ABSTRACT End-to-end automatic driving requires the identification of lane curvature. We proposed a lightweight detection method for the low-light lane curvature based on the Fractional-Order Fusion Model (FFM) to assure real-time performance and increase the reliability of automatic driving in low-light conditions. To begin, the FFM method is introduced to enhance images with low average brightness, fuzzy detail, and a high signal-to-noise ratio. Under low-light conditions, these images cannot clearly express information (such as rainy, snowy, foggy, and other harsh external environments). Then, aiming at the problems of complex network structure, high hardware configuration required for training, and low transmission Frames Per Second (FPS) of real-time detection in the previously proposed YOLOv5, the SETR-C3Block module is proposed. The YOLOv5n is improved by optimizing the configuration of the target detector head and the network’s structure, which solves the problems of low efficiency and redundancy parameters in feature extraction in the network. According to the experimental results on the lane curvature dataset, the mAP@.5:.95 of SETR-YOLOv5n is 87.22%, the transmission FPS of real-time detection is 70.4, and the number of model parameters is only 1.8M. It shows that the SETR-YOLOv5n can meet the lightweight and accuracy requirements of target detection by the mobile terminal or embedded device.

INDEX TERMS Lane curvature detection, automatic driving, lightweight, YOLOv5, fractional-order, image enhancement.

I. INTRODUCTION
With improving people’s living standards and the development of intelligent industry, automatic driving has achieved great success. Autonomous driving has the functions of scene recognition, path planning and motion control, which brings great convenience to driving [1]. Additionally, automatic driving has the potential to significantly cut traffic accident rates and increase driving safety. With the continual advancement of computer vision (CV) technology in recent years, real-time lane curvature automatic detection and recognition based on the CV is now conceivable [2]. However, the storage space and computing power of the onboard platform are generally limited, and there are high requirements for the lightweight level of the detection method [3], [4]. Furthermore, to be employed in natural environments, a vision-based detection model must be resistant to various situations, particularly low-light conditions such as rainy, snowy, foggy, and other hostile external environments. Sometimes, even deep-learning-based CV methods did not perform well in those low-light conditions due to relatively low image quality [5], the YOLOv5 method is difficult to balance these problems.

In recent years, many strategies for improving low-light images have been presented. Histogram Equalization (HE) has been widely used in improving image contrast [6], [7]. Especially when the image contrast is low, this method can make the gray histogram distribution of the image more uniform. However, because HE is a method for increasing
the intensity distribution range of pixels in order to boost the brightness of an image, or to improve the contrast of an image, the enhancement results obtained with HE frequently appear overexposed or distorted [8]. Ibrahim et al. [9] proposed Dynamic Histogram Equalization (DHE), which uses a sliding window strategy to perform HE locally. In this method, the minimum value in the image histogram is used as the image segmentation point. The gray level accumulation frequency in the sub histogram is used as an influencing factor for dividing the new gray range. The sub histograms are equalized in the new gray range to improve the contrast. DHE method can effectively enhance the contrast of the image and prevent the appearance of blurred edges. In addition, some scholars have proposed other methods based on HE, such as weighted threshold HE [10], modified clipped HE [11], and non-parametric modified HE [12], which can enhance various types of images. These methods based on HE have achieved certain results, but they can not avoid the problem that HE methods are sensitive to noise. Another classical nonlinear method is gamma correction, which corrects the brightness deviation by the power-law transformation of the gray value of the input image. The final result is to increase the brightness of the dark area and keep the brightness of the bright area [13]. Later, many methods adopt more global analysis and processing. For example, some scholars design a nonlinear enhancement function based on the local discreteness of wavelet coefficients to enhance the image contrast in the way of statistical modeling [14].

Fractional-order calculus is an important branch of mathematical analysis [15], [16], [17]. Liu et al. [18] introduced fractional-order calculus singular value into face recognition technology, which improved the efficiency and accuracy compared with the traditional recognition technology. Zhang et al. [19] suggested a novel image improvement technique that distinguishes the picture’s smooth region, edge, and texture area using a mix of rough set and particle swarm optimization algorithms. The picture is then enhanced using an adaptive fractional differential filter based on the image segmentation findings. Liu et al. [20] analyzed by selecting multiple images to select the best differential order to enhance the image adaptively. They also gave a template for enhancing the remote sensing image, with a size of $5 \times 5$. It is verified that the proposed enhancement method can effectively enhance the remote sensing image. However, it needs to calculate all fractional orders when calculating the best differential order, which makes the amount of calculation very large. Li et al. [21] proposed a new medical image enhancement method that adjusts the fractional order according to the dynamic gradient feature of the entire image. Zhang et al. [22] presented a rough set and fractional-order differentiator-based picture enhancing method. This image segmentation technique can obtain more picture layers than previous algorithms by focusing on information and maintaining more details. For low-light images, several researchers used various ways to alter the image illumination and proposed the fusion-based image enhancement method [23], which effectively improved the image’s illumination. However, the method’s goal is limited to poorly lighted images, not degraded images or of poor quality. Similarly, Zhang et al. [24] proposed a dual illumination estimation-based automatic image exposure correction system. The multi-exposure image fusion approach transforms an image with both under-exposed and over-exposed parts into a well-exposed image overall. Despite its positive findings, this approach has not been proved to preserve the fine details of medical images necessary for diagnosis. Similarly, more image-enhancing methods based on nonlinear optimization with many constraints have been proposed. Zhou et al. [25] proposed a new optimization strategy for managing gamma values to improve image contrast. A typical gamma correction is used in the enhancing procedure. Imposing many optimization constraints increases the computing complexity of this technique. As a result, we introduced the FFM-based model in this work for upgrading various lane curvature images by enhancing the dark area while keeping the bright area of the input image.

