

Article

A Lightweight Cross-Layer Smoke-Aware Network

Jingjing Wang ¹, Xinman Zhang ^{1,*} and Cong Zhang ²
¹ School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China; wjingjing@stu.xjtu.edu.cn

² AECC Sichuan Gas Turbine Establishment, Mianyang 621000, China; asccwish@163.com

* Correspondence: zxm@xjtu@mail.xjtu.edu.cn

Abstract: Smoke is an obvious sign of pre-fire. However, due to its variable morphology, the existing schemes are difficult to extract precise smoke characteristics, which seriously affects the practical applications. Therefore, we propose a lightweight cross-layer smoke-aware network (CLSANet) of only 2.38 M. To enhance the information exchange and ensure accurate feature extraction, three cross-layer connection strategies with bias are applied to the CLSANet. First, a spatial perception module (SPM) is designed to transfer spatial information from the shallow layer to the high layer, so that the valuable texture details can be complemented in the deeper levels. Furthermore, we propose a texture federation module (TFM) in the final encoding phase based on fully connected attention (FCA) and spatial texture attention (STA). Both FCA and STA structures implement cross-layer connections to further repair the missing spatial information of smoke. Finally, a feature self-collaboration head (FSCHead) is devised. The localization and classification tasks are decoupled and explicitly deployed on different layers. As a result, CLSANet effectively removes redundancy and preserves meaningful smoke features in a concise way. It obtains the precision of 94.4% and 73.3% on USTC-RF and XJTU-RS databases, respectively. Extensive experiments are conducted and the results demonstrate that CLSANet has a competitive performance.

Keywords: smoke detection; CLSANet; cross-layer connection; spatial perception; texture federation; self-collaboration head



Citation: Wang, J.; Zhang, X.; Zhang, C. A Lightweight Cross-Layer Smoke-Aware Network. *Sensors* **2024**, *24*, 4374. <https://doi.org/10.3390/s24134374>

Academic Editors: Dongsheng Zhang and Zhilong Su

Received: 27 May 2024

Revised: 25 June 2024

Accepted: 2 July 2024

Published: 5 July 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Smoke detection, as an early warning of flames, is extremely essential to minimize the damage to ecosystems, public safety, and human life and property caused by fires. Previous smoke detection approaches rely heavily on manual observation or specialized smoke sensors, but they are inevitably constrained by the limited human resources and monitoring range, respectively [1]. In recent years, thanks to the mature installation of surveillance cameras and the Internet of Things (IoT) formed by millions of smart edge devices, vision-based smoke detection, as a study focus, has gradually attracted wide public attention [2].

To implement smoke detection in practical production, many researchers have devoted great efforts and a number of advanced technologies have been developed. Based on the traditional features of smoke, Tian et al. [3] devised an image formation model formulated as convex optimization that solved a sparse representation problem using dual dictionaries for the smoke and background components, respectively. Dimitropoulos et al. [4] presented a new higher-order linear dynamical system (h-LDS) descriptor, which applied higher-order decomposition of the multidimensional image data and dynamic analysis to a video-based smoke identification system. With the development of deep learning, Yar et al. [5] employed a novel vision transformers (ViT) architecture by combining shifted patch tokenisation and local self-attention modules, but it only implements fire scene classification and has no localization capability. Tao et al. [6] gave an adaptive frame selection network (AFSNet) with augmented dilated convolution for smoke detection. The pixels with high responses

and the local regions were taken into account to learn the contribution of each pixel automatically. Cao et al. [7] proposed an enhanced feature foreground network (EFFNet). It utilized efficient branch channels to predict the source mask and bounding boxes of smoke plumes. However, smoke detection is still a challenging task due to its irregular shape, complex state, and appearance which are susceptible to environmental disturbances.

Cross-layer connection can enhance the original features of each layer by enabling information exchange among multiple layers and is progressively applied to visual inspection to strengthen performance. Feature pyramid networks (FPNs) and path aggregation networks (PANs) [8], as classical cross-layer structures, fused semantic and object details over features at different scales in an incremental manner. Moreover, Li et al. [9] presented a cross-layer extraction structure and multi-scale down-sampling network with bidirectional transpose FPN (BCMNet) for smoke detection, and its cross-layer incorporated linear feature multiplexing and receptive field amplification. Long et al. [10] added horizontal and vertical connections to a regularized cross-layer ladder network (RCLN), where all four layers of features were computed by manifold regularization constraints to maximize information transmission. Li et al. [11] devised a cross-layer feature pyramid network (CFPN). It indiscriminately aggregated multi-scale features from different levels to gain access to rich contexts. As we can see, the existing cross-layer connectivity regards each feature map equally, i.e., it conducts the same operations on each layer during fusion but ignores the differences between high- and low-level features, which introduces a certain degree of information redundancy and noise in the extracted features. Considering the variability of smoke, it is tough to design an effective cross-layer network to integrate its features.

