Department of Informatics
Technical University of Munich

Semester Thesis

Object detection in Haze

Supervisor Prof. Dr.-Ing. habil. Alois C. Knoll

Advisor Liguo Zhou, M.Sc.

Author Jingyu Zhang

Date May 16, 2022 in Garching

Disclaimer

I confirm that this Semester Thesis is my own work and I have documented all sources and material used.

Garching, May 16, 2022

(Jingyu Zhang)

Abstract

The existing image acquisition equipment is very sensitive to the interference of the external environment. In the haze environment, the acquired outdoor images are often seriously degraded, mainly manifested as blurred scene feature information, low contrast, and color distortion, which is not conducive to the computer vision system for the real image. Feature extraction, which affects its subsequent analysis, understanding, recognition and other series of processing, greatly reduces the practical application performance of the vision system and limits the application value of the image.

This paper makes some minor improvements based on yolov5x to achieve object detection in foggy environments. This paper tries three aspects: 1. Image enhancement 2. Adding a small target detection layer 3. Add an attention mechanism. compared with the original yolov5x in accuracy. It achieves better accuracy and robustness, and realizes the recognition and positioning of pedestrians, cars, buses, and other objects in a foggy environment.

Zusammenfassung

Die vorhandene Bilderfassungsausrüstung ist sehr empfindlich gegenüber Störungen durch die äußere Umgebung. In der Dunstumgebung sind die erfassten Außenbilder oft ernsthaft verschlechtert, was sich hauptsächlich in verschwommenen Szenenmerkmalsinformationen, geringem Kontrast und Farbverzerrung manifestiert, was für das Computer-Vision-System für das reale Bild nicht förderlich ist. Die Merkmalsextraktion, die ihre anschließende Analyse, ihr Verständnis, ihre Erkennung und andere Verarbeitungsreihen beeinflusst, reduziert die praktische Anwendungsleistung des Bildverarbeitungssystems erheblich und begrenzt den Anwendungswert des Bilds.

Dieses Dokument macht einige kleinere Verbesserungen basierend auf yolov5x, um eine Objekterkennung in nebligen Umgebungen zu erreichen. Dieses Papier versucht drei Aspekte: 1. Bildverbesserung 2. Hinzufügen einer kleinen Zielerfassungsschicht 3. Hinzufügen eines Aufmerksamkeitsmechanismus. verglichen mit dem ursprünglichen yolov5x in der Genauigkeit. Es erreicht eine bessere Genauigkeit und Robustheit und realisiert die Erkennung und Positionierung von Fußgängern, Autos, Bussen und anderen Objekten in einer nebligen Umgebung.

Contents

Chapter 1

Introduction

Haze weather not only affects people's health, but also brings inconvenience to traffic. Due to the large amount of pollutant particles contained in the smog weather, the visibility is reduced, the electronic eyes at the intersection are affected, and various violations cannot be detected in time, resulting in various traffic accidents. Moving target detection in haze weather is the basis for subsequent target recognition and behavior analysis.

Based on physical scattering models [13, 15, 17], the dehazing process is usually expressed as

$$I(x) = J(x) \cdot t(x) + A \cdot (1 - t(x)) \tag{1.1}$$

where I(x) and J(x) are the hazy image and the clean image respectively. A is the global atmospheric light, and t(x) is the medium transmission map, which is affected by the unknown depth information. For the most previous dehazing methods they are firstly preferred to estimate the transmission map t(x) or the atmospheric light A, then going to recover the final target image J(x). However, the first step is usually not recommended, because both the transmission map t(x) and the atmospheric light A are often unknown in the real scenarios.

Figure 1.1: The top images detected by FFA-Net,the middle images detected by AddLayer, the bottom images detected by CBAM.

