IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 November 2023, accepted 12 December 2023, date of publication 18 December 2023,
date of current version 18 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3344666

RESEARCH ARTICLE

Research on Real-Time Detection Algorithm for Pedestrian and Vehicle in Foggy Weather Based on Lightweight XM-YOLOViT

HUIYING ZHANG, YIFEI GONG, FEIFAN YAO, AND QINGHUA ZHANG
College of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin City, Jilin 132022, China

Corresponding author: Huiying Zhang (yingzi1313@163.com)

This work was supported by the Science and Technology Development Plan of Jilin Provincial Department of Science and Technology under Grant 20220508145RC.

ABSTRACT A novel XM-YOLOViT real-time detection algorithm for pedestrians and vehicles in foggy weather based on YOLOV5 framework is proposed, which effectively solves the problems of dense target interference and obscuration by haze, and the detection effect in complex foggy environments is improved. Firstly, Inverted Residual Block and MobileViTV3 Block are introduced to construct XM-net feature extraction network, secondly, EIOU is used as a location loss function and a high-resolution detection layer is added in the Neck region. In terms of data, a nebulization method is designed to map images from fogless space to foggy space based on the atmospheric scattering model and the dark channel prior. Finally, the validity on four datasets under different foggy environments is verified, respectively. The experimental results show that the accuracy, recall and mAP of the XM-YOLOViT model are 54.95%, 41.93% and 43.15% respectively, and with an F1-Score of 0.474, which is 3.42%,7.08%,7.52% and 13.94% improved, the model parameter reduction of 41.7% to 4.09M, the FLOPs is 25.2G and detection speed is 70.93 FPS compared to the baseline model. The XM-YOLOViT model has better performance than the advanced YOLO detectors, the F1-Score and mAP are improved by 5.57% and 3.65% compared with YOLOv7-tiny, and 2.38%, 2.37% respectively compared with YOLOv8s. Therefore, the XM-YOLOViT algorithm proposed in this article has high detection accuracy and an extremely lightweight structure, which can effectively improve the efficiency and quality of detection tasks for UAV in foggy weather, especially for extremely small targets. Our source code is available at: https://github.com/AFeiV8/XM-YOLOViT.

INDEX TERMS Fog detection, XM-YOLOViT, XM-net, high-resolution layer, nebulization method, lightweight structure, tiny object.

I. INTRODUCTION

In recent years, UAV is widely used to carry out tasks that people not easily accomplish because of small size and fast movement. The detection target is seriously obscured by haze under the foggy weather, which has a great impact on recognition and detection, therefore, the foggy target detection technology is difficulty of UAV vision tasks. However, it is significant because it can collect crucial analytical data for the analysis of foggy road conditions, foggy traffic management, foggy rescue, and other work. In order to obtain a better viewing field, the UAV needs to fly at a certain height and adjust its flight height at any time. Therefore, the scale of targets changes dramatically, and most targets are small and densely distributed, which brings certain difficulties to UAV target detection. Therefore, a more effective fog detection algorithm is urgently needed.

In recent years, significant progress has been made in object detection technology. Researchers have proposed many methods based on deep learning to solve the problem of fog target detection. Researcher has proposed designing a disparity reduction module based on defogging technology [1],

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

[2], [3], [4], [5]. In [6] and [7], researchers designed a style transfer module based on generative adversarial networks to reduce differences. In [8], [9], and [10], researchers utilized the domain adaptive approach to study the alignment relationship between the source domain and the target domain to achieve disparity reduction. Scholars have proposed using more complex structures to enhance the feature extraction ability of the model and improve detection accuracy [11], [12], [13].

Although these works have achieve good results, it is impossible to deploy on UVA and other mobile devices because of the introduction of artificial prior knowledge or the complexity of the architecture, resulting in the algorithm being complex, the model magnitude being too large and not optimized for detection of small and dense objects. In this paper, a lightweight and efficient object detection algorithm based on CNN and Transformer is proposed to detect pedestrians and vehicles under fog conditions for UAVs. These primary innovations are as follows:

1) The Inverted Residual Block and MobileViTV3 Block are used to design XM-net as the backbone for the XM-YOLOViT model to enhance the feature extraction and global modeling ability.

2) The EIoU Loss is used as the localization loss function to accelerate convergence and improve detection accuracy.

3) The detection layer with higher resolution can enhance the multi-scale processing ability of the model, and the influence of the drastic change of target scale during UAV flight is suppressed, thereby, the detection accuracy is improved.

4) In terms of data, a fogging algorithm is designed based on the atmospheric scattering model and the dark channel prior which mapping images to foggy spaces, to obtain more natural foggy images for training and testing the high-performance model proposed in this paper.

The remainder of the paper is organized as follows: The related work is presented in section II. Section III describes the design process and architectural details of XM-YOLOViT. Section IV introduces the fog detection algorithm and its implementation in detail. Section V describes the related experimental tests and analyzes the experimental results. Section VI discusses the limitations of the work and presents future work. Section VII is the conclusion of this paper.

II. RELATED WORKS

The related work in this paper covers the current state of object detection, foggy object detection, model lightweight, and small dense objects. Representative works in each field are reviewed as below.

A. OBJECT DETECTION

At present, one-stage (SSD series [14], [15], [16] and YOLO series [17], [18], [19]) and two-stage (Faster-RCNN [20]) are the main target detectors. The two-stage network has high detection accuracy, but the real-time performance is poor due to a large amount of computation. One-stage algorithm with better real-time performance is simple and efficient. Among them, YOLOv5 is the most commonly used, but the detection accuracy is low in complex environments, especially for small targets, dense targets, and extreme weather conditions with severe occlusion.

B. OBJECT DETECTION IN FOGGY WEATHER

In foggy weather, severe degradation of image quality or severe occlusion of targets can have a significant impact on target detection. Researchers have proposed many solutions based on deep learning, which can be divided into two categories.

In the first type of research, the first class of research the artificial prior knowledge or deep learning module is introduced to reduce the difference between the source and target domains before performing the detection task. Some researchers used physical models to represent foggy images. He et al. [1] proposed a single-image defogging method based on the dark channel prior, Zhu et al. [2] proposed a fast single-image defogging method based on the color attenuation prior. The image becomes clear by defogging and the image quality is improved through the above technology, but the original texture and color of the object may be damaged to some extent. Liu et al. [3] proposed to add a dark channel defogging algorithm based on YOLOv7. The efficiency of dark channel defogging algorithm is effectively improved by down-sampling and up-sampling, and the head of ECA module is added to the network to improve the accuracy of target classification and regression. Dong et al. [4] proposed a multi-scale enhancement feature fusion defogging network based on U-Net structure, which skillfully combines enhancement strategies and back projection techniques for image defogging. The IDOD-YOLOv7 model is suggested by Qiu et al. [5] which is based on IDOD module and YOLOV7 module for joint learning, and the IDOD module is responsible for image defogging and image enhancement to improve detection accuracy. An improved YOLOv5 algorithm is presented by Zhai et al. [21], the brightness and contrast of the image are adjusted by the improved adaptive histogram equalization method in the process of image preprocessing, which highlights the detailed information of vehicle image markers and improves the detection accuracy of a single image. Liu et al. [22] used GCANet to defog the image and enhance the representation of the object boundary, then the YOLOV5 model is used to realize the detection. Several studies have used generative adversarial networks for image preprocessing. Shan et al. [6] proposed a UDA model that uses the generation of adversarial networks for image translation, which enhanced the anti-interference ability of the detection network. a new generative adversarial network based on CycelGAN is designed by Guo et al. [7] to achieve style translation between foggy images and normal images before performing the detection task. Some studies that use domain adaptive methods to study the alignment relationship between source and target domains, Hu et al. [8] proposed DAGL-

Faster, which enhances Faster-RCNN with multiple domain classifiers, these classifiers assist the network in extracting features that are invariant to the domain difference between the source domain (typical weather) and the target domain, and the consistency regularization is introduced to optimized the detection performance, Sindagi et al. [9] defined a new prior adversarial loss based on prior knowledge to supervise the adaptive process, which mitigating the effect of weather on the detection performance, Liu et al. [10] proposed a domain adaptive model called IA-YOLO and a differentiable image processing (DIP) module is designed, which can adaptively learn the brightness, color, tone, and weather features of an image. The interference of weather information in the image can be suppressed and the potential information can be recovered after processing by DIP module.

In another class of research, the feature extraction ability and the detection accuracy of the model are improved by designing the complex structure. Meng et al. [11] introduced SwinFocus based on YOLOv5 to enhance the feature extraction capability of the original algorithm and added a decoupling head to the model, which effectively improves the detection performance of fuzzy and small targets under foggy conditions. To improve the robustness and detection performance of the network, Fang et al. [12] proposes the ODFC-YOLO model which adds a cross-stage partial decoder at the mid-end of the backbone to reduce the difference between fuzzy and clear images, meanwhile, the GCEE module is used to construct global contextual features and remote dependencies. Wang et al. [13] proposed a foggy day detection algorithm based on YOLOv5, the parameterized ResNeXt model is used as the backbone, meanwhile, the FEM module is designed to extract more useful features using the attention mechanism.

C. MODEL LIGHTWEIGHT AND SMALL DENSE OBJECT

Due to the limited computing resources and storage space of UAV or other mobile devices, more lightweight and efficient algorithms are needed to realize target detection. To reduce the computational amount of the standard convolution, Howard et al. [23] proposed the Depthwise Separable Convolution structure, which separates the standard convolution into Depthwise Convolution (DW convolution) and Pointwise Convolution (PW convolution). Han et al. [24] proposed the Ghost module, which is used to generate more feature mappings from low-cost operations, thereby the computational overhead is reduced and the representation capability of the model is improved. These structures are usually used for lightweight work of models [25], [26], [27], [28], [29] because they can be used as plug-and-play single-layer structures and lightweight backbones.

In some studies, optimization has been done in the direction of model lightweight and density target detection. Yang et al. [30] suggested a new detection head to increase the model detection accuracy for small targets, K-means++ algorithm to optimize the scale of the initial anchors, the GhostNet module is used to replace the relevant convolution in YOLOv5 to create lightweight models. Wang et al. [31] proposed a lightweight target detection algorithm based on YOLOv4, which used MobileNetv3 as a backbone to reduce the model parameters, then used inflated convolution instead of the SPP structure, Dcn-Dw structure to replace convolution operation in PAN-Net, and finally introduced CBAM module before the Head. Li et al. [32] proposed an improved lightweight dense pedestrian detection algorithm which the GhostNet is used as the backbone to reduce the number of parameters and the amount of computation. In the front part, CBL module is replaced by GSCV module, CSP module is replaced by VoV-GSCSP module, which improves the overlap of prediction frames in dense scenes. Although these works can effectively reduce the model parameters and the lightweight model is achieved, the improvement of detection accuracy is less, even at the expense of accuracy.

