forests

MDPI

Article

SWVR: A Lightweight Deep Learning Algorithm for Forest Fire Detection and Recognition

Li Jin [1], Yanqi Yu [1], Jianing Zhou [1], Di Bai [2,*], Haifeng Lin [1,*] and Hongping Zhou [1]

[1] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China
[2] College of Information Management, Nanjing Agricultural University, Nanjing 210037, China
* Correspondence: baidi000@njau.edu.cn (D.B.); haifeng.lin@njfu.edu.cn (H.L.); Tel.: +86-25-8542-7827 (H.L.)

Abstract: The timely and effective detection of forest fires is crucial for environmental and socio-economic protection. Existing deep learning models struggle to balance accuracy and a lightweight design. We introduce SWVR, a new lightweight deep learning algorithm. Utilizing the Reparameterization Vision Transformer (RepViT) and Simple Parameter-Free Attention Module (SimAM), SWVR efficiently extracts fire-related features with reduced computational complexity. It features a bi-directional fusion network combining top-down and bottom-up approaches, incorporates lightweight Ghost Shuffle Convolution (GSConv), and uses the Wise Intersection over Union (WIoU) loss function. SWVR achieves 79.6% accuracy in detecting forest fires, which is a 5.9% improvement over the baseline, and operates at 42.7 frames per second. It also reduces the model parameters by 11.8% and the computational cost by 36.5%. Our results demonstrate SWVR's effectiveness in achieving high accuracy with fewer computational resources, offering practical value for forest fire detection.

Keywords: forest fire detection; WIOU; RepViTBlock; SimAM; VoVGSCSP

Citation: Jin, L.; Yu, Y.; Zhou, J.; Bai, D.; Lin, H.; Zhou, H. SWVR: A Lightweight Deep Learning Algorithm for Forest Fire Detection and Recognition. *Forests* **2024**, *15*, 204. https://doi.org/10.3390/f15010204

Academic Editor: José Aranha

Received: 8 December 2023
Revised: 16 January 2024
Accepted: 18 January 2024
Published: 19 January 2024

1. Introduction

One of nature's most precious resources, forests are essential to ecological processes including water conservation, pollution control, climate regulation, and wildlife protection [1]. On the other hand, forests provide human beings with a variety of resources, such as timber, economic plants, food, etc., which have great economic value [2]. The serious effects of forest fires include soil erosion, the destruction of vegetation resources, and the degradation of air quality, affecting the living environment and economy of human beings and even threatening people's lives, safety, and health [3]. The early stages of a fire, marked by the initiation of smoke and flames, are critical for fire warnings [4]. Therefore, the early detection of forest fires has become an urgent necessity.

Traditional methods of detecting forest fires typically involve manual detection, lookout tower observations, sensor detection, satellite remote sensing, and computer vision. Manual detection and watchtower detection require a significant amount of manpower when observing surrounding fires and can be constrained by space and time, and, most importantly, they can pose a threat to personal safety. In addition, the detection results often fall short of expectations. Sensors have a limited detection range and are susceptible to environmental noise, making it challenging to accurately detect forest fires [5]. Moreover, sensor detection comes with high maintenance costs and is prone to issues such as power shortages and communication interruptions. Satellite remote sensing is commonly used for macroscopic monitoring, but the image pixels representing forest fires are typically limited in number and size, posing a challenge for accurate detection and making it difficult to detect localized areas and early-stage small fires in a timely manner. It is also often hindered by issues like orbital paths and cloud cover, limiting its ability to meet the time and accuracy requirements of fire detection. Thanks to the rapid development of electronic and imaging technologies, computer vision-based techniques for forest fire detection have

gained significant popularity and widespread application in recent years. We primarily use optical and thermal imaging sensors. These sensors are integral to the identification and analysis of fire characteristics in various environments. These methods typically rely on the color characteristics, geometric shapes, and movement trajectories of flames or smoke for fire detection [6].

Prema et al. [7] carried out color filtering in the YUV color space to offer an early forest fire detection technique, utilizing an SVM classifier to train on extracted spatiotemporal features such as wavelet energy. They analyzed multiple features of fires for early forest fire detection. However, the effectiveness in detecting smoke color regions was not satisfactory. Calderara et al. [8] employed a novel Bayesian method to detect fire regions in scenes. They applied wavelet transform coefficients and color information on these images to construct a statistical model of image energy decay using a temporal Gaussian mixture. This clarification specifies the type of imagery, namely RGB, used in their method. By employing a color blending approach that combines reference fire colors with input frames, this model successfully accomplishes swift fire detection while minimizing the occurrence of false alarms. Yang et al. [9], considering prior knowledge of forest fire issues, introduced L0 constraints into the fire class and proposed a kernel-based L1 regularized PreVM fire detection method, exhibiting reduced error warning rates and increased fire detection rates. In summary, traditional image-based detection methods have made progress, primarily relying on extracting image features and training classifiers for fire detection. However, they exhibit poor resistance to interference, with shadows, occlusions, and lighting changes negatively impacting the effectiveness of fire detection.

Over the past few years, with the rapid advancement of artificial intelligence, deep learning has made remarkable strides in the field of forest fire detection. When compared to traditional methods, deep learning-based algorithms for forest fire detection are able to automatically acquire a deeper understanding of the advanced and intricate features in image and video data, with good robustness and accuracy and significantly reduced labor costs, and are widely used in a variety of complex scenarios. Convolutional neural networks (CNNs) are powerful deep learning models that excel in automatically extracting and learning relevant object features through the training of large volumes of image data. They have found extensive application in fire detection tasks.

In order to detect fires early, Huang et al. [10] presented the Wavelet-CNN approach, which combines a CNN with spectrum analysis. Using this technique, spectral characteristics are fed into several CNN layers, improving the detection with backbone networks MobileNet v2 (MV2) and ResNet50. Barmpoutis et al. [11] utilized the Faster R-CNN model to identify candidate fire regions, employing multidimensional texture analysis to enhance the detection accuracy. However, this approach increases the computational complexity.

Muhammad et al. [12] introduced a framework based on a CNN and IoMT for early fire detection, using AlexNet as the baseline architecture. This method focuses on adjustments to the fully connected layer to suit specific datasets, aiming to reduce false fire alarms. However, challenges remain in terms of eliminating these false alarms and improving the model accuracy in diverse natural environments [13].

Further advancements have been made by Abdusalomov et al. [14] and Reis et al. [15], concentrating on long-range small fire detection and training forest fire images acquired by UAVs, respectively. Rostami et al. [16] developed the MultiScale-Net network for active fire detection under challenging geographic and lighting conditions.

Despite these advancements, the complexity of these models often hinders their practical deployment. Efforts to create lightweight models, as seen in the works of Li et al. [17], Lv et al. [18], and Huang et al. [19], focus on reducing the model parameters and improving the feature extraction capabilities. Zhang et al. [20] achieved this by integrating Channel Shuffle into SqueezeNet's fire module.

However, there are limitations in comprehensive fire detection, and weather conditions such as fog and snow can represent an obstacle to network recognition.

To summarize, deep learning-based systems for the detection of fires have achieved notable success. However, when deploying these algorithms for forest fire detection on edge devices, our research team conducted an extensive analysis of recent studies and real-world applications, focusing on the performance of existing models in diverse environments, and found that several challenges persist:

1. Quantized models achieve high accuracy in forest fire detection, but they typically have a large number of variables, demanding large storage and computational resources, making it challenging to deploy them on low-power platforms with limited resources, such as drones and remote sensing devices.
2. There are strict real-time criteria for forest fire detection, and it is challenging to integrate the weighted model across different devices, it is unable to detect fires quickly, and the model is less flexible.
3. Lightweight models with smaller sizes often offer faster inference speeds. However, they may struggle to capture complex features in images, such as those with leaves, cloud cover, or other obstructions, and further improvements are needed to enhance the accuracy.

Therefore, the foci and challenges in detecting forest fires include achieving lightweight models while ensuring the detection accuracy of the model. In order to tackle the aforementioned obstacles, this paper suggests a forest fire detection model called SWVR.

It can be applied for real-time forest fire monitoring when installed on fire towers or used for fire intensity observation when integrated into firefighting cameras.

The key contributions of this study are as follows.

1. In order to lessen the computational complexity of the model, we create a new feature extraction network. It combines the efficient architecture of the lightweight Vision Transformer (ViT) with the traditional lightweight convolutional neural network (CNN) by using consecutive CBS modules and stacked RepViTBlock modules. Therefore, it outperforms common modules like ElAN and C3 in terms of its lightweight design, improving the efficiency and efficiency of the model's detection.
2. To address the issue of insufficient feature extraction for small forest fire images, in this paper, the feature extraction network incorporates the SimAM technique. The SimAM mechanism directly derives multidimensional weights for the feature maps based on an energy function. This increases the network's capacity for feature extraction, which raises the detection accuracy of the model. Importantly, this improvement is achieved without the need for additional parameters, effectively enhancing the model's rate of inference.
3. In order to address the challenge of large model sizes that make it difficult to deploy them on edge devices, this paper adopts a bi-directional fusion backbone network that mixes top-down and bottom-up approaches. The main backbone network is connected with the feature fusion network using GSConv, enabling cross-layer connections. Additionally, the VOVGSPSP module is utilized to effectively combine characteristics at different levels under the constraint of smaller parameters and computational requirements. This promotes information flow within the model and drastically lowers the quantity of model parameters. As a result, the suggested strategy offers a compromise between model size and performance, increasing the deployment's viability on edge devices.
4. To address the issue of low-quality samples in the forest fire detection dataset, which can impact the model's generalization performance, this paper makes use of the Wise Intersection over Union (WIoU) loss function during training. The WIoU loss function introduces a non-monotonic dynamic focus mechanism by estimating the outlierness of anchor boxes. This mechanism allows the model to possess higher tolerance for low-quality examples, thereby improving the model's overall capacity for generalization. By incorporating the WIoU loss function, the model becomes more robust to variations and noise in the dataset, leading to improved performance in real-world situations.