Compared with other methods, end-to-end automatic driving dramatically reduces the hardware cost and the research difficulty [26]. This method can also obtain universality in different scenarios with the help of dataset diversity. However, at present, end-to-end automatic driving still lacks practicability and reliability and is dependent on the performance of the onboard computer [27]. Therefore, the application of this technology has high requirements for the lightweight and detection accuracy of the target detection method [28]. YOLOv5 is widely used in CV tasks such as target detection and has achieved great success. However, due to the limitation of storage space and hardware performance, the storage and calculation of the YOLOv5 network model on vehicle equipment is still a considerable challenge. In 2021, YOLOv5 released the 6.0 Version, and this new version integrates many new functions based on YOLOv5 5.0 Version, streamlines and fine-tunes the network structure, and introduces the YOLOv5n, which maintains the depth of YOLOv5s and reduces the width parameter from 0.5 to 0.25. After this operation, the total parameters are reduced by 75%, from 7.3M to 1.8M. The number of Floating-Point Operations Per Second (FLOPs) is reduced by 72%, from 17.0G to 4.7G, which is very suitable for the onboard computer environment. However, YOLOv5n still has some problems, such as low detection accuracy and poor convergence. At the same time, under the condition of low light, the overall clarity of the image is low, the edge is fuzzy, and the details and texture structure of the image are not well preserved, which has seriously affected YOLOv5n the detection effect. As a result, it is essential to integrate the fractional-order picture enhancement method with the lane curvature detection method based on improved YOLOv5n to increase the real-time and reliability of automatic driving while lowering the hardware and software costs of automatic driving.
A. CONTRIBUTION

According to the above research results, this paper proposes a lightweight low-light lane curvature detection method SETR-YOLOv5n based on FFM. This method, based on the YOLOv5n lightweight target detection model, combines the FFM image enhancement method and the new network module SETR-C3 in this paper. It optimizes the network structure, which solves the problem of the accuracy of the traditional lightweight target detection algorithm that is not high due to low-light and a poor environment. We think this research is significant because target detection, as an auxiliary task for reinforcement learning, has been paid more and more attention in the research of autonomous driving. This paper will be most helpful to researchers who study the application of the combination of target detection and reinforcement learning in autonomous driving in low-light conditions. This paper’s contributions can be summarized as follows:

1) To overcome the bottleneck of the model in low-light image detection performance and provide high-quality images for image processing and analysis, FFM for low-light image enhancement is introduced to enhance the quality of images. Compared to the previously proposed integer-order methods [29], this method preserves the visual appearance and gets more detailed semantic features in low-light areas.

2) This paper proposes a lightweight model for low-light lane curvature detection based on FFM named SETR-YOLOv5n, which improves the network structure based on YOLOv5n. In the feature extraction stage, the SETR-C3 module, which is proposed in this paper embedded with a weighted feature fusion method and two attention mechanisms is used to replace the C3 module, which better meets the needs of the YOLOv5n network detection layer; In the part of the detection head, through the cluster analysis of the lane curvature dataset, the distribution of anchor frame width and height is obtained. The detection head is simplified and optimized to reduce the YOLOv5n’s parameters, improve the efficiency of target detection, and better detect the lane curvature in a complex road environment.

3) Aiming at the problem of the unbalanced number of pictures in each category in the lane curvature dataset, the original 8000 images are expanded with the tools of Albumentations. The images are expanded to more than 10000 by the image enhancement tool, and the enhanced lane curvature dataset is established.

4) The method proposed in this paper is a lightweight target detection method based on pure vision under low-light conditions, which is more economical and practical than the traditional target detection method under low-light conditions using lidar or infrared thermal imager and can meet the lightweight and accuracy requirements of target detection method by the mobile terminal or embedded device. In the later chapters of this paper, we discuss the possibility of combining the method proposed in this paper with deep reinforcement learning so that the method proposed in this paper can be applied in practice later by researchers.

B. ORGANIZATION

The main contents of each section of this paper are as follows: Section II describes the composition and background of SETR-YOLOv5n. In Section III, the effectiveness of the method is illustrated by a series of comparative experiments. In Section IV, the influence of the SETR-YOLOv5n component on performance and the future application of SETR-YOLOv5n in automatic driving are discussed. Finally, Section V summarizes the full text and draws a conclusion.

II. THE METHODS

A. FRACTIONAL-ORDER FUSION MODEL BASED ON RETINEX

1) RELATED WORK

Image is an important way to record and transmit information. Due to the influence of many reasons, the definition and quality of the image will gradually decline in image transmission, which leads to some difficulties when people transmit pictures for analysis many times. Therefore, image enhancement has become an important part of image processing in this case. The image sensor is the main source of input datasets for various optical imaging devices, computer image processing systems, and auto-drive systems. In the process of automatic driving, due to the influence of weather, exposure conditions, and other factors, the image sensor’s visibility and contrast of the image output will be reduced, resulting in poor image quality and affecting the effect of automatic driving. Low-light images have low contrast, concentrated gray level range, and low image quality, which seriously affects the effect of target detection [30]. Therefore, improving the low-light image quality is significant in practical applications.

To model the human visual perception system, the Retinex hypothesis is developed [31]. The image perceived by the observer is primarily regarded as an image with multiplicative noise in this theory. As a result, the illumination map is a multiplicative noise with a gradual transformation that is generally uniform. The Retinex theory works by estimating the noise in distinct pixels in a picture and then removing the noise to get the original reflection image. Furthermore, the lighting schemes of the three channels of the color image must be assumed to be identical. The theory argues that the observed color image may be broken down into two components: reflectivity and illumination. The following is a representation of the mathematical expression:

\[ S = R \cdot L. \]  

(1)

where \( S \) and \( R \) are the captured image and the reflectance, respectively. \( L \) represents illumination, and \( \cdot \) represents element-wise multiplication.

Fractional-order calculus was born in 1695. Leibniz, a German mathematician, thought about what the expression meaning is when the derivative-order becomes \( \frac{1}{2} \). The evolution of fractional-order calculus theory is a topic of nearly
exclusive interest for a few mathematicians and theoretical physicists, with a 300-year history [32]. Compared with integer-order calculus, fractional-order calculus expands the order of operation [33]. Although implementation complexity is relatively high, it has a higher degree of freedom and flexibility than integer-order calculus. There are mostly unstable signals with various characteristics in modern signal analysis and processing. These signals have the characteristics of nonlinearity, noncausality and non-stationary. Therefore, a variety of methods in fractional-order calculus method are suitable for processing this kind of signal. The most commonly used fractional calculus definitions are Grünwald-Letnikov, Riemann–Liouville, and Caputo. This paper uses the Grünwald-Letnikov (G-L) fractional calculus definitions as it has the advantages of converting to convolution in numerical implementation and is suitable for image signal processing. The G-L definition is used as follows:

\[
G^\alpha L D\frac{f(x)}{a} = \lim_{N \to \infty} \left\{ \frac{1}{N^\alpha} \sum_{k=0}^{N-1} \left[ \Gamma(\alpha) \right] \cdot f \left( x - k \left( \frac{a}{N} \right) \right) \right\}, \quad (2)
\]

where the gamma function is defined as \( \Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1} dx \).

