Attention Enhancement of YOLO for Vehicle Detection

Caixiao Ouyang [1,*,†], Hu Jiwei[2,†] , Youyuan She[1,†] and Chunzhi Wang[3,†]

[1] *Wuhan Vocational College of Sortware and Engineering, Wuhan 430205, China*
[2] *Wuhan Fiberhome Technical Services Co., Ltd., Wuhan 430205, China;*
[3] *Hubei University of Technology, Wuhan 430068, China*

Abstract

Vehicle detection and recognition is an important research. An attention and feature fusion target detection algorithm based on the improved YOLOv4 algorithm is proposed to achieve a more effective screening of vehicle targets in traffic scenes. Considering the cost deployment problem of traffic recognition algorithms, this paper uses YOLOv4 as the base architecture, firstly, the lightweight DenseNet is used as the backbone feature extraction network; secondly, effective channel attention (ECA) and Adaptive Spatial Feature Fusion (ASFF) are used to enhance the PANet structure with attention-guided fusion; in addition, the weight ratio of the loss function is optimized and the mosaic method is used for training enhancement.

Keywords

Target detection; vehicle detection; YOLOv4; feature fusion; attention mechanism; lightweighting

1. Introduction

YOLOv1 [1] achieves real-time performance of 155 fps. The algorithm divides the network into multiple grids, and each grid is responsible for predicting only the location and class of targets whose centers fall on that grid. This was followed by the SSD [2] and YOLOv2 [3], both of which improved detection accuracy and speed. However, the accuracy of these algorithms is still relatively limited, especially for small targets. YOLOv3 [4] uses an Anchor-based approach that allows targets at different scales to be preassigned a close detection frame form, although YOLOv3 uses MSE as the border regression loss function, which makes YOLOv3's localization of targets not precise. RetinaNet [5] analyzes the category imbalance problem existing in the first stage of network training and proposes Focal loss that can automatically adjust the weights according to the Loss size, making the

BAIT'2024: The 1st International Workshop on "Bioinformatics and applied information technologies", October 02-04, 2024, Zboriv, Ukraine

[*] Corresponding author.

[†] These authors contributed equally.

ouyangcaixiao@163.com (C. Ouyang); hujiwei@fiberhome.com(J. Hu); yongyuanshe@163.com (Y. She); chunzhiwang@hbut.edu.cn(C. Wang)

0009-0004-9837-7136 (C. Ouyang); 0000-0002-9622-744X(J. Hu); 0009-0004-4925-2206 (Y. She); ; 0000-0002-6742-3644 (C. Wang) ;

training more focused on difficult samples. Yolov4 introduces the SPP module [6], Mish [7] activation function, etc., to improve the performance of the network.

With the development of deep learning algorithms, multi-target and multi-scale detection in complex environments, severe partial occlusion of vehicles, and high requirements for computing hardware are in the focus of research [8].

FPN [9] is a network for solving multi-scale detection problems. It uses a pyramid structure to make features flow between vertical and horizontal and propagates semantic information between multiple layers to build multi-scale features. However, FPN does not handle the difference of information at different levels reasonably, and the operation of fused features is obtained by summing the higher-level features with the next level directly after sampling, which limits the self-learning of features. Therefore, recently appeared works to optimize and improve FPN. For example, PANet [10] adds an extra top-down path to the original structure and adopts a channel superposition when fusing features, which both uses new feature information and ensures the preservation of original features. In addition, the attention mechanism (AM) is gradually becoming a popular method to improve detection performance. Various attention modules, used as a plug-and-play component, bring good performance improvements at an acceptable model complication. They select from the channels or spatial dimensions of the model and filter out the feature information that is more interesting and better matches the detection target.

This paper proposes a vehicle detection algorithm based on feature fusion and attention enhancement, which can purposefully alleviate the problems of missed detection, false detection, and accuracy degradation caused by detection scale or occlusion while reducing the complexity of the model. The main work of this paper is as follows:

1. DenseNet [11] with lower complexity is used as the backbone network of the detection model.
2. Introducing effective channel attention (ECA) [12] attention network, filling in the structure between the backbone and neck layer to achieve a smooth transition of features and selection of channel information.
3. Improving the network structure of the feature pyramid, adding Adaptive Spatial Feature Fusion (ASFF) [13] fusion module based on PANet.