This paper's remaining sections are arranged as follows. In Section 2, the SWVR forest fire detection model is introduced. Section 3 discusses the dataset utilized in the experiments and the assessment metrics for model performance are described. Section 4 outlines the configuration and arrangement of the experiment, including the platform used and the training parameters. It also validates the capabilities of the RepViTBlock module, SimAM attention module, VoVGSCSP module, and WIoU loss function in recognizing forest fires. In Section 5, we investigate and discuss the conclusions of the experiment. In Section 6, we summarize the full text.

2. Materials and Methods

2.1. Dataset

The effective evaluation of model performance in fire detection in forests depends on a high-quality dataset [21]. The data for this study were collected over a period of six months and carefully selected and annotated. Firstly, we used web crawlers to obtain images of wildfires in different situations; secondly, we used sensors, cameras, and other devices to collect real-time data. By combining data from these two sources, we created a rich dataset of forest fires. We fused image data from multiple sources to create a homemade dataset containing 7061 forest fire images based on different perspectives of the forest background (Table 1).

Table 1. Quantity of pictures in each set.

Dataset	Train	Valid	Test	Sum
Quantity	5648	706	707	7061

Our dataset includes various types of fire incidents, such as ground fires caused by combustible materials like dry leaves and twigs. These ground fires typically have a lower flame height, smaller burning area, and fewer pixels and are challenging to detect. Additionally, there are vertical fires that spread along the tree trunk into the canopy. Furthermore, we have examples of surface fires caused by grassland, shrubs, and low-height trees, which have the characteristic of rapid fire spreading and require real-time detection. Different perspectives and distances were considered when capturing these fire images, predominantly consisting of RGB imagery. Various examples of fire incidents, depicted in Figure 1, are captured in this format, providing a clear representation of the fires in natural color spectra.

We utilized labeling software (Labelimg, v1.8.6) to annotate the images, including labels with information such as the coordinates, length, and height. Once the annotation process was completed, we converted the XML label format to the TXT format, resulting in the final dataset. To ensure the random and unbiased division of the dataset into training, validation, and test sets, we utilized a simple random sampling method. This process involved using a random number generator to select images for each set without any specific criteria, thus ensuring that each image had an equal likelihood of being chosen for any of the sets. The dataset was subsequently divided into training, validation, and test sets, with a distribution ratio of 8:1:1. This method helped in maintaining the randomness and diversity of the dataset, crucial for the robustness of the model.

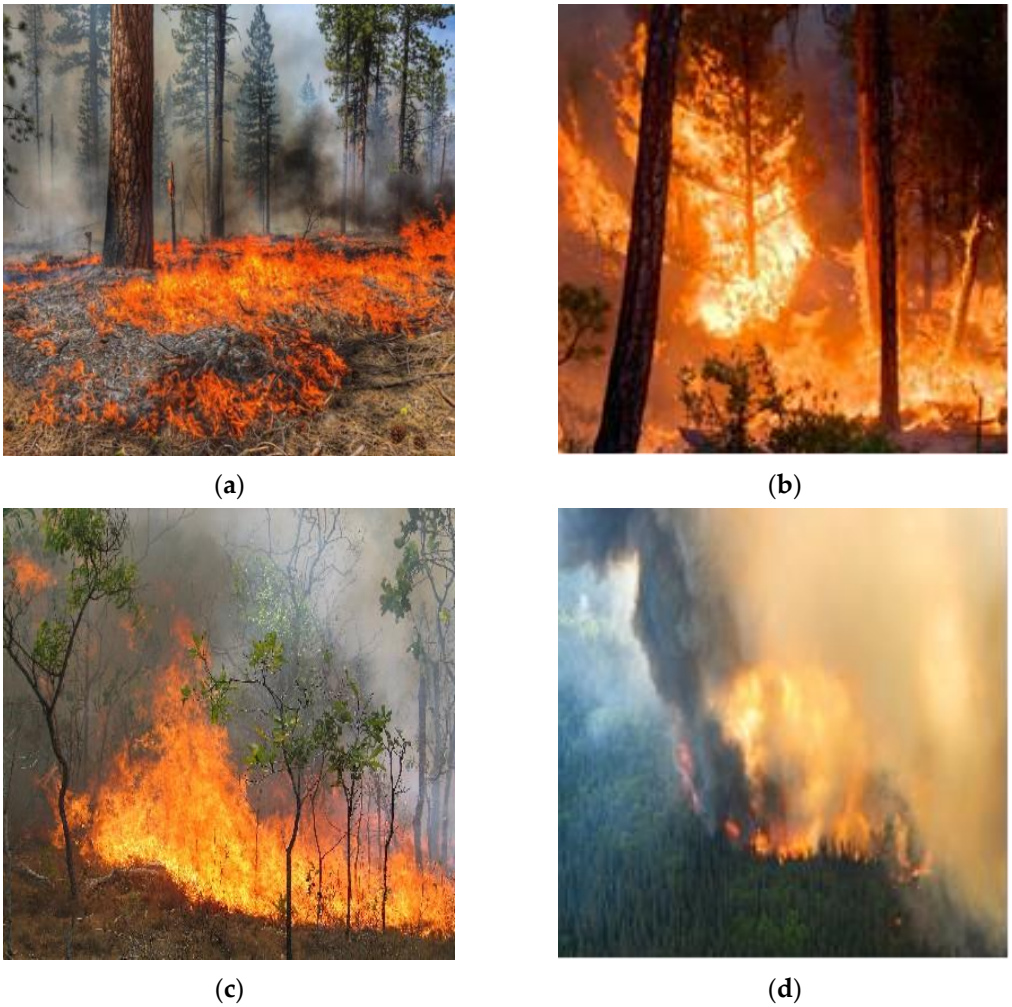

Figure 1. Typical forest fire pictures from the dataset: (**a**) subterranean fire; (**b**) vertical fire; (**c**) surface fire; (**d**) long-range capture of forest fire.

2.2. Reparameterization Vision Transformer Block (RepViTBlock)

Considering the real-world implementation of our model, especially in vast and complex forest regions, the efficient processing of a substantial quantity of fire photographic information is imperative. A low-power consumption model is crucial for devices powered by batteries, such as drones, as it can effectively prolong the flight time and provide extended detection capabilities. To further enhance the lightweight design of the model, this paper utilizes the RepViTBlock module [22] in the backbone network.

Traditional lightweight networks such as AlexNet [23], VGGNet [24], MobileNet [25], and SqueezeNet [26] employ techniques like reducing the kernel size, depthwise separable convolution, compression algorithms, and channel shuffling to reduce the computational complexity and parameter count. They aim to achieve real-time inference on devices with limited computational resources. However, these networks often face challenges in balancing a lightweight design and accuracy and may not provide sufficient performance in handling complex tasks such as object detection and semantic segmentation.

In contrast, RepViT is an innovative lightweight network that combines the efficient architecture of the lightweight Vision Transformer with traditional lightweight CNNs, primarily based on the RepViTBlock module. This integration enhances the model's performance and efficiency on mobile devices.

The effectiveness of ViT primarily stems from its universal token mixer and channel mixer architecture, known as the MetaFormer architecture. Building upon this, we dissected the block structure of a standard lightweight MobileNetV3-L model to obtain the

RepViTBlock module. In the architecture of our model, we have strategically repositioned the deep convolutional layers to a higher level in the network. This upward shift allows for the earlier processing of deep convolutional features. In parallel, we have moved the position of the squeeze and excitation (SE) attention mechanism, placing it in front of the deep filters. This adjustment improves the model's ability to facilitate the interaction of spatial information. This allows the SE mechanism to more effectively weigh the importance of different channels in the feature maps, leading to improved spatial feature representation and recognition performance in the context of forest fire detection. Additionally, to further enhance the model's performance, we have used structural reparameterization to introduce a multi-branch topology for the deep filters during training. Through these steps, the token mixer and channel mixer were successfully separated, resulting in the RepViTBlock module. During the inference process, the multi-branch topology of the token mixer was consolidated into a single deep convolution, effectively reducing the computational and memory costs, decreasing the latency, and thus improving the speed of model inference, particularly advantageous for mobile devices such as drones. The proposed RepViTBlock structure is depicted in Figure 2.

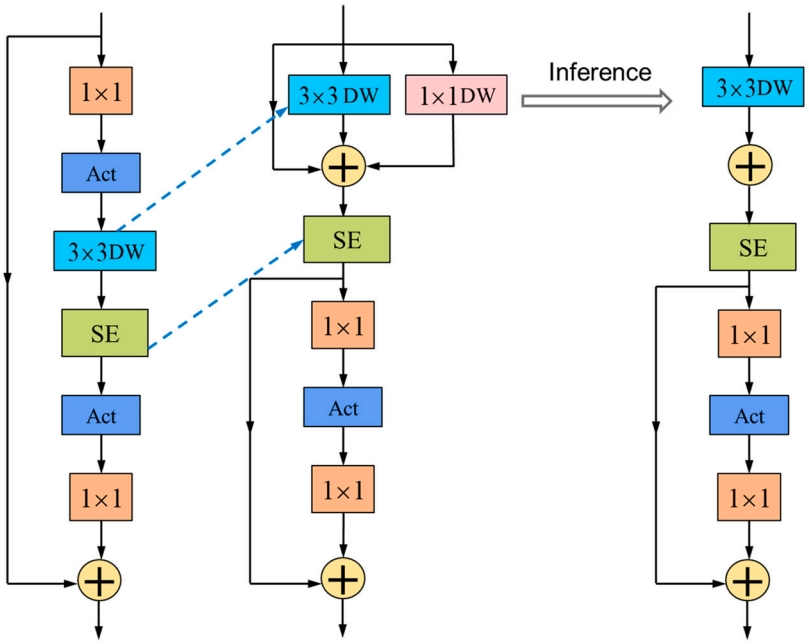

Figure 2. Structure of the RepViTBlock module.