In recent years, the research and application of fractional-order calculus have made rapid progress in signal analysis and processing. As a two-dimensional signal, combining the image signal and fractional-order calculus theory has also attracted more and more scholars’ attention. The research and application of fractional-order calculus theory in two-dimensional image signals mainly focus on image enhancement, image denoising, image segmentation, edge detection, digital watermarking, and so on. In recent years, fractional-order calculus has been widely used to describe some nonclassical phenomena in engineering applications. Dai et al. [29] Proposed the FFM method using fractional-order calculus. Compared with integer-order calculus, it can better preserve the texture details of the image and suppress the noise. The energy function is modeled as:

\[
E_x(R) = \left\| \left( D_x^{1,1} R \right)^{v_2} \right\|_1 + \left\| \left( D_y^{1,1} R \right)^{v_2} \right\|_1, \quad (3)
\]

where \( \| \cdot \|_1 \) designates \( l_1 \) norms, and \( v_1, v_2 \) are fractional parameters.

2) THE IMPLEMENTATION OF FFM

The following is more information about the illumination priors:

1) The illumination map for the three channels in RGB images is the same. [34].
2) \( R \) is limited to the unit interval, and \( S \leq L \) can be obtained according to Equation (1).
3) \( L \) and \( S \) should be close enough [35].

In general, we set the optimization objectives as:

\[
E_5(L, R) = \| L \cdot R - S \|_2^2 + \alpha E_1(L) + \beta E_2(R) + \gamma \| L - \hat{L} \|_2^2, \quad (4)
\]

where \( \alpha, \beta \) and \( \gamma \) are the parameters.

As shown in Figure 1, the final enhancement result of the image is obtained through the fusion process. The fusion method can be expressed as follows:

\[
J(x, y) = \sum_{i=1}^{k} W_i(x, y) f \left( E_i(x, y) \right), \quad (5)
\]

where \( W_i \) is the weight of the i-th image. The principle block diagram of the FFM is shown in Figure 1.

The weight formula is as follows:

\[
W_1 = \frac{1}{1 + e^{-4lT_i + 2}}, \quad (6)
\]

\[
W_2 = e^{-\frac{(y - 0.5)^2}{0.25^2}}, \quad (7)
\]

\[
W_3 = \frac{1}{1 + e^{-4lT_i - 2}}, \quad (8)
\]

where \( l \) is the corresponding illumination.

According to the article [29], the n-th \( R \) can be got:

\[
V_{r_n} = (M_r + \beta N_r)^{-1} \left( V_{l_{n-1}} \cdot V_{s} \right), \quad (9)
\]

where \( M_r \) is the diagonal matrix formed by \( \left( V_{l_{n}} \cdot V_{l_{l}} \right) \), \( N_r = D_x^{1,1} W_1 D_y^{1,1} + D_y^{1,1} W_1 D_x^{1,1} \).

It is identical to the R sub-problem, the n-th \( L \) can be got:

\[
V_{l_n} = (M_l + \alpha N_l + \lambda I)^{-1} \left( V_{r_{n-1}} \cdot V_{s} + \lambda V_{j} \right), \quad (10)
\]

where \( M_l \) is the diagonal matrix formed by \( \left( V_{r_{n-1}} \cdot V_{r_{n-1}} \right) \), \( N_l = \left( D_x^{1,1} U_1 D_x^{1,1} + D_y^{1,1} U_1 D_y^{1,1} \right) \), and \( I \) is the identity matrix.

3) EXPERIMENT AND ANALYSIS

This research employs Information Entropy (IE) to objectively evaluate the improved picture in order to understand the small variations in the image more logically. Entropy is an estimate of the amount of information that will be available before the result is known. Information is the information provided by a given event, and entropy is an estimate of the amount of information that will be available before the result is known. The average quantity of information in a picture is reflected by image information entropy, which is a statistical form of feature. The image’s IE is then represented as follows:

\[
H = -\sum_{i=1}^{L} p(a_i) \log_2 p(a_i), \quad (11)
\]

where \( a_i \) is the random output signal of the image.
Seven low-light images are tested to validate the performance of the suggested model, as shown in Figure 2. In visual comparison, it can be seen that FFM can significantly enhance low visibility images.

As shown in Table 1, it can be seen that compared with several mainstream image enhancement algorithms, the FFM method can effectively improve the contrast of low-light images.

TABLE 1. Comparison between mainstream image enhancement methods.

| Dataset            | Image | CRM [37] | JIEP [36] | FFM  |
|--------------------|-------|----------|----------|------|
| Ours               | Image1| 7.0484   | 7.3495   | 7.6914|
|                    | Image3| 7.6579   | 7.7116   | 7.8453|
|                    | Image4| 7.5918   | 7.5429   | 7.6116|
|                    | Image5| 6.1425   | 6.2980   | 6.4426|
|                    | Image6| 6.9701   | 6.9371   | 7.1521|
|                    | Image2| 7.6279   | 7.0749   | 7.7291|
|                    | Image7| 7.2744   | 7.3245   | 7.5245|
| RGB-Thermal        | Image1| 0.8555   | 0.9247   | 0.9601|
|                    | Image3| 0.8422   | 0.8859   | 0.9462|
|                    | Image4| 0.7278   | 0.9039   | 0.9433|
| Nighttime-Datasets | Image5| 0.7819   | 0.8976   | 0.9670|
|                    | Image6| 0.7790   | 0.8841   | 0.9990|
|                    | Image2| 0.8388   | 0.8972   | 0.9548|
|                    | Image7| 0.8470   | 0.8520   | 0.9446|

Feature Similarity Index Measure (FSIM) is one standard for image quality evaluation based on Human Visual System (HVS). Its proposal is a new standard inspired by Structure Similarity Index Measure (SSIM). It is a dimensionless quantity with a value between 0 and 1. The larger the value, the better the image restoration quality. The calculation of the feature similarity index needs to consider two factors: one is phase consistency and the other is gradient amplitude. These two elements fuse to characterize the local quality of the image. After obtaining the local quality map, the phase consistency is used again as the weight function to deduce and calculate the final quality of the whole image. The formula of the FSIM:

$$FSIM(x, \hat{x}) = \frac{\sum_{x \in \Omega} S_L(x) \cdot PC_m(x)}{\sum_{x \in \Omega} PC_m(x)},$$  \hspace{1cm} (12)

where $\Omega$ represents the whole image domain, $S_L(x)$ represents the similarity between the original image $x$ and the reconstructed image $\hat{x}$, and $PC_m(x)$ represents the maximum value of phase consistency.