2. Materials and Methods

2.1. Libertinus fonts for Linux Related Materials

2.1.1. One-stage Target Detection

The YOLO series algorithm innovates on the detection principle of the Faster Region-based CNN (R-CNN) series by abandoning the RPN approach and using regression to obtain the coordinate information of the bbox. YOLOv1 is an one-stage target detection algorithm. This algorithm was quickly deployed in many real-world projects due to the dramatic increase in detection speed. Many one-stage target detection algorithms have emerged since then [14].

YOLOv4 consists of the CSPDarknet53 backbone network, SPPNet, PANet feature fusion network, and the YOLO-Head detection head module, that is used in YOLOv3. It is shown in Figure 1.

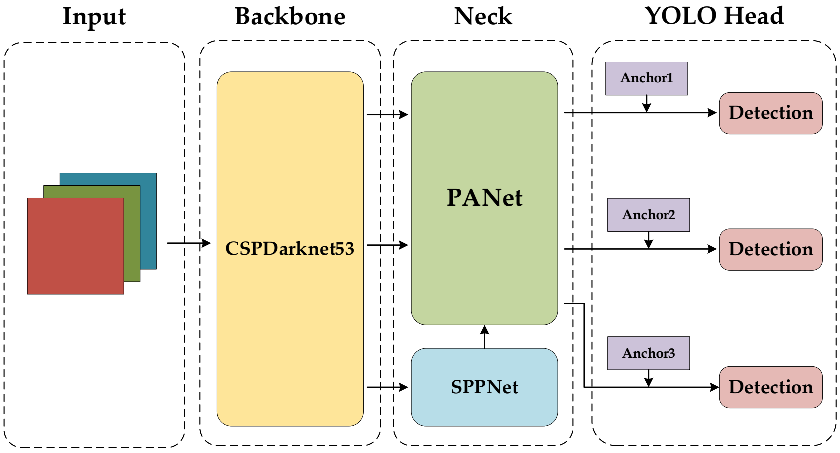


Figure 1 ：YOLOv4 network structure.

CSPDarknet53 is an improvement on Darknet53, which uses the CSPNet structure and applies a more extensive residual structure to reduce the information loss during training and further enhance the learning ability of the network. The activation function Leaky ReLU is replaced by the Mish function, whose upward unbounded property avoids model saturation due to numerical capping. In addition, its micro design for negative values brings better gradient flow. The Mish smoothed activation function ensures better accuracy and generalization.

Between the backbone network and the detection head is the Neck layer, which is composed of the SPP (Spatial Pyramid Pooling) module and the PANet module. The output of the backbone network is adjusted by the convolutional layer and used as the input of the SPP module. The SPP outputs the input data after doing maximum pooling and data stacking at different scales, and is adjusted by the convolutional layer and used as the input of the PANet network together with the two intermediate layers of the backbone network. PANet does further fusion of three sets of feature maps at different scales by some convolution, upsampling, downsampling and data stacking to enhance the perceptual field of feature maps at different scales and output three layers of data information.

The YOLO-Head in the detection layer receives the input from PANet and performs the final prediction process. The YOLO-Head with three a priori frames each will predict three feature maps with scales of 13X13, 26X26, and 52X52, respectively, and based on the a priori frame analysis information, the preliminary prediction frame will be output after non-maximum suppression.

In this paper, we improve the training and inference speed of the one-stage detection algorithm by modifying the backbone network of the model, based on the YOLOv4 algorithm, and improve the model structure using the AM and feature fusion module to enhance the detection performance of the algorithm.

2.1.2. Feature Parymid Network (FPN)

Feature Pyramid Representation (FPN) addresses the challenge of scale variation in target detection. Its structural layer design allows the model to better utilize the feature information extracted from the backbone network.

Initial target detection, either one-stage or two-stage, is usually performed with an external detection head after the feature map is output at the last layer of the last stage of Backbone. This approach is called the single-stage object detection algorithm. However, in this algorithm, the scale of the last output feature map of the backbone is too different from the input image, which is easy to cause information loss, especially the detection capability of small targets is insufficient. Subsequent studies found that the single-stage target detection algorithm cannot effectively transfer the information of various scales in the original image. Therefore, later target detection algorithms gradually developed into a feature pyramid network (FPN) using multi-scale, multi-stage feature maps to enhance the characterization ability of the model.

The FPN evolved through continuous iterations and can be divided into four models, as shown in Figure 2.