There are some similar studies on small and dense object detection by UAV. Based on YOLOv5s, a small target detection algorithm [33] for UAV is proposed by multi-scale feature fusion, improved ASFF and adding CBAM module before backbone network and each prediction network. The detection performance has been improved, but the detection accuracy is still not high enough, especially in extreme weather. Zhu et al. [34] proposed a new detection head TPH for UAV to small target detection based on YOLOv5. The Transformer is integrated into the C3 module, and a tiny object detection head is added. However, the design for TPH is rather heavy, which will seriously affect the real-time performance.

III. DESIGN OF XM-YOLOViT MODEL

The XM-YOLOViT model based on the YOLOv5 architecture is shown in Fig. 1. The overall architecture consists of three parts: Backbone, Neck, and Head. Backbone (XM - Net) performs the feature extraction, the Neck part consists of SPPF, FPN, and PAN, which is able to enhance multi-scale capability of the model, and Head is responsible for target prediction and localization.

The main components of YOLOv5s are shown in Fig. 2, where CBS is a three-layer structure of ConvBNSiLU. Among them, the BottleNeck1 is used only for C3 Blocks in CSP-Backone (named backbone in YOLOv5 in this way in the paper), whereas the BottleNeck2 is used for the other C3 Blocks.

A. DESIGN OF XM-NET ARCHITECTURE

A hierarchical feature extraction network called XM-net with strong extraction capability is proposed, and the problems of feature extraction and insufficient global modeling capability of CSP-Backbone are solved. The down-sample and mapping structures are designed based on the Inverted Residual Block [35] and MobileViTV3 Block [36]. The basic feature extraction unit is composed of down-sampling and mobile-vit3block, which is used repeatedly in the feature extraction process. The structure of XM-net is shown in Fig. 3.

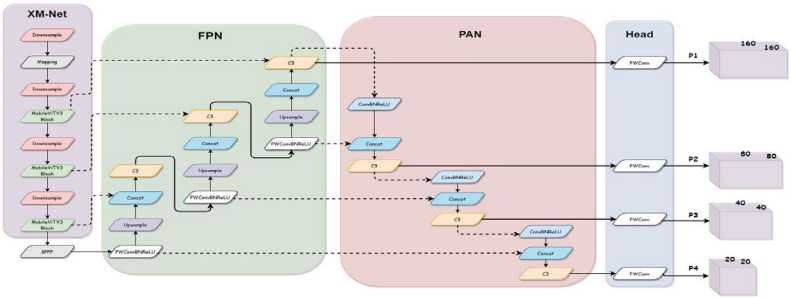

FIGURE 1. XM-YOLOViT architecture diagram.

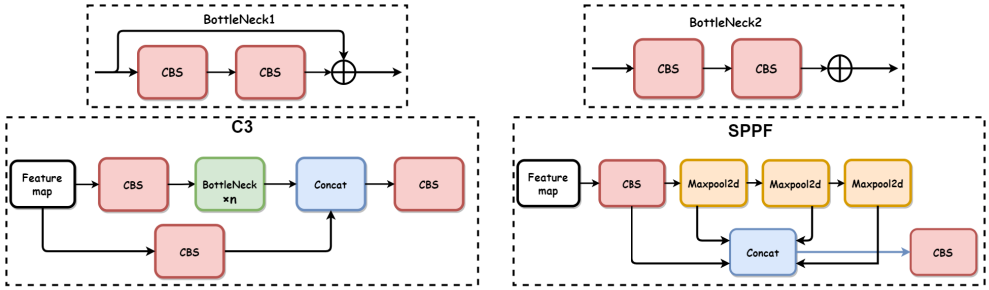

FIGURE 2. Main components of YOLOv5s.

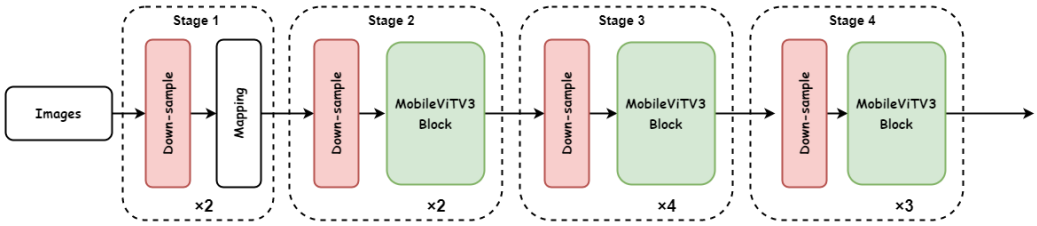

FIGURE 3. Structure of XM-net.

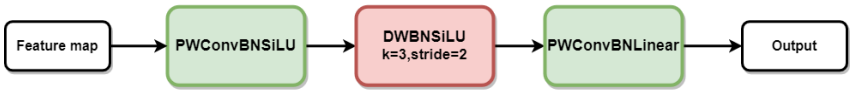

FIGURE 4. Structure of down-sampling layer.

FIGURE 5. Architecture diagram of mapping.

B. DESIGN OF DOWN-SAMPLE AND MAPPING

Down-sampling layer is mainly used to down-sample the feature map. In the first Down-sampling layer, a CBS module with convolutional kernel size of 6 × 6 and step size of 2 is used to obtain a broader receptive field. All the remaining down-sampling layers use Inversed Residual Blocks with

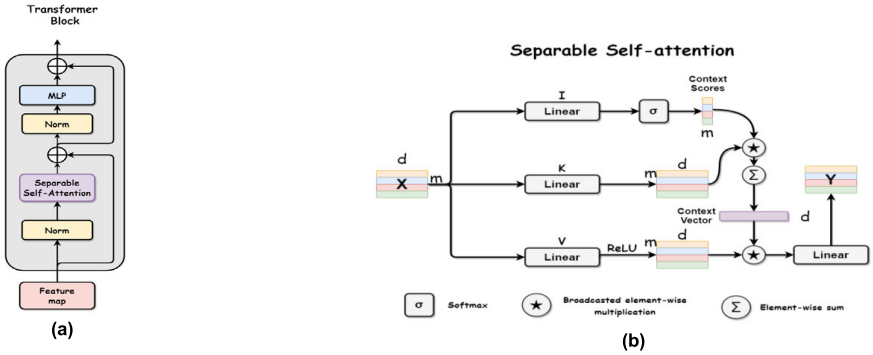

FIGURE 6. (a) Transformer blocks used in MobileViTV3 (b) Structure of separable self-attention.

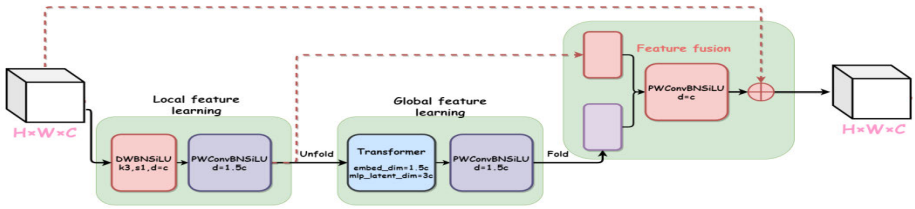

FIGURE 7. Structure of MobileViTV3 block.

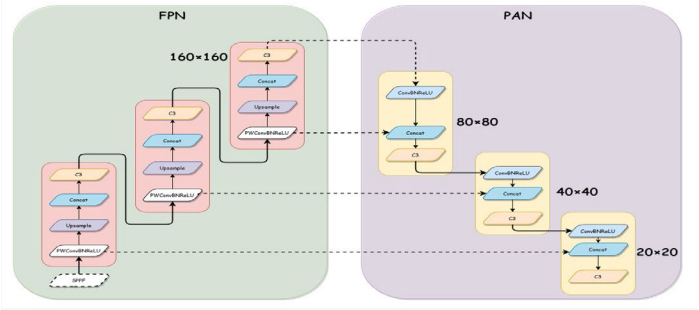

FIGURE 8. Structure of FPN and PAN.

step size of 2, and the down-sampling structure is shown in Figure 4.

The Inverted Residual Block is a lightweight structure. First, the input features are projected into the high-dimensional space using a linear combination of PW convolutional learning input channels, and then the input features are extracted by DW convolution, which are input into the PW convolution module and projected into a low-dimensional space. A linear activation layer is used after the final PW convolution to prevent nonlinear activation functions from seriously destroying information in low-dimensional features. Mapping is mainly used to adjust the number of channels to map the features to the high dimensional space. The Mapping layer is implemented by Inverted Residual Block with a stride of 1 and its structure is shown in Fig. 5.

The Mapping includes a shortcut branch which is different from the Inverted Residual Blocks with step size of 2. The shortcut branch is used only when the stride of the DW convolutional layer is 1 and the input features have the same shape as the end features since it is an additive operation at the end.

C. MobileViTV3 BLOCK

The three components of MobileViTV3 Block are local feature learning, global feature learning and feature fusion. The local feature learning part combines DW convolution and PW convolution to encode the local information of the input features, and then learn the linear combination of the input channels and project them to the high-dimensional space. The global feature learning part first uses the Unfold layer to divide the features into patches and convert them from image form to sequence form, followed by the Transformer Blok to encode the relationships between the patches for global interaction modeling, and then the Fold layer is used to reduce the features to image form.

A SSA (Separable Self-attention) [37] with linear temporal complexity and relative lightweight is employed in

the Transformer Block when global feature learning is performed, as shown in Fig. 6(a). The Separable Self-attention employs the three-branch to process input feature X as shown in Fig. 6 (b). In branch I, each token$\in \mathbb{R}^d$ in X is first mapped to a scalar using the linear layer to get $X_I \in \mathbb{R}^m$, and then, the $c_s \in \mathbb{R}^m$ is obtained by Softmax operation. The broadcast mechanism as well as element-by-element multiplication is used to calculate c_v and the final result. In MSA (Multi-head Self-attention) [38], it is necessary to multiply each feature information in the input token with each information in the key token to obtain the attention matrix, the time complexity is $O(k^2)$. However, in SSA, the input token is first mapped into a scalar and then multiplied with each message in the key token using the broadcast mechanism, and with a time complexity is $O(k)$. Therefore, the memory space is saved and the computational efficiency is improved by SSA.

Feature fusion is the fusion of local and global features and shortcut branches. To obtain the final output of MobileViTV3 Block, the local and global learning features are first combined and mapped to the same dimension as the input features using PW convolution, finally, additively fused with shortcut branches. Structure of MobileViTV3 Block is shown in Fig 7.

D. HIGH RESOLUTION DETECTION LAYER

In this paper, a PFN and PAN structure with three detection layers is designed. A new detection layer with higher resolution (160 × 160) is added to the three detection layers, which preserves more original image information and enhances the multi-scale processing capability of the model, reduces the influence of the target scale change during UAV flight and is beneficial to the detection of the minimal target.