Next, we mainly optimized the network structure from four aspects, the stem, downsampling layers, classifier, and overall stage proportions, aiming to substantially enhance the model's performance.

1. Convolutional extractor for shallow networks: In the shallow networks, we used early-stage convolutions as the stem to address the issues of optimization sensitivity and training recipe sensitivity.
2. Employing a separate and deeper downsampling layer: By adjusting the channel dimension using a 1×1 convolution, we formed a feedforward network by connecting the inputs and outputs of two 1×1 convolutions through a residual connection. We also added a RepViTBlock module to further deepen the downsampling layer, reducing the information loss resulting from the resolution reduction.
3. Utilizing a simpler classifier: Compared to MobileNetV3-L, we employed only a global average pooling layer and a linear layer as the classifier, reducing the latency.
4. Optimization of overall stage proportions: Based on empirical observations, we divided the ratios of block quantities in the model's four stages as 1:1:7:1, achieving a deeper layout.

The proposed network structure of RepViT based on the RepViTBlock module is illustrated in Figure 3 below.

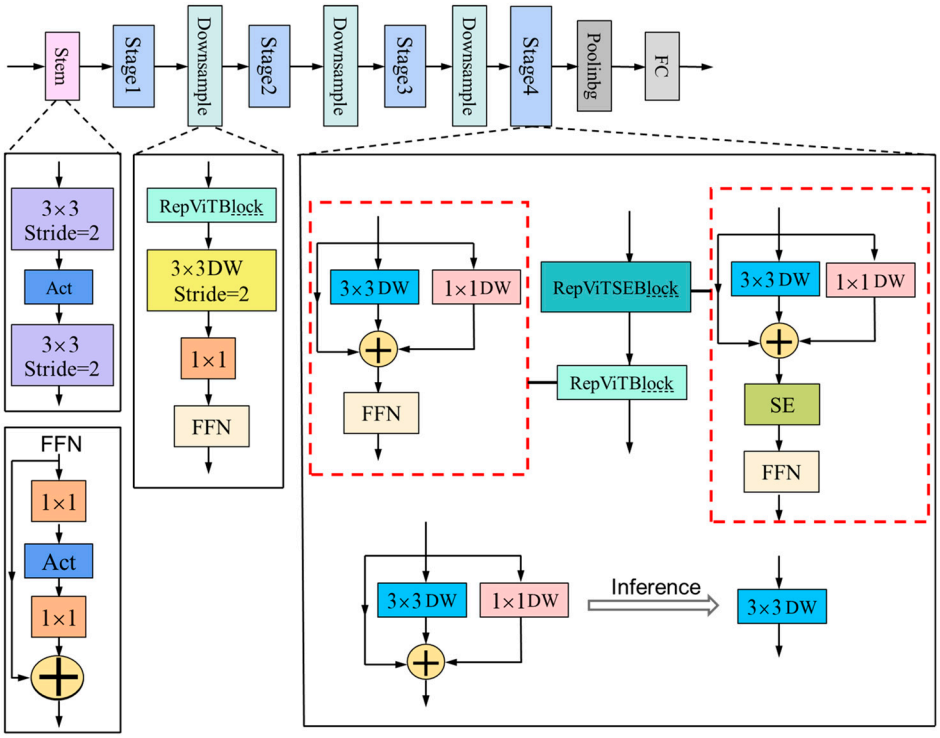

Figure 3. RepViT network structure.

We noticed that the RepViT network architecture primarily relied on the optimized RepViTBlock module, which has demonstrated impressive performance. Therefore, in this article, the RepViTBlock module is integrated into the backbone network. It combines the efficient architecture of the lightweight ViT with a traditional lightweight CNN, resulting in a lighter module that effectively reduces the computational complexity. This reduction in complexity is beneficial in reducing the energy consumption and enabling real-time monitoring and control to promptly detect and suppress forest fires.

2.3. Simple Parameter-Free Attention Mechanism (SimAM)

In forest environments, factors such as fog, snow, and varying lighting conditions can introduce significant noise and interference, leading to reduced accuracy in fire detection. To effectively mitigate the impact of these irrelevant factors and improve the real-time detection, this paper integrates the Simple Parameter-Free Attention Mechanism (SimAM) [27] into the network for feature extraction. Traditional channel attention mechanisms and spatial attention mechanisms can only differentiate and refine features in a single dimension, while treating other dimensions equally. They generate one-dimensional weights for channels and two-dimensional weights for spatial dimensions based on feature maps, and then expand them to form an attention feature map. However, this approach limits the feature extraction capability of the network.

Compared to traditional attention mechanisms, SimAM is a simple yet highly effective lightweight attention module that treats all dimensions equally. It is capable of deriving three-dimensional weights for the mapping of features, with no additional parameters required. SimAM considers the correlation between spatial and channel factors without augmenting the original network with additional parameters. It improves the capability of the network to represent features effectively.

In neuroscience, neurons that exhibit different firing patterns from their surrounding neurons are defined as information-rich neurons, often demonstrating significant spatial

inhibition effects. These neurons are considered to be more important. To identify neurons with spatial inhibition effects, we introduced an energy function for each neuron and obtained a rapid analytical solution. By minimizing the energy function, weight values for each neuron are obtained, which quantify the degree of linear separability among neurons. The defined energy function is represented by Equation (1):

$$e_t(w_t, b_t, y, x_i) = (y_t - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_0 - \hat{x}_i)^2 \tag{1}$$

Here, $\hat{t} = w_t t + b_t$ and $\hat{x}_i = w_t x_i + b_t$ represent, respectively, the linear transformations of t and x_i. Moreover, t and x_i represent the target neuron as well as other neurons within the same channel of the input feature $X \in R^{C \times H \times W}$, where C, H, and W represent the height, number of channels, and breadth, respectively. The index i represents the index of the neuron in the spatial dimension, and $M = H \times W$ indicates the quantity of neurons present in this channel. The weights and biases of the transformations are represented by w_t and b_t. The true values of the target neuron and other neurons, denoted as y_t and y_0, respectively, are simplified using binary labels (-1 and 1). The final energy function is obtained by incorporating a regularization term λ and is represented by Equation (2):

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1 - (w_t x_i + b_t))^2 + \Big(1 - (w_t t + b_t)\Big)^2 + \lambda {w_t}^2 \tag{2}$$

The rapid analytical solutions for w_t and b_t in Equation (2) are obtained as follows:

$$w_t = -\frac{2(t - u_t)}{(t - u_t)^2 + 2{\sigma_t}^2 + 2\lambda} \tag{3}$$

$$b_t = -\frac{1}{2}(t + u_t)w_t \tag{4}$$

The equation $u_t = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i$ represents the mean value of neural units on the channel. ${\sigma_t}^2 = \frac{1}{M-1}\sum_{i=1}^{M-1}(x_i - u_t)^2$. σ_t represents the variance of the neural units $x_i (i = 1, 2, 3 \cdots M-1)$ on the channel. Finally, by simplification, the minimum energy function is derived, as shown in Equation (5).

$$e_t^* = -\frac{4\big(\hat{\sigma}^2 + \lambda\big)}{(t - \hat{u})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{5}$$

Equation (5) indicates that a lower value of e_t^* corresponds to greater distinction in the vicinity of the target neuron t and the surrounding neurons. Consequently, it is assigned a higher weight, indicating its greater significance in visual processing. In this context, a larger value of $\frac{1}{e_t^*}$ implies higher importance for each neuron. Therefore, the importance of each neuron can be obtained by evaluating $\frac{1}{e_t^*}$.

In the aforementioned feature extraction network based on RepViTBlock, there are still limitations in extracting features from small and complex forest fire images. In order to tackle this problem, this work presents the SimAM module after the RepViTBlock module in the backbone network. The SimAM module is based on neuroscience and defines an energy function for each neuron, which effectively helps the model to learn the features of forest fire images, particularly for small targets. As a result, non-target background interference can be effectively suppressed, leading to enhanced precision in detection with the model. Additionally, the SimAM module is inserted subsequent to the RepViTBlock module in the network design.

2.4. VoVGSCSP

Forest fire detection typically requires models with a high recognition speed and accuracy. However, considering that networks with strong non-linear expressive power

often have a greater quantity of model parameters, they need significant computational resources and are challenging to deploy on resource-limited embedded platforms such as drones or observation towers. Therefore, it is crucial to ensure model performance while keeping it lightweight. This research proposes a bi-directional fusion network that utilizes GSConv to accomplish this objective. The network combines both top-down and bottom-up approaches to facilitate feature fusion. By leveraging the GSConv operations, the network effectively combines information obtained from multiple levels of the feature hierarchy.

Standard convolutions (SC) have good feature extraction and fusion capabilities. However, using too many standard convolutions for feature extraction leads to the large accumulation of parameters. Depthwise separable convolutions (DSC) reduce floating-point operations and model parameters, but they may result in the loss of significant channel information. To overcome these limitations, Li H et al. [28] introduced the GSConv module, depicted in Figure 4.