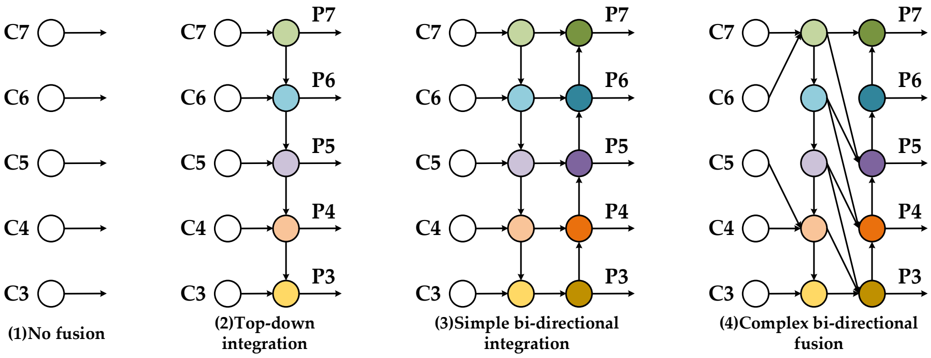

C7
C6
C5
C4
C3
P7
P6
P5
P4
P3
(1)No fusion
(2)Top-down integration
(3)Simple bi-directional integration
(4)Complex bi-directional fusion

Figure 2：Various FPN modes

1. A typical representative of fusion-free and at the same time utilizing multi-scale features is the SSD algorithm, which directly predicts objects of different sizes from the feature maps outputted by different stages.
2. There are many classical models that use algorithms with top-down fusion approach, such as Faster RCNN, Mask RCNN [15], Yolov3, RetinaNet, etc. They use the same kind of FPN models, and the difference is that feature maps of different scales are involved in feature fusion.
3. PANet proposes a top-down model followed by an additional bottom-up secondary fusion, which can be called a bidirectional fusion structure. YOLOv4 uses a fine-tuned version of PANet, which makes feature fusion not additive, but feature stacking.
4. The proposed PANet proved the effectiveness of bidirectional fusion, introduced more complex bidirectional fusion structures, such as NAS-FPN [16] and BiFPN [17].

Various FPNs are designed to maximize the utilization of the multi-scale feature maps from backbone, and its optimization leads to significant improvement of object detection. Therefore, the algorithms in this paper in concert with the fusion of PANet and ASFF to enhance the reuse and extraction of feature maps and avoid the loss of effective information [18, 19].

2.1.3. Attention Mechanisms

The AM focuses on local information while suppressing distracting information [20]. From a mathematical point of view, AMs provide a weight-based model to perform operations. The process of extracting image features from feature maps in a NN is seen to vary in the degree to which different feature maps provide overall information [21]. The AM uses the network layer to calculate the weight values corresponding to the relevant feature maps, and then applies these weights to the feature maps, so that the feature maps with a large role in extracting information become somewhat more influential on the overall [22]. The AMs can currently be classified into following types: channel AMs, spatial AMs, and mixed spatial and channel AMs.

2.1.3.1. Spatial AM

Not all regions in an image are equally important, only the task-relevant regions are important. The spatial attention model is to find the most important parts of the network for processing.

The Spatial Transformer Network (STN) [23] is a spatial-based Attention by learning the shape change of the input so as to accomplish preprocessing operations suitable for a specific task. The ST module consists of the Localisation net, the Grid generator and Sample. The Localisation net determines the parameter θ of the input required transformation. The Grid generator finds the mapping T(θ) of the output to the input features by θ and the defined transformation. The Sample combines the location mapping and transformation parameters to select the input features and combine them with bilinear interpolation for the output.

2.1.3.2. Channel AM

For a set of images processed by the CNN, its effective information can be extracted from two dimensions. One dimension is the scale of the image in space, that is, the length and width. The other dimension is the channel information. Therefore, Attention based on channel orientation is also common.

SENet (Sequeeze and Excitation Net) [24] is a channel type Attention model, which automatically enhances or suppresses channels after model learning by modeling the importance of each feature channel. It divides a bypass branch after the normal convolution operation, and this branch is compressed and fully connected to obtain a set of weight values. By applying this set of weights to each of the original feature channels, the importance of the different channels can be learned.

2.1.3.3. Fusion of spatial and channel AMs

CBAM (Convolutional Block Attention Module) [25] is a representative network that combines spatial and channel AMs. It uses a channel-then-space approach for collocation, so that the model models the important information of channel and spatial locations separately.

