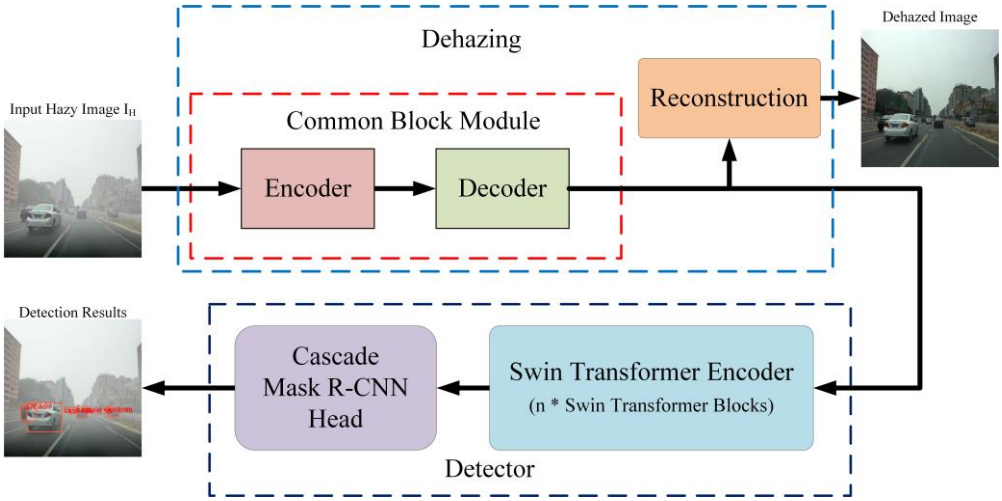

Figure 3. The architecture of our method. The framework consists of two subnetworks: the dehazing network and the detection network.

The workflow of the Swin Transformer encoder is to divide the input image $H \times W \times 3$ into a set of non-overlapping patches through the patch partition, where the size of each patch is 4×4, the feature dimension is $4 \times 4 \times 3$, and the quantity is $H/4 \times W/4$. Then, after changing the feature dimension of the divided patch to $4 \times 4 \times C$ through a linear embedding, it is sent to multiple Swin Transformer blocks to achieve global multi-scale feature learning. After that, multiple patch merging layers are used to build hierarchical feature maps. Finally, it is sent to the regression head for object positioning and regression.

In the training stage, the visual Transformer encoder pre-trained on ImageNet is used for feature extraction, the dimension parameter C is set to 96, and the number of Transformer blocks n is set to 6. The Cascade Mask R-CNN regression module includes a classifier h_x and a regressor f_x, where L_{cls} and L_{loc} are the classification and localization losses, and in each training stage t, the IoU threshold is optimized, and the optimized cascade loss, which is described as:

$$L\left(x^t, g\right) = L_{cls}\left(h_t\left(x^t\right), y^t\right) + \lambda\left[y^t \geq 1\right] L_{loc}\left(f_t\left(x^t, b^t\right), g\right) \tag{8}$$

where $b^t = f_{t-1}\left(x^{t-1}, b^{t-1}\right)$, g is the ground truth box of x^t, $\lambda = 1$ is the trade-off coefficient, [·] is the index function, y^t is the label, and the cascade loss ensures that the effective training of the detector is continuously improved for the detection effect of the position. In inference, by applying the same cascading process, the quality of the hypothesis will also be sequentially improved, so as to improve the detection effect. In label prediction, the distinction between the object and the background is solved by the IoU index. If it is higher than the threshold u, then the image block x is responsible for the prediction of the object. Assuming that the category label of x is a function of u, the inference is made according to u:

$$y = \begin{cases} g_y, & \mathrm{IoU}(x, g) \geq u \\ 0, & \mathrm{IoU}(x, g) < u \end{cases} \tag{9}$$

where g_y is the ground truth box position label, and g is the real category. The task of the regressor is to use the regressor $f(x, b)$ to return a candidate box b to the position of the real object box g. A box contains the four coordinates of (b_x, b_y, b_w, b_h), and the loss of the regressor is set as:

$$R_{loc}[f] = \sum_{i=1}^{N} L_{loc}\left(f\left(x_i, b_i\right), g_i\right) \tag{10}$$

where L_{loc} is the loss of L_2, x_i is the network input, y_i is the category number, N is the batch size, and i is the coordinate (x, y, w, h) of the regression box.