In this paper, we propose a lightweight cross-layer smoke-aware network (CLSANet) with only 2.38 M, and it shares information across different levels with varying preferences to enhance feature propagation and improve real-world smoke detection. To the best of our knowledge, a significant difference between CLSANet and existing approaches is that CLSANet does not have an equal but a different focus on the treatment of high- and low-level features when deploying cross-layer connectivity. Specifically, we only embed the spatial details derived from the low layer and the semantic features developed from the high layer in each fusion stage to preserve the favorable information. Following this principle, the design of our CLSANet performs biased cross-layer cooperation of features between the high and low levels, which effectively removes redundancy and retains meaningful details. We compare the smoke detection results of different algorithms, including SSD [12], RetinaNet-50 [13], RetinaNet-101 [13], EfficientDet [14], YOLOv5 [15], YOLOX [16], EdgeYOLO-Tiny [17], and CLSANet, as shown in Figure 1. To facilitate comparison, the original images are presented in the first row, the second row gives the detection results of the existing algorithms, and the third row provides the performance of our CLSANet. When confronted with smoke in various forms, such as light smoke, fuzzy smoke, or interference from smoke analogs, existing methods suffer from inaccuracies such as missed or false detection. It can be found that the proposed CLSANet shows the most sensitive and accurate detection of smoke, indicating the superiority of our biased cross-layer design. The main contributions of this paper are as follows:

1. A lightweight cross-layer smoke-aware network (CLSANet) which is only 2.38 M is designed. Considering the variable shape and appearance of smoke, it integrates and distributes cross-scale features with distinct preferences and encourages dynamic communication across layers.
2. We propose a spatial perception module (SPM) to enable low-level spatial information to guide the refinement of semantic features in deep layers. The separate features of each layer can thus access both spatial and semantic details, resulting in a decreased loss of important information during the multi-layer progressive fusion.
3. We propose a texture federation module (TFM) as the final feature encoding. Spatial texture attention (STA) is designed to filter shallow information, and at the high layers,

we devise fully connected attention (FCA), which breaks the convolutional constraint on spatial location with the assistance of full connection.

4. We propose a feature self-collaboration head (FSCHead). The low-level features that keep the spatial details are only for localization while the deep maps which refine the semantic information after multi-computation are merely for classification. FSCHead can remove redundancy in such a very concise way.

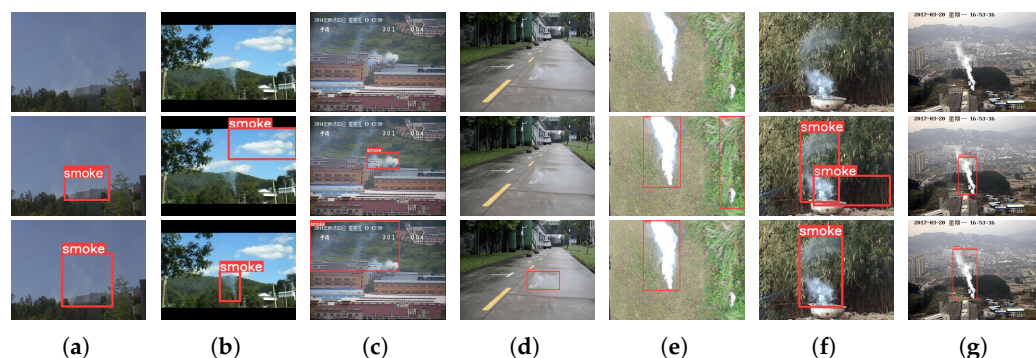


Figure 1. Examples of the input images and their detection results of different algorithms, which are (a) SSD [12], (b) RetinaNet-50 [13], (c) RetinaNet-101 [13], (d) EfficientDet [14], (e) YOLOv5 [15], (f) YOLOX [16], and (g) EdgeYOLO-Tiny [17], respectively. Each column includes the input image and its associated detection results obtained by the existing approach and our CLSNet. It is apparent that the conventional methods inevitably suffer from false and missed detection, but our proposed CLSNet can distinguish interference and detect precise smoke areas, which demonstrates the effectiveness of our cross-layer design.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 introduces the details of our method. Experiments and analyses are discussed in Section 4. Finally, the paper is concluded in Section 5.

2. Related Work

Early methods based on manual observation and sensors have limited monitoring scope. With the evolution of smart mobile devices, machine vision-based smoke detection has been researched widely in recent years. Especially for neural network-based smoke detection, a series of novel model frameworks and advanced strategies have emerged.

