Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.046125

ARTICLE

Target Detection Algorithm in Foggy Scenes Based on Dual Subnets

Yuecheng Yu[1,*], Liming Cai[1], Anqi Ning[1], Jinlong Shi[1], Xudong Chen[2] and Shixin Huang[1]

[1]School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang, 212100, China

[2]Information Department, China Merchants Heavy Industry, Nantong, 226116, China

*Corresponding Author: Yuecheng Yu. Email: zhjyuyuecheng@163.com

Received: 19 September 2023 Accepted: 11 December 2023 Published: 27 February 2024

ABSTRACT

Under the influence of air humidity, dust, aerosols, etc., in real scenes, haze presents an uneven state. In this way, the image quality and contrast will decrease. In this case, It is difficult to detect the target in the image by the universal detection network. Thus, a dual subnet based on multi-task collaborative training (DSMCT) is proposed in this paper. Firstly, in the training phase, the Gated Context Aggregation Network (GCANet) is used as the supervisory network of YOLOX to promote the extraction of clean information in foggy scenes. In the test phase, only the YOLOX branch needs to be activated to ensure the detection speed of the model. Secondly, the deformable convolution module is used to improve GCANet to enhance the model's ability to capture details of non-homogeneous fog. Finally, the Coordinate Attention mechanism is introduced into the Vision Transformer and the backbone network of YOLOX is redesigned. In this way, the feature extraction ability of the network for deep-level information can be enhanced. The experimental results on artificial fog data set FOG_VOC and real fog data set RTTS show that the map value of DSMCT reached 86.56% and 62.39%, respectively, which was 2.27% and 4.41% higher than the current most advanced detection model. The DSMCT network has high practicality and effectiveness for target detection in real foggy scenes.

KEYWORDS

Target detection; fog target detection; YOLOX; twin network; multi-task learning

1 Introduction

As one of the important tasks in the field of computer vision, target detection has been widely used in various fields, such as autonomous driving, intelligent transportation, intelligent ships, and so on [1]. Currently, as the mainstream methods, target detection based on deep learning is mainly divided into two categories, including two-stage detection methods and one-stage detection methods [2]. For the two-stage methods, the feature is usually extracted by using the preset candidate boxes, and then the detection task is realized by classification. The representative methods include Mask RCNN [3], Fast RCNN [4], and Faster RCNN [5]. In contrast to two-stage methods, The end-to-end deep learning model is adopted in one-stage methods. The corresponding representative methods include include YOLO series [6–9], SSD [10], and RetinaNet [11], etc. For most target detection methods, using clustering to determine the sizes of bounding boxes is a common approach. Among them, For

most target detection methods, it is a common method to determine the size of the candidate box by clustering, and the three-point center clustering method is one of the common clustering methods [12].

In recent years, fog target detection has gradually become one of the important tasks of intelligent navigation and intelligent ships. When the above target detection model is directly applied to clean images, it can often achieve better results. However, the visibility and contrast of foggy images are significantly lower than that of clean images. Consequently, when these models are directly applied to images in foggy scenes, it is difficult for these models to detect objects in images [13–14]. Currently, there are three common foggy target detection methods. The first type of method belongs to the direct detection method. Firstly, the detection model is pre-trained on the general data set, and then the haze data set is used to fine-tune the model parameters to improve the adaptability of the model in the haze scene [15]. The second kind of method is often designed using the framework of joint optimization [16]. At first, the image restoration network was used to remove the fog from the image to enhance the visibility of the target, and then the target detection task in the fog and haze scene was realized with the help of the joint detection network. Based on this idea, Li et al. [17] used a lightweight CNN to generate clean images and combined them with the Faster R-CNN, which significantly improved target detection in foggy scenes. Similarly, the enhanced dehazing detection network proposed by Li [18] uses multi-scale Retinex for image restoration and then combines with the YOLO network to complete the detection task. The third type of method completes the detection of fog targets by designing a multi-task learning framework [19]. Relying on the backbone network and recovery sub-network, the method of Dual-subnet Network (DSNet) realizes the joint learning of three learning tasks, namely, visibility enhancement, object classification, and object localization, under the same framework [20]. In this way, fog target detection is transformed into a joint optimization problem of multi-task learning, which improves the generalization ability of the model.

Although the direct detection method alleviates the problem of insufficient data sets in the real fog and haze scene, it fails to eliminate the impact caused by fog and haze in essence. Joint optimization methods show that a clean image is conducive to the classification and localization of the target. However, in inhomogeneous foggy scenes, haze elimination often depends on a large dehazing network. In addition, as shown in Fig. 1, when the image contains dense fog or uneven haze, irreversible image quality loss will lead to errors in the image restored by the model. These errors will also bring additional burden to the subsequent detection network. The method of DSNet adopts a multi-task learning framework to realize the simultaneous training of multiple learning tasks under the same framework, which improves the overall robustness of the model. However, in the detection stage, the backbone detection network of DSNet needs to complete the feature extraction of dehazing and detection at the same time. This is bound to reduce the ability of the model to extract deep semantic features in the image.

Two kinds of dehazing detection methods, including joint optimization and multi-task learning, provide us with new inspiration. Intuitively, the clean image is beneficial to improve the fog detection performance of the model, while the collaborative training under the multi-task framework is beneficial to the interaction of information between different learning tasks. Thus, as shown in Fig. 2, a Dual Subnet Based on Multi-task Collaborative Training (DSMCT) is proposed in this paper. DSMCT is composed of a detection sub-network based on YOLOX and a dehazing sub-network based on GCANet. When training the network, the dehazing network as a teacher network can supervise and detect the learning of the network. Because the total loss of the model considers the loss of the dehazing module and the loss of the detection network at the same time, it promotes the detection branch to focus more on the feature extraction of clean information in the fog image. Considering the shortcomings of GCANet in the perception of complex foggy images, the deformable convolution

module is introduced into GCANet to enhance the ability of the network to extract the low-level features of images [21]. In addition, multi-task collaborative training will hurt the extraction of deep semantic information. To alleviate these adverse effects, Vision Transformer (Vits) [22] is added to the backbone network, and the corresponding attention mechanism is designed to improve the insufficient attention of the Vits module to feature location.