Classifier $h(x)$ assigns an image block x to one of the $M+1$ classes. The extra class represents the background class. The loss of classifier is set as:

$$R_{cls}[h] = \sum_{i=1}^{N} L_{cls}\left(h\left(x_i\right), y_i\right) \tag{11}$$

where L_{cls} is the cross-entropy loss, and $h(x)$ is the m-dimensional estimation of the category posterior distribution.

4. Experiment

In this algorithm, the two modules of dehazing and object detection are combined to realize the end-to-end multi-task learning. Multi-task learning refers to learning multiple related tasks at the same time, and what is learned from one task can benefit others. In the field of computer vision, many multi-task learning methods have been proposed and proved to be effective. The detection performance of the model is improved by the joint optimization of dehazing and object detection. Moreover, in order to measure the performance of the algorithm in this paper through experiments, an object detection dataset in the haze is constructed. By evaluating the prediction quality and quantity of the synthetic haze dataset (Haze-Car) and the natural haze dataset (Real Haze-100), the proposed algorithm is compared with other advanced object detection methods. Experimental results of object detection under severe weather conditions are summarized in this section.

4.1. Dataset

4.1.1. Dehazing Dataset

The dehazing model is trained by using the synthetic haze dataset Realistic Single Image Dehazing (RESIDE) [22]. This dataset includes 2061 real outdoor images from real-time weather in Beijing and corresponding depth maps, as shown in Figure 4. The

atmospheric scattering model describes the equation for obtaining the hazy image degradation model as follows:

$$I(x) = J(x)t(x) + \alpha(1 - t(x)) \tag{12}$$

$I(x)$ is the hazy picture; $J(x)$ is the restored real scene picture; α is the global atmospheric light value, which represents the influence of other light paths in the atmospheric environment on the observation direction, generally a global constant; and $t(x)$ is the medium transmission, which describes the ability of light to penetrate haze, and generally takes the value is between 0 and 1.

When the atmosphere is homogeneous, the medium transmission can be expressed as:

$$t(x) = e^{-\beta d(x)} \tag{13}$$

β is the atmospheric scattering coefficient. When the atmosphere is uniform, β is a fixed value for the whole image at a certain moment. $d(x)$ is the distance from the scene object to the sensor, that is, the scene depth. As the scene depth d increases, the scene brightness decays exponentially.

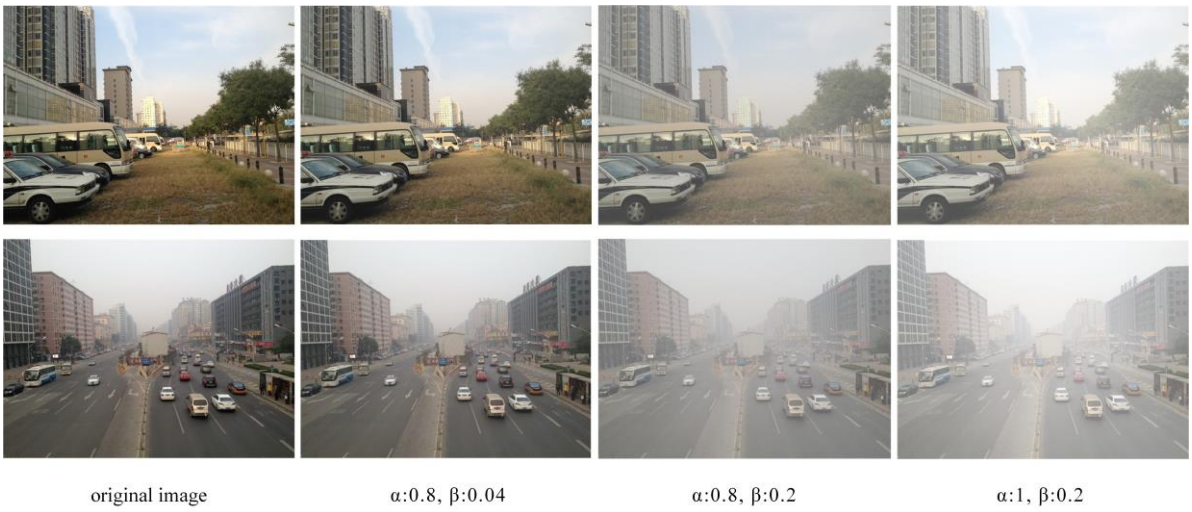

Figure 4. An example of the RESIDE dataset.