As shown in Fig. 8, the cascaded unit in FPN structure includes adjustment channels, up-sampling and feature stitching, and finally a feature map with a scale of 160 × 160 is obtained. Because of the symmetry between the FPN and the PAN, the same three-layer cascade structure is used in the PAN, and the CBS module in the cascade unit is used to down-sample the image features, and the feature maps with three scales of 80 × 80, 40 × 40, and 20 × 20 at the end of the PAN will be obtained. Four detection layers with various resolutions will be obtained after processing by the Head unit.

E. LOSS FUNCTION

In this paper, the classification accuracy and positioning accuracy of the model are jointly optimized by using classification loss, confidence loss and positioning loss. Binary Cross Entropy Loss (BCE Loss) is used for classification loss and confidence loss, EIoU Loss [39] is used for localization loss to improve the detection accuracy. The formulas of the three loss functions are as follows:

$$L_{\text{cls}}(O, P) = -\lambda_{\text{cls}} \frac{\sum_{i \in pos} \sum_{j \in cls} (O_{ij} \ln \sigma(P_{ij}) + (1 - O_{ij}) \ln \sigma(1 - P_{ij}))}{N_{pos}} \tag{1}$$

$$L_{\text{conf}}(P_{obj}, C) = -\lambda_{\text{conf}} \frac{\sum_i P^i_{obj} \ln \sigma(C_i) + (1 - P^i_{obj}) \ln \sigma(1 - C_i)}{N} \tag{2}$$

$$\sigma(x) = Sigmoid(x) \tag{3}$$

In Equation (1), N_{pos} denotes the number of all positive samples, $O_{ij} \in \{0, 1\}$ denotes the presence or absence of the *jth* target type in prediction box *i*, P_{ij} denotes the result predicted by the model. In equation (2), P^i_{obj} denotes the presence or absence of the target in box *i*, C_i denotes the confidence value of the target in the prediction box *i* for the model.

$$L_{loc} = \lambda_{loc} \left[1 - IoU + \frac{\rho^2(b^p, b^{gt})}{(w^c)^2 + (h^c)^2} + \frac{\rho^2(w^p, w^{gt})}{(w^c)^2} + \frac{\rho^2(h^p, h^{gt})}{(h^c)^2} \right] \tag{4}$$

$$IoU = \frac{GT \cap P}{GT \cup P} \tag{5}$$

In Equation (4), ρ^2 denotes the square of the Euclidean distance, b^p and b^{gt} represent the coordinates of the center point of the prediction and reality boxes, respectively, w^p and h^p are the widths and heights of the predicted box, w^{gt} and h^{gt} are the widths and heights of the real box, w^c and h^c represent the width and height of the minimum bounding rectangle formed by the prediction box and the real box. The meaning represented by each symbol is shown in Fig. 9.

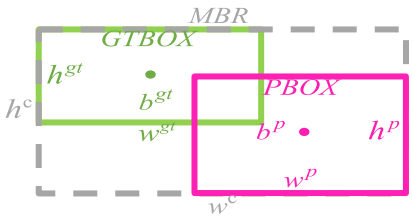

FIGURE 9. Meaning of various symbols for EIoU.

The different λ_i in Equation (1), Equation (2) an Equation (4) denote the balance coefficients of the three losses, and the total loss can be expressed as:

$$L = L_{\text{cls}} + L_{\text{conf}} + L_{loc} \tag{6}$$

IV. DESIGN OF FOGGING ALGORITHM

In this paper, a mapping relation is designed based on the atmospheric scattering model [40] and the dark channel prior [1], which maps the fog-free image data into the fogged space, and the parameters are fine-tuned according to the actual effect to enhance the data diversity.

A. ATMOSPHERIC SCATTERING MODEL

The atmospheric scattering model is a mathematical model that describes the scattering of light as it passes through the

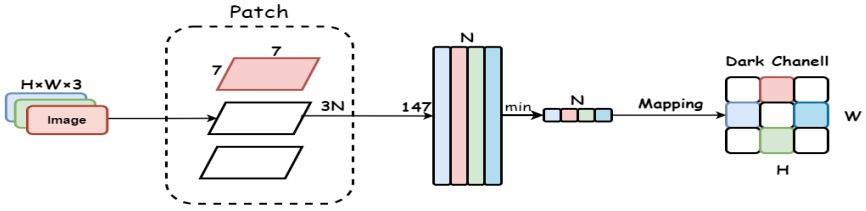

FIGURE 10. Schematic diagram of dark channel calculation process.

Algorithm 1 Fogging Algorithm

while i ≤ number of batches **do**

Step I : Get Dark Channel

Require: b, the batch size. H and W, the image height and width

- Pad images: panding, and Kernel initialization: kernel size, kernel size = K, padding=K//2
- Use Kernel to divide images into multiple patches: shape (b, 3, H, W) → (b, 147, H×W)
- Perform minimum value operation on all patches to obtain dark channel: shape (b, 147, H×W) → (b, 1, H×W)

Step II : Calculate Atmospheric Value

Require: R, the sampling ratio.

- Sample a certain proportion of dark pixels in dark channel: shape (b, 1, H×W) → (b, 1, H×W×R)
- Find the locations of these dark pixels on the original image: shape (b, 3, H×W×R)
- Perform a maximum value operation on these pixels and average the three channels to obtain the atmospheric light value: shape (b, 3, H×W×R) → (b, 3, 1) → (b,1)

Step III : Mapping images using physical models

- Pixel normalization: Pixel / 255
- Use formula (2) (3) (4) to map the image to obtain the foggy image: the experimental section shows in detail all parameters in equation.

i = i + 1

end while

atmosphere, the expression is expressed as:

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{7}$$

In Equation (7), $I(x)$ represents a foggy image, $J(x)$ represents fog-free image, $t(x)$ is the atmospheric transmittance, and A is the atmospheric light value. Therefore, when $t(x)$ and A are known, the image is mapped from fog-free space to foggy space. The following equations can be used to estimate atmospheric transmittance when the atmosphere is uniform, $t(x)$ is expressed as:

$$t(x) = e^{-\beta d(x)} \tag{8}$$

$$d(x) = s - \frac{\rho(L, C)}{20} \tag{9}$$

In Equation (8), x is the atomization center, β is the atmospheric scattering coefficient, and $d(x)$ is the scene depth of radiation that can be calculated by Equation (9). In Equation (9), ρ is the Euclidean distance between the two data, s is the dimensions of the fogging space, L is the height and width information of the original image, C is the coordinates of the center point of the fogging space.

B. DARK CHANNEL PRIOR

In most non-sky regions for the fogless images, pixels have extremely low intensity (dark pixels) in at least one of the three RGB color channels. The channel composed of these dark pixels is called a dark channel, defined as:

$$J^{dark}(x) = \min_{c\in\{r,g,b\}}(\min_{y\in\Omega(x)}(J^c(y))) \tag{10}$$

In Equation (10), J^c is color channel of J, $\Omega(x)$ is a local patch centered on x. Assuming $J(x)$ is an outdoor image without fog, the intensity of J^{dark} is extremely low and tends to zero except for the sky region based on the dark channel priority theory.

C. ESTIMATING THE ATMOSPHERIC LIGHT VALUES

In this paper, the value A of atmospheric light value is estimated from real fog images used dark-channel prior. The

steps for obtaining the atmospheric light value are as follows: (1) calculate the dark channel, (2) sample the brightest dark pixels that are proportional to the number of image pixels, (3) The sampled dark pixel corresponds to the original image, and the average value of the three RGB channels of that pixel is the atmospheric light value A.

The entire image is first divided into multiple patches of 7×7 when calculating the dark channel. The dark channel is obtained by flattening all channels in the image together with all patches to obtain N (number of patches) group vectors of $P \in \mathbb{R}^{147}$. Since the use of sliding windows similar to convolutional operations to partition the image into patches, N is equal to the length of the image multiplied by its width. And then performing a matrix minimization operation on $\Phi \in \mathbb{R}^{147\times N}$, as shown in (11):

$$J^{dark}(x) = \min_{y\in\Phi(x),d=0}(y) \tag{11}$$

In Equation (11), $\Phi(x)$ represents a matrix $\Phi \in \mathbb{R}^{147\times N}$ that contains all patch pixels of three channels, d is the dimension in which matrix minimization operation is performed. Calculation process of dark channel is shown in Fig. 10.

When sampling dark pixels, the brightest dark pixel in the dark channel is taken. Most of the image resolution is 1920×1080 for the collected real fog image, therefore, the number of dark pixel sampling is modified. Selecting 0.9% of the number of pixels in the original image as the number of samples, that is, the first 0.9% of the brightest dark pixels in all dark channels are sampled, therefore, the operation of sampling dark pixels is completed. Finally, the sampled dark pixels corresponding to the same position of the original image are selected and the maximum value of the sampled pixels is limited to 220 (0.89). The atmospheric light value A is determined by averaging the three channels of RGB at that pixel position. The process is shown in Fig. 11. Pseudo code of the algorithm in this section is shown in Algorithm 1.

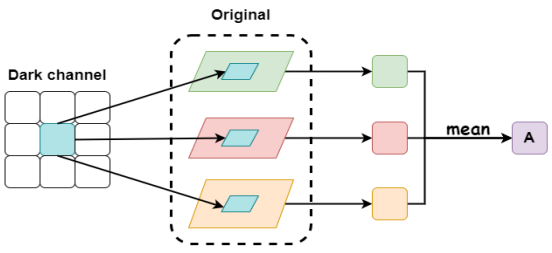

FIGURE 11. Process of calculating A.

V. EXPERIMENTAL TESTING AND VALIDATION

To verify the performance of the proposed algorithm, the ablation experiments are carried out in four different fogging spaces, moreover, the comparative experiments are carried out for XM-YOLOViT and real-time detector in YOLO series. Then the global modeling ability of XM-Net is visualized to show the powerful global modeling ability of XM-Net. Finally, a comparative experiment is conducted on the visual detection effects of XM-YOLOViT and YOLOv5s under foggy conditions. The experimental platform and training settings are shown in Table 1, all models in the experiment are trained with the same settings, the same configuration as the default hyper-parameter and training strategy of YOLOv5 are used.

TABLE 1. Experimental configuration table.

Configuration	Model or version
system	Windows 11
CPU	12^{th} Gen Intel(R) Core(TM) i7-12700H
GPU	NIVIDA RTX3050 Mobile (4G)
framework	Pytorch 1.13.1
input size	640
batch size	4
epoch	150
optimizer	SGD

TABLE 2. The number of images in each datasets.