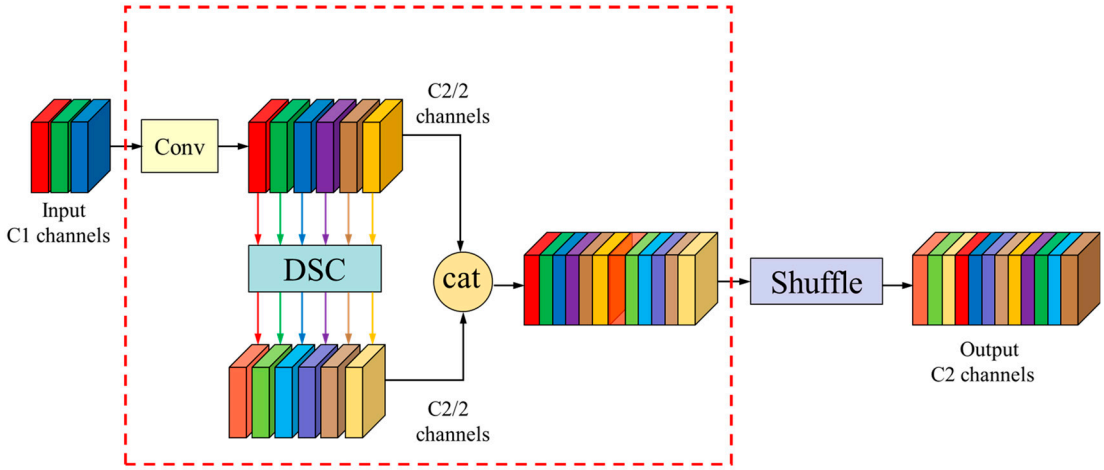

Figure 4. Structure of the GSConv module.

GSConv is a module that combines standard convolution and depthwise separable convolution to fuse them, compressing channels to reduce the computational complexity. As shown in the diagram, the input information undergoes a standard convolution, which compresses the channel number to half of the output channels, denoted as C2/2. It then passes through a separable convolution with depth while maintaining the same quantity of channels. The results of both convolutions are concatenated using the Concat operation. Finally, the shuffle module is applied to shuffle the information, allowing the data produced by the standard convolution to penetrate each and every piece of data generated by the depthwise separable convolution, achieving channel information interaction and restoring the channel number to C2, resulting in the output of GSConv.

It is important to note that using GSConv during feature extraction can lead to the disconnection of hidden connections between certain channels, resulting in the loss of semantic information and potentially affecting the final predictions. Therefore, to better extract features, we still employ standard convolution (SC) in the feature extraction layer.

If GSConv is used throughout the entire model, it would lead to an increased network depth, which can impede the flow of data and significantly increase the inference time. To further accelerate the inference speed while maintaining accuracy, we propose the GS bottleneck module based on GSConv. Building upon this, we construct the VoVGSCSP module, as illustrated in Figure 5.

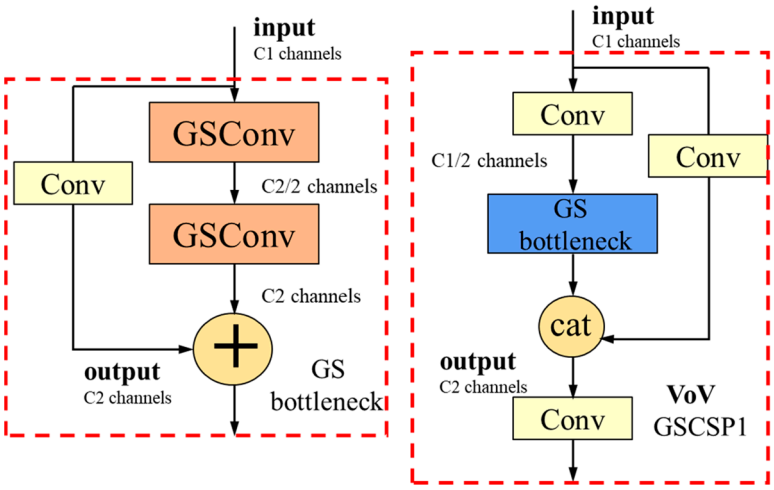

Figure 5. Structure of the VoVGSCSP module.

During feature fusion in the neck layer, as semantic information propagates downwards, the width and height reduce while the number of channels increases. When reaching the GS bottleneck, they reach their maximum and minimum values, respectively, becoming sufficiently elongated and requiring no further transformation. Therefore, at this stage, the repeated information in the feature maps processed by GSConv is reduced, and no further compression is needed. Hence, in this work, the VoVGSCSP module is employed in the layer of the neck of the network to swap out the original ELAN block. This effectively minimizes the network structure and computational complexity without sacrificing the accuracy.

To further facilitate the fusion of deep-level information, in this study, a bottom-up feature fusion network is shown. This network complements the top-down fusion by integrating features from lower levels of the hierarchy. The VoVGSCSP module is retained to improve the model's detection performance. Through the blend of bottom-up and top-down fusion mechanisms, the model achieves comprehensive feature fusion and an improved ability to detect forest fire targets accurately.

2.5. Wise Intersection over Union (WIoU)

Having a clearly specified bounding box loss function is essential in enhancing the effectiveness of forest fire detection models. The complete IoU (CIoU) loss function accelerates the regression speed of predicted bounding boxes by incorporating the aspect ratio between the ground truth and forecasted boxes. However, in the context of forest fire detection, the presence of low-quality data images is inevitable. In such cases, geometric measures like the aspect ratio and distance increase the severity of the penalty for these poor-quality samples.

In order to decrease interference during training and lessen the penalty on geometric measures when anchor boxes and target boxes are aligned well, this research presents the Wise-IoU v3 [29] loss function for forest fire detection. Compared to its predecessor, Wise-IoU v3 attenuates the penalty on geometric factors by dividing the minimal bounding box's height and breadth. Building upon the boundary box loss with a two-layer attention mechanism in Wise-IoU v1, Wise-IoU v3 avoids incorporating the aspect ratio into the loss function. By estimating the outlierness of anchor boxes and defining a dynamic non-monotonic focusing mechanism, high-quality anchor boxes with a modest β are assigned small gradient gains, permitting anchor boxes of moderate quality to be the focus of the regression procedure. Simultaneously, low-quality anchor boxes with a large β are given minimal gradient gains, which effectively reduces the penalty on anchor box regression for low-quality cases.

Firstly, the computation formula for the attention-based boundary box loss function, $Wise - IoU\ v1$, is as follows:

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \tag{6}$$

$$R_{WIoU} = exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{\left(W_g{}^2 + H_g{}^2\right)^*}\right) \tag{7}$$

In the given context, L_{WIoUv1} represents the boundary box loss of $ise - IoU\ v1$, L_{IoU} represents the IoU boundary box loss, and R_{WIoU} denotes the penalty term. W_g and H_g indicate the minimum enclosing box's width and height, whereas the coordinates of the ground truth box's center point are represented by x_{gt} and y_{gt}, and the coordinates of the predicted box's center point are represented by x and y. Additionally, W_g and H_g are the length and breadth of the minimum enclosing box, and the asterisk $(*)$ indicates the separation of W_g and H_g from the computational graph.

Then, the outlierness is defined to describe the anchor box quality, and it is computed using the following formula:

$$\beta = \frac{L^*{}_{IoU}}{\overline{L_{IoU}}} \in (0, +\infty) \tag{8}$$

In this context, β represents the outlierness, and $\overline{L_{IoU}}$ denotes the sliding average value.

Finally, by utilizing the outlierness β to construct a non-monotonic focusing coefficient γ, we obtain the computation formula for the $Wise - IoU\ v3$ boundary box loss function as follows:

$$L_{WIoUv3} = \gamma L_{WIoUv1} \tag{9}$$

$$\gamma = \frac{\beta}{\delta \alpha^{\beta - \delta}} \tag{10}$$

The parameters α and δ are adjustable hyperparameters. When β equals δ, γ is set to 1. When the outlierness satisfies $\beta = C$ (where C is a constant), the anchor box will receive the highest gradient gain. In this study, α and δ are set to 1.9 and 3, respectively.

In summary, when faced with low-quality wildfire samples, the model utilizing Wise-IoU v3 as the loss function allocates smaller gradient gains. This lessens the harm caused by subpar anchor boxes while enabling the model to increase its localization abilities for typical anchor boxes. As a result, the model exhibits higher tolerance towards poor-quality illustrations, thereby enhancing the model's total capacity for generalization.

2.6. SWVR

The detailed architecture of the SWVR network is illustrated in Figure 6. This visualization provides a comprehensive overview of our model's structure, showcasing the various layers and mechanisms employed, including the placement of the Squeeze-and-Excitation (SE) attention mechanism and the arrangement of the convolution layers.

The input module primarily consists of mosaic data augmentation and adaptive anchor box computation. Mosaic data augmentation creates new training samples by concatenating multiple images, thereby raising the data's complexity and variety. It randomly selects a primary image and then randomly selects portions from other images to be copied and pasted onto the primary image, forming a composite image. Additionally, the coordinates and class labels of the target boxes need to be adjusted to fit the changes in the composite image. By utilizing Mosaic data enhancement, the diversity of the data is enhanced, facilitating the model's understanding of the contextual relationships between target objects and their surrounding environments. Furthermore, Mosaic data augmentation increases the quantity of training examples, lowering the possibility of overfitting to certain photos and enhancing the model's training stability. This paper employs adaptive anchor box computation, which adjusts the size and anchor box aspect ratio determined by the distribution and qualities of the dataset's features and objectives. This approach increases the precision and flexibility in object detection. Traditional object detection methods typically

use fixed-size and fixed-aspect-ratio anchor boxes, which may not be able to effectively accommodate various types of objects. Adaptive anchor box computation, on the other hand, utilizes clustering analysis that considers both the size and aspect ratio distribution of the targets. By identifying a collection of anchor boxes that more closely resemble the real items in the dataset, this approach aids in improving the detection performance.