As shown in Table 2, we can see that the FFM method can effectively enhance the image with little distortion and extract the image’s semantic features more comprehensively.

TABLE 2. Comparison between related image enhancement algorithms.

| Dataset         | Image  | CRM [37] | JIEP [36] | FFM  |
|-----------------|--------|----------|----------|------|
| Ours            | Image1 | 0.8555   | 0.9247   | 0.9601|
| RGB-Thermal     | Image3 | 0.8422   | 0.8859   | 0.9462|
| Nighttime-Datasets | Image4| 0.7278   | 0.9039   | 0.9433|
| RGB-Thermal     | Image5 | 0.7819   | 0.8976   | 0.9670|
| Nighttime-Datasets | Image6| 0.7790   | 0.8841   | 0.9990|
|                 | Image2 | 0.8388   | 0.8972   | 0.9548|
|                 | Image7 | 0.8470   | 0.8520   | 0.9446|

B. YOLOv5n METHOD

YOLOv5 series network models include 5 network models of different sizes: s, m, l, x, and n. Among them, the YOLOv5n network model is the latest YOLOv5 series network model [38]. The YOLOv5n network model, on the one hand, has a high detection accuracy and a quick reasoning speed. The weight file of the YOLOv5n network model, on the other hand, is modest, about 75% less than that of YOLOv5s, indicating that YOLOv5n is ideal for deployment to embedded devices for real-time detection. Because the
model’s accuracy, real-time, and lightweight are very important for the accuracy and efficiency of automatic driving lane curvature target detection, this research aims to improve the automatic driving lane curvature detection method based on YOLOv5n. In this study, 15 different targets need to be identified, and the recognition model has high requirements for real-time detection and lightweight performance. Therefore, this study comprehensively considers the accuracy, detection efficiency, and parameters scale of the recognition model and improves the design of the lane curvature target detection method based on YOLOv5n.

YOLOv5n network model is mainly composed of a backbone module and head module. Backbone, as a feature extraction module, is mainly composed of the C3 and the SPPF modules. Specifically, the most important module of the backbone network is the C3 module. This module is improved on the basis of the BottleneckCSP module. Compared with the BottleneckCSP module, it has reduced one convolution layer in the Bottleneck structure, which can reduce the number of parameters and storage space of the method and significantly improve the running speed of the method with a slight loss of detection accuracy.

YOLOv5n replaces the Focus module with a convolution layer, replaces the SPP module with the SPPF module, and changes several super parameters to effectively reduce the number of parameters in the YOLOv5 and develop a lightweight target detection model. After these improvements, the number of parameters and FLOPs of YOLOv5n are greatly reduced with a small loss of accuracy (see Table 3).

**TABLE 3. Comparison between mainstream YOLOv5 detection methods.**

| Model     | mAP@0.5 | mAP@0.5:0.95 | Parameters | FLOPs |
|-----------|---------|--------------|------------|-------|
| YOLOv5n   | 28.4%   | 46.0%        | 1.9MB      | 4.5G  |
| YOLOv5s   | 37.2%   | 56.0%        | 7.2MB      | 16.5G |
| YOLOv5n   | 45.2%   | 63.9%        | 21.2MB     | 49.0G |
| YOLOv5l   | 48.8%   | 67.2%        | 46.5MB     | 109.1G|
| YOLOv5x   | 50.7%   | 68.9%        | 86.7MB     | 205.7G|

**C. SENet MODULE**

In each convolution process, some complex interference information will inevitably be distributed on some channels, resulting in network performance degradation. Attention mechanism has been widely used in neural networks [39]. With the in-depth study of the channel attention mechanism, adjust the channel weight of each channel information, give different weights to each channel information, and screen the channel information according to the weight, which can effectively mitigate the impact of interference information in the complex automatic driving environment. As shown in Figure 3, The typical representative of channel attention is Squeeze-and-Excitation Networks (SENet).

In Figure 3, the input characteristic diagram U, U has C channels, and the space size of each channel is \(H \times W\). The global average pool is performed for each channel, and the calculation formula of channel weight \(Z_C\) is shown in formula:

\[
Z_C = F_{sq}(U) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U(i, j),
\]

where output \(Z_C\) is a one-dimensional array with length \(c\), representing the weight obtained by compressing the channel. The activation function is shown as formula:

\[
S_C = F_{ex}(Z_C, W) = \sigma(W_2 \delta (W_1, Z_C)),
\]

where: the dimension of \(S_C\) is \(1 \times 1 \times C\), and \(c\) corresponds to the generated channel attention weight. The channel attention weights need to be obtained through the previous full connected hierarchy and nonlinear learning; The dimension of \(W_1\) is \(C \times c\), the dimension of \(W_2\) is \(C \times \frac{c}{r}\), and \(r\) is the shrinkage coefficient. The shrinkage layer is composed of two fully connected layers.

Finally, the input channel is weighted and adjusted. The channel attention weighting formula is:

\[
V = S_C \otimes F_{scale} (U, S_C),
\]

where the symbol \(\otimes\) represents element by element multiplication.