Classification	Number
Train	4615
Validate	1320
Test-L	660
Test-M	660
Test-H	660
Test-Rand	660

A. DATASETS

It is impossible to find a publicly available dataset that can satisfy the requirements of UAV shooting, real foggy environment, minimal targets and dense targets at the same time. To reduce the cost of experiments and to compensate for the shortcomings of current public datasets, the algorithm proposed in this paper is used to generate image data to meet the research needs on the basis of high-quality UAV data sets. At the same time, the diversity of the dataset can be flexibly enhanced and the test on multiple datasets in different foggy environments for the model can be realized. In this paper, the large-scale UAV image dataset which collected and published by the AISKYEYE team in the Machine Learning and Data Mining Laboratory of Tianjin University [41] is used, that is VisDrone2019-DET.The original dataset is processed to meet the research requirements.

First, the night-time images and their labeling data in the datasets are deleted. Secondly, the object categories are adjusted so that all samples are eventually divided into pedestrian, car, and LV (light-duty vehicles). Finally, the number of images in the training set, verification set and test set are adjusted, the final number of images in the training set is

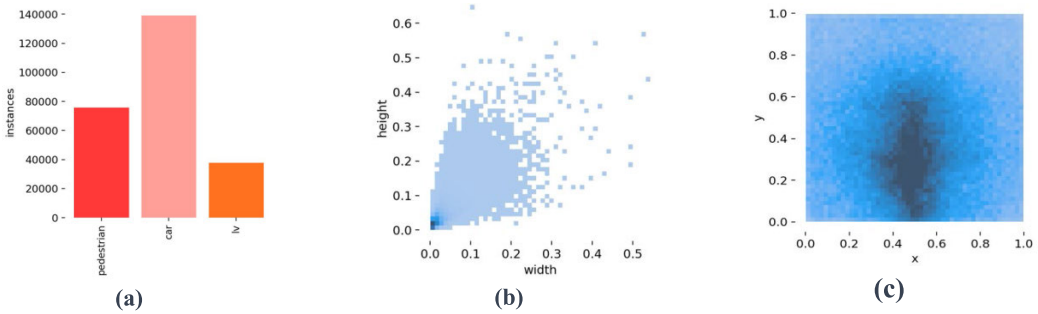

FIGURE 12. Visualization results of the analysis of the dataset. (a)Distribution of object categories in the dataset. (b) Distribution of object centroid locations. (c) Distribution of object sizes.

4615, which is divided according to the ratio of 7:2:1. The exact number of training sets, validation sets and different test sets are shown in Table 2. Visualization results are shown in Fig. 12.

As shown in Fig. 12, the number, size, and location distribution of labeled boxes in the datasets are visualized. The datasets used in this experiment contain a large number of labeled samples and most of the samples are extremely small in scale, therefore the detection task is extremely difficult.

B. FOGGING PARAMETERS AND EFFECTS

Under the condition of single variable, a large number of experiments show that the effect is the best when the size of atomization space is set to 55, and several groups of fogging parameters with better atomization effect are determined through experiments, the best atomization parameters of verification set and test set are shown in Table 3 (the value A in the table is its coefficient relative to 255), optimum atomization parameters for the validation and test sets are shown in Table 4. In Table 3 and Table 4, A denotes the atmospheric light value, β is the parameter controlling the fog concentration, the larger b the greater the fog concentration.

TABLE 3. Best fogging parameters in training set.

Classification	Number	A	β	Center
L	1500	0.72	0.7	(0.5,0.5)
M	1500	0.72	0.1	(0.5,0.5)
H	1615	0.65	0.078	(0.5,0.33)

TABLE 4. Mapping parameters for validation and test set.

Datasets	A	β	Center
Val	0.65 or 0.72	0.07∼0.105	(0.5,0.33) or (0.5,0.5)
Test-L	0.72	0.07	(0.5,0.33)
Test-M	0.72	0.100	(0.5,0.33)
Test-H	0.65	0.078	(0.5,0.33)
Test-Rand	0.65 or 0.72	0.07∼0.105	(0.5,0.33) or (0.5,0.5)

The test set is mapped to light fog space, medium fog space, dense fog space and random fog space to test the performance of the model in different environments. The effect images of different fogging spaces are shown in Fig. 13. It can be seen that the atomization effect of the three spaces is very good, and some targets in medium fog and dense fog space no longer visible to the naked eye, which accords with the required effect of the study. With the increase of fog concentration, the difficulty of detection task also increases.

C. ALGORITHM TESTING AND RESULT ANALYSIS

YOLOv5s is used as the baseline model. The evaluation metrics of the model are P (Precision), R (recall rate), mAP@0.5 (mean Average Precision, IOU is 0.5), FLOPs and FPS, which are widely used in the target detection field. The formulas for P, R, F1-Score and mAP are as follows:

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$R = \frac{TP}{TP + FN} \tag{13}$$

$$F1 - Score = 2 \times \frac{\mathrm{P} \times R}{P + \mathrm{R}} \tag{14}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{1} P(R_i) d(R_i) \tag{15}$$

In Equations (12) and (13), *TP* (True Positive) indicates that the positive sample is correctly classified as a positive sample. *FP* (False Positive) indicates that negative samples are incorrectly classified as positive samples. *FN* (False Negative) indicates that the positive samples are wrongly classified as negative samples. N is the number of categories detected, in this article $N = 3$. FLOPs indicate the complexity of the model, FPS represents number of images detected by model per second.

1) ABLATION EXPERIMENTS

Ablation experiments are performed on the baseline model and several improved models to observe the performance improvement of the models with each optimization scheme. The XM-net is used as the Backbone for some optimization schemes due to the significant architectural differences

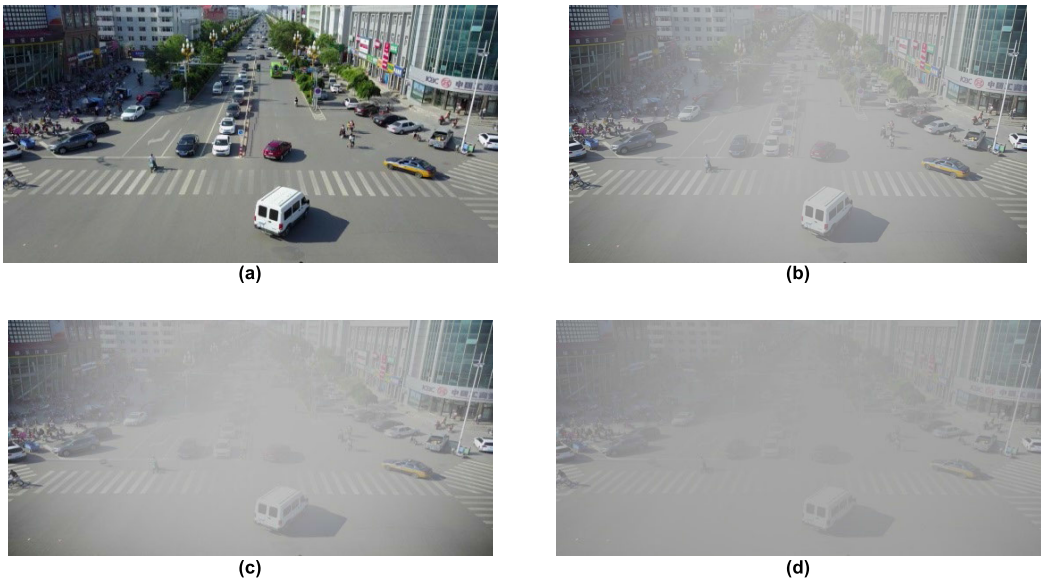

FIGURE 13. (a) original image, (b), (c), and (d) corresponding three mapping spaces L, M, and H.

TABLE 5. Ablation experimental results.

Model	XM-Net	EIoU Loss	Multi-Layer	Params (M)	P (%)	R (%)	F1-score	mAP (%)
Baseline				7.01	51.53	34.85	0.416	35.63
M1	√			3.97	53.35	38.15	0.445	39.25
M2	√	√		3.97	53.33	38.78	0.450	39.53
M3	√		√	4.09	**56.28**	40.45	0.470	42.55
XM-YOLOViT	√	√	√	**4.09**	54.95	**41.93**	0.474	**43.15**

TABLE 6. DEtaied layer information of XM-net.

	Output	Kernel	Channel	ED-TF	D-MLP	N-TFB	Patch
Image	640×640	—	3	—	—	—	—
Stage1-1	320×320	6×6	32	—	—	—	—
Stage1-2	160×160	3×3	64	—	—	—	—
Stage2	80×80	3×3	96	144	288	2	2×2
Stage3	40×40	3×3	128	192	384	4	2×2
Stage4	20×20	3×3	160	240	480	3	2×2

between CSP-Backbone and XM-net. The average values of the test obtained in four different fogging spaces are taken as the final test results and the ablation experimental results are shown in Table 5.

In the M1 scheme, the proposed XM-net is used as the Backbone. The test results show that the performance of the model is improved, the accuracy rate is increased by 1.82%, the recall rate is increased by 3.3%, the F1-score is increased by 6.97%, the mAP improvement of 3.62% with a parameter reduction of 43.3%. Although the accuracy is reduced by 0.02% compared to the M1 scheme. For the M2 scheme, the recall and the mAP increased by 0.63% and 0.28% respectively, which used EIoU as the positioning loss function based on the M1 scheme. M3 scheme adds a

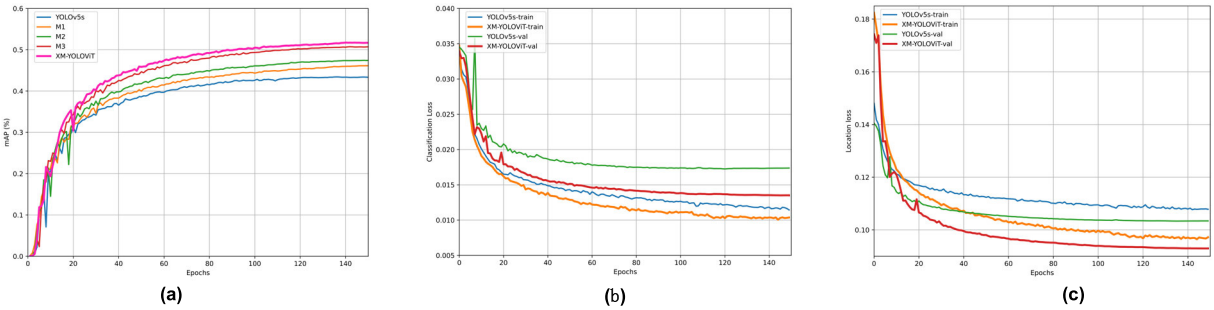

FIGURE 14. (a) Mean average precision for each model. (b) Classification loss of each model (c) Location loss of each model.