Where β represents the decay coefficient, that is, the haze density, $\beta \in \{0.04, 0.06, 0.08, 0.1, 0.12, 0.16, 0.2\}$, including seven categories, α denotes the global atmospheric light, $\alpha \in \{0.8, 0.85, 0.9, 0.95, 1\}$, including five categories, there are 72,135 synthetic haze images after synthesis.

4.1.2. Detection Dataset

There are few public datasets dedicated to haze detection. In order to conduct haze vehicle detection experiments, 6000 RGB images containing vehicles were collected from three public haze datasets, O-HAZE [44], RESIDE [18], and Foggy Cityscapes [45], with a minimum size of 640 × 480 and a maximum size of 2048 × 1536. O-HAZE is the first database of outdoor scenes, composed of pairs of real hazy and corresponding haze-free images. O-HAZE contains 45 different outdoor scenes depicting the same visual content recorded in haze-free and hazy conditions, under the same illumination parameters. The REISDE training set contains 13,990 synthetic hazy images, generated using 1399 clear images from existing indoor depth datasets NYU2 [46] and Middlebury stereo [47]. We synthesize 10 hazy images for each clear image. Foggy Cityscapes derives from Cityscapes

[48] and consists of a large and diverse set of urban street hazy scenes. Then, according to the annotation protocol of Cityscapes, the vehicle class is labeled car, forming a new dataset named Haze-Car. The number of car instances is 72,743. This dataset is used for training, which is divided into three parts: training, validation, and test sets with a ratio of 7:1:2. In order to test the real scene, 100 real haze images were collected from UA-DETRAC [49], RADIATE [50], and the Internet and labeled for evaluation, named Real Haze-100. The dataset is summarized in Table 1.

Table 1. Dataset summary.

Dataset	Total Image	Car Instance	Source	Quantity	Train	Val	Test
Haze-Car	6000	72,743	O-HAZE	2628	2628	0	0
			RESIDE	1531	1531	0	0
			Foggy Cityscapes	1841	0	641	1200
Real Haze-100	100	1114	UA-DETRAC	64	0	0	64
			RADIATE	36	0	0	36

4.2. Experiment Environment

The detailed information of the experimental environment is as follows:

Hardware environment: Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz, 32GB RAM, NVIDIA GeForce RTX 2080Ti GPU, Ubuntu 18.04 OS.

Software environment: The parallel computing framework versions are CUDA10.1, Python3.7, OpenCV3.4, PyTorch1.6, MMDetection2.12.0, and mmcv full1.3.4.

4.3. Evaluation

The average precision (AP) value [51], as a commonly used evaluation index in the field of object detection, can comprehensively reflect the performance of the model. The AP value is an integral of the precision–recall (P-R) curve over the recall rate based on precision. The P-R curve can be obtained by using the recall ratio as the horizontal axis and the precision ratio as the vertical axis. A higher AP value indicates a higher identification accuracy.

Precision (P) refers to the ratio of the true prediction bounding boxes in all prediction bounding boxes, as shown in Equation (14):

$$P = \frac{TP}{TP + FP} \tag{14}$$

Recall (R) refers to the radio of the true bounding boxes in all ground truths, as shown in Equation (15):

$$R = \frac{TP}{TP + FN} \tag{15}$$

TP represents true positive, that is, the number of objects correctly detected by the model; FP represents false positive, that is, the number of objects incorrectly detected by the model; and FN represents false negative, that is, the number of objects missed by the model.

4.4. Experiment Results on Dehazing

In the training of the dehazing model, 512 × 512 synthetic haze images are used to train the model from scratch as a training set. The Adam optimizer is used to set the initial learning rate as 1×10^{-4}, the learning rate attenuates 0.5 every two epochs, and the batch size is 8. Training on eight NVIDIA GeForce RTX 2080Ti GPUs for 50 epochs took 15 h.

The image quality peak signal-to-noise ratio (PSNR) of the test set increased from 15.2 to 24.5.

Figure 5 shows the images after dehazing. Figure 5a,b, respectively, represent the dehazing effect on synthetic and real hazy days. For real haze images collected, the dehazing effect perceived by human eyes is not as clear as the synthetic data test set. The reason is that the image size is different from the training input data. However, the role of the dehazing model is not only to generate haze-free images as the input of the detection subnet but also to learn the visibility enhancement task through the encoder–decoder module to generate clean features, which is conducive to the final detection accuracy.