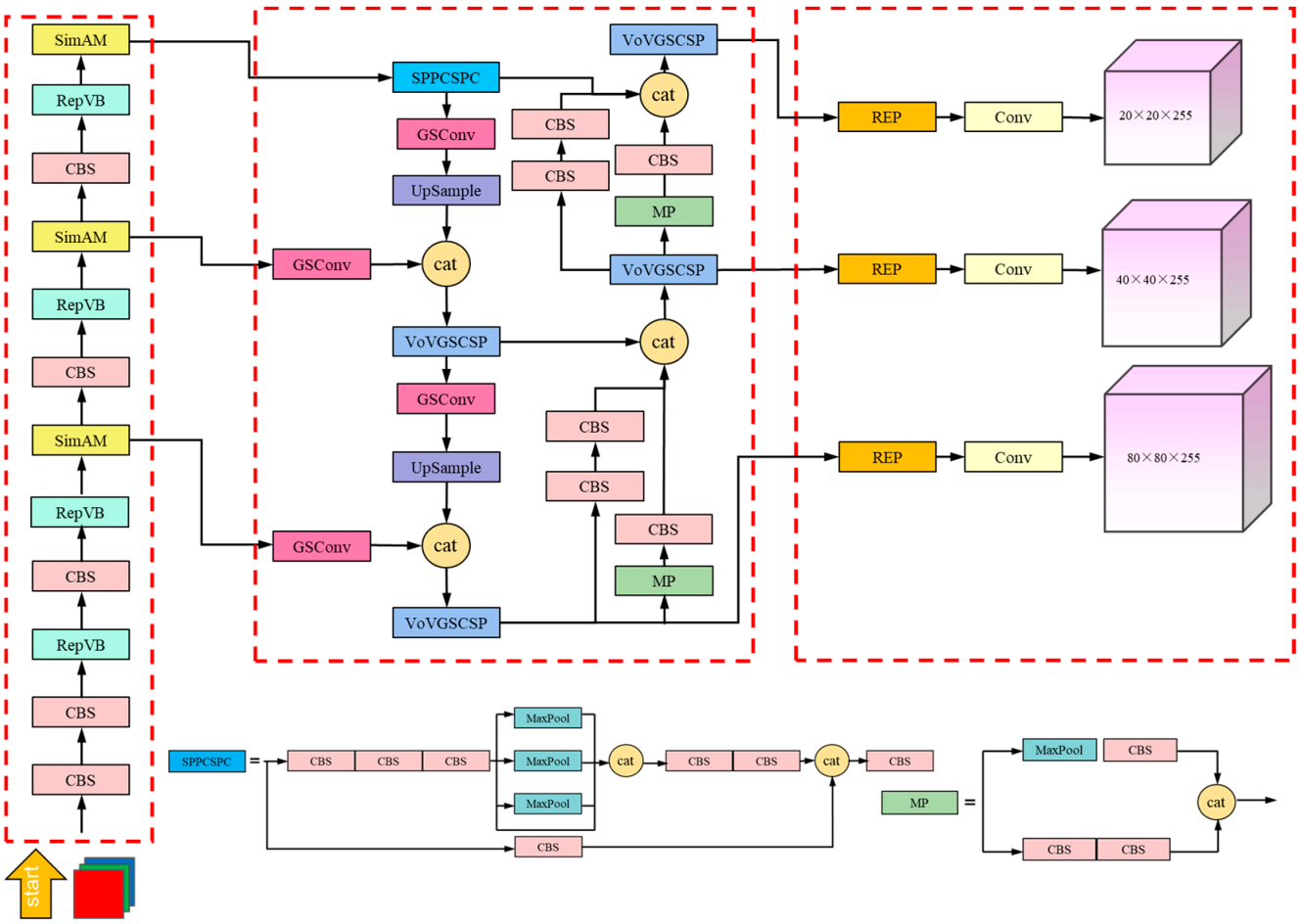

Figure 6. Architecture of the SWVR network.

By removing low- to high-level information from the input picture, the feature extraction layer is essential to the learning process of the model. In the SWVR model proposed in this paper, the first 12 layers are composed of a series of convolutional block sequences (CBS), RepViTBlock modules, and SimAM modules. Among them, the RepViTBlock module adopts a multi-branch topology with token mixers during training. By introducing multiple parallel attention heads, the module enhances the model's focus on important information within the input sequence. During inference, these attention heads are merged into a single deep convolution, effectively reducing the computation and memory costs. Starting from the fifth layer, the output of the RepViTBlock module is augmented with the SimAM attention mechanism. This addition strengthens the capacity of the model to extract characteristics related to small forest fire targets, addressing the issue of local information loss in the RepViTBlock module. The SimAM module helps in capturing relevant features and compensating for any loss in local information, therefore enhancing the model's ability to identify tiny forest fire targets.

The feature fusion component of the SWVR model combines network fusion from the top down and bottom up, which is achieved through modules such as VoVGSCSP, GSConv, and Upsampling. Initially, the feature maps undergo multi-scale information fusion using Spatial Pyramid Pooling Concatenation and Spatial Pyramid Convolution (SPPCSPC).

This step helps to integrate features at different scales. Then, the upper-level information extracted by GSConv is upsampled and mixed with the feature data from the backbone. This joint input is then passed through the VoVGSCSP module, enabling the effective fusion of information from different levels. By employing the VoVGSCSP, GSConv, and Upsampling modules, the SWVR model achieves the efficient fusion of information coming from both the top down and the bottom up, enhancing the overall feature representation and enhancing the model's capacity to gather thorough data for the detection of forest fires.

In the detection layer, the obtained large, medium, and small feature maps are individually processed for object recognition. The REPConv operation is applied, which not only minimizes the computing cost during inference but also boosts the resilience of the learned data. Afterwards, the resulting images are fed through the Non-Maximum Suppression (NMS) algorithm to traverse the candidate bounding box results and obtain the final detection of forest fire targets. Finally, based on the bounding box loss function's anticipated outcomes, namely the Weighted IoU (WIoU) Loss, the positions and sizes of the candidate boxes are adjusted to more accurately localize the forest fire targets. By incorporating REPConv, NMS, and the WIoU Loss, the SWVR model achieves the precise detection and localization of forest fire targets, while also improving the computational efficiency and training data robustness.

3. Methods for Evaluation

Model Evaluation

To assess the model's efficacy in detecting forest fires, this paper selects a set of metrics from three aspects: accuracy, light weight, and real-time capability. Among them, precision, recall, AP@0.5, and AP@ [0.5:0.95] are used to assess the model's accuracy, while the parameters and giga floating-point operations per second (GFLOPs) are used for lightweight evaluation, and the frames per second (FPS) is employed to assess the performance in real time.

1. *AP*: *AP* represents the model's detection accuracy. It is the area under the *PR* curve, with precision and recall as the horizontal and vertical coordinates; its computation necessitates both of these qualities. The formula that follows illustrates this:

$$AP = \int_0^1 P(R)d(R) \tag{11}$$

where P and R represent the proportions of correct predictions in the predicted positive samples and actual positive samples, respectively. The calculation formulas are shown as Equations (12) and (13):

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

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

TP indicates the quantity of fire samples that were accurately identified as fires, *FP* represents the quantity of non-fire samples that were mistakenly identified as fires, the number of non-fire samples that were accurately identified as such is denoted by *TN*, and *FN* represents the quantity of fire samples that were mistakenly identified as non-fire. The values of *TP* and *FP* are dependent on the intersection over union (IoU) and the specified threshold. *TP* is the quantity of bounding boxes where the IoU exceeds the threshold, whereas *FP* denotes the quantity of bounding boxes where the IoU falls below the threshold. The intersection's ratio to the union between the ground truth bounding box and the anticipated bounding box is known as IoU.

2. Parameters: The quantity of parameters in a model is referred to as its parameters, including convolutional parameters, fully connected parameters, BatchNorm parameters, embedding parameters, and so on. The ultimate size of the model is directly impacted by the number of parameters.

3. *GFLOPs*: *GFLOPs* represents the quantity of floating-point operations executed in a second in billions. It serves as a measure of algorithmic complexity. The calculation formula is as follows:

$$GFLOPs = \left(2C_iK^2 - 1\right)HWC_0 \tag{14}$$

where C_i and C_0 represent the quantity of channels for input and output, H and W reflect the feature map's size, and K denotes the size of the kernel.

4. *FPS*: *FPS* is a measure of how many photos can be processed or found in a second, reflecting the speed of detection. The calculation formula is as follows:

$$FPS = \frac{N}{T} \tag{15}$$

where N denotes the whole quantity of pictures detected and T denotes the overall amount of time needed to test every image.

4. Results

4.1. Training and Hyperparameter Configuration

The software and hardware configurations of the paper's experimental platform are presented in Table 2. The pertinent parameters needed to train the algorithm to detect forest fires are shown in Table 3.

Table 2. Experimental conditions.