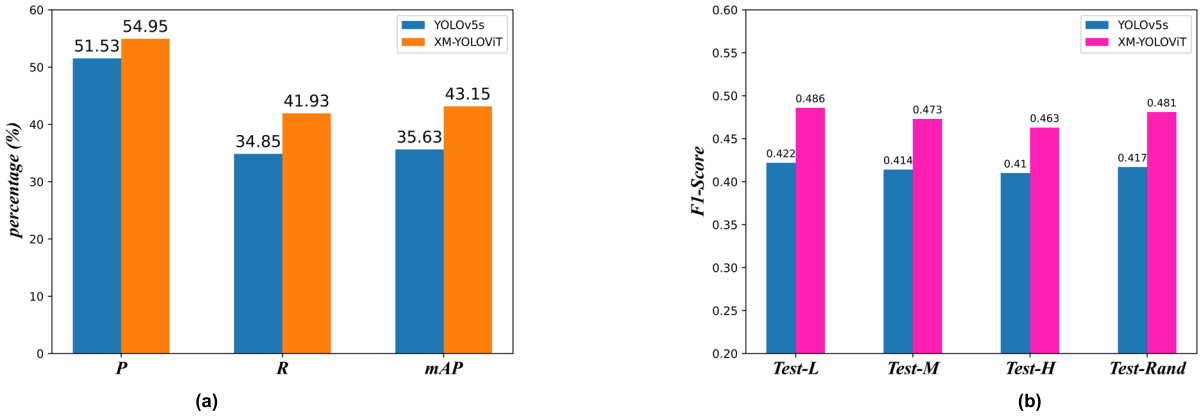

FIGURE 15. (a) Bar chart of P, R, and mAP. (b) Bar chart of F1-Score.

TABLE 7. Model performance testing and comparison.

Datasets	P (%)		R (%)		F1-score		mAP(%)	
	Baseline	Ours	Baseline	Ours	Baseline	Ours	Baseline	Ours
Test-L	52.1	**56.0**	35.6	**42.9**	0.422	**0.486**	36.3	**44.2**
Test-M	51.6	**54.7**	34.6	**41.6**	0.414	**0.473**	35.5	**42.9**
Test-H	51.2	**53.6**	34.0	**40.7**	0.410	**0.463**	34.8	**41.9**
Test-Rand	51.2	**55.5**	35.2	**42.5**	0.417	**0.481**	35.9	**43.6**

higher resolution detection layer on the basis of M1 scheme, the precision, recall, mAP of the M3 scheme are improved by 2.93%, 2.3% and 3.9% compared with the M1 scheme, respectively. The XM-YOLOViT model has an accuracy improvement of 3.42%, a recall improvement of 7.08%, the F1-Score improvement of 14.42% and mAP improvement of 7.52% compared to the baseline model, the performance has been comprehensively improved. The detailed layer information of XM-Net is shown in Table 6, Kernel represents the kernel size of the convolution layer in each stage, ED-TF denotes the embedding dimension of the Transformer Block, D-MLP denotes the input dimension of the MLP structure, and N-TFB denotes the number of Transformer Blocks in the MobileViTV3 Block.

MAP curve and Loss curve are shown in Fig. 14. In Fig. 14(a), the mAP curves for each model tended to stabilize as the number of training sessions increased. In the convergence stage, the mAP curves of each experimental scheme are better than the baseline model, and the XM-YOLOViT scheme being the best among them.

As shown in Fig.14(b) an Fig.14(c), the training and verification loss curves of the baseline model and XM-YOLOViT model show a downward trend with the increase of training times, and tend to be stable in the middle and late stages of training, finally reaching convergence. The loss of XM-YOLOViT has a more drastic decreasing trend, and the final loss value is lower compared with the baseline model. It can be seen from the verification loss curve that there is no overfitting phenomenon in the process of model training. As a result, the XM-YOLOViT has stronger fitting and generalization capabilities than the baseline model.

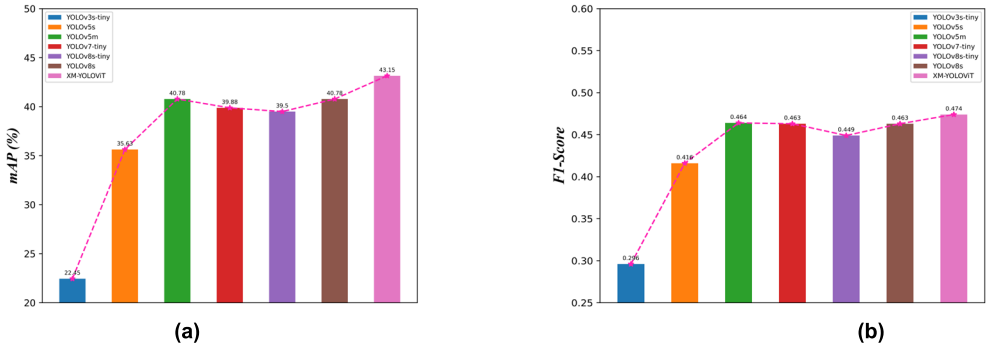

FIGURE 16. (a) Bar chart of mean average precision of each model (b) Bar chart of F1-Score of each model.

The test results for XM-YOLOViT and baseline model under the four fogging spaces are shown in Table 7. From Table 7, it can be seen that the precision, recall, F1-Score and mAP of the XM-YOLOViT model are improved by 3.1%, 7.3%, 15.17% and 7.9% respectively in the light fog space compared to the baseline model. In the medium fog space, the performance of XM-YOLOViT is improved by 2.4% in precision, 7.0% in recall, 14.25% in F1-Score and 7.4% in mAP, although the accuracy of both models has decreased, the performance of XM-YOLOViT is still greatly improved compared with the baseline model. Under the dense fog space, the mAP of the baseline model is already below 35%, whereas the XM-YOLOViT model has an accuracy improvement of 4.3%, a recall improvement of 6.7%, F1-Score improvement of 12.93% and mAP improvement of 7.1%. In the random fog space, the performance of XM-YOLOViT model is far superior to the baseline model. Therefore, the proposed algorithm achieves an average improvement of 3.42% in accuracy, 7.08% in recall, 15.35% in F1-Score and 7.52% in mAP compared to the baseline model in four different difficulty testing tasks. The visualization results of XM-YOLOViT and baseline model in four testing tasks are shown in Fig. 15.

As can be seen from Fig.15, the comprehensive performance of XM-YOLOViT model is ahead of the baseline model in the four different difficulty test tasks.

The model properties of the baseline model and XM-YOLOViT are shown in Table 8. From the test results, it can be seen that the training time for XM-YOLOViT with the new architecture increased by 23%, and FLOPs increased by 59% compared to the baseline model. The detection speed of XM-YOLOViT lags behind the baseline model. However, at present, most of the high-quality videos shot by drones are 60FPS, the XM-YOLOViT model with a detection speed of 70.93FPS has been able to carry out real-time detection on most of the mobile device images. As a whole, XM-YOLOViT gives up excessive detection speed in exchange for a huge improvement in detection accuracy performance, which is what we expect.

TABLE 8. Model level and detection speed.

Model	Size (MB)	Training duration（h）	FLOPs (G)	FPS
Baseline	13.7	7.9	15.8	125
XM-YOLOViT	8.72	10.2	25.2	70.93

2) COMPARATIVE EXPERIMENTS

XM-YOLOViT, YOLOv3 [18], YOLOv7 [19], and YOLOv8 algorithms are selected for comparative experiments. The experimental models are set to the same order of magnitude as the baseline model to reduce the influence of parameter differences between different versions of the detector. The tiny model is used for YOLOv3 and YOLOv7, the width multiple is adjusted to 0.9 and 1.06, respectively. The width multiple is adjusted to 0.383 for YOLOv8s and named YOLOv8s-tiny. During the experiment, the input image size is set to 640, the confidence threshold is set to 0.001, and the IoU threshold is set to 0.6. To evaluate the performance of the model more objectively, the test results are calculated using the same way as the calculation in Table 5, the comparison results are shown in Table 9.

According to the experimental results, the algorithm proposed in this paper has the smallest weight file level and the least model parameters, the recall rate of 41.93%, the F1-Score of 0.474, and the mAP of 43.15%, all of which are higher than the rest of the algorithms in the comparative experiments. Although the precision is slightly lower than that of the YOLOv5m and the YOLOv8s, the detection performance of XM-YOLOViT is better than that of YOLOv5m and YOLOv8s as shown by the F1-Score and mAP. Therefore, XM-YOLOViT is ahead of the rest of the algorithms in the experiment in terms of model magnitude as well as detection effect. The histograms and line chart of mAP and F1-Score for each model in the comparison experiment are shown in Fig. 16.

As can be seen from Fig.16, the mAP and F1-Score of XM-YOLOViT are higher than all the algorithms in the comparison experiment, and they all maintain a leading advantage of more than 2% compared to YOLOv8s. Fig. 17 shows the PR (Precision-Recall) curves of each algorithm on different datasets. Although the XM-YOLOViT are slightly lower than

TABLE 9. Results of comparative experiments.

Model	Size (MB)	Params (M)	FLOPs (G)	P (%)	R (%)	F1-score	mAP (%)	FPS
YOLOv3(tiny)	13.7	7.15	**11.2**	37.15	24.13	0.296	22.45	**250.21**
YOLOv5s	13.7	7.01	15.8	51.53	34.85	0.416	35.63	125.00
YOLOv5m	40.2	20.86	47.9	56.00	39.63	0.464	40.78	115.52
YOLOv7(tiny)	13.3	6.87	16.1	54.28	40.43	0.463	39.88	24.92
YOLOv8s(tiny)	13.6	7.04	19.3	56.65	37.23	0.449	39.50	135.13
YOLOv8s	21.4	11.12	28.4	**57.53**	38.73	0.463	40.78	121.95
XM-YOLOViT	**8.72**	**4.09**	25.2	54.95	**41.93**	0.474	**43.15**	70.93

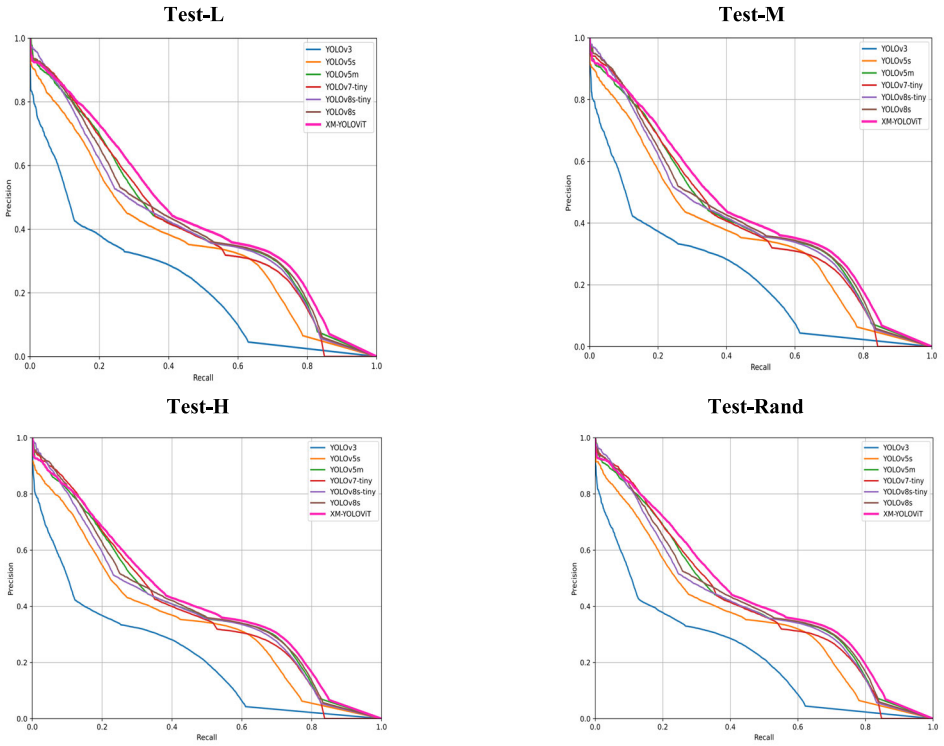

FIGURE 17. PR curves for all algorithms on different datasets.