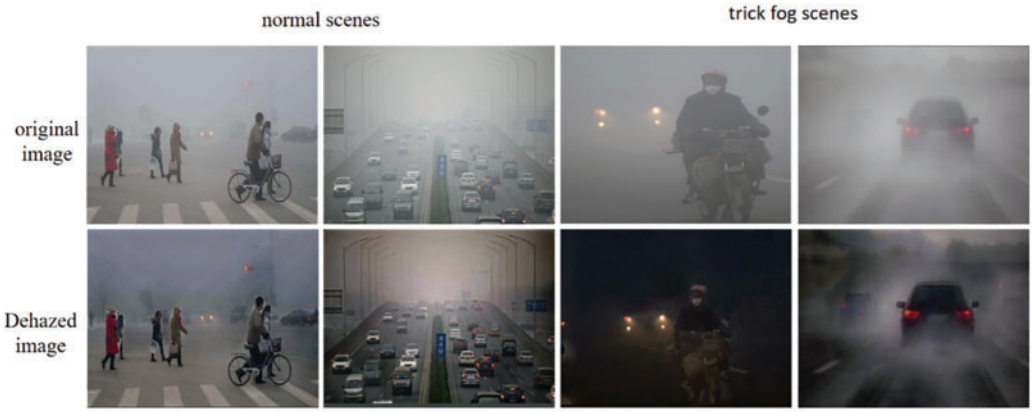

Figure 1: Dehazing visualization

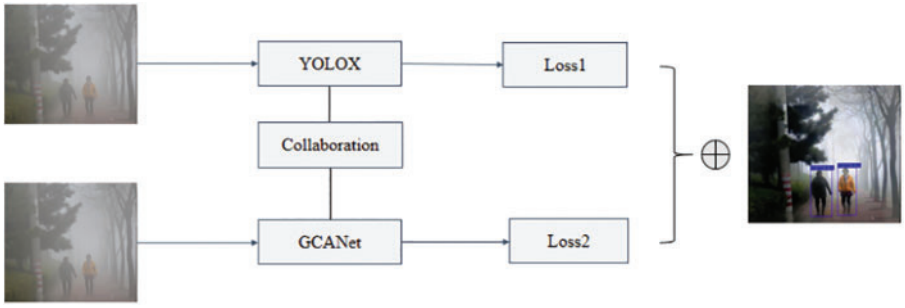

Figure 2: Schematic diagram of collaborative training

2 Related Works

2.1 Target Detection Algorithm

For the target detection algorithm based on deep learning, the two-stage method needs to first divide the image into different candidate regions, and then obtain the detection results based on regional feature recognition [2]. Although this kind of algorithm has relatively high detection accuracy, the complexity of the calculation process affects the real-time performance of the algorithm [3]. In contrast, the one-stage algorithm only needs one feature extraction to achieve target detection, which can better balance the detection speed and detection performance [6]. For target detection in foggy scenes, it generally needs to rely on the dehazing network to extract the clean information in the image, and then run the detection network to complete the target detection. Therefore, when the two-stage target detection method is directly applied to target detection in a fog scene, it is difficult to obtain satisfactory algorithm efficiency.

As a one-stage target detection network, YOLOX adopts CSPDarknet53 as its backbone network, which can greatly reduce the amount of calculation while fully ensuring the ability of feature representation. In addition, the SPP structure of YOLOX uses pooling kernels of different sizes. In this way, YOLOX can better extract features of different receptive fields, so it can better adapt to the recognition of targets with different scales. Compared with other versions of YOLO models, YOLOX adopts the design with anchor-free. This makes the algorithm unnecessary to predefine the size proportion of the detection box and improves the adaptability of the anchor box to various target shapes. In foggy scenes, the target proportion tends to change greatly due to the occlusion of fog and haze, which is often difficult for conventional anchor boxes. YOLOX with anchor-free is more suitable for foggy scenes.

2.2 Transformer Network

Transformer [23] is a neural network model composed of an encoder and a decoder. Originally, applied to machine translation tasks. As a natural language processing model, the encoder and decoder of the Transformer are composed of several multi-head attention modules. With the help of the multi-head attention module, the input word vector group is mapped to the specific Q, K, and V feature spaces, and then the attention scores of all word vectors can be calculated according to formula (1). As shown in formula (1), to prevent gradient explosion, after calculating the correlation degree of Q and K, it is necessary to divide by the square root of the dimension, i.e.,

$$attention\,(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Compared with the CNN network, the Transformer can focus on more important information when calculating the correlation degree between word vectors using the self-attention mechanism. In recent years, Transformer has also been applied in the field of computer vision. Dosovitskiy et al. [24] split the image into 16 × 16-sized vectors and then completed the classification task with the help of the self-attention mechanism in the Transformer. As a lightweight transformer network, mobilevit [25] not only gives full play to CNN's advantages in spatial induction bias but also skillfully combines the transformer's ability to process global information. Chen et al. [26] designed a serialized Transformer to expand low-resolution 2D images into 1D vectors. In this way, by ignoring the two-dimensional structure information in the image and taking the 1D vector after dimension reduction as the input of the network, image classification using a Transformer is realized. It can be seen that when using a Transformer for image processing tasks, slicing or dimensionality reduction can not only avoid the possible memory explosion caused by a Transformer but also give full play to the advantages of a Transformer in feature extraction.

2.3 Collaborative Training

The cooperation of multiple models in collaborative training can improve the overall performance of the algorithm [27]. In the process of collaborative training, although each model focuses on different tasks, the algorithm will rely on the feedback of the total loss of the model to update the gradient. Consequently, collaborative training can realize mutual promotion among multiple tasks in the process of learning. Alex et al. [28] believed that the strength of model performance depends on the relative weights of each task loss. This means that weighting each item of the loss function is conducive to the balance between different learning tasks. Jiang et al. [29] proposed a method for model sharing and collaboration among multiple devices, which can improve the overall performance of the model during training. Fan et al. [30] proposed a group collaborative training method to improve the detection

accuracy of common salient targets in different scenes by using consistent information at the group level. It can be seen that collaborative training needs to design different collaboration mechanisms according to different learning tasks.