Besides these, there are many other AMs related to research [26, 27]

2.2. The Proposed Method

Figure 3 shows the architecture of the proposed algorithm. It takes one-stage target detection algorithm YOLOv4 as the reference architecture and divides the algorithm framework into four parts: data pre-processing and input, backbone network, FPN structure and prediction network. The pre-processed images are sent to the backbone network, which adopts a lightweight DenseNet structure consisting of different numbers of Dense Blocks and Transition Layers. Depending on the number of sub-module overlays, the backbone network extracts the feature information at different scales and passes it into the FPN network. Before this information is passed into SPPNet and PANet, the feature information will be further filtered and refined by three ECA attention modules. Then the information output from the bidirectional fusion-type network PANet is fed into the complex fusion network ASFF, which makes the feature map information at different scales form the interaction. Finally, the information extracted from the ASFF network is fed into the YOLO detection head, and the prediction results of the image are obtained after the information decoding and other operations. Next, the backbone network, FPN structure and loss function of the algorithm in this paper are described in more detail, respectively.

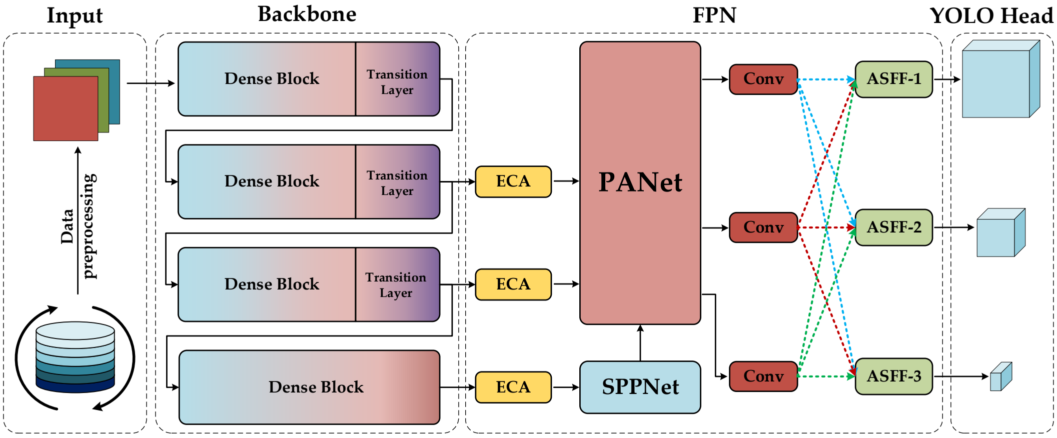


Figure 3 :The structure of the proposed algorithm

2.2.1. Lightweighting Of The Backbone

The lightweight network DenseNet is integrated and bridged with the original YOLOv4 to achieve faster, more accurate, and less computationally intensive target detection results. Specifically, the backbone network is replaced with DenseNet-121, and the rest of the architecture is optimized on the basis of YOLOv4.

As another type of CNN with deeper layers, it has the following advantages:

1. Fewer number of parameters compared to ResNet.

2. More emphasis and encouragement on feature reuse.
3. The network is easier to train and has some regularization effect.
4. The problems of gradient vanishing and model degradation are alleviated.

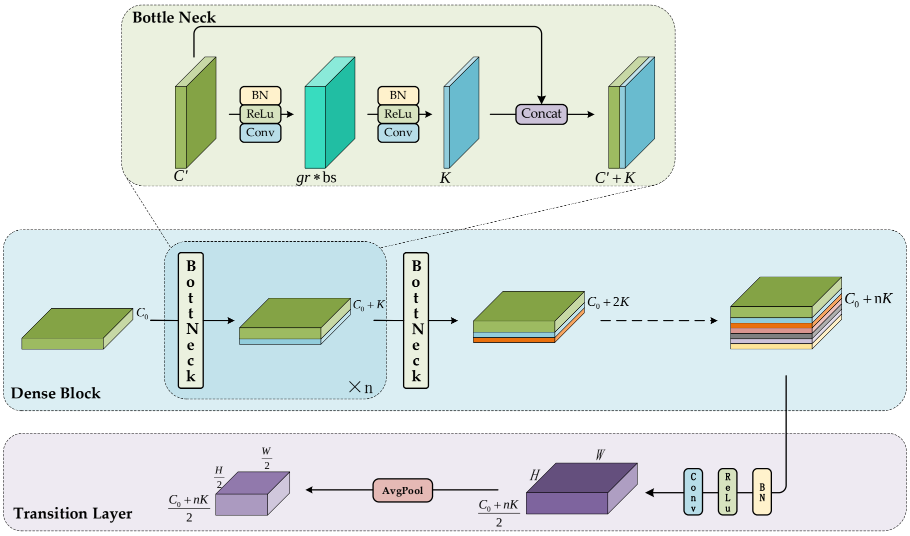


Figure 4 ：DenseNet main structure

DenseNet is mainly composed of Dense Blocks and Transition Layers.