TABLE 10. Test results of advanced algorithms on different datasets.

Datasets	P (%)			R (%)			mAP(%)		
	YOLOv7-tiny	YOLOv8s	Ours	YOLOv7-tiny	YOLOv8s	Ours	YOLOv7-tiny	YOLOv8s	Ours
Test-L	56.5	**58.6**	56.0	40.6	39.3	**42.9**	40.6	41.4	**44.2**
Test-M	53.4	**56.9**	54.7	40.4	38.7	**41.6**	39.6	40.7	**42.9**
Test-H	52.3	**56.5**	53.6	40.1	37.9	**40.7**	39.0	39.9	**41.9**
Test-Rand	54.9	**58.1**	55.5	40.6	39.0	**42.5**	40.3	41.1	**43.6**

others in the comparison experiments when the recall is very low, the PR curves of XM-YOLOViT are higher than that other models as the recall rate increased, which shows that the XM-YOLOViT has better prediction ability than other algorithms.

The test results of XM-YOLOViT, YOLO7 and YOLO8s on different data sets are shown in Table 10. As a whole, the test results are in line with expectations, and the detection accuracy decreases with the increase of task difficulty. In terms of details, YOLOv8s is more advantageous in terms of detection accuracy. The detection accuracy of the proposed algorithm is slightly lower than YOLOv8s, which is only 0.5% lower than YOLOv7-tiny in light fog, and higher than that of YOLOv7-tiny in other environments. In terms of the

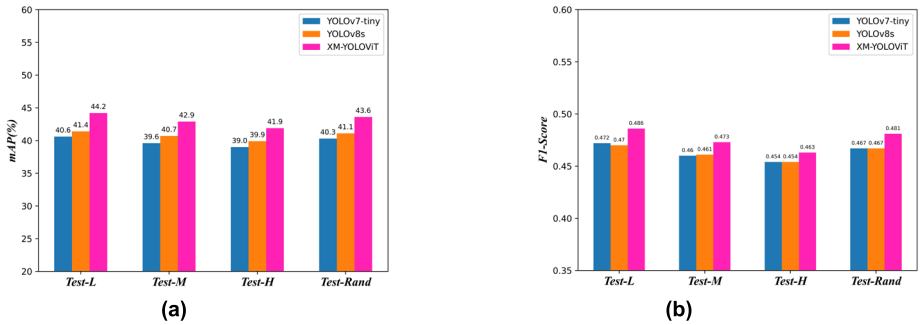

FIGURE 18. Histograms of mAP and F1-Score for advanced algorithms.

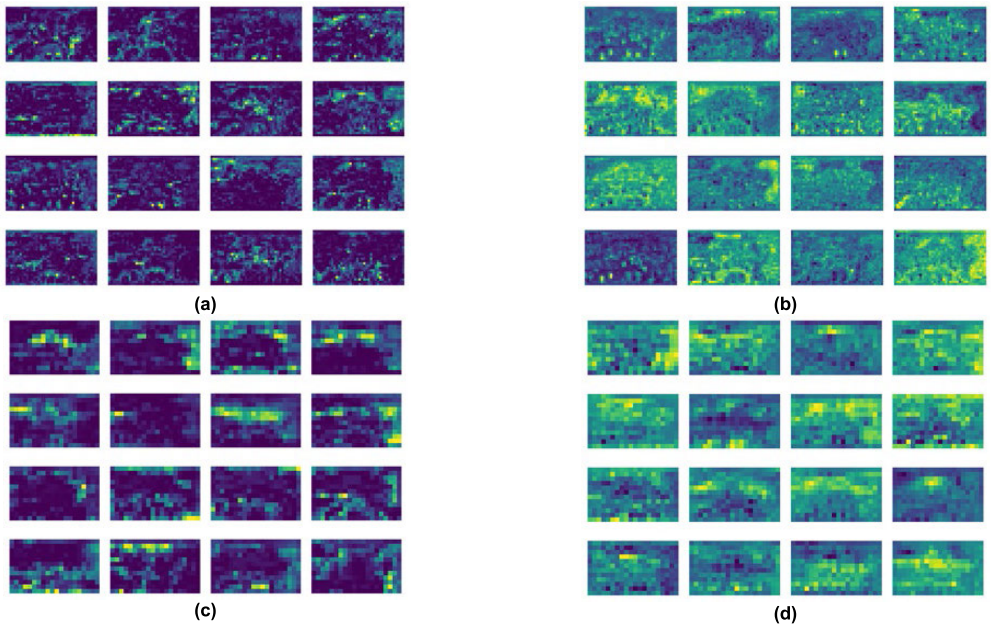

FIGURE 19. (a) (c) Visualized the last two layers of C3 Blocks in CSP Backbone. (b) (d) Visualized the last two layers of MobileViTV3 Block in XM-net.

TABLE 11. Average precision for each algorithm.

Model	Pedestrian (%)	Car (%)	Lv (%)
YOLOv3(tiny)	7.22	51.43	8.68
YOLOv5s	16.65	71.30	18.83
YOLOv5m	22.13	77.01	23.13
YOLOv7(tiny)	18.80	75.25	25.73
YOLOv8s(tiny)	16.95	76.05	25.48
YOLOv8s	18.20	77.30	**26.72**
XM-YOLOViT	**24.7**	**78.88**	26.10

recall and the mAP, XM-YOLOViT takes an absolute lead over YOLOv7-tiny and YOLOv8s, and YOLOv8s has a lower recall than YOLOv7-tiny on four different datasets. Compared to YOLOv7-tiny, recall of XM-YOLOViT increased by 2.3%, 1.2%, 0.6% and 1.9% respectively, the mAP increased by 3.6%, 3.3%, 2.9% and 3.3% respectively on four different datasets. The recall has increased by 3.6%, 2.9%, 2.8% and 3.5%, mAP has increased by 2.8%, 2.2%, 2.0% and 2.5% respectively compared with YOLOv8s.

The visualization of F1-Score and mAP of these three algorithms on different datasets is shown in Fig. 18. It is not difficult to find that both mAP and F1-Score of the algorithms proposed in this paper are higher than those of the current advanced algorithms in the same family in the same series on different datasets.

Table 11 shows the average detecting precision of each algorithm in the comparison experiments for different categories on the four datasets. The data in Table 9 is the average value of the test results on the four datasets. It is obvious from Table 11 that XM-YOLOViT model has better detection capacity in the Pedestrian class than other algorithms. Although, it is extremely tiny and difficult to detect for Pedestrian class targets, with the powerful global modeling and feature extraction capabilities provided by XM-Net and the multi-scale processing capabilities of the subsequent network for XM-YOLOViT model, which is ahead of the current advanced detectors in detecting small and extremely small

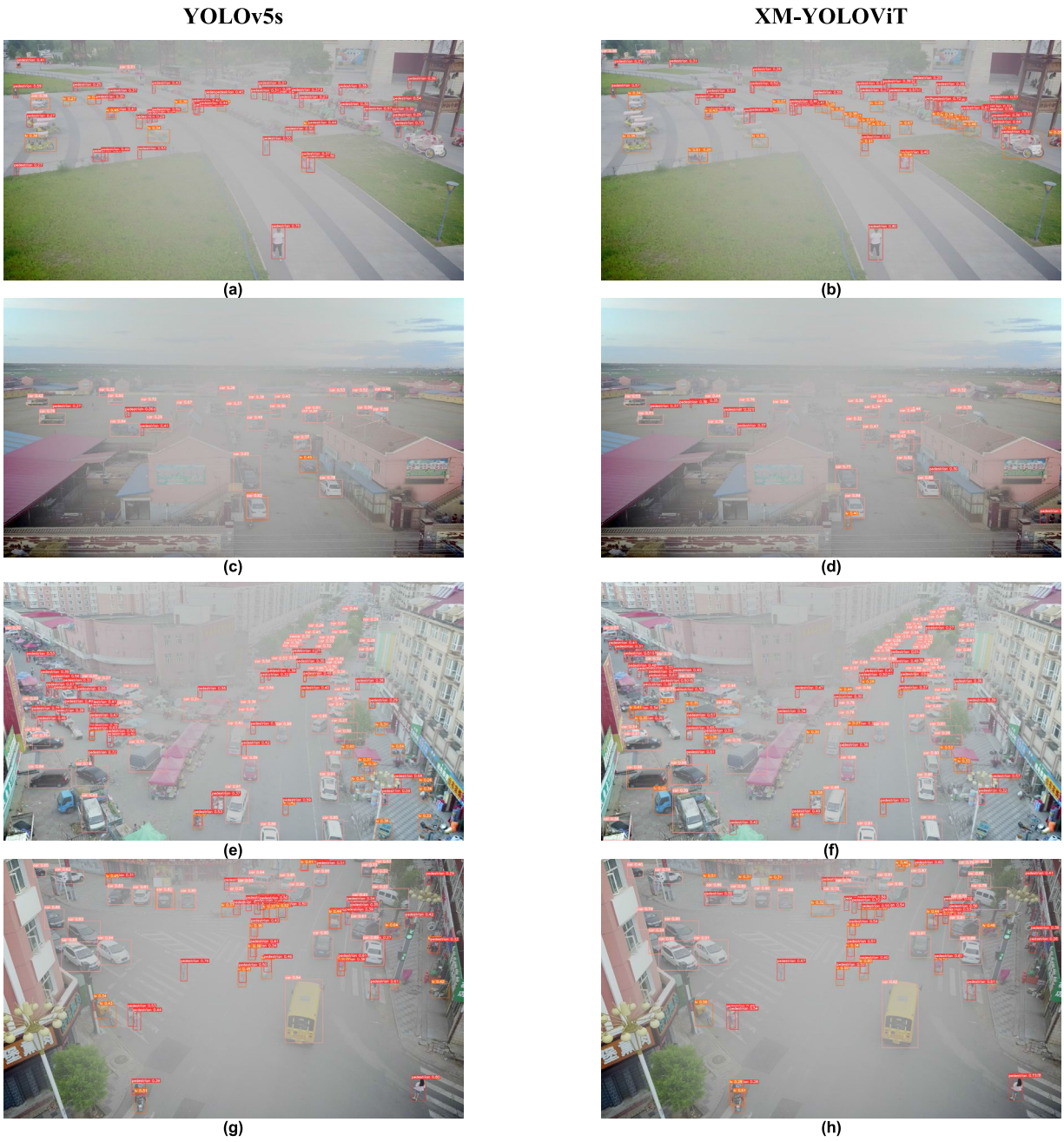

FIGURE 20. Visualization results of XM-YOLOViT and YOLOv5s Test-L.

targets. As a result, it is capable of more tasks and is more practical.