**D. MHSA MODULE**

The road traffic environment of automatic driving is very complex and changeable, and there are a lot of occlusions or lack of vision. Multi-Head Self-Attention (MHSA) module can deal with these problems well. As shown in Figure 4, the...
MHSA layer introduces relative position-coding, which considers not only the content information but also the relative distance between elements in different positions, effectively correlates the information and position perception between objects, and improves the detection accuracy of medium and large targets [40]. The following stages are involved in calculating the attention value: To derive the weight, first compute the similarity between the generated attention-related query vector $Q$ and each key vector $K$. Secondly, the obtained weights are normalized by using the softmax function; Finally, the weight and the corresponding value vector $V$ are weighted and summed to obtain the final attention value. MHSA splices several calculated self-attention heads in series, called Multi-heads that can be expressed as:

$$\text{MultiHead}(Q, K, V) = \text{Concat} (\text{head}_1, \text{head}_2, \ldots, \text{head}_h) W^{\text{out}},$$

where the Concat refers to the splice operation, and $W^{\text{out}}$ refers to the matrix for linear transformation. The initial self-attention result achieved by utilizing the scale’s dot product attention is $\text{head}_h$, which is written as:

$$\text{head}_h = \text{Attention}(Q_h, K_h, V_h) = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) V_h,$$

The variance of the dot product of $Q_h$ and $V_h$ alleviates the gradient disappearance problem of softmax. Vectors $Q_h, K_h$ and $V_h$ are given as follows:

$$Q_h = IW_h^Q,$$

$$K_h = IW_h^K,$$

$$V_h = IW_h^V.$$

### E. PROPOSED METHOD

1) SETR-C3 MODULE

The cross-scale connection structure in the C3 module overcomes the gradient disappearance problem of backpropagation and the neural network degradation problem in the process of YOLOv5n training [41], as well as the efficiency of training deep network. However, the contributions of high-level and low-level semantic features in target detection are different [42]. Inspired by the efficientdet method [43], we add a Weighted Feature Fusion (WF) mechanism in the SETR-C3 module. The formula of the WF mechanism can be expressed as follows:

$$O = \sum_i \frac{\omega_i}{\epsilon + \sum_i \omega_i} \cdot I_i,$$

where: $O$ corresponds to the output of the weighted feature fusion mechanism, $\omega$ represents the learnable weight of the semantic feature, $I$ represents the semantic feature, $i$ represents the sequence number of the input edge of this node, and $\epsilon$ is a positive constant to ensure that the denominator of the formula of the weighted feature fusion mechanism is non-zero.

Any C3 module of YOLOv5n can be replaced by the SETR-C3 module proposed by us, which can effectively obtain deeper feature information and rich semantic information. However, in order to introduce the attention mechanism without changing the backbone network so that the pre-trained weight in the public dataset can be used for migration learning and reduce the training time of the network, this design replaces some C3 modules in the YOLOv5n network module with the SETR-C3 module (see Figure 5) based on optimizing the YOLOv5n network model. In order to effectively extract the contour features of the detected lane and obtain the more detailed features of the lane curvature, the SENet module and MHSA module are introduced. The network structure is reasonably adjusted according to the respective focus areas of the SENet module and MHSA module to improve the network recognition accuracy and enhance the ability of network global dependency modeling. At the same time, according to the anchor box size clustering results of the enhanced lane curvature dataset, we simplify the small target detection head in the YOLOv5n network module to fast the running speed with a slight loss of detection accuracy and anti-interference.

2) OPTIMIZATION OF THE MODEL STRUCTURE OF THE SETR-YOLOv5n

The 8000 original images in the dataset used in the SETR-YOLOv5n network model evaluation were intercepted from the video we took while driving in the Shiboyuan tourist area of Shenyang. On this basis, aiming at the problem of the unbalanced number of pictures in each category in the lane

![FIGURE 5. The structure of the SETR-C3.](image1)

![FIGURE 6. The expansion effect of some image datasets.](image2)
curvature dataset, the original 8000 images are expanded by using the enhancement tool, and the images are expanded to more than 10000 by rotating, cutting, random erasing and adjusting the contrast of the images. The expansion effect of some images is shown in Figure 6. 1500 images from 10000 images are randomly selected as the test dataset, 1600 images from 10000 images are randomly selected as the verification set.

![Distribution of anchor box](image1)

(a) Distribution of anchor box (b) Anchor box width distribution and height distribution

**FIGURE 7.** Statistics of lane curvature dataset anchor frame.

The production of the dataset anchor box is closely related to the fitting degree of YOLOv5 network module training, the generalization ability of the network, and the accuracy of prediction results. The anchor box statistics of all marked lane angles are shown in Figure 7.

The structure of the SETR-YOLOv5n network module we designed is shown in Figure 8. According to the statistical clustering results of the center point position, length, and width of all anchor boxes in the dataset, the center point position of the anchor box in the lane curvature dataset shows regional aggregation, and the size of the anchor box is large. To improve the detection efficiency of the SETR-YOLOv5n method, we simplified the target detection head masked with red background according to the characteristics of the lane curvature dataset. In addition, based on the idea of the residual network, we added three connections represented by red lines and introduced the feature information of the shallow layer of the model into the deep layer of the neck network. These connections can enhance the backpropagation ability of the gradient, avoid the disappearance of the gradient and reduce the loss of lane feature information.

**II. DETECTION MODEL TRAINING**

**A. PERFORMANCE METRICS AND EXPERIMENTAL PLATFORM**

The evaluation index system of this experiment includes average accuracy, recall and accuracy, as shown in Equation (22), (23) and (24). The closer the mAP value is to 1, the better the overall performance of the model [44]. There are 15 types of lane curvatures in the dataset used in this study, so the mAP calculation is the average of the fifteen Average Precision (AP) of 15 lane curvatures. Its value is the area surrounded by the recall and accuracy curve.

\[
\text{Precision} = \frac{TP}{TP + FP} \times 100\%, \quad (22)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \times 100\%, \quad (23)
\]

\[
mAP = \frac{\sum_{\hat{C}} \text{AP}({\hat{C}})}{\hat{C}} \times 100\%, \quad (24)
\]

where TP counts the set of lane curvature images that the method can correctly detect, FP counts the set of lane curvature images that the method incorrectly recognizes. FN indicates the counts of lane curvature images that the method does not detect. \(\hat{C}\) indicates the counts of lane line curvature image categories. According to the performance of the computer, the training of SETR-YOLOv5n must take a long time, so the performance of computer hardware has a significant impact on the training time of the SETR-YOLOv5n. The running equipment of this experiment is configured as follows: NVIDIA 3070ti graphics card, AMD 5600x CPU, 32GB Corsair memory.