The dense block is composed of several bottle necks. Each block uses the same number of output channels, and then uses a loop to connect the input and output of each block in the channel dimension. The structure of the bolt neck is shown in the upper part of Figure 4.

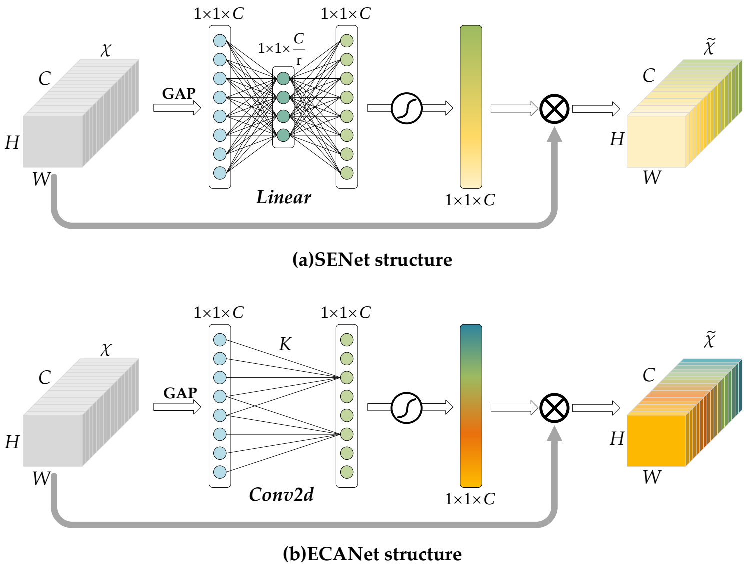
BN-ReLU is placed before the convolution module for processing. Each Bottle Neck contains two convolutions, the first one is a 1*1 convolution, which has 4k output channels. Here, k is a feature map growth factor, which is the number of feature maps contributed by each Bottle Neck. The second 3*3 convolution has k output channels. Finally, the input of the module and the output of the 3*3 convolution are concat stacked to obtain the overall number of output channels of the module as C`+k.

The Dense Block structure is shown in the middle part of Fig. 4. It consists of several Bottle Necks. The number of input channels of the whole Dense Block is C0. Since the output of Bottle Neck stacks, the output and input of the final convolutional structure in its interior, the number of feature channels will be increased by k for each Bottle Neck that passes through it. Therefore, the number of final output feature maps of a Dense Block composed of n Bottle Neck is C0+nk. The input of each Bottle Neck is a stack of all the outputs of its preceding layers.

The Transition Layer controls the model complexity. Its structure is shown in the bottom of Fig. 4. Since the number of channels increases with each Dense Block connection, its overuse will result in an overly complex model. Therefore, the Transition Layer first reduces the number of channels by a 1×1 convolution layer, and then to compress the height and width of the feature map, an average pooling layer with stride=2 is used for downsampling, which further reduces the model complexity.

2.2.2. Citation of Attentional Mechanisms

Among the types of attention modules, channel AMs have great potential in im-proving the performance of deep CNNs. However, there are a large number of AMs developing more complex attention modules to achieve better performance, which will inevitably increase the complexity of the model. To strike a balance between model complexity and performance, this paper refers to an effective channel attention module (ECA) that contains only a small number of parameters while delivering significant performance gains.



(a)SENet structure

(b)ECANet structure

Figure 5：SENet and ECANet structures

SE-Net is the basis of ECA-Net optimization and its structure is shown in Figure 5(a). Global average pooling is first performed separately for each input channel, followed by two fully connected layers using different activation functions. This computational process causes the channel features to be mapped from high to low and then to high dimensions. This dimensionality reduction operation reduces the complexity of the model, but it also cause the loss of critical information.

ECA-Net empirically shows, by observing SE-Net and improving it, that avoiding dimensionality reduction is important for learning channel attention and that proper cross-channel interaction can increase model complexity only slightly while maintaining performance. Its structural design is shown in Figure 5(b).