In this paper, we propose 3 new ideas based on YOLOv5, the detection result is in the figure 1.1. Compared with YOLOv4 [2], YOLOv5 [9] has the following advantages: faster. On the YOLOv5 Colab notebook, run Tesla, P100, we see inference time of just 0.007 seconds per image, which translates to 140 frames per second (FPS), which is more than 2x faster than YOLOv4; Higher precision. In Roboflow's test on the blood cell count and detection (BCCD) dataset, a mean precision (mAP) of about 0.895 was achieved with only 100 epochs of training. It is true that EfficientDet and YOLOv4 have comparable performance, but it is very rare to see such a comprehensive performance gain without any loss in accuracy; Smaller size. The weights file for YOLOv5 is 27 megabytes. The weight file of YOLOv4 (using Darknet architecture) is 244 MB. YOLOv5 is nearly 90% smaller than YOLOv4! This means that YOLOv5 can be more easily deployed to embedded devices. We summarize the contributions of our work as follows:

1.Image enhancement, that is, image preprocessing, deblurring the image, and then process this image with yolov5x for training

2. Adding a small target detection layer on yolov5 by down sampling to a larger multiple, increasing Feature layer extraction

3. Add an attention mechanism to yolov5, modify the backbone based on the original feature layer, increase the weight of the target feature, and reduce or ignore the influence of the background.

1.1 Related work

Dehazing methods can be mainly divided into two types: one is based on the atmospheric degradation model, which uses neural networks to estimate the parameters in the model. Most of the early methods are based on this idea; the other is to use the input fog image, directly output to get the image after dehazing.

Dehazing Algorithm Based on Image Restoration To make up for the information lost during corruption, many traditional methods are used. [1, 7] Given a hazy image I(x), most dehazing algorithms try to estimate t(x) and A. However, estimating the transmission map from a hazy image is an ill-posed problem generally. Early prior-based methods try to estimate the transmission map by exploiting the statistical properties of clear images, such as dark channel prior [8] and color-line prior [3]. The principle of dark channel prior is: In most non-sky local areas, some pixels will always have at least one-color channel with very low values. In other words, the minimum value of the light intensity in this area is a small number. There are three main factors that cause low and medium channel values in dark primary colors in real life: a) shadows from glass windows in cars, buildings, and cities, or the shadows of natural landscapes such as leaves, trees, and rocks; b) brightly colored objects or surfaces , some of the three channels of RGB have very low values (such as green grass/trees/plants, red or yellow flowers/leaves, or blue water); c) objects or surfaces with darker colors, such as Dark grey tree trunks and stones. In short, natural scenes are full of shadows or colors, and the dark primary colors of the images of these scenes are always very gray. Unfortunately, these image priors are easily inconsistent with the practice, which may lead to inaccurate transmission approximations. Thus, the quality of the restored image is undesirable.

Dehazing algorithm based on deep learning To address this problem, Convolutional Neural Networks (CNNs) have been used to estimate transmissions [4, 19, 22] or directly predict sharp images [10, 12, 20]. These are efficient and outperform prior-based algorithms with significant performance improvements. Such methods can be mainly divided into two categories. The first category is still based on the atmospheric degradation model, which uses neural networks to estimate the parameters in the model. Most of the early methods are based on this idea. The second type is to use the input foggy image to directly output the dehazed image, which is often called end2end in deep learning.

This paper [16] proposes a dehazing method for ultra-high-definition images through multi-guided bilateral learning. The crux of the approach is to use a deep CNN to build an affine bilateral grid, an efficient feature storage container that preserves the detailed edges and textures in the image. At the same time, multiple guidance matrices are established to assist the affine bilateral mesh to recover high-quality features, providing rich color and texture information for image dehazing.

Quantitative and qualitative results show that the proposed network outperforms state-of-the-art dehazing methods in terms of accuracy and inference speed (125 fps) and produces visually surprising results on real-world 4K haze images the result of.