(a) Synthetic haze image dehazing effect

(b) Real haze image dehazing effect

Figure 5. Visual results of dehazing.

4.5. Comparison Experiment with Other Detection Algorithms

4.5.1. Training Process

In the training of the detector, for the initialization of the model, the parameters trained by the dehazing model in Section 4.4 and the Swin-T [10] model in Swin Transformer are used. AdamW is used as the optimizer, the initial learning rate is set to 1×10^{-3}, and the linear learning rate decay strategy is adopted. The weight decay is set to attenuate 0.1 per 20,000 iterations, and each image is randomly cropped, filled, and flipped horizontally. After 220,000 iterations on 8 RTX 2080Ti graphics cards, the fine tuning took 4 h, in which the loss dropped to about 0.39, and the AP value of the test set reached 91%. The loss function curve is shown in Figure 6.

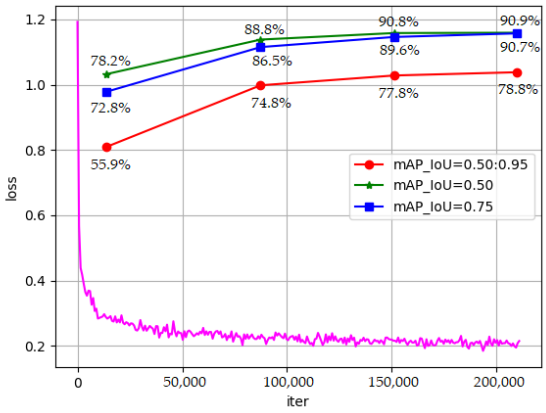

Figure 6. Training process and loss changes.

4.5.2. Comparison of Detection Results

Figure 7 shows the experimental results of six detection algorithms, including the method in this paper and the most advanced detection model SSD [14], Faster-RCNN [9], YOlOv4 [52], EfficientDet [53], and Swin Transformer [10]. Figure 7a–c represents the test sets of Haze-Car, and Figure 7d,e represents images of Real Haze-100 in natural haze days. According to the visualized detection results, the method proposed in this paper has good detection performance for long-distance small vehicles with low visibility in haze days.

Figure 7. Qualitative comparison of detection results of our method and different algorithm models on the synthetic haze image Haze-Car (**a**–**c**) and the natural haze image Real Haze-100 (**d**,**e**).

For the effect of haze detection, SSD, Faster-RCNN, and YOLOV4 algorithms have poor detection performance. It can be clearly seen in Figure 7 that vehicles with clear near distance can be detected, while vehicles with dense long-distance and small objects often miss detection. The precision and recall rate of EfficientDet and Swin Transformer is significantly better than the first three detection methods, and EfficientDet's detection performance for occluded areas and small objects is inferior to Swin Transformer. For example, in the fourth row of Figure 7, the pink circle in the images in Figure 7a,d,e represents that the vehicles at the occluded area are not detected, and the black circles in the images in Figure 7b,c represent the area where the small objects are not detected, and the brown circle in Figure 7a,d indicates that the gray vehicle is not detected. As for the method proposed in this paper, after adding the dehazing module on the basis of Swin Transformer, the detection in hazy scenes shows better performance. For instance, in the sixth row, the dense areas in the yellow circles in the images in Figure 7b,c, and small fuzzy objects in the blue circles in the images in Figure 7a,c,d,e can be accurately detected. Table 2 is the confusion matrix of test sets Haze-Car and Real Haze-100, from which we can see that our algorithm has fewer false detections than missed ones.

Table 2. Confusion matrix of test sets Haze-Car and Real Haze-100.

Test Set	Real	Predict	
		Car	**Back_Ground**
Haze-Car	car	13,290	986
	back_ground	328	0
Real Haze-100	car	917	158
	back_ground	39	0

In order to quantitatively evaluate the performance of the algorithm in this paper, 1200 images, including vehicle categories in the dataset Haze-Car (as shown in Figure 7a–c) and 100 real haze images in self-calibrated Real Haze-100 (as shown in Figure 7d,e), were used for testing, and the results are shown in Table 3. In terms of detection accuracy, the algorithm in this paper is 12.4% higher than the vehicle detection algorithm based on CNN in the public dataset Haze-Car and 30.1% higher in the Real Haze-100. This is thanks to Transformer's attentional mechanism to capture global context information, thereby establishing multi-scale object dependence and extracting more powerful image features. In addition, by adding the image dehazing module, the detection accuracy of Swin Transformer in synthetic haze and real haze test sets is improved by 3.3% and 12.5% respectively, reflecting the performance advantages of this method in real haze days.