On the left is the feature of the original input image, which is first subjected to global average pooling (GAP) [28] to obtain a 1×1×C feature map, on which ECA obtains the local cross-channel interaction by fast one-dimensional convolution of size K, where the parameter K can be generated by an adaptive function based on the size of the input channel C, which represents the local coverage of the cross-channel interaction. After that, a Sigmoid function is used to generate the weight share of each channel, and then the original input features are combined with the channel weights to obtain the features with channel attention. The network constructed with this module makes it easier to extract discriminative features of images based on channel dimensionality.

To avoid the consumption of large computational resources due to manual adjustment, the size of the parameter k can be generated adaptively by a function with the convolution kernel k calculated as:

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right| odd \tag{1}$$

where |t|odd denotes the odd number of t-nearest neighbors, γ is set to 2, and b is 1. From (1), it is clear that the communication range of the high-dimensional channel is longer, while the communication range of the low-dimensional channel is relatively contracted.

In this paper, three ECA layers are inserted at the connection between Backbone and Neck of the model to avoid dimensionality reduction while better bridging the two components, making the feature transfer of the model more efficient and preventing the disappearance of feature information. At the same time, the ECA layer allows the model to focus on more critical features and suppress unnecessary features, which improves the detection accuracy.

2.2.3. Spatially Adaptive Fusion Of Feature Layers

In general, the lower level features of the network contain more location information and the higher level features contain more semantic information. The PANet structure is used in YOLOV4 to further fuse and output the higher and lower level features. After downsampling, the network does bidirectional propagation and then upsampling, and fuses the information from the same level downsampling by lateral connection, and then sends the feature information of different scales to different detectors.

However, the PANet connection simply stacks the top-down and bottom-up layers of information together, and there is a lack of communication between the layers to transfer the information. To more fully utilize the semantic information of the high-level features and the fine-grained features of the underlying features, this paper introduces a new feature fusion method, Adaptive Spatial Feature Fusion (ASFF), in the proposed algorithm.

ASFF can enhance the extraction capability of PANet and can fuse the information of multiple feature layers simultaneously. Its idea is to adaptively adjust the spatial weights of each scale features in fusion by learning. Its underlying structure is shown in Figure 6.

Figure 7 shows the operation of layers in ASFF. First, X1, X2 and X3 are derived from the feature information at different scales of level1, level2 and level3 output in PANet, respectively.

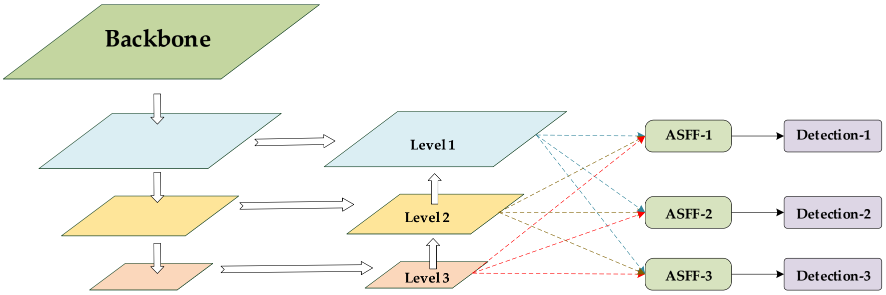


Figure 6：ASFF schematic

The ASFF-3 is an example of a convolution with the kernel of 3*3, the step size of 2, and a padding of 1. The X2 is scaled down to the same value as X3 with equal number of channels, and is denoted as level_1_resized. The number of channels and dimensionality of level_1_resized, level_2_resized and X3 are the same. Finally, level_1_resized, level_2_resized, and X3 are multiplied by α, β, and γ, respectively, and the values are summed, and the number of channels is adjusted by a final convolutional layer to obtain a new feature layer with multi-layer perceptual field fusion. The formula is expressed as follows:

$$y_{ij}^{l}=\alpha_{ij}^{l}\cdot X_{ij}^{1\to l}+\beta_{ij}^{l}\cdot X_{ij}^{2\to l}+\gamma_{ij}^{l}\cdot X_{ij}^{3\to l} \tag{2}$$

where yijl represents the new feature map of a layer obtained by ASFF, αijl, βijl, and γijl represent the weight parameters learned through the three feature layers, and αijl+βijl+γijl=1 is guaranteed by the Softmax function.

where y_{ij}^{l} represents the new feature map of a layer obtained by ASFF, α_{ij}^{l}, β_{ij}^{l}, and γ_{ij}^{l} represent the weight parameters learned through the three feature layers, and $\alpha_{ij}^{l}+\beta_{ij}^{l}+\gamma_{ij}^{l}=1$ is guaranteed by the Softmax function.