Experimental Environment	Details
Programming Language	Python 3.8
Operating system	Windows 11
Deep learning framework	PyTorch 1.11.0
GPU	RTX 3080/10 GB
Cuda	CUDA:11.3
CPU	Intel(R)Platinum 8255C (40 GB)

Table 3. Model training parameters.

Training Parameter	Details
epochs	500
batch size	8
img size (pixels)	640 × 640
initial learning rate	0.01
optimization algorithm	SGD
momentum	0.937

4.2. Module Effectiveness Analysis

To better validate the influence of the suggested module improvements on the detection performance and demonstrate the advantages of the SWVR model, we carried out a series of comparative and ablation experiments using the same dataset. In the experiments, we trained and tested the models using a standard image size of 640 × 640. The stochastic gradient descent (SGD) optimizer was utilized with an initial learning rate of 0.01 and we trained them from scratch for 500 epochs.

4.2.1. Effectiveness of SimAM

To assess the effectiveness of the proposed framework, we introduced various modules, such as SE [30], CBAM [31], SimAM [27], GAM [32], and EMA [33], into the feature extraction layer. The test findings are showcased in Table 4.

Table 4. Comparison of different attention mechanisms.

Module	AP@0.5/%	FPS	Params	GFLOPs
Origin	73.7	74.07	9.14 M	26.0
CBAM	71.9	50.25	9.21 M	26.1
SE	76.93	68.4	9.21 M	26.1
GAM	75.5	49.75	23.89 M	73.2
SimAM	77.8	51.8	9.14 M	26.0

The "origin" model refers to the baseline model without the addition of any attention mechanisms.

The results indicate that different attention mechanisms focus on different relevant parts of the input, resulting in variations in model performance and effectiveness compared to the baseline model. CBAM considers both spatial and channel dimensions of information. However, it may excessively focus on noise in the input, leading to incorrect weight allocation and a relative decrease in AP@0.5 in this experiment. SE adaptively adjusts the weights of different channels, thereby enhancing the attention to important channels and effectively improving the detection accuracy. Although the FPS of the SE module is slightly reduced compared to the baseline, it is not significant. Additionally, in comparison to the baseline model, the model's computational complexity and number of parameters rise with the addition of the SE module. The GAM module efficiently gathers global contextual data and results in a small improvement in AP@0.5 compared to the baseline. However, the GAM module has more computational complexity and a greater number of model parameters. SimAM does not demand more parameters to derive 3D attentional weights for the feature maps, but treats all dimensions equally by constructing an energy function to mine the significance of neurons. This effectively enhances the model's capacity to retrieve features, while allowing the number of parameters to be decreased relative to the original. The results show that the value of AP increases by 4.1% relative to the original, and the FPS is second only to that of SE, while its computational complexity is not elevated. Therefore, the SimAM module has been selected to improve the ability of the model to extract features, effectively addressing the dual objectives of lightweight deployment and high precision of detection within the framework of wildfire detection. This deliberate decision aligns with the imperative of meeting the stringent requirements of computational efficiency while optimizing the precision of wildfire incident identification and mitigation in practical field scenarios.

4.2.2. Effectiveness of WiseIoU

In this trial, we assessed the effectiveness of the baseline model by applying four different functions: DIoU [34], MPDIoU [35], WiseIoU [29], and CIoU [34]. These functions were utilized to measure the intersection over union (IoU) between the ground truth annotations and the anticipated bounding boxes. By comparing the results obtained using these different functions, we aimed to assess their effectiveness in improving the accuracy of the model's object detection capabilities. The outcomes are presented in Figure 7. From the experimental findings, it can be observed that DIoU, WiseIoU, and MPDIoU demonstrate improved calculation accuracy compared to CIoU. Among them, WiseIoU exhibits the most significant enhancement, with a 4.4% improvement over CIoU. Therefore, in order to further improve the accuracy of wildfire detection, WiseIoU has been selected as the loss function in this experiment.

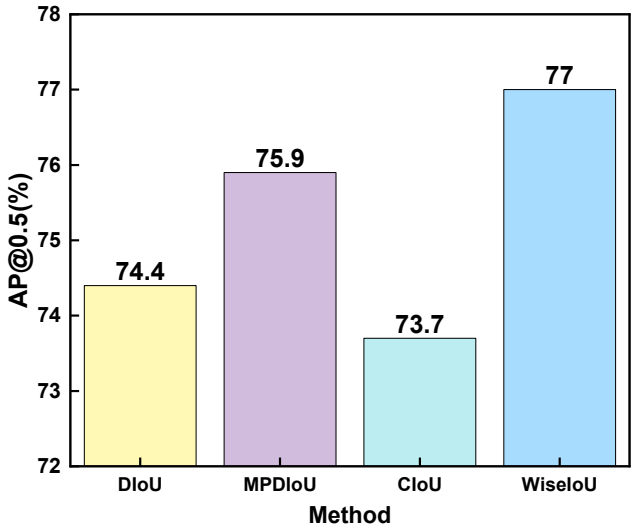

Figure 7. Comparison of different loss functions.

4.2.3. The Effectiveness of RepViTBlock

In this part, we incorporated Pconv [36], PPLC [37], EMO [38], and RepViTBlock into the model and compared their performance. The experimental outcomes are displayed in Table 5 and Figure 8. The observation was made that both Pconv and PPLC exhibited a noticeable decrease in detection accuracy compared to ELAN. EMO demonstrated detection accuracy slightly lower than RepViTBlock, but its computational complexity increased by 35% when compared to ELAN. On the other hand, RepViTBlock showcased a remarkable 3.6% improvement in accuracy over ELAN, with a simultaneous reduction in computational complexity by 24.2% and only a marginal increase in model parameters. Consequently, these findings highlight that RepViTBlock strikes a superior balance between precision and computational intricacy when compared to the aforementioned lightweight networks, as well as ELAN.

Table 5. Comparison of different modules in the backbone network.

Module	AP@0.5/%	Params/M	GFLOPs
ELAN	73.7	9.14	26.0
Pconv	67.6	8.0	20.9
PPLC	70.84	8.9	24.5
EMO	75.3	7.16	35.1
RepViTBlock	76.4	9.3	19.7

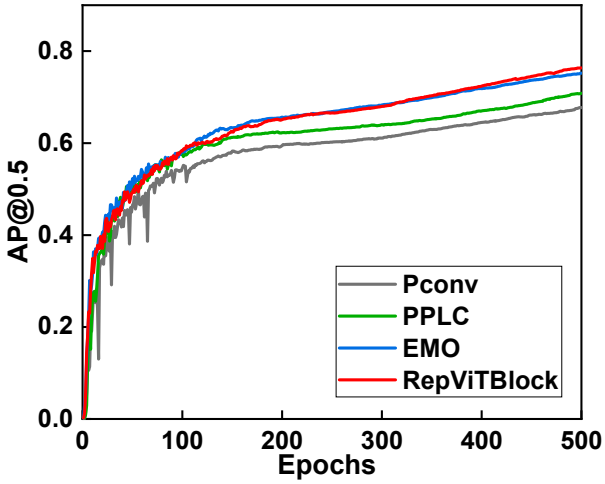

Figure 8. Comparison of AP@0.5 of different modules in the backbone network.

4.3. Ablation Experiment

To confirm the efficacy of the introduced modules in the final model, this study conducted ablation experiments on the same test dataset. The experiments involved systematically removing specific modules and analyzing their impact on the model's performance. Table 6 displays the outcomes of these tests, allowing for a comparative analysis of the impact of every module on the overall functionality of the model.

Table 6. Comparative data of ablation experiments.

Method	P/%	R/%	AP@0.5/%	AP@0.5:0.95/%	FPS	Params/M	GFLOPs
Baseline	76	66.8	73.7	45.5	39.6	9.14	26.0
Baseline + SimAM	79.9	71.4	77.8	48.2	51.8	9.14	26.0
Baseline + WiseIoU	78	70.2	77.0	47.6	38.9	9.14	26.0
Baseline + RepVB	77.3	70.0	76.4	47.3	32.9	9.3	19.7
Baseline + VoVGSCSP	76.1	71.4	76.8	47.8	33.3	7.81	22.7
Baseline + VoVGSCSP + WIoU	78.2	70.0	77.9	49.8	36.5	7.98	23.4
Baseline + VoVGSCSP + RepVB	77	70.4	77.2	49.1	35.8	8.03	16.5
Baseline + VoVGSCSP + RepVB + SimAM	79.3	70.1	78.4	50.22	42.4	8.04	17.5
SWVR	80.5	71.2	79.6	51.6	42.7	8.06	16.5

The "baseline" model refers to the model without the addition of any blocks.

Based on the information in the table, it can be observed that the introduction of SimAM and WiseIoU primarily increases the detection accuracy of the model, with almost no changes in model parameters and computational complexity compared to the baseline model. The incorporation of RepViTBlock yields a significant decrease in computational complexity while concurrently improving the detection accuracy by 2.7%. Similarly, the inclusion of GSconv enhances the model's non-linear expression capabilities through the integration of DWConv layers and shuffle operations. This adaptation results in a reduction in model parameters from 9.14 M to 7.81 M and a corresponding decrease in GFLOPs by 3.3. Furthermore, GSconv demonstrates a commendable 3.1% enhancement in detection accuracy in contrast to the baseline model. These compelling outcomes underscore the efficacy of these modules in achieving a lightweight model without compromising the detection accuracy, thereby enabling the wildfire detection system to process data swiftly and efficiently.