3 Approach

As shown in Fig. 3, DSMCT is a dual sub-network of joint dehazing mode. In general, DSMCT includes two sub-networks, namely, the detection sub-network and the dehazing sub-network. YOLOX is the core of the detection sub-network, while GCANet is the core of the dehazing sub-network. As a dehazing network, GCANet pays more attention to feature extraction of clean information. For this reason, in the training phase, GCANet will be considered a teacher network for monitoring YOLOX to extract features. With the help of the influence of GCANet on the total loss of the model, the back-propagated gradients will make YOLOX focus more on extracting the features of clean information, to improve the detection performance of YOLOX in foggy scenes by using the dehazing sub-network. When designing DSMCT, Dconv operation is added at the back of the GCANet network to enhance the network's ability to capture details in foggy images. In addition, when DSMCT performs collaborative training, the ability of the model to extract depth feature information will degenerate. Therefore, we have incorporated Transformer into YOLOX's backbone network. Furthermore, to improve the obvious shortcomings of the Transformer in capturing spatial information, we integrate the Coordinate Attention model into the DSMCT.

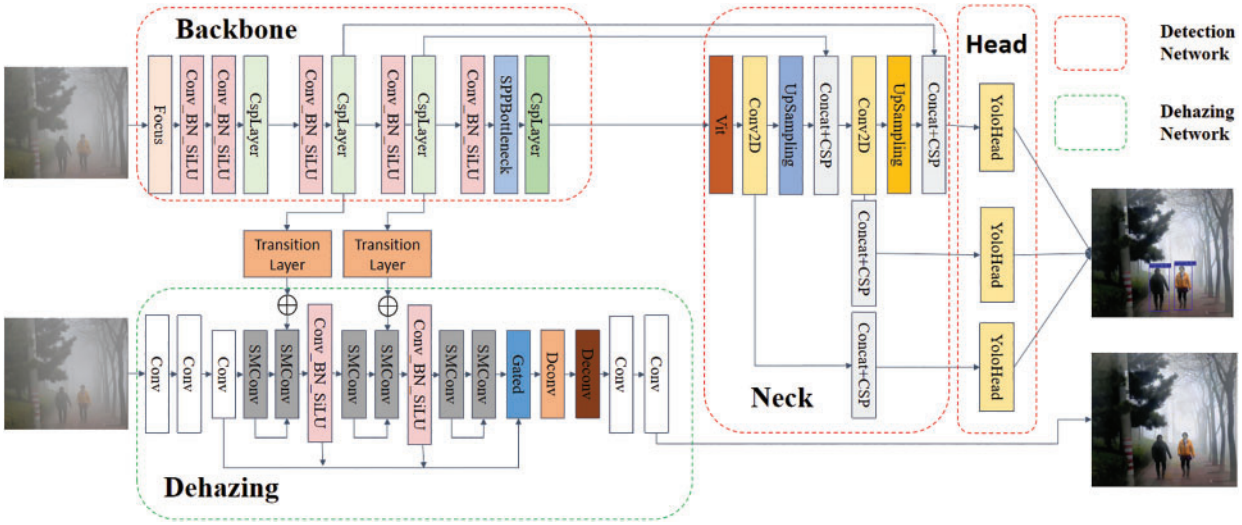

Figure 3: Framework of DSMCT

3.1 Joint Detection Network

In this paper, GCANet is used as a teacher network to supervise the feature extraction of YOLOX, and the ability to extract clean information features using YOLOX is improved by employing back-propagated gradients. Different from the single network model, when designing the joint optimization method of the twin network, we also need to consider the position of the cascade and the weight ratio of each task. There is more pixel-level information in the shallow network, while there is more semantic information in the deep network. If the cascade position is set too shallow, the impact of backpropagation on the network will be small. On the contrary, if the cascade position is set too deep,

it will be difficult for deep-level semantic information to assist in the restoration of pixel-level images. If the quality of the restored image is too poor, the above cascade operation is adverse to the detection network to extract target features and reduce the detection accuracy of the model. Therefore, after comparing the cascading operations between different locations, it is determined to adopt the feature information of the middle layer for cascading operations.

In addition, dehazing is a pixel-level operation with high precision. When the dehazing network and the detection network are cascaded, the unbalanced feature extraction between the detection network and the dehazing network will make it difficult for the dehazing network to restore the fog-free image, which will greatly reduce the detection accuracy. Therefore, the Transition Layer is set at the cascade position of DSMCT, which can be used to solve the problem of the unbalanced number of cascaded channels. Transition Laye contains several operations such as subsampling, normalization, and activation. Its function is to concatenate the channels between two modules of the same size. In addition, to avoid the problem of fine-grained refinement between different tasks, the idea of transfer learning is adopted. After pre-training the dehazing network separately, the results are used to initialize the network weight during joint optimization of the dual subnetworks.

When using the DSMCT for target detection, the losses of two modules will be measured, namely the dehazing loss from the dehazing module and the accuracy loss from the target detection module. When the model is trained by collaborative training, the total loss of the model is obtained by integrating the losses of each module. As shown in formula (2), the total loss of the model includes the regression loss of the boundary box, the confidence loss of the target, the classification loss of the target, and the image restoration loss. The experimental results show that there is an imbalance between image restoration loss and detection loss. Therefore, when calculating the total loss of the model, enlarge the dehazing loss by λ times and then participate in the calculation to balance the weight ratio of dehazing and detection tasks. In formula (2), λ is the amplification coefficient of the dehazing loss, and its value is related to the dataset.

$$\text{total\_}loss = iou\_loss + obj\_loss + cls\_loss + \lambda \cdot fog\_loss \tag{2}$$

As shown in formula (3), the loss of image restoration is measured by MAE loss, where $f(x_t)$ is the pixel value of the haze-free image, y_t is the pixel value of the dehazing image, and n is the number of pixels.

$$fog\_loss = \frac{1}{n}\sum_{t=1}^{n} |y_t - f(x_t)| \tag{3}$$

3.2 Dehazing Network

As an end-to-end dehazing network that can aggregate context information, GCANet gradually restores fog-free images. Since the process of feature extraction of the GCANet network is similar to that of the detection network, we associate the feature information extracted from the backbone network of YOLOX with the middle layer information of GCANet on the channel. In this way, not only will there be no unequal feature information on both sides of the encoder and decoder, as in the U-net network, but also the sufficient integration of feature information between the dehazing network and YOLOX can be realized. For this reason, the interactive information between the two networks by backpropagation, the purpose of YOLOX paying more attention to the clean information in the fog image in the process of feature extraction can be realized.