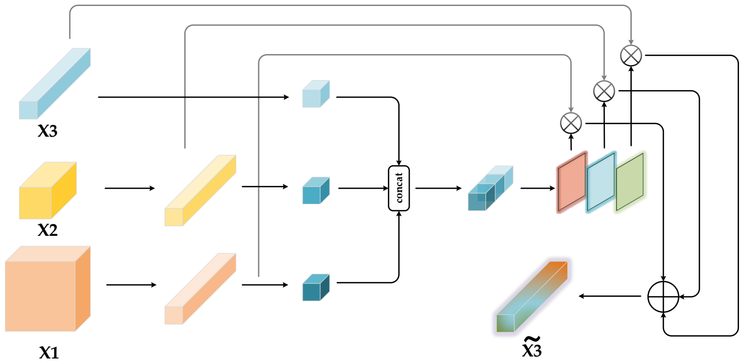


Figure 7 :ASFF specific operations

2.2.4. Design of the loss function

The loss function contains three components: confidence error Lconf classification error Lcls, and regression frame prediction error Lloc [29]. CIoU loss was used in the regression frame prediction error. CIoU is based on IoU, GIoU, and DIoU, and the CIOU takes into account three geometric factors, which are overlap area, centroid distance, and aspect ratio [30].

$$
\begin{aligned}
L_{conf} = & -\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\left[\overline{C_i^j}\log\left(C_i^j\right)+\left(1-\overline{C_i^j}\right)\log\left(1-C_i^j\right)\right]- \\
& \lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{noobj}\left[\overline{C_i^j}\log\left(C_i^j\right)+\left(1-\overline{C_i^j}\right)\log\left(1-C_i^j\right)\right]
\end{aligned}
\tag{3}
$$

$$
L_{cls} = -\sum_{i=0}^{S^2} I_{ij}^{obj} \sum_{c\in classes}\left\{\overline{P_i^j}(c)\log\left[P_i^j(c)\right]+\left[1-\overline{P_i^j}(c)\right]\log\left[1-P_i^j(c)\right]\right\}
\tag{4}
$$

$$
CIoU\left(X,Y\right) = IoU\left(X,Y\right) - \frac{\rho^2\left(X_{ctr},Y_{ctr}\right)}{m^2} - nv
\tag{5}
$$

where S^2 is the number of grids, B is the number of prediction frames in each grid, I_{ij}^{obj}, I_{ij}^{noobj} are the indicated values of the prediction frames containing and not containing the target, $\bar{C}$ is the confidence true value, C is the prediction confidence, λ_{noobj} is the penalty weight factor, $\bar{P}(c)$ is the actual probability that the target in the cell belongs to category c, P(c) is the probability that the prediction is of category c, wgt, hgt are the width and height of the true frame, respectively, IoU(X , Y) is the intersection ratio of the predicted frame X to the real frame Y, ρ2(Xctr, Yctr) is the Euclidean distance between the center point of the predicted frame and the real frame, m is the diagonal distance of the minimum closed region containing both the predicted and real frames, u is the balance adjustment parameter, and v is the parameter measuring the consistency of the aspect ratio.

To balance the loss sensitivity of different detection scales, in this paper, the three prediction heads in the network structure are multiplied with different weights when calculating the total loss. The weights assigned to Yolo Head1, Yolo Head2, and Yolo Head3 are 0.4, 1.0, and 4.0, respectively [31].

3. Conclusion

This paper focuses on the One-Stage target detection method which has higher requirements for detection speed and deployment cost. It helps cameras in traffic scenes to recognize vehicle information and perform vehicle model discrimination. A lightweight target detection algorithm based on attention and feature augmentation is proposed to address the problem of the demand for vehicle detection in smart city construction. The complexity of the algorithm is strictly controlled. The proposed algorithm uses YOLOv4 as the base architecture: (i) significantly reduces the number of model parameters by replacing the DenseNet, which has excellent performance, as the backbone feature extraction network;v(ii) reconstructs the existing FPN network module, uses the ECA attention structure for the transition and transfer of feature information between Backbone and Neck, as well as adds the information cross-fusion function before the final detection layer of the network of the ASFF structure; (iii) while optimizing in terms of the loss function and image preprocessing.