DANet [6] is a classic network applying self-Attention, which introduces a self-attention mechanism to capture feature dependencies in spatial and channel dimensions, respectively. Dual Attention Network (DANet) to adaptively integrate local features and global dependencies. Two types of attention modules are attached on top of conventional dilated FCNs [14] to model semantic interdependencies in spatial and channel dimensions, respectively.

GCANet [5] is to directly learn the residual between the original image and the fog image. smooth dilated convolution, used in place of the original dilated convolution, eliminates gridding artifacts. He proposed two ideas:1. smooth dilated convolution, used in place of the original dilated convolution, eliminates gridding artifacts. 2. gated fusion sub-network, which is used to fuse features of different levels, which is good for both low-level tasks and high-level tasks. Because fusing different-level functions is often beneficial for both low-level and high-level tasks. Therefore, the paper further proposes a gated subnet to determine the importance of different levels and fuse them according to their corresponding importance weights.

However, deep learning-based methods need many real hazy images and their haze-free images used for training. On the other hand, acquiring many real images in the real world is not available. Therefore, most dehazing models resort to training on synthetic hazy datasets. But models learned from synthetic data often do not generalize well to real data due to domain transfer issues.

Chapter 2

Methods

2.1 Fusion Feature Attention Network: FFA-Net

As shown in Fig 2.1, the input of the FFA network [18] is a blurred image, which is passed to a shallow feature extraction part, and then fed into N group structures with multi-hop connections, which are combined by our proposed feature attention module. The output features of N group structures are fused together. Finally, these features are passed to the reconstruction part and the global residual learning structure, resulting in a fog-free output.

Furthermore, each group structure combines B basic block structures with local residual learning, each of which incorporates skip connections and feature attention (FA) modules. FA is an attention mechanism structure consisting of channel attention and pixel attention.

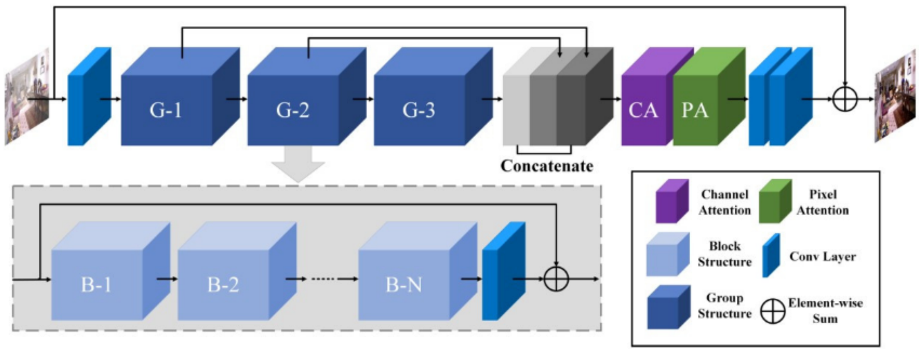

Figure 2.1: The fusion feature attention network (FFA-Net) architecture

Feature Attention (FA) Most image dehazing networks treat channel and pixel features equally and cannot handle images with uneven haze distribution and weighted channels. The feature attention proposed in this paper (Fig. 2.2) consists of channel attention and pixel attention, which can provide additional flexibility when dealing with different types of information.

FA treats different features and pixel regions unequally, which can provide additional flexibility when dealing with different types of information and can expand the representational capabilities of CNNs. A key step is how to generate different weights for each channel and pixel feature.

Channel Attention (CA) Channel attention mainly focuses on different channel features with completely different weighting information for DCP. The key to implementing CA: 1x1 convolution.