Table 3. Comparison results of different detection algorithm performance.

Test Set	Model	Backbone	AP/%	Time/ms
Haze-Car	SSD	VGG-16	43.8	48.2
Haze-Car	Faster R-CNN	VGG-16	47.6	201.4
Haze-Car	YOLOV4	CSPDarknet53	66.4	25.8
Haze-Car	EfficientDet	efficientnet-B1	78.6	19.3
Haze-Car	Swin Transformer	Swin-T	87.7	65.6
Haze-Car	**Ours**	**Dehaze+Swin-T**	**91.0**	**70.1**
Real Haze-100	SSD	VGG-16	39.3	48.2
Real Haze-100	Faster R-CNN	VGG-16	40.5	201.4
Real Haze-100	YOLOV4	CSPDarknet53	45.4	25.8
Real Haze-100	EfficientDet	efficientnet-B1	52.2	19.3
Real Haze-100	Swin Transformer	Swin-T	69.8	65.6
Real Haze-100	**Ours**	**Dehaze+Swin-T**	**82.3**	**70.1**

5. Conclusions

This paper proposes an end-to-end vehicle detection model based on Swin Transformer for vehicle detection in hazy scenes. Firstly, a dehazing network is designed by using Swin Transformer blocks for encoding and decoding. Next, the dehazing module and the Swin Transformer detection module are fused. Then, the transfer learning method is used to train the final end-to-end hazy vehicle detection model. Finally, comparative experiments are conducted on the self-made datasets to prove the effectiveness of each module and the whole framework in real haze scenes. In addition, compared with Swin Transformer, our model has higher detection accuracy, but the speed is slightly reduced. In the future, we will continue to study the real-time performance of the model. This research has certain theoretical and practical significance, and can also be extended to other application fields, such as pedestrian detection, military border warning, and so on.

Author Contributions: Conceptualization, Z.S. and C.L.; methodology, Z.S. and C.L.; software, G.X.; validation, Z.S. and G.X.; formal analysis, Z.S.; investigation, H.Q.; resources, C.L. and H.Q.; data curation, G.X.; writing—original draft preparation, Z.S.; writing—review and editing, Z.S. and G.X.; visualization, G.X.; supervision, H.Q.; project administration, C.L.; funding acquisition, H.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the National Key R&D Program of China (No. 2018YFC0809700).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dong, J.; Pan, J. Physics-based feature dehazing networks. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 188–204.
2. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Processing Syst.* **2016**, *29*, 379–387.
3. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
4. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
5. Le, T.-H.; Jaw, D.-W.; Lin, I.-C.; Liu, H.-B.; Huang, S.-C. An efficient hand detection method based on convolutional neural network. In Proceedings of the 2018 7th International Symposium on Next Generation Electronics (ISNE), Taipei, Taiwan, 7–9 May 2018; pp. 1–2.
6. Li, B.; Peng, X.; Wang, Z.; Xu, J.; Feng, D. End-to-end united video dehazing and detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
7. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
8. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Processing Syst.* **2015**, *28*, 1137 - 1149.
10. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
11. Cai, Y.; Luan, T.; Gao, H.; Wang, H.; Chen, L.; Li, Y.; Sotelo, M.A.; Li, Z. YOLOv4-5D: An effective and efficient object detector for autonomous driving. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–13.
12. Zhao, Z.-Q.; Zheng, P.; Xu, S.-T.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.

14. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
15. He, K.; Sun, J.; Fellow; Tang, X. Single Image Haze Removal Using Dark Channel Prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 2341–2353.
16. Zhu, Q.; Mai, J.; Shao, L. A Fast Single Image Haze Removal Algorithm Using Color Attenuation Prior. *IEEE Trans. Image Processing* **2015**, *24*, 3522–3533.
17. Berman, D.; Treibitz, T.; Avidan, S. Single Image Dehazing Using Haze-Lines. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 720–738.
18. Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; Wang, Z. Benchmarking Single Image Dehazing and Beyond. *IEEE Trans. Image Processing* **2018**, *28*, 492–505.
19. Li, P.; Tian, J.; Tang, Y.; Wang, G.; Wu, C. Deep Retinex Network for Single Image Dehazing. *IEEE Trans. Image Processing* **2020**, *30*, 1100–1115.
20. Cai, B.; Xu, X.; Jia, K.; Qing, C.; Tao, D. Dehazenet: An end-to-end system for single image haze removal. *IEEE Trans. Image Processing* **2016**, *25*, 5187–5198.
21. Li, B.; Peng, X.; Wang, Z.; Xu, J.; Feng, D. Aod-net: All-in-one dehazing network. In Proceedings of the IEEE international Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4770–4778.
22. Zhang, H.; Patel, V.M. Densely connected pyramid dehazing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3194–203.
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Processing Syst.* **2017**, *30*, 6000–6010.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
26. Li, Y.; Zhang, K.; Cao, J.; Timofte, R.; Van Gool, L. Localvit: Bringing locality to vision transformers. *arXiv* **2021**, arXiv:2104.05707.
27. Liu, Y.; Sun, G.; Qiu, Y.; Zhang, L.; Chhatkuli, A.; Van Gool, L. Transformer in convolutional neural networks. *arXiv* **2021**, arXiv:2106.03180.
28. Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; Shlens, J. Stand-alone self-attention in vision models. *Adv. Neural Inf. Processing Syst.* **2019**, *32*, 68–80.
29. Vaswani, A.; Ramachandran, P.; Srinivas, A.; Parmar, N.; Hechtman, B.; Shlens, J. Scaling local self-attention for parameter efficient visual backbones. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12894–12904.
30. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
31. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318.
32. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Chongqing, China, 9–11 July 2021; pp. 10347–10357.
33. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv* **2021**, arXiv:2105.05537.
34. Wu, B.; Xu, C.; Dai, X.; Wan, A.; Zhang, P.; Yan, Z.; Tomizuka, M.; Gonzalez, J.; Keutzer, K.; Vajda, P. Visual transformers: Token-based image representation and processing for computer vision. *arXiv* **2020**, arXiv:2006.03677.
35. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
36. Jiang, X.; Xiao, Z.; Zhang, B.; Zhen, X.; Cao, X.; Doermann, D.; Shao, L. Crowd counting and density estimation by trellis encoder-decoder networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6133–6142.
37. Liang, D.; Chen, X.; Xu, W.; Zhou, Y.; Bai, X. TransCrowd: Weakly-supervised crowd counting with transformers. *Sci. China Inf. Sci.* **2022**, *65*, 1–14.
38. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image transformer. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4055–4064.
39. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
40. Zheng, M.; Gao, P.; Zhang, R.; Li, K.; Wang, X.; Li, H.; Dong, H. End-to-end object detection with adaptive clustering transformer. *arXiv* **2020**, arXiv:2011.09315.
41. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 568–578.

42. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the design of spatial attention in vision transformers. *Adv. Neural Inf. Processing Syst.* **2021**, *34*, 9355–9366.
43. Tong, T.; Li, G.; Liu, X.; Gao, Q. Image Super-Resolution Using Dense Skip Connections. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
44. Ancuti, C.O.; Ancuti, C.; Timofte, R.; De Vleeschouwer, C. O-haze: A dehazing benchmark with real hazy and haze-free outdoor images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 754–762.
45. Sakaridis, C.; Dai, D.; Van Gool, L. Semantic foggy scene understanding with synthetic data. *Int. J. Comput. Vis.* **2018**, *126*, 973–992.
46. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. In Proceedings of the European Conference on Computer Vision, Florence, Italy 7–13 October 2012; pp. 746–760.
47. Scharstein, D.; Szeliski, R. High-accuracy stereo depth maps using structured light. In Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 18–20 June 2003.
48. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
49. Wen, L.; Du, D.; Cai, Z.; Lei, Z.; Chang, M.-C.; Qi, H.; Lim, J.; Yang, M.-H.; Lyu, S. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Comput. Vis. Image Underst.* **2020**, *193*, 102907.
50. Sheeny, M.; De Pellegrin, E.; Mukherjee, S.; Ahrabian, A.; Wang, S.; Wallace, A. RADIATE: A radar dataset for automotive perception in bad weather. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 3 May–5 June 2021; pp. 1–7.
51. Padilla, R.; Netto, S.L.; Silva, E. A Survey on Performance Metrics for Object-Detection Algorithms. In Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niteroi, Brazil, 1–3 July 2020.
52. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
53. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.