**B. COMPARISON WITH RELATED METHODS**

In order to further verify the detection performance of the method, SETR-YOLOv5n is compared with the YOLOv5n and several representatives YOLOv5 and their improved methods YOLOv5x, YOLOv5s, Improved YOLOv5m, Improved YOLOv5l, and YOLOv5-Liteg in terms of accuracy and speed. Table 4 and Figure 10 show that the SETR-YOLOv5n detection approach provides the best precision. At the same time, the performance of SETR-YOLOv5n is significantly better than the other object detection methods. Compared with the baseline YOLOv5n, the SETR-YOLOv5n method improves the two metrics of precision and mAP@0.5:0.95 by 6.29% and 15.72%. Overall, our proposed model achieves the best performance among all models and significantly outperforms the lightweight object detection model YOLOv5-Liteg, with a 3.15 FPS improvement, 3.5MB reduction in parameters, and a 20.06% improvement in mAP@0.5:0.95, which reflects the excellent performance of our lightweight method when performing real-time object detection tasks. However, due to the addition of the SETR-C3 module, which increases the calculation load, and the FFM method is used as image preprocessing, the FPS of SETR-YOLOv5n is slightly lower than YOLOv5n.

As shown in Figure 9, several images are selected to compare the YOLOv5n with the improved method in this paper. It can be seen that both methods can effectively detect the lane curvature target. However, the positioning accuracy of YOLOv5n for the lane curvature target with sparse features, low visibility, and unclear targets is poor, which is easy to cause missed detection. The lane curvature target is easy to be affected by the noise in the image, resulting in a high false alarm rate. The improved method reduces the missed detection rate of lane curvature and improves the detection accuracy. The reason is that the FFM is introduced to
FIGURE 8. The network module of the SETR-YOLOv5n. Compared to the original YOLOv5n, there are three improvements in the architecture. First, the detection head with red background is simplified, the model parameters are reduced, and the detection efficiency is improved. Second, it combines shallow semantic features with deep semantic features, and the FFM is introduced to preprocess the dataset pictures. Third, the C3 module is improved to improve its ability for feature extraction.

FIGURE 9. Comparison of experimental results.

TABLE 4. Comparison between related detection algorithms.

| Model                  | Precision | Recall | mAP@0.5 | mAP@0.5:0.95 | FPS   | Parameters | FLOPs |
|------------------------|-----------|--------|---------|---------------|-------|------------|-------|
| YOLOv5 Liteg           | 92.18%    | 94.62% | 96.93%  | 67.97%        | 67.28 | 5.3MB      | 15.7G |
| YOLOv5s                | 95.22%    | 95.68% | 98.50%  | 76.09%        | 76.92 | 7.3MB      | 17.0G |
| Improved YOLOv5m [2]   | 90.84%    | 99.50% | 99.15%  | 68.64%        | 60.21 | 21.3MB     | 59.3G |
| Improved YOLOv5l [2]   | 91.19%    | 99.61% | 99.17%  | 69.62%        | 51.86 | 47.2MB     | 109.7G|
| YOLOv5s                | 99.47%    | 98.89% | 99.47%  | 90.73%        | 31.45 | 87.7MB     | 218.8G|
| YOLOv5n                | 93.13%    | 97.74% | 97.66%  | 72.31%        | 71.41 | 1.8MB      | 4.7G  |
| SETR-YOLOv5n           | 99.42%    | 98.90% | 99.45%  | 88.03%        | 70.43 | 1.8MB      | 4.3G  |

preprocess the detected image, which improves the visibility of the image under low-light conditions. Meanwhile, SETR-YOLOv5n improves the feature extraction module, which can extract more comprehensive lane curvature target features to locate the lane curvature target accurately.

IV. DISCUSSION
A. OPTIMIZATION OF THE TARGET DETECTION HEAD
In this experiment, the relationship between the performance of the YOLOv5n network model and the optimization of the target detection head is analyzed. Optimizing the target detection head plays an essential role in the YOLOv5n algorithm. However, different parts of the optimization of the target detection head have different effects on the effect of YOLOv5n target detection. In order to further analyze the performance of the SETR-YOLOv5n target detection algorithm proposed in this paper, we performed ablation experiments on the lane curvature dataset. We set the same parameters for each variable in the experiment to ensure fairness. Moreover, the YOLOv5n network module, the YOLOv5n network module with small target detection head deleted (YOLOv5ntypeB), and the YOLOv5n network
module with small and medium target detection head deleted (YOLOv5ntypeC) are trained and verified on the lane curvature dataset, including parameters, FLOPs, precision, recall, mAP@0.5, and mAP@0.5:0.95 are the evaluation indicators. The lane curvature dataset is used for training by observing the four index curves of Precision, Recall, mAP@0.5, and mAP@0.5:0.95 of the model in the training process, the relationship between the performance of the YOLOv5n model and the target detection head is analyzed, as shown in Figure 11. When training on the dataset, when the model training is iterated to 500 steps, The YOLO5n network with the small target detection head removed (YOLOv5ntypeB) is ahead of the other two network models in the above four indicators.

According to the performance comparison and analysis of the above models, the number of FLOPs and parameters of YOLOv5ntypeB and YOLOv5ntypeC network models are less than YOLOv5n, which will save a lot of memory space. Although YOLOv5ntypeB has slightly more parameters than YOLOv5ntypeC, its precision, recall, mAP@0.5, and mAP@0.5:0.95 are higher than YOLOv5ntypeC. It can be concluded that the performance of the YOLOv5ntypeB network model is better than that of the YOLOv5n and YOLOv5ntypeC network models.