We already know that convolution can extract spatial tiles around each square of the input tensor and apply the same transformation to all tiles. The extreme case is that the extracted

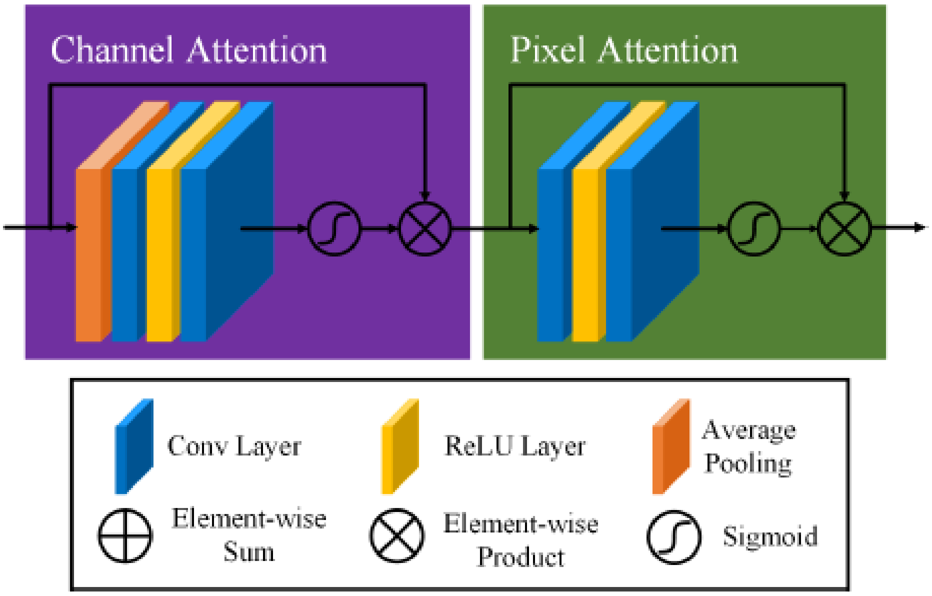

Figure 2.2: Feature attention module

tile contains only one square. This is the convolution operation equivalent to passing each square vector through a Dense (fully connected) layer: it computes features that mix the information in the input tensor channels, but not across the space together (as it only looks at one square at a time). This 1x1 convolution [also called pointwise convolution] is featured in the Inception module, and it helps to distinguish channel feature learning from spatial feature learning. Pixel Attention (PA) Considers the uneven distribution of haze on different image pixels, the network is made to pay more attention to informative features, such as dense haze pixels and high-frequency image regions

Similar to CA, we use ReLu and sigmoid activation functions to feed the input, directly into two convolutional layers. The shape changes from C × H × W to 1 × H × W.

sigmoid function:

$$S(x) = \frac{1}{1+e^{-x}} \tag{2.1}$$

The sigmoid function compresses any value into the [0, 1] interval, and its output value can be regarded as a probability value. Often used in binary classification problems and regression problems. Disadvantages: The two ends of the function curve are relatively flat, soft saturation occurs, and the gradient disappears easily. ReLu function:

$$f(x) = max(0, x) \tag{2.2}$$

The ReLU function is actually a piecewise linear function, which turns all negative values into 0, while the positive values remain unchanged. This operation is called one-sided suppression. That is: in case the input is negative, it will output 0, then the neuron will not be activated. This means that only part of the neurons will be activated at the same time, making the network very sparse, which in turn is very computationally efficient.

2.2 Add a small target detection layer

One of the reasons why the YOLOv5 small target detection effect is not good is because the size of the small target sample is small, and the down sampling multiple of yolov5 is relatively large, and it is difficult for the deeper feature map to learn the feature information of the small target. The layer performs detection after splicing the shallow feature map with the deep feature map.

The implementation of this method is very simple and effective. It only needs to modify the model file yaml of yolov5 to add a small target detection layer. However, after adding the detection layer, the problem is that the amount of calculation increases, resulting in a decrease in the speed of inference detection. However, for small targets, it has a good improvement.

Difficulties of small target detection:

1. The aspect ratio of the detection frame is variable, and the missed detection rate is high

2. The target scale distribution is uneven, small target samples are scarce, and the missed detection rate is high

3. Small targets have low resolution, few features, and high false detection rate

4. The target scale span is large, multiple scales coexist, and the missed detection rate is high

The main reason for the poor detection effect of small objects is the problem of small object size. Taking the input 608×608 of the network as an example, the down sampling in yolov5 is used 5 times, so the final feature map size is 19×19, 38×38, 76×76.