The structure of the aggregate module of GCANet is shown in Fig. 4. In the process of restoring complex foggy images, GCANet first uses a series of Smooth Dilated Convolutions (SDConv) to extract features and then fuses the feature information. To enhance the representation ability of the model, the fusion of feature information extracted by SDConv at different depths is assisted by the Gate Fusion sub-network (GFSnet). Finally, the aggregated feature layer is restored to the same size as the original image through the deconv module. The receptive field R of a regular convolution is defined by formula (4), where R defines the size of the receptive field and the offset of the corresponding position.

$$R = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\} \tag{4}$$

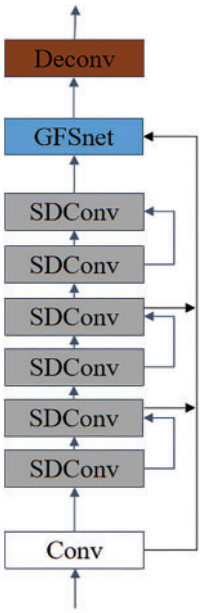

Figure 4: Information aggregation module of GCANet

Regular convolution is shown in formula (5), where p_0 represents each position on the output feature map, p_n represents each value in the convolution kernel corresponding to the receptive field R space, w represents the feature layer, and x is the convolution kernel.

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \tag{5}$$

Compared with regular convolutions, SDConv has a larger receptive field, which can retain more spatial information. However, if SDConv is used too much, it will inevitably lead to the loss of image detail information. This means that only using multiple SDConv to extract features in GCANet is not necessarily advantageous for dehazing tasks in complex scenes.

As shown in formula (6), deformable convolution is a convolution operation with an offset, where p represents the offset. Compared with regular convolution, deformable convolution can effectively capture more edge information. In the convolution operation, the deformable convolution can realize the offset in the X and Y directions after applying a convolution with 2N channels to the input feature

map.

$$y(p_0) = \sum_{pn \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{6}$$

It can be seen that deformable convolution can help the network better adapt to the changes in the shape and edge of specific objects. For this reason, as shown in Fig. 5, the Deconv operation is added after the module of GFSnet. As a result, the offset of deformable convolution in the process of feature extraction can alleviate the loss of image details caused by the SDConv operation and improve the smoothness of the image. Between the dehazing network and the detection network, the feature map processed by the Transition Layer needs to be channel-spliced with the GCANet middle layer. After convolution, normalization, and activation, the spliced feature map is restored to the same size and number of channels as before.

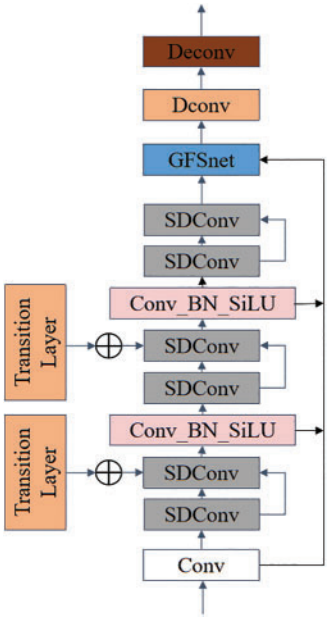

Figure 5: Module with deformable convolution module

3.3 Feature Enhancement Module Integrated with Transformer

The DSMCT consists of a dehazing sub-network and a detection sub-network, and YOLOX is still the core of the detection sub-network. When YOLOX is directly applied to fog images, fog will also be extracted as a class feature. This means that haze as a disturbance feature will reduce the detection ability of YOLOX to real targets. According to the performance that GCANet focuses on extracting clean information, the DSMCT is trained in a collaborative way to help YOLOX reduce the extraction of haze information. To perform cooperative training between the two subnetworks, YOLOX needs to transfer some features extracted in the training process to GCANet. When minimizing the total loss of DSMCT, the information fed back by the loss function can affect the feature extraction of YOLOX to a certain extent, which makes YOLOX more inclined to extract clean information in foggy images. At this time, if the backbone network remains unchanged while the backbone network adds the operation of extracting pixel-level information, this will inevitably weaken the network's ability to understand

deep semantic information. In the subsequent process of target classification, if the model cannot fully understand the deep semantic information, the accuracy of target classification will inevitably decrease.

The standard YOLOX network captures image features by convolution operation, but the size of the convolution kernel limits its capture of global information, and cannot handle the fine interaction between global information and local information. As a deep network based on a self-attention mechanism, the Transformer can not only obtain the global features of images but also enhance the understanding of image semantic information based on capturing local features in different channels. As shown in Fig. 6, when designing DSMCT, a Transformer was added to the bottom of the YOLOX backbone network. It is worth noting that the size of the bottom feature layer of the backbone feature extraction network in YOLOX is $20 \times 20 \times 1024$, while the input of the Transformer is a one-dimensional vector. Therefore, to realize the combination of Transformer and YOLOX, it is necessary to flatten each layer of the bottom features of the backbone network into a vector.

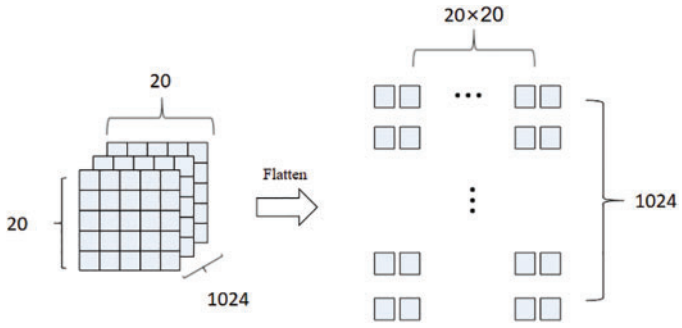

Figure 6: Schematic diagram of incorporating the transformer module

However, the two-dimensional feature itself contains certain spatial location information. If two-dimensional features are mapped directly to one-dimensional features, the spatial information contained in two-dimensional features will be lost. As a spatial attention mechanism, Coordinated Attention [31] can enhance the integration of spatial information in the network. To make up for the deficiencies of the Transformer in capturing spatial information, Coordinate Attention is introduced into the Transformer. Consequently, the module of Coordinate Attention Transformer (CATF) is designed. As shown in Fig. 7, the input of CATF comes from the bottom layer of the backbone network of YOLOX with a size of 20×20. At the same time, the CA module is used to strengthen the spatial information in each channel. When these enhanced spatial features are expanded into one-dimensional vectors, the Transformer can be used to encode the position of all vectors. Furthermore, the features extracted by the residual edge and the multi-head attention module in the transformer are stacked to prevent the gradient from disappearing.