D. VISUALIZATION OF GLOBAL MODELING CAPABILITIES

The visualization experiments on the global modeling capabilities of XM-net and CSP-Backbone are conducted. For the last two layers of the C3 Block in the CSP-Backbone and the MobileViTV3 Block in the XM-net, sixteen channels in each feature layer are randomly chosen for visualization. Visualization results are shown in Fig 19.

As shown in Fig. 19, the higher the brightness of a pixel, the higher the model's attention to that pixel. The test results show that the overall brightness of the visualized image for each channel in Fig. 19(b) and Fig. 19(d) is higher than that of Fig. 19(a) and Fig. 19(c), which indicate that the CSP-Backbone only focuses on local features in most cases, whereas the attention range of context information occupies a large portion of the whole image for XM-net. Therefore, its global modeling capability for XM-net is greatly improved.

E. DETECTION EFFECT AND ANALYSIS

To further validate the performance of the XM-YOLOViT model on target detection for UAV in foggy weather, the visualization experiments of XM-YOLOViT and baseline

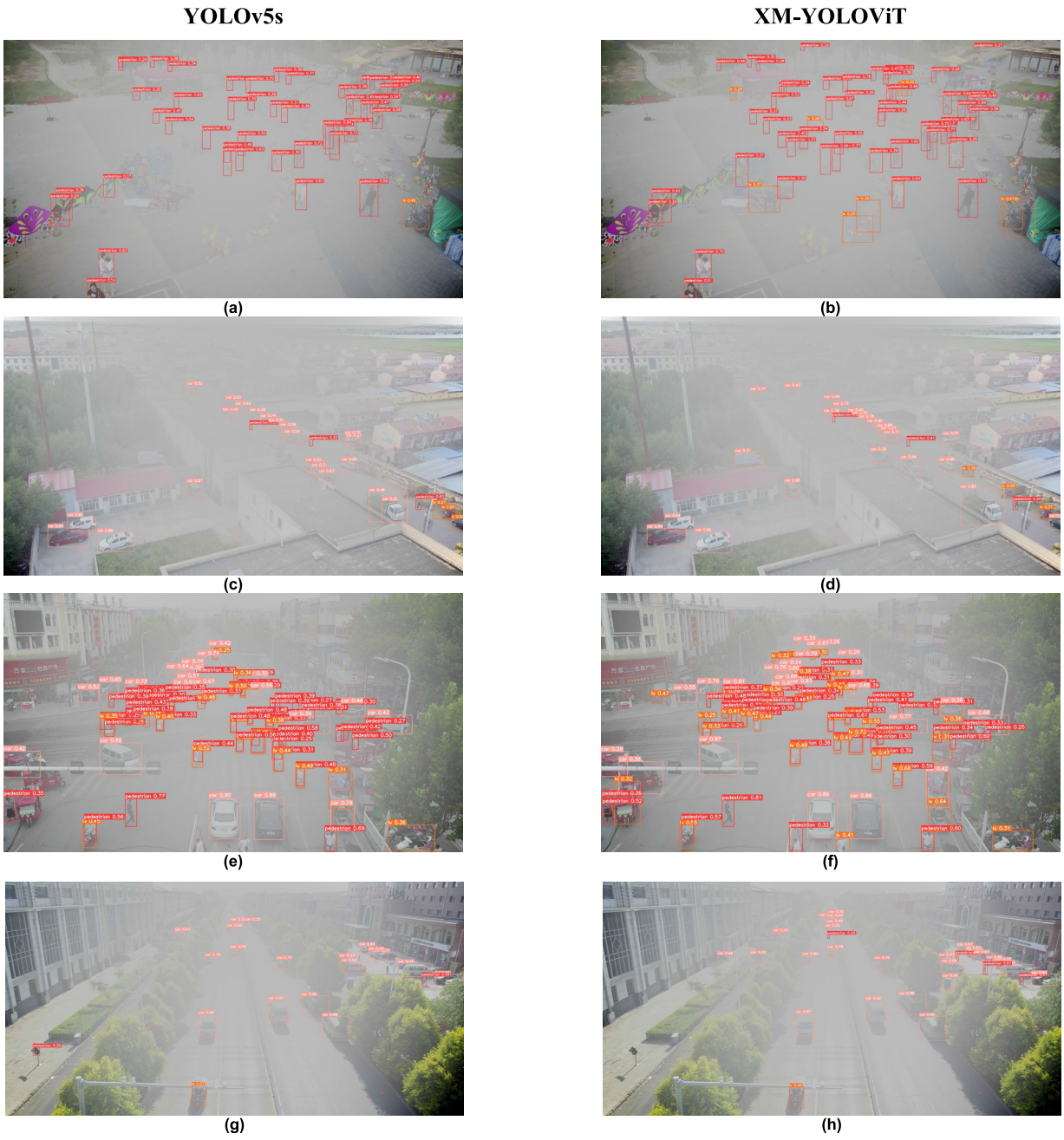

FIGURE 21. Visualization results of XM-YOLOViT and YOLOv5s on Test-M.

model on different datasets are implemented. The performances of the two models on the non-high density fog dataset are shown in Fig. 20 and Fig. 21. It is not difficult to find that XM-YOLOViT has greater detail processing ability than YOLOv5s and that is particularly evident on high density fog datasets. Therefore, the visual test results of the two models on the high concentration fog dataset are shown in Fig. 22 and Fig. 23, and the visual differences between the two models are analyzed in detail.

As shown in Fig. 22, the XM-YOLOViT model has almost the same detection effect as the baseline model when detecting large-scale close-range targets. However, the algorithm proposed in this paper is obviously better in detecting small and long-distance targets. As shown in Fig. 22(K) and Fig. 22(L), the Yolov5s algorithm does not detect any target, whereas XM-YOLOViT can detect the target accurately while XM-YOLOViT still accurately detect the target under the condition of severe occlusion and minimal target. Therefore, XM-YOLOViT algorithm has a strong ability to capture minimal targets. The four regions selected in Fig. 22 are magnified and the detection effect is shown in Fig. 23.

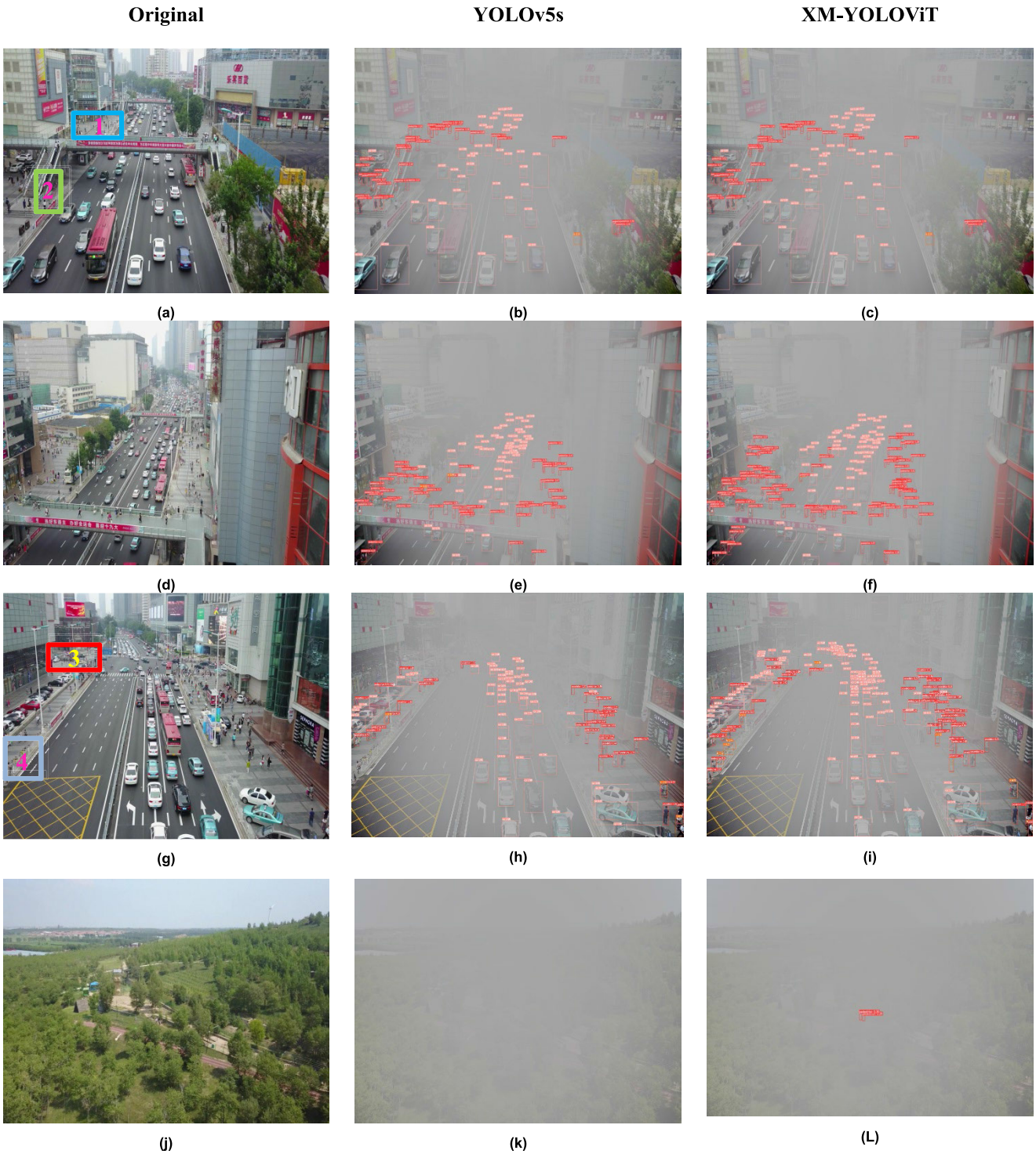

FIGURE 22. Visualization results of XM-YOLOViT and YOLOv5s on Test-H.