Among the three feature maps, the largest 76×76 is responsible for detecting small targets, and corresponding to 608×608, the receptive field of each feature map is 608/76=8*8 size.

That is, if the width or height of the target in the original image is less than 8 pixels, it is difficult for the network to learn the feature information of the target. In addition, many images have a large resolution. If the down sampling is simply performed, the multiple of down sampling is too large, and it is easy to lose data information. However, if the multiple is too small, network forward propagation needs to save many feature maps in memory, which greatly exhausts GPU resources, and is prone to memory explosion, making it impossible to train and reason normally. In this case, the segmentation method can be used. The large image is first divided into small images, and then each small image is detected. Many cars in the middle area are detected:

However, this method has advantages: Accuracy The small image after segmentation is input into the target detection network, and the lower limit of the minimum target pixel will be greatly reduced. For example, it is divided into 608×608 size and sent to the network with the input image size of 608×608. According to the above calculation method, on the original image, the small target with length and width greater than 8 pixels can learn features.

2.3 CBAM:Convolutional Block Attention Module

The attention mechanism mimics the way people look at objects: by taking a simple glance, analyzing an important part of the image, and then focusing on that location, rather than giving equal attention to all areas of the frame.

CBAM [21] learns the Attention map of What (note what) and Where (note where) in the channel and spatial through the feature map extracted from the image before. The basic principle is as follows:

The core is the green box Channel Attention Module (channel attention mechanism) and

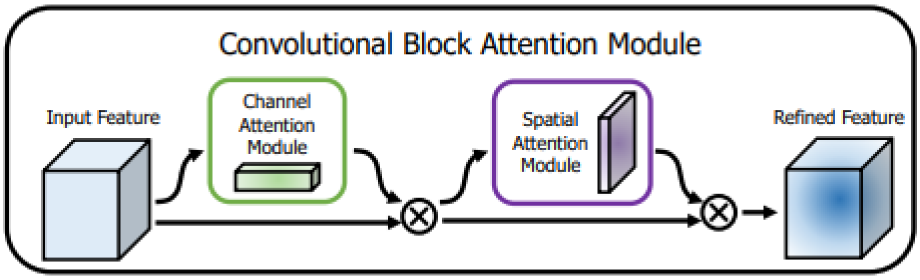

Figure 2.3: CBAM module

the purple box Spatial Attention Module (spatial attention mechanism) in the figure 2.3.

$$F' = M_c(F) \bigotimes F$$
$$F" = M_s(F') \bigotimes F' \tag{2.3}$$

$\bigotimes$ means element-wise multiplication.The original CNN feature F, the shape is W×H×C, the attention weight $M_c(F)$ is obtained after the channel attention mechanism Mc, and the shape is 1×1×C, $M_c(F)$ and F are multiplied to obtain the feature F ';F' goes through a spatial attention mechanism Ms to the attention weight $M_s(F')$, the Shape is W×H×1, $M_s(F')$ and F' are multiplied to get the final features F", F" The Shape is W×H×C. After CBAM, the shape of the feature has not changed, so CBAM can be inserted after the original CNN feature, and the network does not need to be changed.

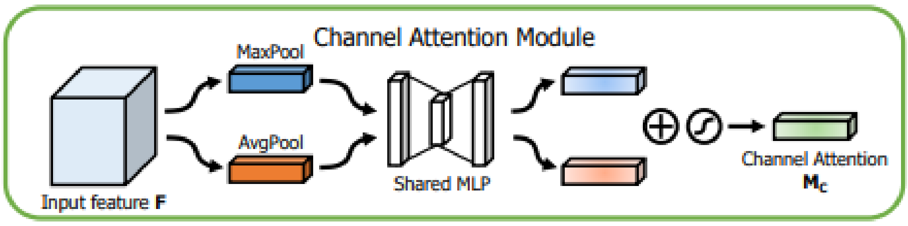

Figure 2.4: Channel Attention Module

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))$$
$$M_c(F) = \sigma(W_1(W_0(F^c_{Avg})) + (W_1(W_0(F^c_{Max})) \tag{2.4}$$

where σ denotes the sigmoid function. Note that the MLP weights, W_0 and W_1, are shared for both inputs and the ReLU activation function is followed by W_0.