References

[1] W J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

[2] Liu W, Anguelov D, Erhan D, et al, "SSD: Single Shot MultiBox Detector," 2016 Computer Vision and Pattern Recognition (ECCV), 2016, doi: 10.1007/978-3-319-46448-0_2.

[3] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6517-6525, doi: 10.1109/CVPR.2017.690.

[4] Fang, M. T., Chen, Z. J., Przystupa, K., Li, T., Majka, M., & Kochan, O. (2021). Examination of Abnormal Behavior Detection Based on Improved YOLOv3. Electronics 2021, 10, 197.

[5] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 318-327, 1 Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[6] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015, doi: 10.1109/TPAMI.2015.2389824.

[7] Misra D. Mish: A self regularized non-monotonic neural activation function[J]. arXiv preprint arXiv:1908.08681, 2019, 4(2): 10.48550.

[8] Q. Ailing and T. Ning, "Fine-grained vehicle recognition method based on improved ResNet," 2020 2nd International Conference on Information Technology and Computer Application (ITCA), 2020, pp. 588-592, doi: 10.1109/ITCA52113.2020.00129.

[9] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936-944, doi: 10.1109/CVPR.2017.106.

[10] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, "Path Aggregation Network for Instance Segmentation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759-8768, doi: 10.1109/CVPR.2018.00913.

[11] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

[12] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," 2020 IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR), 2020, pp. 11531-11539, doi: 10.1109/CVPR42600.2020.01155.

[13] Liu S, Huang D, Wang Y. Learning spatial fusion for single-shot object detection[J]. arXiv preprint arXiv:1911.09516, 2019.

[14] X. Dai et al., "Dynamic Head: Unifying Object Detection Heads with Attentions," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7369-7378, doi: 10.1109/CVPR46437.2021.00729.

[15] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[16] G. Ghiasi, T. -Y. Lin and Q. V. Le, "NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7029-7038, doi: 10.1109/CVPR.2019.00720.

[17] M. Tan, R. Pang and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10778-10787, doi: 10.1109/CVPR42600.2020.01079.

[18] Y. Gong, X. Yu, Y. Ding, X. Peng, J. Zhao and Z. Han, "Effective Fusion Factor in FPN for Tiny Object Detection," 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1159-1167, doi: 10.1109/WACV48630.2021.00120.

[19] Xiong, G., Przystupa, K., Teng, Y., et al. (2021). Online measurement error detection for the electronic transformer in a smart grid. *Energies*, *14*(12), 3551.

[20] Jiang, K., Zhang, C., Wei, B., Li, Z., & Kochan, O. (2024). Fault diagnosis of RV reducer based on denoising time–frequency attention neural network. Expert Systems with Applications, 238, 121762.

[21] Xu, X., Przystupa, K., & Kochan, O. (2023). Social Recommendation Algorithm Based on Self-Supervised Hypergraph Attention. Electronics, 12(4), 906.

[22] Y. Wang and A. Zell, "Yolo+FPN: 2D and 3D Fused Object Detection With an RGB-D Camera," 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 4657-4664, doi: 10.1109/ICPR48806.2021.9413066.

[23] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[J]. Advances in neural information processing systems, 2015, 28.

[24] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132-7141, doi: 10.1109/CVPR.2018.00745.

[25] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[26] F. Wang et al., "Residual Attention Network for Image Classification," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6450-6458, doi: 10.1109/CVPR.2017.683.

[27] Deng J, Cheng L, Wang Z. Attention-based BiLSTM fused CNN with gating mechanism model for Chinese long text classification[J]. Computer Speech & Language, 2021, 68: 101182.

[28] Lin M , Chen Q , Yan S . Network In Network[J]. Computer Science, 2013.

[29] N. K. Kim and H. K. Kim, "Polyphonic Sound Event Detection Based on Residual Convolutional Recurrent Neural Network With Semi-Supervised Loss Function," in IEEE Access, vol. 9, pp. 7564-7575, 2021, doi: 10.1109/ACCESS.2020.3048675.

[30] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 12993-13000.

[31] Jiang Z, Fan Y. Singularity intensity function analysis of autoregressive spectrum and its application in weak target detection under sea clutter background[J]. Radio Science, 2020, 55(10): 1-8.