To evaluate the network performance, refer to the parameters and FLOPs in the training process. The test results of the three models are shown in Table 5. The test results show that the parameters of YOLOv5n are relatively large, as high as

---

**TABLE 5. Comparison between related detection algorithms.**

| Model            | Precision | Recall | mAP@0.5 | mAP@0.5:0.95 | FPS    | Parameters | FLOPs |
|------------------|-----------|--------|---------|--------------|--------|------------|-------|
| YOLOv5n          | 93.13%    | 95.74% | 97.66%  | 72.31%       | 67.28  | 1.8MB      | 4.7G  |
| YOLOv5ntypeB     | 98.73%    | 98.82% | 99.40%  | 84.46%       | 76.92  | 1.78MB     | 4.2G  |
| YOLOv5ntypeC     | 97.42%    | 98.51% | 99.33%  | 82.96%       | 31.45  | 1.77MB     | 4.2G  |

---

**FIGURE 10.** Comparison of four evaluation index curves of four YOLOV5n methods during trainings.
1.8MB, while the parameters of YOLOv5ntypeB are 1.78MB and the parameters of YOLOv5ntypeC are 1.77MB. For the FLOPs operations, the FLOPs of YOLOv5ntypeB and YOLOv5ntypeC are 10.64% less than YOLOv5n. The above analysis results show that because the lane belongs to large targets, when using the lane curvature data set for testing, the utilization rate of the small target detection head in the SETR-YOLOv5n target detection model is significantly lower than that of other target detection heads.

B. ON THE IMPROVEMENT OF MODEL STRUCTURE

In order to verify the impact of each improved method on the YOLOv5n target detection method, we evaluated each method on the lane curvature dataset data set and designed seven groups of ablation experiments. Using consistent experimental settings, the impact of different methods on the YOLOv5n target detection method is shown in Table 6. Among them, model A integrates the SENet module on the basis of YOLOv5n, model B integrates the MHSA module on
the basis of model A, model C integrates the new connections mechanism on the basis of model B, model D integrates the weight feature fusion mechanism on the basis of model B, model E integrates the new connections mechanism on the basis of model D, and SETR-YOLOv5n adds FFM to preprocess the lane curvature image to be detected on the basis of model E, where “✓” indicates the use of corresponding methods. It can be seen from Table 6 that gradually using five optimization methods will continuously improve the average accuracy, and the average accuracy of the finally improved model has increased by 14.72%. Model A, integrating the SENet module can significantly improve the effect of YOLOv5n target detection. Model A can adjust the channel weight of each channel information, give different weights to each channel information, and filter the channel information according to the weight, which can effectively alleviate the influence of interference information in the complex autonomous driving environment, and significantly improve the recognition success rate of target detection under low-light conditions. The above analysis results show that the light in the driving scenes mainly comes from natural light, and driving scenes often change. So there may reach existence low-light level in driving scenes, which reduces the ability of the network to extract features. In this paper, we use the FFM algorithm to improve the problem of reducing feature extraction ability caused by low-light levels through image enhancement and get good results.

In addition, the camera moves with the vehicle and is accompanied by shaking, so some images are blurred, and the image background changes quickly. MHSA and SENet modules are introduced into the method of this paper to enhance the target contour and texture feature extraction ability. Furthermore, with the vehicle’s movement and the switching of the angle of view, the shape and size of the lane are also changing. Introducing a new connections mechanism enhances the target detection model’s feature extraction and representation ability. The SETR-C3 module of the method in this paper considers more comprehensively the channel attention information and the global self-attention calculation, which obviously improves the detection accuracy. Meanwhile, it can be seen that using the FFM to process the image to be detected can effectively improve detection accuracy. Thus, the target detection ability of the method in this paper on the low-light lane image captured by the camera is improved. The above comparison further verifies the superiority of the method in this paper.

![FIGURE 12. The architecture of the Actor-Critic frame.](image)

C. APPLICATION IN AUTOMATIC DRIVING

Deep reinforcement learning is a combination of deep learning and reinforcement learning. It uses the perception ability of deep learning to solve the modeling problem of strategy function and value function and then uses error backpropagation algorithm to optimize the objective function. At the same time, it uses the decision-making ability of reinforcement learning to define problems and optimize objectives, which can be controlled directly according to the input image. It is an artificial intelligence method closer to human thinking mode.

In the practical application of autonomous driving systems, the deep reinforcement learning method is widely used by researchers, such as the method based on the Actor-Critic frame (see Figure 12). This method can make driving behavior decisions according to the driving environment. However, due to the high training requirements of deep reinforcement learning, the depth of environment perception neural network of deep reinforcement learning cannot be designed too deep, resulting in the limited driving environment state that the reinforcement learning method can perceive under low-light conditions, and it is unable to make driving behavior decisions usually. Currently, the target detection method has been widely used as an auxiliary reinforcement learning task [45], [46], [47]. The SETR-YOLOv5n target detection method proposed in this paper can quickly detect the lane’s position and the lane’s bending angle under low-light conditions, mark it in the video frame and transmit the video frame to the deep reinforcement learning algorithm as an autonomous driving environment. The deep reinforcement learning algorithm can output the corresponding autonomous driving decision according to the video frame output by the SETR-YOLOv5n target detection method.

V. CONCLUSION

Aiming at YOLOv5’s problems of complex structure, too many params, high hardware configuration required for training, the low image signal-to-noise ratio of received images under low-visibility conditions and low FPS, combined with the characteristics of strong regional lane curvature and sparse target distribution in the image of lane curvature dataset, an improved lightweight target detection method SETR-YOLOv5n is proposed. Compared with traditional non-deep learning methods, the target detection method based on deep learning proposed in this paper can extract more abundant target features and improve the efficiency of target detection. Firstly, the FFM for low-light image enhancement is introduced to improve the quality of
images. Secondly, the method optimizes the configuration of the target detection head and proposes a SETR-C3 module embedded with two attention mechanisms to replace some C3 modules to optimize the effect of feature extraction. Such a lightweight model is easier to deploy in the mobile terminal or embedded terminal equipment and has particular practical significance in automatic driving.

The method proposed in this paper can be applied in many fields, such as under the complex war environment, it can help soldiers during the progress of vehicles or be used as a component module of household robot systems. Meanwhile, the ability of automatic driving to defend against attacker attacks and external disturbance should be continuously enhanced to facilitate the promotion and popularization of automatic driving. Future work may adjust the previously proposed model according to the specific application and combine it with a reinforcement learning algorithm to improve the practical application effect of the previously proposed model.