Channel attention module. In convolution, there are usually feature outputs of multiple channels, and the features of some channels will have a greater impact on the final target shown in the figure 2.4 so it is necessary to better focus on these channels. The common practice is global pooling, in CBAM, two pooling of global maximum and global average are used respectively, and the nonlinear feature changes are performed through the fully connected network, the fully connected network is composed of multi-layer perceptron (MLP) with one hidden layer.and the changed network is summed and reactivated to obtain the attention weight.

$$M_s(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)]))$$
$$M_s(F) = \sigma(f^{7\times7}(F^s_{Avg}; F^s_{Max})) \tag{2.5}$$

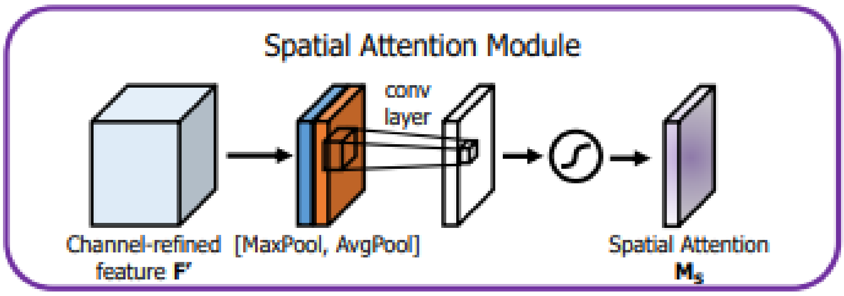

Figure 2.5: Spatial Attention Module

Spatial attention module. Generally, in order to make the model focus more on the more important features of the spatial shape, the model needs to strengthen and weaken the features of certain positions in the space. The structure is shown in the figure 2.5 Try MaxPool and AvePool on the Channel respectively and generate two 2D images: F^s_{Avg} size is W× H×1, F^s_{Max} size is W×H×1. Concat the two outputs on the channel, use the $f^{7\times7}$ kernel for convolution and then activate it with the sigmoid function, which can get the spatial attention weight.

Chapter 3

Experiments

3.1 Datasets.

We randomly choose realword hazy images from the RESIDE dataset [11] for training. The dataset is divided into five subsets, namely, ITS (Indoor Training Set), OTS (Outdoor Training Set), SOTS (Synthetic Object Testing Set), URHI (Unannotated real Hazy Images), and RTTS (real Task-driven Testing Set). We use only the RTTS dataset,for realword dataset, we choose 3457 hazy images for training, 865 hazy images for validation. In the training phase, we randomly crop all the images to 640×640 and normalize the pixel values to [-1, 1].

3.2 Training details.

We evaluate the proposed method against the following state-of-the-art approaches: YOLOV5 [9], dark channel prior [8], GCA-Net, MSBDN-DFF, FFA-Net, we use those methods firstly to preprocess with images,then are trained by YOLOV5, the result is shown in the figure 3.1 2 another methods are AddLayer, CBAM, different with above methods, without preprocessing hazy images can directly be trained, only feature extraction step changes, therefore the last 2 methods have faster speed.

Table 3.1: Detection Result on hazy images.