4 Experiment and Analysis

4.1 Dataset and Experimental Environment

Currently, PASCAL VOC is a relatively common target detection dataset, including 20 object categories. In this dataset, the style span of images is large, which is beneficial to improving the generalization performance of the model. According to the method of depth estimation for a single image in reference [32], the foggy data set, named FOG _VOC, is synthesized by combining the depth estimation on the public data set PASCAL VOC with detailed fogging. The foggy effects on some

images in the FOG_VOC datasets are shown in Fig. 8. In addition, the RTTS dataset contains 4322 pictures of real fog and haze scenes, which is the only real data source currently applied to target detection tasks in fog scenes. To fully verify the effectiveness of the model and its detection effect in the real scene, various models are tested on FOG_VOC and RTTS. When training the models, FOG_VOC is divided into a training set and a test set. The training set contains 13405 images and the test set contains 1657 images.

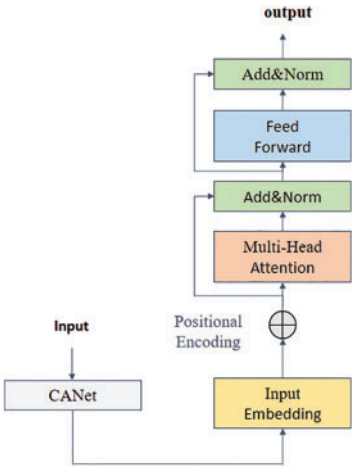

Figure 7: CATF module

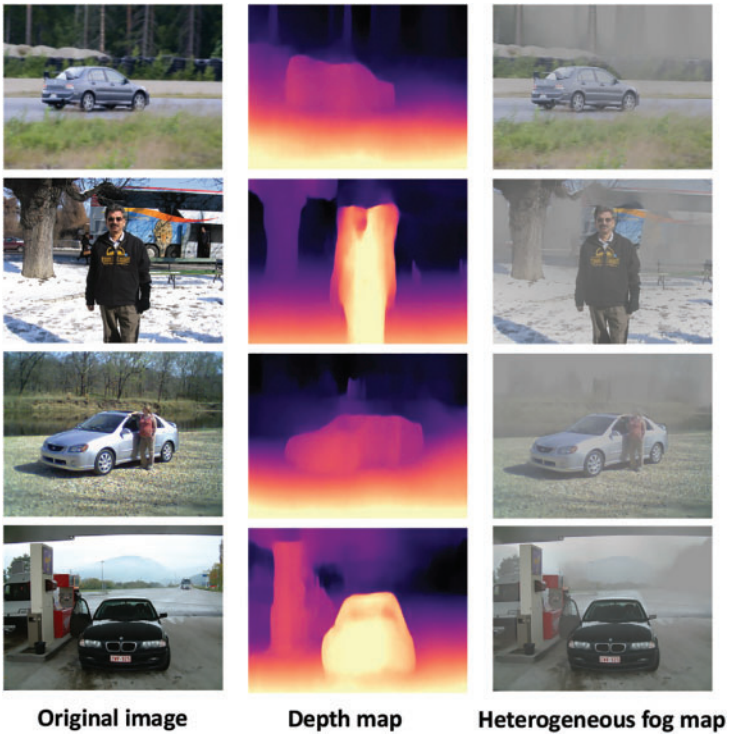

Figure 8: Display of FOG_VOC dataset

In this paper, the GPU model used is RTX 2080 Ti for the training model, and the model code is built based on the Pytorch framework. The random gradient descent (SGD) method was used for model training. In the overall training phase, the learning rate is set to 0.001, the batch value is set to 4, and the number of training iterations is set to 150 epochs.

4.2 Selection of Training Method

Different from the single network target detection model, the training loss of DSMCT includes two parts, namely, dehazing loss and detection loss. If the direct training method is adopted, it is difficult to achieve the simultaneous convergence of the two networks. To ensure that these two parts of the network can work properly, the pre-training weight is used to initialize the model concerning the transfer learning idea. Fig. 9 shows the model losses corresponding to the three training strategies.

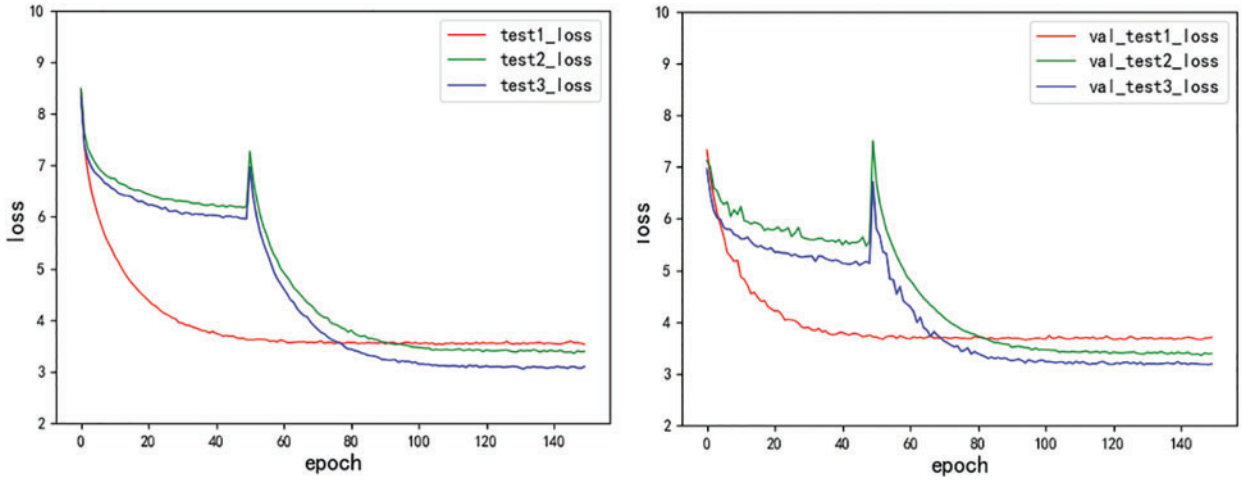

Figure 9: Comparison of model loss

The training strategy adopted by the method of test1 is to directly train the model after initializing the weight of the detection network. As shown in Fig. 9, when using the method of test1 to train the model, the model loss tends to be stable after about 50 iterations. It can be seen that the convergence speed of the model is fast at this time. However, the dehazing image output from the model trained by this method shows that the dehazing branch can not complete the dehazing task well. This means that the auxiliary function of the dual branch network is not effective here. The basic reason is that if only the detection module is initialized, and the initial weight of the dehazing module is a random value, it is difficult to drive the dehazing module to converge in the way of multi-task collaborative training.