REFERENCES
[1] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, “A survey of deep learning techniques for autonomous driving,” J. Field Robot., vol. 37, no. 3, pp. 362–386, 2020.
[2] C. Ye, Y. Wang, Y. Wang, and M. Tie, “Steering angle prediction YOLOv5-based end-to-end adaptive neural network control for autonomous vehicles,” Proc. Inst. Mech. Eng., J. D. Automobile Eng., vol. 236, no. 9, pp. 1991–2011, 2021.
[3] W. Song, Y. Yang, M. Fu, Y. Li, and M. Wang, “Lane detection and classification for forward collision warning system based on stereo vision,” IEEE Sensors J., vol. 18, no. 12, pp. 5151–5163, Jun. 2018.
[4] W. Fang, L. Wang, and P. Ren, “Tinier-YOLO: A real-time object detection method for constrained environments,” IEEE Access, vol. 8, pp. 1935–1944, 2020.
[5] G. Li, Y. Yang, X. Qu, D. Cao, and K. Li, “A deep learning based image enhancement approach for autonomous driving at night,” Knowl.-Based Syst., vol. 213, Feb. 2021, Art. no. 106617.
[6] A. M. Reza, “Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement,” J. VLSI Signal Process. Syst. Signal, Image Video Technol., vol. 38, no. 1, pp. 35–44, 2004.
[7] H. D. Cheng and X. J. Shi, “A simple and effective histogram equalization approach to image enhancement,” Digit. Signal Process., vol. 14, no. 2, pp. 158–170, Mar. 2004.
[8] T. K. Kim, J. K. Paik, and B. S. Kang, “Contrast enhancement system using spatially adaptive histogram equalization with temporal filtering,” IEEE Trans. Consumer Electron., vol. 44, no. 1, pp. 82–87, Feb. 1998.
[9] H. Ibrahim and N. S. P. Kong, “Brightness preserving dynamic histogram equalization for image contrast enhancement,” IEEE Trans. Consumer Electron., vol. 53, no. 4, pp. 1752–1758, Nov. 2007.
[10] Q. Wang and R. K. Tan, “Fast image/video contrast enhancement based on weighted thresholded histogram equalization,” IEEE Trans. Consumer Electron., vol. 53, no. 2, pp. 757–764, May 2007.
[11] Y. Li and H. Zhang, “Modified clipped histogram equalization for contrast enhancement,” in Proc. 13th Int. Conf. Parallel Distrib. Comput., Appl. Technol., Dec. 2012, pp. 653–658.
[12] S. Peddar, S. Tewary, D. Sharma, V. Karar, A. Ghosh, and S. K. Pal, “Non-parametric modified histogram equalisation for contrast enhancement,” IET Image Process., vol. 7, no. 7, pp. 641–652, Oct. 2013.
[13] S.-C. Huang and C.-H. Yeh, “Image contrast enhancement for preserving mean brightness without losing image features,” Eng. Appl. Artif. Intell., vol. 26, nos. 5–6, pp. 1487–1492, May/Jun. 2013.
[14] A. Foi, M. Trimeche, V. Katkovnik, and K. Egiazarian, “Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data,” IEEE Trans. Image Process., vol. 17, no. 10, pp. 1737–1754, Oct. 2008.
[39] Y. Li, J. Zeng, S. Shan, and X. Chen, “Occlusion aware facial expression recognition using CNN with attention mechanism,” *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[40] M. Zhu, Y. Tang, and K. Han, “Vision transformer pruning,” 2021, arXiv:2104.08500.

[41] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[42] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen, “Efficient and accurate arbitrary-shaped text detection with pixel aggregation network,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8440–8449.

[43] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[44] S. Visa, B. Ramsay, A. Ralescu, and E. Van Der Knaap, “Confusion matrix-based feature selection,” in *Proc. CEUR Workshop*, vol. 710, 2011, pp. 120–127.

[45] M.-X. Jiang, C. Deng, Z.-G. Pan, L.-F. Wang, and X. Sun, “Multiobject tracking in videos based on LSTM and deep reinforcement learning,” *Complexity*, vol. 2018, pp. 1–12, Nov. 2018.

[46] M. Jiang, T. Hai, Z. Pan, H. Wang, Y. Jia, and C. Deng, “Multi-agent deep reinforcement learning for multi-object tracker,” *IEEE Access*, vol. 7, pp. 32400–32407, 2019.

[47] X. Tang, J. Chen, K. Yang, M. Toyoda, T. Liu, and X. Hu, “Visual detection and deep reinforcement learning-based car following and energy management for hybrid electric vehicles,” *IEEE Trans. Transport. Electrific.*, vol. 8, no. 2, pp. 2501–2515, Jun. 2022.

**YANG LIU** was born in Liaoning, China, in 1993. He received the B.E. degree in measurement and control technology and instrumentation program control from Shenyang Aerospace University, Shenyang, China, in 2016, and the M.Sc. degree in applied mathematics from Liaoning Technical University, Fuxin, China, in 2020, where he is currently pursuing the Ph.D. degree with the School of Mechanical Engineering and Automation, Northeastern University, Shenyang, China. His current research interests include deep learning and reinforcement learning.

**YONGFU WANG** received the Ph.D. degree in control science and engineering from Northeastern University, Shenyang, China, in 2005. He is currently a Professor with the School of Mechanical Engineering and Automation, Northeastern University. His research interests include unmanned vehicles, intelligent modeling, and control of mechatronic systems.

**QIANGSHENG LI** was with the Dalian Power Plant, Huaneng Power International, Inc., for 23 years, where he is currently an Engineer. His research interests include modeling, optimization, and control for complex industrial processes, intelligent control, and reinforcement learning.

**JIAN WANG** is currently an Engineer with the Dalian Power Plant, Huaneng Power International, Inc. His research interest includes modeling, optimization, and control for complex industrial processes.

---

**YANG LI** was with the Dalian Power Plant, Huaneng Power International, Inc., for 26 years, where he is currently an Engineer. His research interests include modeling, optimization, and control for complex industrial processes, intelligent control, and deep learning.

**QIANGSHENG LI** was with the Dalian Power Plant, Huaneng Power International, Inc., for 23 years, where he is currently an Engineer. His research interests include modeling, optimization, and control for complex industrial processes, intelligent control, and reinforcement learning.

**JIAN WANG** is currently an Engineer with the Dalian Power Plant, Huaneng Power International, Inc. His research interest includes modeling, optimization, and control for complex industrial processes.