Module	Detector	mAP@.5by train	mAP@.95by train	mAP@.5by test
-	YOLOV5x	0.7687	0.5147	0.768
Dark channel prior	YOLOV5x	0.7843	0.5253	0.777
GCA-Net	YOLOV5x	0.7802	0.5156	0.766
MSBDN-DFF	YOLOV5x	0.7245	0.4894	0.703
FFA-Net	YOLOV5x	0.7875	0.5332	0.785
AddLayer	YOLOV5x	0.7903	0.5205	0.782
CBAM	YOLOV5x	0.790	0.5335	0.788

As a conclusion, We can see from the above table 3.1 that the FFA-Net, AddLayer and CBAM models have a certain improvement in accuracy compared with the original yolov5x, and the above algorithms can adapt to better prediction and calibration targets in foggy environments Objects, while AddLayer and CBAM omit the image dehazing process, and weight the extracted feature layers with different weights, which is conducive to selectively ignoring the fog and irrelevant objects in the background, but the structure of the neural

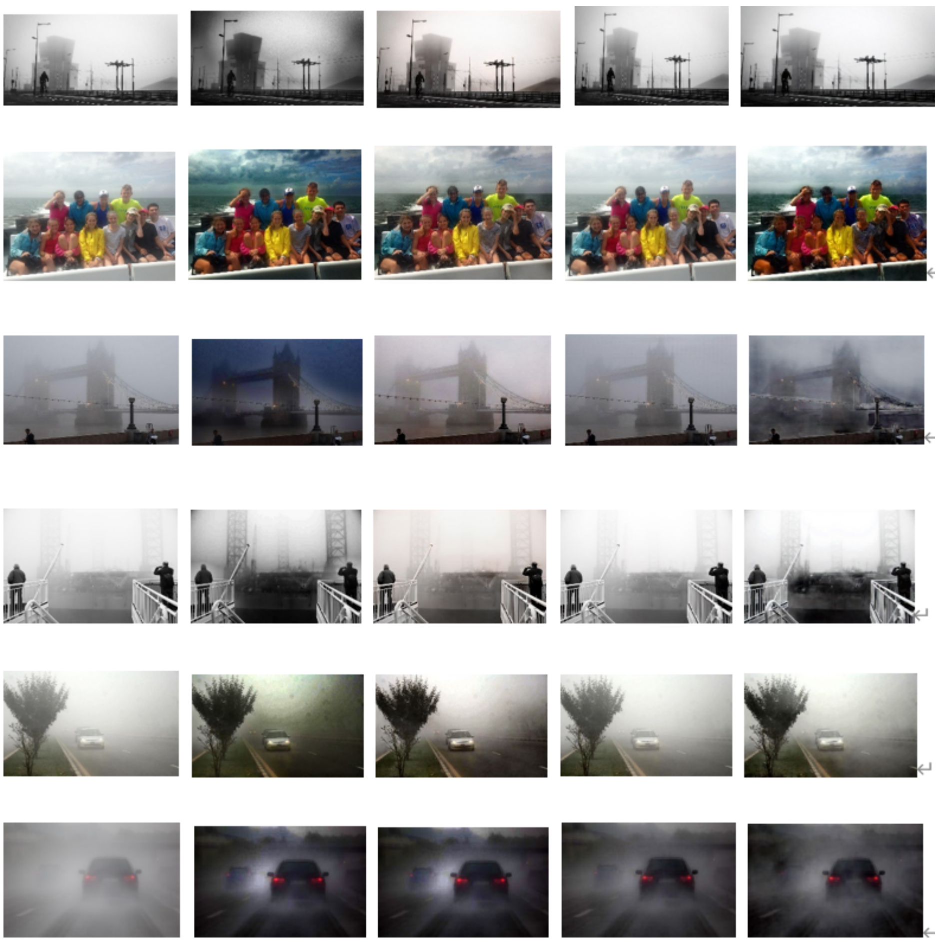

Figure 3.1: (1) origin (2) dark channel prior (3) GCA-Net (4) MSBDN (5) FFA-Net

network has not changed. What's more, the parameters have not increased too much, so that the training time of the model is no different from the original yolov5x, but the detection effect has been improved to a certain extent.