The method of test2 improves the training strategy of the test1 method and adopts the strategy of initializing dehazing and detecting the initial weights of two subnetworks at the same time. Its training strategy is to initialize the initial weights of the two subnetworks at the same time. As shown in Fig. 9, when the model is trained with the method of test2, the total loss of the model is optimized. From the dehazing image output by the dehazing branch, the dehazing effect is improved compared with the output image of the method of test1, but it is still not clear enough. After analyzing the total loss, dehazing loss, and detection loss of the model, it is found that the proportion of the dehazing loss value in the total loss is far less than that of the detection loss value. This means that in the process of updating model parameters, the algorithm is more inclined to update the detection sub-network

to reduce the loss value of the detection network. In this case, the algorithm is bound to weaken the update of the dehazing subnet. This will reduce the dehazing performance of the network.

The method of test3 further improves the training strategy of the test2 method. From the experimental results of the test2 method, it can be seen that the dehazing loss value is far less than the detection loss value, which is the prime reason for the poor dehazing effect of the model. This means that balancing the loss values of the dehazing and detection modules will help improve the performance of the model. Therefore, based on the data set used in this experiment, the magnification of the dehazing loss value is determined through experiments. It should be noted that this magnification is the coefficient defined in Eq. (2). The experimental results show that the total loss of the model is more balanced when the coefficient is 5 on the experimental data set. As shown in Fig. 9, after training the model with the test3 method, the total loss of the model is further reduced and the double branch structure is balanced. From the dehazing image output from the dehazing branch, the dehazing effect is improved.

4.3 Experimental Verification and Performance Analysis

To verify the effectiveness of the model in this paper, DSMCT is compared the performance with typical foggy target detection models such as YOLOX, YOLOV7, UHD-YOLOX, IA-YOLOX, AECR-YOLOX, AODNet and DS-Net, respectively on synthetic dataset and real-world datasets. In the above model, YOLOX and YOLOV7 adopt the direct detection methods, UHD-YOLOX, IA-YOLOX, AECR-YOLOX, and AODNet adopt the joint optimization methods, while DS-Net adopts the multi-task learning method. Table 1 shows the test results of the above models on the FOG_VOC dataset, and Table 2 shows the test results of these models on the RTTS dataset. It should be noted that to more intuitively compare the detection results in real scenes, Table 2 only shows the detection results of five categories on the RTTS dataset.

Table 1: Detection precision comparison on FOG_ VOC

	Category					mAP
	Car	Person	Motorbike	Bicycle	Bus	
YOLOX [7]	75.07	77.06	67.65	74.90	83.14	75.57
YOLOV7 [8]	79.08	**88.08**	**84.61**	80.29	89.42	84.29
UHD-YOLOX [33]	80.24	83.58	77.63	76.81	88.93	81.43
IA-YOLOX [34]	77.54	79.18	69.51	75.68	84.22	77.22
AECR-YOLOX [35]	81.27	76.44	70.85	73.47	82.43	76.89
AODNet [17]	77.16	80.25	73.15	74.41	80.68	77.13
DS-Net [20]	80.62	79.44	71.39	77.64	83.28	78.47
DSMCT	**87.44**	84.38	82.88	**83.93**	**94.19**	**86.56**

Fog_VOC data set is a kind of data set with artificial fog. The complexity of haze images in the data set is much lower than that of foggy images in the real scene. As shown in Table 1, the detection accuracy of the joint optimization method is generally higher than that of the method directly using YOLOX for target detection. This shows that these methods of "defogging+detection" can significantly improve the detection accuracy. YOLOV7 is a kind of large-scale detection model, and its parameter scale is far larger than YOLOX. The large-scale parameters in YOLOV7 ensure that

the model can obtain high detection accuracy in low-complexity haze scenes. However, the original YOLOX is a lightweight detection model, which leads to the detection accuracy of each category of YOLOX on Fog_VOC is significantly lower than that of other detection models. This means that when lightweight networks are applied to target detection in foggy scenes, dehazing processing will directly affect the detection performance of the model. It can be seen that the core idea of the "defogging + detection" method is to give full play to the image enhancement of the dehazing network, to improve the accuracy of model detection.

Table 2: Detection precision comparison on RTTS

	Category					mAP
	Car	Person	Motorbike	Bicycle	Bus	
YOLOX	64.61	78.04	44.33	45.63	34.26	52.24
YOLOV7	51.83	80.65	52.18	49.02	29.65	52.67
UHD-YOLOX	64.37	78.27	**55.15**	53.46	38.66	57.98
IA-YOLOX	60.93	70.65	40.19	41.55	25.82	47.83
AECR-YOLOX	61.49	76.23	45.78	44.28	25.74	50.69
AODNet	61.24	76.21	37.35	40.74	25.46	48.21
DS-Net	65.56	80.12	45.21	50.49	30.48	54.37
DSMCT	**74.18**	**84.86**	54.40	**57.90**	**40.60**	**62.39**

As shown in Table 1, The detection accuracy of DSMCT on the FOG_VOC dataset is as high as 84.38%, which is significantly higher than the dehazing accuracy of other types listed in Table 1. This is because DSMCT adopts a collaborative training mode and accomplishes two tasks of defogging and detection in parallel. This training method is conducive to the detection network learning more clean information in foggy scenes and is not affected by the dehazing effect. So, the detection accuracy of the model is improved without increasing complexity.

Compared with the experimental results shown in Table 1, the detection accuracy of almost all models in Table 2 has decreased significantly. The reason is that RTTS is a real-world dataset, and FOG_ VOC is a synthetic dataset. Compared with synthetic foggy images, foggy images in real scenes are more complex. In addition, because the above models are trained by artificial hazing images when these models are applied to target detection in real foggy scenes, There are often obvious domain shifts. As a result, the overall detection accuracy of all models has declined significantly.

In the above methods of "defogging+detection", the detection accuracy of AODNet is the lowest. This is because the model of AODNet adopts a lightweight dehazing network. When it is applied to a real environment with a high complexity of haze, it is difficult to obtain high-quality restored images. In comparison, UHD-YOLOX has higher detection accuracy. This is because UHD-YOLOX adopts a dehazing network with outstanding dehazing performance to ensure the quality of restored images. Therefore, when the dehazing network and YOLOX are jointly optimized, high-quality information restoration can help the feature extraction of the network. As a single-stage direct detection model, the detection performance of YOLOX and YOLOV7 does not depend on the dehazing image. However, for the method of "defogging+detection", the detection task depends on the quality of the image restored by the dehazing network. When these methods are used in real scenes, due to the high complexity of

haze, high-quality restored images cannot be obtained in the dehazing phase, which is bound to harm subsequent detection tasks.