Subsequently, we proceeded to incorporate additional modules on the foundation of GSconv. The usage of WiseIoU proved advantageous in boosting the model's capability to accurately locate standard anchor boxes. Consequently, the precision of the detection witnessed a significant improvement of 3.3%. However, it is worth noting that there was a slight trade-off in terms of the detection speed, which experienced a marginal decrease. We introduced RepViTBlock into the backbone network, combining the efficient architecture of the lightweight Vision Transformer (ViT) with traditional lightweight convolutional neural networks (CNNs). This integration resulted in a significant reduction of 6.3% in GFLOPs, while simultaneously improving the model's AP@0.5/% value by 2.7%. Notably, this approach further reduced the model's computational complexity. By leveraging the strengths of both the lightweight ViT and CNNs, RepViTBlock achieves a more efficient and effective model for computer vision tasks. Finally, under the premise of no significant changes in computational complexity and model specifications, we introduced the SimAM attention mechanism. The inclusion of SimAM resulted in a noteworthy improvement of 3.3% in both the precision (P) and recall (R) metrics. Moreover, the AP@0.5 experienced a substantial increase of 4.7%. Furthermore, the model's FPS demonstrated a notable improvement of 2.8, with only a slight increase in Params and GFLOPs as trade-offs. These findings highlight the effectiveness of incorporating SimAM, as it leads to significant performance enhancements without compromising the computational complexity and model size. This indicates that SimAM can assist the network in gathering fire data more effectively within complex settings, without increasing the model parameters and

computational complexity, thereby improving the generalization ability and precision of the model.

In conclusion, the suggested SWVR model attains a remarkable 79.6% AP@0.5., operating at a frame rate of 42.7 frames per second, with only 16.5 GFLOPs and 8.06 million parameters. In contrast to the baseline model, the AP@0.5 is improved by 5.9%, demonstrating its effectiveness in enhancing the detection accuracy. Additionally, the model exhibits a significant FPS improvement of 3.1, indicating an enhanced inference speed. Furthermore, the model's Params and GFLOPs are decreased by 1.08 and 9.5. This suggests that with fewer computing resources, our suggested model may achieve high detection performance. Indeed, the improved SWVR network model not only strengthens the detection accuracy but also offers the advantage of being more lightweight, making it appropriate for installation on equipment for the monitoring of forest fires, such as unmanned aerial vehicles (UAVs) and observation towers.

4.4. Comparative Experiment

On the premise of keeping the experimental environment unchanged, the SWVR model was compared with the most representative object detection algorithms currently available: YOLOv7 [39], SSD [40], Faster R-CNN [41], RetinaNet [42], Grid R-CNN [43], YOLOv7x [39], and YOLOv7-tiny [39]. Moreover, the validation set's top performers were chosen for testing. The results are shown in Table 7.

Table 7. Comparative data of different detection algorithms.

Method	AP@0.5/%	Params/M	GFLOPs	FPS
YOLOv7	73.7	9.14	26.0	39.6
YOLOv7x	81.79	70.8	188.9	29.04
Yolov7-tiny	68.83	6.01	13.2	70.3
Faster R-CNN	62.38	41.2	206.66	38.5
RetinaNet	65.78	19.61	153.79	48.5
SSD	80.2	23.75	342.75	95.3
Grid R-CNN	70.14	64.24	340.14	26.8
SWVR	79.6	8.06	16.5	42.7

The YOLOv7x model attains 81.79% detection accuracy, which is relatively high. However, it has a high parameter count of 70.8 million and GFLOPs of 188.9. On the other hand, YOLOv7-tiny has significantly fewer parameters and lower computational complexity, with 6.01 million parameters and 13.2 GFLOPs. However, its detection accuracy is only 68.83%, indicating lower model precision. YOLOv7 achieves an AP@0.5 of 73.7% with 9.14 million parameters and 26.0 GFLOPs.

Compared to the aforementioned popular object detection algorithms, the suggested improvement method in this paper exhibits significant advantages. SWVR achieves accuracy of 79.6% for forest fire detection. As opposed to YOLOv7, Faster R-CNN, RetinaNet, and Grid R-CNN, SWVR improves the detection accuracy by 5.9%, 17.22%, 13.82%, and 9.46%, respectively. It is evident that this model significantly enhances the detection accuracy, while the other networks experience noticeable decreases in performance.

Faster R-CNN achieves the lowest detection accuracy of 62.38%, with a parameter count of 41.2 million. This represents an increase in parameters of 32.06 million over the baseline model. It has a GFLOPs value of 206.66, which is increased by 694.84% compared to the baseline. RetinaNet and SSD have high FPS values of 48.5 and 95.3, respectively, indicating a significant advantage in terms of inference speed. However, their parameter counts are 19.61 million and 23.75 million, respectively, representing an increase of 114.5% and 159.8% compared to YOLOv7. On the other hand, the SWVR model has a parameter count of only 8.03 million, which is a decrease of 12.14% compared to the baseline. RetinaNet and SSD exhibit a significant increase in GFLOPs, indicating higher computational complexity. Furthermore, although SSD achieves accuracy of 80.2%, its

parameter count and computational complexity far exceed those of the baseline model, making it unsuitable for lightweight requirements. In contrast to the reference model, Grid R-CNN has a much larger parameter count and computational complexity, resulting in decreased detection accuracy and FPS.

Compared to the current, most representative object detection algorithms, although the SWVR model does not exhibit a significant improvement in inference speed, it is still capable of meeting practical application requirements. Furthermore, it achieves a 3.1 improvement compared to the baseline model. Regarding the precision of detection, parameter count, and computational complexity, the SWVR model performs well.

In summary, the SWVR algorithm effectively balances a lightweight design with model performance, and, overall, it outperforms some common object detection algorithms.

4.5. Verification in Different Scenarios

We tested the baseline and the proposed SWVR model on forest fires in different scenarios. Some visual results of the tests are as follows: the left image shows the original model, and our upgraded model is seen in the right image.

As shown in Figure 9, when there are substantial flames in the middle and later phases of the flame, compared to the original model, our model has a closer detection frame to the target and a higher confidence level.

(**a**) (**b**)

Figure 9. Typical fire detection results. (**a**) Baseline's detection box cannot fully cover the flame; (**b**) our model's detection box can cover it well.

As shown in Figure 10, when monitoring ground fires with multiple targets, the original model detected three fires with the problem of missed detection, indicating its weak ability to extract flame features, while our model detected all fires.

(**a**) (**b**)

Figure 10. Detection of multi-target fires. (**a**) The baseline model can only detect three fire targets; (**b**) our model has completely detected 9 fire targets.

As shown in Figure 11, in remote forest fire images, the original model cannot detect some smaller flames well, while our model can detect all flames, which is conducive to the timely detection and extinguishing of fires.

(**a**) (**b**)

Figure 11. Detection of small target fires. (**a**) The baseline model only detects three fire targets; (**b**) our model successfully detected 9 small target fires.

In addition, as shown in Figure 12, this model performs well in detecting background interference such as leaf occlusion and smoke. The baseline model may have false positives, while our model can correctly detect all flames, indicating the good anti-interference performance of our model.

(**a**) (**b**)

Figure 12. Fire detection situation with trees and smoke blocking. (**a**) The baseline model mistakenly detects trees as fires in complex environments; (**b**) our model can accurately detect all fire targets in thick smoke.

In summary, the SWVR forest fire detection algorithm suggested in this article effectively overcomes the problems of small flame targets and background interference and can accurately detect flames more comprehensively.

5. Discussion

Forest fires have become a serious natural disaster that threatens ecosystems, infrastructure, and human lives. The monitoring and timely detection of forest fire occurrences are crucial. Traditional image recognition algorithms based on manually defined features exhibit poor anti-interference performance and high false detection rates, while object detection algorithms based on CNNs offer high accuracy but come with a large number of model parameters. To overcome these obstacles, we have developed a lightweight, high-precision, and real-time system, SWVR. The system is deployed on computing devices that meet the computational requirements, with a focus on reducing the algorithm parameters and computational complexity. This enables us to identify forest fires in real time with the maximum precision.

Firstly, within the network for feature extraction, we utilize the RepViTBlock module. This module is an improvement over the standard lightweight MobileNetV3-L block structure. By shifting the depthwise separable convolutions upwards, we successfully separate the token mixer and channel mixer to construct the RepViTBlock. Additionally, we introduce structural reparameterization to effectively reduce the computational and

memory costs. Furthermore, in order to further minimize the model's parameters, we unify the kernel size of the convolutions in MobileNetV3-L to 3×3. This change allows us to achieve a lightweight model and accelerate inference while ensuring its accuracy, greatly reducing the model's reliance on hardware environments.

Then, to address the variations in flame size and form in forest fire detection scenarios, as well as the susceptibility to factors like weather, lighting, and background noise, we introduce the parameter-free Similarity Attention (SimAm) module into the feature extraction network. This module assigns higher weights to neurons carrying features related to forest fires, enhancing the model's resistance to interference and its feature extraction capabilities. Furthermore, at the beginning of a forest fire, the flames are typically smaller, especially in images captured from a long distance or with a wide field of view. The introduced SimAm attention module can effectively enhance the focus on smaller flames without increasing the quantity of network parameters.

Additionally, inside the layer of feature fusion, we make use of the lightweight convolution GSConv. This convolution method combines depthwise convolution (DWConv) and standard convolution (SConv) effectively, along with the shuffle operation, resulting in a significant reduction in model parameters. Furthermore, the introduced Voxel-Wise, Group-Wise, and Channel-Wise Spatial Pyramid (VoVGSCSP) module in the feature fusion layer enhances the capability for feature fusion. This module further enhances the precision of the model and its real-time performance by capturing and incorporating contextual data on many scales.