3.3 Conclusion.

In this paper,we add FFA-Net, AddLayer and CBAM models to yolov5x, which improves the accuracy of detecting objects in fog, making it adaptable to some extreme weather. For common objects, such as bicycle, bus, car, motorbike, person, it can be very good is detected. In the future, these methods can be applied to the autonomous driving industry, which will play a positive role in avoiding obstacles and pedestrians and avoiding vehicle collisions in foggy environments.

Bibliography

[1] Berman, D., Avidan, S., et al. “Non-local image dehazing”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1674–1682.

[2] Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. “Yolov4: Optimal speed and accuracy of object detection”. In: *arXiv preprint arXiv:2004.10934* (2020).

[3] Bui, T. M. and Kim, W. “Single image dehazing using color ellipsoid prior”. In: *IEEE Transactions on Image Processing* 27.2 (2017), pp. 999–1009.

[4] Cai, B., Xu, X., Jia, K., Qing, C., and Tao, D. “Dehazenet: An end-to-end system for single image haze removal”. In: *IEEE Transactions on Image Processing* 25.11 (2016), pp. 5187–5198.

[5] Chen, D., He, M., Fan, Q., Liao, J., Zhang, L., Hou, D., Yuan, L., and Hua, G. “Gated context aggregation network for image dehazing and deraining”. In: *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2019, pp. 1375–1383.

[6] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., and Lu, H. “Dual attention network for scene segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3146–3154.

[7] Hautière, N., Tarel, J.-P., and Aubert, D. “Towards fog-free in-vehicle vision systems through contrast restoration”. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2007, pp. 1–8.

[8] He, K., Sun, J., and Tang, X. “Single image haze removal using dark channel prior”. In: *IEEE transactions on pattern analysis and machine intelligence* 33.12 (2010), pp. 2341–2353.

[9] Jocher, G. *ultralytics/yolov5*. https://github.com/ultralytics/yolov5. Oct. 2020.

[10] Li, B., Peng, X., Wang, Z., Xu, J., and Feng, D. “Aod-net: All-in-one dehazing network”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 4770–4778.

[11] Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., and Wang, Z. “Benchmarking single-image dehazing and beyond”. In: *IEEE Transactions on Image Processing* 28.1 (2018), pp. 492–505.

[12] Li, R., Pan, J., Li, Z., and Tang, J. “Single image dehazing via conditional generative adversarial network”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8202–8211.

[13] Li, Y., You, S., Brown, M. S., and Tan, R. T. “Haze visibility enhancement: A survey and quantitative benchmarking”. In: *Computer Vision and Image Understanding* 165 (2017), pp. 1–16.

[14] Long, J., Shelhamer, E., and Darrell, T. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[15] McCartney, E. J. “Optics of the atmosphere: scattering by molecules and particles”. In: *New York* (1976).

[16] Meng, G., Wang, Y., Duan, J., Xiang, S., and Pan, C. “Efficient image dehazing with boundary constraint and contextual regularization”. In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 617–624.

[17] Narasimhan, S. G. and Nayar, S. K. “Vision and the atmosphere”. In: *International journal of computer vision* 48.3 (2002), pp. 233–254.

[18] Qin, X., Wang, Z., Bai, Y., Xie, X., and Jia, H. “FFA-Net: Feature fusion attention network for single image dehazing”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 11908–11915.

[19] Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., and Yang, M.-H. “Single image dehazing via multi-scale convolutional neural networks”. In: *European conference on computer vision*. Springer. 2016, pp. 154–169.

[20] Ren, W., Ma, L., Zhang, J., Pan, J., Cao, X., Liu, W., and Yang, M.-H. “Gated fusion network for single image dehazing”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3253–3261.

[21] Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. “Cbam: Convolutional block attention module”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[22] Zhang, H. and Patel, V. M. “Densely connected pyramid dehazing network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3194–3203.