The model of DSMCT proposed in this paper adopts the cooperative training mode of defogging network and detection network. This means that the detection network does not completely depend on the output of the dehazing network. Only the dehazing network needs to provide some weights during training to assist the parameter learning of the detection network. Thus, the influence on the downstream detection task caused by the failure of the dehazing phase is avoided. As shown in Table 2, the detection accuracy of DSMCT is slightly lower than that of UHD-YOLOX in the category of motorbike, and the detection accuracy of other categories is higher than that of other methods. It can be seen that DSMCT has better adaptability in real haze scenes.

To directly display the detection effect of DSMCT, the visualization effect of DSMCT, YOLOX, UHD-YOLOX, YOLOV7, and other models is shown in Fig. 10. As shown in Fig. 10, compared with other models, the recognition effect of DSMCT in the case of haze has been significantly improved, effectively avoiding the omission of target detection. As shown in Fig. 10a, DSMCT can effectively detect all bicycles and people in the foggy image, while the other three models have missed detection to some extent. As shown in Figs. 10b and 10c, the DSMCT successfully detected the vehicles hidden behind the haze, while other models were unable to detect the corresponding targets. When the haze scene is complex, the restoration performance of the dehazing network is limited. In addition, the lack of feature information on the target itself will lead to more noise in image restoration. Fig. 10d shows the detection effect of dense targets in foggy images. As shown in the experimental results, DSMCT can still accurately detect vehicles under haze, while other models will miss detection.

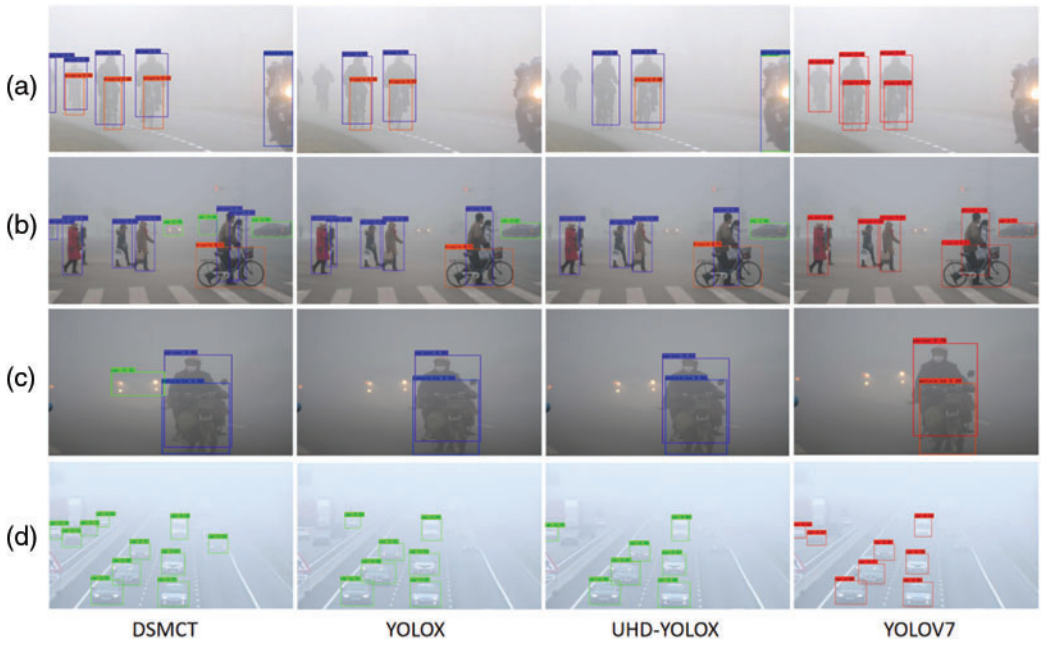

Figure 10: Comparison of visual effects of different models

4.4 Ablation Experiment

To verify the effectiveness of each module, relevant ablation experiments were conducted on RTTS. In the ablation experiments, the basic detection network was YOLOX, and the dehazing network was GCANet. The experimental results are shown in Table 3, the detection accuracy of the original YOLOX in the real fog scene is only 52.2%. Benefiting from the help of dual subnets, the detection accuracy increased by 3.6% after adding GCANet. Then, the deformable convolution module and transformer module were added to the model, and the detection accuracy was improved by 5.0% and 7.2% respectively compared with the original YOLOX. The experimental results show that the addition of deformable convolution improves the dehazing ability of the original GCANet in the real fog scene, and the transformer module further improves the feature extraction ability of the model. Finally, after integrating all the improved schemes, the detection accuracy of the model increased by 10.1%, and the overall accuracy reached 62.3%

Table 3: Abalation experiments on RTTS dataset

YOLOX	GCANet	Deformable convolution	Transformer	mAP/%
√				52.2
√	√			55.8
√	√	√		57.2
√	√		√	59.4
√	√	√	√	62.3

5 Conclusion

In this paper, a dual subnet based on multi-task collaborative training (DSMCT) for target detection in foggy scenes is proposed. In the proposed method, a dehazing network, GCANet, is used as the supervisory network for an object detector, YOLOX, to enhance the extraction of clean information in foggy scenes in order to improve detection accuracy. With the help of multi-task collaborative training method, the dehazing loss of the dehazing sub-network is used to promote the detection network to extract clean information in the foggy image more effectively. In addition, to improve the ability of the dehazing sub-network to extract the underlying features of images, an information aggregation module integrating deformable convolution is designed in GCANet. At the same time, the CATF module is designed based on the integration of Vision Transformer and attention mechanism, which ensures the ability of the backbone network in YOLOX to extract deep semantic information according to collaborative training mode. The experimental results on a synthetic hazy dataset (FOG_VOC) and a realistic hazy dataset (RTTS) show that DSMCT has higher detection accuracy and better adaptability in real foggy scenes compared with existing foggy target detection models. However, since it is impossible to simultaneously take both foggy and non-foggy images in the same scene, the existing model training dataset can only be generated by artificially adding fog. Therefore, we will focus on the research of high-quality artificial fogging algorithms and efficient collaborative training methods in the subsequent work.