As shown in Fig. 23, XM-YOLOViT can detect many objects in the four locations that YOLOv5s cannot. For example, more pedestrians on the footbridge and on the ground are detected for XM-YOLOViT model in Region 1. In Region 2, the bicycles (light vehicles) are detected for XM-YOLOViT model and cannot be detected by YOLOv5s model. In Region 3, the distant targets that are heavily obscured by haze are detected for XM-YOLOViT model. In Region 4, the bicycles (light vehicles) that are obscured by the traffic fence are detected for XM-YOLOViT model. From the above experimental results, it can be seen that the interference caused by complex background information such as density and

FIGURE 23. Comparison of the detection results of the four selected regions.

occlusion can be significantly suppressed for XM-YOLOViT model, and the missed detection of targets can be greatly avoided, therefore, the accuracy of target detection is greatly improved compared to the baseline model. In conclusion, the XM-YOLOViT model has good detection performance in extreme environments and accurately real-time detection can be achieved.

VI. DISCUSSION

Although the XM-YOLOViT model has an absolute advantage in detection performance, the XM-YOLOViT model is hybrid architecture of CNN and Transformer, which is not pure convolution architecture. It is well known that models based on Transformer architectures are often difficult to train. Although the optimized Transformer structure is already quite lightweight, most of the current hardware devices are not optimized for Transformer architectures, which results in the training time and forward inference time for such models being longer than that of the pure convolutional architecture. During the experiment, the training time of XM-YOLOViT increased by about 23% and the detection speed decreased by 43% compared with the baseline model.

At present, the detection speed of XM-YOLO will not affect the real-time performance, but it is still expected to continue to optimize the architecture to shorten the training time and improve the detection speed in our future work.

At the same time, we will also deploy XM-YOLOViT to mobile devices such as UAV, and optimize the algorithm according to the test results.

VII. CONCLUSION

To solve the problem that the target is heavily obscured by smog, the scale of the target is small and the change of the target is violent, an XM-YOLOViT model based on YOLOv5 framework is proposed for pedestrian and vehicle detection. The lightweight of the model is realized and the detection accuracy is improved by the hybrid architecture of CNN and Transformer model is used. In order to obtain a better image dataset, an atomization method is designed to map fog-free images from fog-free space to foggy space based on the atmospheric scattering model and dark channel prior. Experimental results show that the XM-YOLOViT detection algorithm has a significant performance improvement compared to the baseline model, the precision, the recall, the F1-Score, the mAP are improved by 3.42%, 7.08%, 13.94% and 7.52%, respectively, the model parameter is reduced by 41.7%. And the detection effect is better than YOLOv7-tiny and YOLOv8s.The F1-Score and the mAP for the XM-YOLOViT model are improved by 5.57% and 3.65% respectively compared to YOLOv7-tiny, and improved by 2.38% and 2.37% compared to YOLOv8s. The foggy detection algorithm proposed in this paper has high detection accuracy and an extremely lightweight structure, which provides a novel method for detection in complex foggy weather. At the same time, because the model is very lightweight, it can be easily deployed on UAVs, so it can better perform foggy road condition analysis, foggy traffic management, foggy rescue, etc., and it is very important to collect more accurate analysis data for these tasks.

REFERENCES

[1] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[2] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3522–3533, Nov. 2015.

[3] X. Liu and Y. Lin, "YOLO-GW: Quickly and accurately detecting pedestrians in a foggy traffic environment," *Sensors*, vol. 23, no. 12, p. 5539, Jun. 2023.

[4] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M.-H. Yang, "Multi-scale boosted dehazing network with dense feature fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2154–2164.

[5] Y. Qiu, Y. Lu, Y. Wang, and H. Jiang, "IDOD-YOLOV7: Image-dehazing YOLOV7 for object detection in low-light foggy traffic environments," *Sensors*, vol. 23, no. 3, p. 1347, Jan. 2023.

[6] Y. Shan, W. F. Lu, and C. M. Chew, "Pixel and feature level based domain adaptation for object detection in autonomous driving," *Neurocomputing*, vol. 367, pp. 31–38, Nov. 2019.

[7] Y. Guo, R. L. Liang, Y. K. Cui, X. M. Zhao, and Q. Meng, "A domain-adaptive method with cycle perceptual consistency adversarial networks for vehicle target detection in foggy weather," *IET Intell. Transp. Syst.*, vol. 16, no. 7, pp. 971–981, 2022.

[8] M. Hu, Y. Wu, Y. Yang, J. Fan, and B. Jing, "DAGL-faster: Domain adaptive faster r-CNN for vehicle object detection in rainy and foggy weather conditions," *Displays*, vol. 79, Sep. 2023, Art. no. 102484.

[9] V. A. Sindagi, P. Oza, R. Yasarla, and V. M. Patel, "Prior-based domain adaptive object detection for hazy and rainy conditions," in *Computer Vision—ECCV*. Glasgow, U.K.: Springer, 2020, pp. 763–780.

[10] W. Liu, G. Ren, R. Yu, S. Guo, J. Zhu, and L. Zhang, "Image-adaptive YOLO for object detection in adverse weather conditions," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 2, pp. 1792–1800.

[11] X. Meng, Y. Liu, L. Fan, and J. Fan, "YOLOv5s-fog: An improved model based on YOLOv5s for object detection in foggy weather scenarios," *Sensors*, vol. 23, no. 11, p. 5321, Jun. 2023.

[12] W. Fang, G. Zhang, Y. Zheng, and Y. Chen, "Multi-task learning for UAV aerial object detection in foggy weather condition," *Remote Sens.*, vol. 15, no. 18, p. 4617, Sep. 2023.

[13] H. Wang, Y. Xu, Y. He, Y. Cai, L. Chen, Y. Li, M. A. Sotelo, and Z. Li, "YOLOv5-fog: A multiobjective visual detection algorithm for fog driving scenes based on improved YOLOv5," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[14] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.

[15] Z. Li and F. Zhou, "FSSD: Feature fusion single shot multibox detector," 2017, *arXiv:1712.00960*.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Computer Vision—ECCV*. Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.

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

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

[19] 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. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[20] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[21] Y. Zhai, W. Zeng, and N. Li, "A novel detection method using YOLOv5 for vehicle target under complex situation," *Traitement du Signal*, vol. 39, no. 4, pp. 1153–1158, Aug. 2022.

[22] Z. Liu, S. Zhao, and X. Wang, "Research on driving obstacle detection technology in foggy weather based on GCANet and feature fusion training," *Sensors*, vol. 23, no. 5, p. 2822, Mar. 2023.

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[24] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1577–1586.

[25] H. Wang, H. Qian, S. Feng, and S. Yan, "CALYOLOv4: Lightweight YOLOv4 target detection based on coordinated attention," *J. Supercomput.*, vol. 79, no. 16, pp. 18947–18969, Nov. 2023.

[26] S. Liang, R. Chen, G. Duan, and J. Du, "Deep learning-based lightweight radar target detection method," *J. Real-Time Image Process.*, vol. 20, no. 4, p. 61, Aug. 2023.

[27] P. Ding, H. Qian, J. Bao, Y. Zhou, and S. Yan, "L-YOLOv4: Lightweight YOLOv4 based on modified RFB-s and depthwise separable convolution for multi-target detection in complex scenes," *J. Real-Time Image Process.*, vol. 20, no. 4, p. 71, Aug. 2023.

[28] C. Liu, X. Wang, Q. Wu, and J. Jiang, "Lightweight target detection algorithm based on YOLOv4," *J. Real-Time Image Process.*, vol. 19, no. 6, pp. 1123–1137, Dec. 2022.

[29] Z. Cao, L. Fang, Z. Li, and J. Li, "Lightweight target detection for coal and gangue based on improved YOLOv5s," *Processes*, vol. 11, no. 4, p. 1268, Apr. 2023.

[30] R. Yang, J. Zhang, X. Shang, and W. Li, "Lightweight small target detection algorithm with multi-feature fusion," *Electronics*, vol. 12, no. 12, p. 2739, Jun. 2023.

[31] L. Wang, Q. Ni, C. Chen, and H. Yang, "Lightweight target detection algorithm based on improved YOLOv4," *IET Image Process.*, vol. 16, no. 14, pp. 3805–3813, Dec. 2022.

[32] M. Li, S. Chen, C. Sun, S. Fang, J. Han, X. Wang, and H. Yun, "An improved lightweight dense pedestrian detection algorithm," *Appl. Sci.*, vol. 13, no. 15, p. 8757, Jul. 2023.

[33] S. Shen, X. Zhang, W. Yan, S. Xie, B. Yu, and S. Wang, "An improved UAV target detection algorithm based on ASFF-YOLOv5s," *Math. Biosciences Eng.*, vol. 20, no. 6, pp. 10773–10789, 2023.

[34] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2778–2788.
[35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[36] S. N. Wadekar and A. Chaurasia, "MobileViTv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features," 2022, *arXiv:2209.15159*.
[37] S. Mehta and M. Rastegari, "Separable self-attention for mobile vision transformers," 2022, *arXiv:2206.02680*.
[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[39] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," 2021, *arXiv:2101.08158*.
[40] E. J. McCartney, *Optics of the Atmosphere: Scattering by Molecules and Particles*. New York, NY, USA: Wiley, 1976.
[41] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang, and L. Bo, "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 213–226.

HUIYING ZHANG received the Ph.D. degree from the Changchun University of Technology, in 2017. Currently, she is with the School of Information and Control Engineering, Jilin Institute of Chemical Engineering, China. Her main research interests include spatial optical communication modulation reception technology, visible light communication, and localization technology.

YIFEI GONG is currently pursuing the bachelor's degree in engineering with the School of Information and Control Engineering, Jilin Institute of Chemical Engineering. His main research interests include artificial intelligence and computer vision and image processing.

FEIFAN YAO received the bachelor's degree from Weifang Medical University, in 2022. He is currently pursuing the master's degree in engineering with the School of Information and Control Engineering, Jilin Institute of Chemical Engineering. His research interests include image processing and generative artificial intelligence.

QINGHUA ZHANG received the bachelor's degree from the Harbin Institute of Petroleum, in 2022. He is currently pursuing the master's degree in engineering with the School of Information and Control Engineering, Jilin Institute of Chemical Engineering. His research interests include image processing, communications, and signal processing.