Finally, building upon the aforementioned improvements, we utilize WiseIoU to optimize the bounding box regression loss function, further enhancing the model's convergence speed and detection accuracy. WiseIoU is based on an ever-changing non-monotonic focal mechanism that uses "outlierness" as a substitute for the intersection over union (IoU) to assess the anchor boxes' quality. It uses a prudent gradient gain allocation approach to lessen the superior anchor boxes' competition and mitigate the adverse gradients caused by low-quality fire annotations. This approach addresses the issue of poor overall performance in detecting low-quality examples within the dataset. It is essential in improving the model's detection accuracy. By incorporating WiseIoU, the model benefits from a more effective and robust loss function, resulting in enhanced convergence and more precise identification outcomes.

While our research represents a significant advance, we acknowledge certain limitations. The specific types of images used for training, whether RGB or multispectral, are not explicitly specified in this study. We intend to address this gap in future research. Additionally, the performance of the algorithm in scenarios with potential sources of error, such as sunlight, has not been explicitly tested and will be a focal point of our future investigations. In terms of practical applications, our future work will explore the algorithm's integration with RGB cameras and other devices, aiming to enhance its real-world applicability to national parks, forest plantations, and different related environments under risks of extreme fire occurrence. For instance, such instruments could be installed in existing fire towers and poles.

6. Conclusions

The conclusive experimental findings show that the proposed SWVR model achieves an AP@0.5 of 79.6, with a frame rate of 42.7. Notably, the model achieves these results with a low computational cost, as indicated by the GFLOPs of only 16.5 and a small parameter count of 8.06 million. In contrast to the baseline model and other common detection algorithms, it has achieved comprehensive improvements in all indicators and has significant advantages regarding its light weight and detection performance, making it appropriate for deployment on miniaturized embedded systems based on drones. We also note that while the detection accuracy of the SWVR model is greatly enhanced in comparison to the baseline, there is still room for further improvement. In scenarios where there are bright spots caused by sunlight reflection or artificial lighting, false alarms

may occur due to the similarity in shape and color between flames and these bright spots. In the future, it would be beneficial to collect a diverse range of images in complex environments to improve the model's performance in learning the distinctive features of flames. By associating the morphological features of flames with semantic information in the environment, we can enhance the model's recognition capability for flames of different shapes and sizes.

Author Contributions: L.J. was responsible for the program design and drafting of the initial manuscript. Y.Y. and J.Z. assisted with data collection and analysis. D.B. and H.Z. revised the initial manuscript. H.L. designed the project and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Research and Development Plan of Jiangsu Province (Grant No. BE2021716).

Data Availability Statement: All data generated or presented in this study are available upon request from the corresponding author. Furthermore, the models and code used during the study cannot be shared at this time as the data also form part of an ongoing study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Salesa, D.; Cerdà, A. Soil erosion on mountain trails as a consequence of recreational activities. A comprehensive review of the scientific literature. *J. Environ. Manag.* **2020**, *271*, 110990. [CrossRef] [PubMed]
2. Bohn, F.J.; Huth, A. The importance of forest structure to biodiversity–productivity relationships. *R. Soc. Open Sci.* **2017**, *4*, 160521. [CrossRef]
3. Baldrian, P.; López-Mondéjar, R.; Kohout, P. Forest microbiome and global change. *Nat. Rev. Microbiol.* **2023**, *21*, 487–501. [CrossRef]
4. Kinaneva, D.; Hristov, G.; Raychev, J.; Zahariev, P. Early forest fire detection using drones and artificial intelligence. In Proceedings of the 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 20–24 May 2019; pp. 1060–1065.
5. Dogra, R.; Rani, S.; Sharma, B. A review to forest fires and its detection techniques using wireless sensor network. In *Advances in Communication and Computational Technology: Select Proceedings of ICACCT 2019*; Springer: Singapore, 2021; pp. 1339–1350.
6. Kataev, M.Y.; Kartashov, E.Y. Computer Vision Method for Forest Fires Detection Based on RGB Images Obtained by Unmanned Motor Glider. *Light Eng.* **2021**, *29*, 71–78. [CrossRef] [PubMed]
7. Emmy Prema, C.; Vinsley, S.S.; Suresh, S. Multi feature analysis of smoke in YUV color space for early forest fire detection. *Fire Technol.* **2016**, *52*, 1319–1342. [CrossRef]
8. Calderara, S.; Piccinini, P.; Cucchiara, R. Vision based smoke detection system using image energy and color information. *Mach. Vis. Appl.* **2011**, *22*, 705–719. [CrossRef]
9. Yang, X.; Hua, Z.; Zhang, L.; Fan, X.; Zhang, F.; Ye, Q.; Fu, L. Preferred Vector Machine for Forest Fire Detection. *Pattern Recognit.* **2023**, *143*, 109722. [CrossRef]
10. Huang, L.; Liu, G.; Wang, Y.; Yuan, H.; Chen, T. Fire detection in video surveillances using convolutional neural networks and wavelet transform. *Eng. Appl. Artif. Intell.* **2022**, *110*, 104737. [CrossRef]
11. Barmpoutis, P.; Dimitropoulos, K.; Kaza, K.; Grammalidis, N. Fire detection from images using faster R-CNN and multidimensional texture analysis. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8301–8305.
12. Muhammad, K.; Ahmad, J.; Baik, S.W. Early fire detection using convolutional neural networks during surveillance for effective disaster management. *Neurocomputing* **2018**, *288*, 30–42. [CrossRef]
13. Chen, G.; Cheng, R.; Lin, X.; Jiao, W.; Bai, D.; Lin, H. LMDFS: A Lightweight Model for Detecting Forest Fire Smoke in UAV Images Based on YOLOv7. *Remote Sens.* **2023**, *15*, 3790. [CrossRef]
14. Abdusalomov, A.B.; Islam, B.M.S.; Nasimov, R.; Mukhiddinov, M.; Whangbo, T.K. An improved forest fire detection method based on the detectron2 model and a deep learning approach. *Sensors* **2023**, *23*, 1512. [CrossRef] [PubMed]
15. Reis, H.C.; Turk, V. Detection of forest fire using deep convolutional neural networks with transfer learning approach. *Appl. Soft Comput.* **2023**, *143*, 110362. [CrossRef]
16. Rostami, A.; Shah-Hosseini, R.; Asgari, S.; Zarei, A.; Aghdami-Nia, M.; Homayouni, S. Active fire detection from landsat-8 imagery using deep multiple kernel learning. *Remote Sens.* **2022**, *14*, 992. [CrossRef]
17. Li, W.; Zhang, L.; Wu, C.; Cui, Z.; Niu, C. A new lightweight deep neural network for surface scratch detection. *Int. J. Adv. Manuf. Technol.* **2022**, *123*, 1999–2015. [CrossRef] [PubMed]
18. Lv, Y.; Liu, J.; Chi, W.; Chen, G.; Sun, L. An inverted residual based lightweight network for object detection in sweeping robots. *Appl. Intell.* **2022**, *52*, 12206–12221. [CrossRef]

19. Huang, J.; He, Z.; Guan, Y.; Zhang, H. Real-time forest fire detection by ensemble lightweight YOLOX-L and defogging method. *Sensors* **2023**, *23*, 1894. [CrossRef] [PubMed]
20. Zhang, J.; Zhu, H.; Wang, P.; Ling, X. ATT squeeze U-Net: A lightweight network for forest fire detection and recognition. *IEEE Access* **2021**, *9*, 10858–10870. [CrossRef]
21. Muhammad, K.; Khan, S.; Elhoseny, M.; Ahmed, S.H.; Baik, S.W. Efficient fire detection for uncertain surveillance environment. *IEEE Trans. Ind. Inform.* **2019**, *15*, 3113–3122. [CrossRef]
22. Wang, A.; Chen, H.; Lin, Z.; Pu, H.; Ding, G. Repvit: Revisiting mobile cnn from vit perspective. *arXiv* **2023**, arXiv:2307.09283.
23. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25, Proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012, 3–6 December 2012, Lake Tahoe, NV, USA*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2012.
24. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
25. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
26. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
27. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 11863–11874.
28. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles. *arXiv* **2022**, arXiv:2206.02424.
29. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
31. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*; Springer: Cham, Switzerland, 2018; pp. 3–19.
32. Liu, Y.; Shao, Z.; Hoffmann, N. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv* **2021**, arXiv:2112.05561.
33. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
34. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12993–13000.
35. Siliang, M.; Yong, X. MPDIoU: A Loss for Efficient and Accurate Bounding Box Regression. *arXiv* **2023**, arXiv:2307.07662.
36. Chen, J.; Kao, S.H.; He, H.; Zhuo, W.; Wen, S.; Lee, C.H.; Chan, S.H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031.
37. Cui, C.; Gao, T.; Wei, S.; Du, Y.; Guo, R.; Dong, S.; Lu, B.; Zhou, Y.; Lv, X.; Liu, Q.; et al. PP-LCNet: A lightweight CPU convolutional neural network. *arXiv* **2021**, arXiv:2109.15099.
38. Zhang, J.; Li, X.; Li, J.; Liu, L.; Xue, Z.; Zhang, B.; Jiang, Z.; Huang, T.; Wang, Y.; Wang, C. Rethinking Mobile Block for Efficient Attention-based Models. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 1389–1400.
39. Wang, C.; Bochkovskiy, A.; Liao, H.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
40. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part I*; Springer: Cham, Switzerland, 2016; pp. 21–37.
41. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28, Proceedings of the 29th Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2015.

42. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
43. Lu, X.; Li, B.; Yue, Y.; Li, Q.; Yan, J. Grid R-CNN. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7363–7372.