Acknowledgement: Thanks to Professor Li Yongzheng's team for their helpful discussion on research background and help on experimental equipment. Thanks to all the reviewers and editors for their helpful suggestions.

Funding Statement: This work was jointly supported by the Special Fund for Transformation and Upgrade of Jiangsu Industry and Information Industry-Key Core Technologies (Equipment) Key Industrialization Projects in 2022 (No. CMHI-2022-RDG-004): "Key Technology Research for Development of Intelligent Wind Power Operation and Maintenance Mothership in Deep Sea".

Author Contributions: Yuecheng Yu and Liming Cai jointly discussed the research content. Yuecheng Yu designed the overall architecture and experimental plan of the model, and wrote and revised the paper. Liming Cai implemented the algorithm and experimental plan, and proposed further modification suggestions for the algorithm based on the experimental results. Anqi Ning has been engaged in the preprocessing of experimental data. Anqi Ning, Jinlong Shi, Xudong Chen and Shixin Huang participated in the analysis of experimental results and the discussion of model optimization. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The download address of the data set of this paper is: https://universe.roboflow.com/test-mdnu9/rtts.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] L. He *et al.*, "End-to-end video object detection with spatial-temporal transformers," in *Proc. ACM*, Chengdu, Sichuan, China, 2021, pp. 1507–1516.

[2] L. Jiao *et al.*, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019. doi: 10.1109/ACCESS.2019.2939201

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, Venice, Italy, 2017, pp. 2961–2969.

[4] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, Santiago, Chile, 2015, pp. 1440–1448.

[5] 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, 2017. doi: 10.1109/TPAMI.2016.2577031

[6] J. Redmon and A. Farhadi, *YOLOv3: An Incremental Improvement*. Los Alamos, New Mexico, USA: National Laboratory, 1991. doi: 10.48550/arXiv.1804.02767

[7] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, *YOLOv4: Optimal Speed and Accuracy of Object Detection*. Los Alamos, new mexico, USA: National Laboratory, 1991. doi: 10.48550/arXiv.2004.10934

[8] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, *YOLOX: Exceeding YOLO Series in 2021*. Los Alamos, New Mexico, USA: National Laboratory, 1991. doi: 10.48550/arXiv.2107.08430

[9] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. CVPR*, Vancouver, Canada, 2023, pp. 7464–7475.

[10] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. ECCV*, Amsterdam, Netherlands, 2016, pp. 21–37.

[11] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. ICCV*, Venice, Italy, 2017, pp. 2980–2988.

[12] Y. Tang, J. Huang, W. Pedrycz, B. Li, and F. Ren, "A fuzzy clustering validity index induced by triple center relation," *IEEE Trans. Cybern.*, vol. 53, no. 8, pp. 5024–5036, 2023. doi: 10.1109/TCYB.2023.3263215

[13] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proc. ICCV*, Seoul, Korea, 2019, pp. 7314–7323.
[14] L. Liu *et al.*, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020. doi: 10.1007/s11263-019-01247-4
[15] J. S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data. Eng.*, vol. 22, no. 10, pp. 1345–1359, 2009. doi: 10.1109/TKDE.2009.191
[16] B. Li *et al.*, "Reside: A benchmark for single image dehazing," *IEEE Trans. Image. Process.*, vol. 28, no. 1, pp. 492–505, 2018. doi: 10.1109/TIP.2018.2867951
[17] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, *An All-in-One Network for Dehazing and Beyond*. Los Alamos, New Mexico, USA: National Laboratory, 1991. doi: 10.48550/arXiv.1707.06543
[18] W. Li, "Vehicle detection in foggy weather based on an enhanced YOLO method," *J. Phys. Conf. Ser.*, vol. 2284, no. 1, pp. 012015, 2022. doi: 10.1088/1742-6596/2284/1/012015
[19] Y. Y. Liao, S. Kodagoda, Y. Wang, L. Shi, and Y. Liu, "Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks," in *Proc. ICRA*, Stockholm, Sweden, 2016, pp. 2318–2325.
[20] C. S. Huang, H. T. Le, and D. W. Jaw, "DSNet: Joint semantic learning for object detection in inclement weather conditions," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2623–2633, 2020. doi: 10.1109/TPAMI.2020.2977911
[21] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. ICCV*, Venice, Italy, 2017, pp. 764–773.
[22] D. Zhou *et al.*, *DeepViT: Towards Deeper Vision Transformer*. Los Alamos, New Mexico, USA: National Laboratory, 1991. doi: 10.48550/arXiv.2103.11886
[23] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, Montreal, Canada, 2017, pp. 6000–6010.
[24] A. Dosovitskiy *et al.*, *An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale*. Los Alamos, New Mexico, USA: National Laboratory, 1991. doi: 10.48550/arXiv.2010.11929
[25] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. ICCV*, Kigali, Rwanda, 2021.
[26] M. Chen *et al.*, "Generative pretraining from pixels," in *Proc. PMLR*, Princeton, NJ, USA, 2020, pp. 1691–1703.
[27] L. Chen, L. Wu, R. Hong, K. Zhang, and M. Wang, "Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach," in *Proc. AAAI*, New York, USA, 2020, pp. 27–34.
[28] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. CVPR*, Salt Lake City, UT, USA, 2018, pp. 7482–7491.
[29] L. Jiang, Z. Zhou, T. Leung, L. Li, and F. Li, "MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proc. of PMLR,* Stockholmsmässan, Stockholm, Sweden, 2018, pp. 2304–2313.
[30] Q. Fan *et al.*, "Group collaborative learning for co-salient object detection," in *Proc. CVPR*, 2021, pp. 12288–12298.
[31] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. CVPR*, Online, 2021, pp. 13713–13722.
[32] C. Godard, O. Mac Aodha, M. Firman, and J. G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. ICCV*, Seoul, Korea, 2019, pp. 3828–3838.
[33] L. Shetty, "Non homogeneous realistic single image dehazing," in *Proc. WACV*, Waikoloa, HI, USA, 2023, pp. 548–555.
[34] W. Liu *et al.*, "Image-adaptive YOLO for object detection in adverse weather conditions," in *Proc. AAAI*, Vancouver, Canada, 2022, pp. 1792–1800.
[35] H. Wu *et al.*, "Contrastive learning for compact single image dehazing," in *Proc. CVPR*, 2021, pp. 10551–10560.