2.1. Smoke Detection Algorithm

There are two main categories of prevalent smoke detection approaches. One is traditional descriptors based on hand-crafted features, and the other relies on deep learning [18]. Appana et al. [19] detected smoke based on its dispersion, color, and semi-transparency characteristics extracted from optical flow patterns and spatiotemporal energy. Filonenko et al. [20] took full advantage of shape and color information to evaluate the probability that a pixel belonged to the fluid smoke region. Prema et al. [21] examined distinctive texture attributes of smoke regions which were extracted by the co-occurrence of hamming distance-based local binary pattern (CoHDLBP) and co-occurrence of local binary pattern (CoLBP), including homogeneity, energy, correlation, and contrast. With the popularity of deep convolutional networks, smoke detection based on them gradually started to be explored. Hashemzadeh et al. [22] introduced a hybrid algorithm based on deep learning and spatiotemporal features of smoke. The extracted moving regions were individually submitted to a customized convolutional neural network. Tao et al. [23] presented an attention-aggregated attribute-aware network (AANet) to merge spatiotemporal and context information and decode video attributes. Gu et al. [24] proposed a deep dual-channel neural network (DCNN). The first subnetwork extracted details of smoke and the base information was captured in the second subnetwork, such as texture and contours, respec-

tively. Almeida et al. [25] designed a lightweight convolutional neural network (CNN) for wildfire detection combined with edge computing devices. Mukhiddinov et al. [26] gave a wildfire smoke detection system using unmanned aerial vehicle (UAV) images based on an optimized YOLOv5. Saydirasulovich et al. [27] introduced the BiFormer attention mechanism in an enhanced YOLOv8 model to obtain accurate smoke detection in UAV images. Munsif et al. [28] designed an efficient deep learning model of only 9 M that can be easily deployed on resource-constrained devices. Tao et al. [29] devised a channel-enhanced spatiotemporal network (CENet) for recognizing industrial smoke emissions. A new loss function and several channel-enhanced modules ensured sufficient supervision information. Chen et al. [30] put forward an end-to-end deep neural network called DesmokeNet to remove the smoke by a two-stage recovered pipeline. Deep learning-based methods have achieved better performance than traditional techniques for smoke detection. However, due to the randomness of smoke shapes, intraclass variations, obstructions, and clutters, the pursuit of elaborate network structures while ignoring the essential attributes of smoke restricts the network's generalization ability.

2.2. Cross-Layer Application Algorithm

Cross-layer modules can acquire the nonlocal associations of each layer and further strengthen the smoke-relevant feature through the aggregation and balance between the global and local contexts [31]. Li et al. [32] designed a smoke recognition model based on a detection transformer (DETR). Cross-layer deformable attention was used in the encoder-decoder structure to speed up the convergence process. Tao et al. [33] proposed a smoke density estimation network (SDENet) to explicitly analyze channel interdependencies and decode four-layer features. Its progressive cross-layer connectivity primarily relied on attention mechanisms. Zhang et al. [34] designed a multi-scale convergence coordinated pyramid network with mixed attention and fast-robust NMS (MMFNet). It combined dual attention and coordinated convergence module to refine the pyramid network and then smoke features were organized at multiple scales. Zhan et al. [35] devised an adjacent layer composite network based on a recursive feature pyramid with deconvolution and dilated convolution and global optimal non-maximum suppression (ARGNet) to monitor smoke. It improved FPN into a recursive pyramid with deconvolution and dilated convolution (RDDFPN) to achieve cross-layer multi-scale fusion. Yuan et al. [36] stacked some encoder-decoder structures and presented a wave-shaped neural network (W-Net). Cross-layer connections were employed between the peaks and troughs which contained abundant localization and semantic information, respectively. Tao et al. [37] presented a dual-branch smoke segmentation model SmokeSegger coupling a Transformer branch and a CNN branch. There are four cross-connections between the two branches to enhance the expression of global and local content. Yuan et al. [38] proposed a classification-assisted gated recurrent network (CGRNet) for smoke segmentation, where a multi-scale context contrasted local structure (MCCL) and a dense pyramid pooling module (DPPM) took similar cross-layer processing of feature maps for each layer via conventional convolution. In addition, Song et al. [39] introduced a cross-layer semantic guidance module (CSGM) to guide the shallow feature layer via deep-level information and then gave a cross-layer semantic guided network (CSGNet) based on YOLOv6. Zhang et al. [40] designed an efficient detection network based on cross-layer feature aggregation (CFANet). It integrated multi-scale features based on the avoidance of semantic gaps and compensated for the defect of layer-by-layer feature transfer that only focuses on the previous layer. It is obvious that existing cross-layer connections are used to complement global and local features and facilitate multi-scale detection, including attention mechanisms and pyramid networks.

In this paper, considering that the shallow layers contain rich spatial details and the deep layers extract precise semantic information after more computation, we propose a lightweight CLSNet by biased cross-layer connections between low-level texture features and high-level classification information.

3. Proposed Method

CLSANet considers the essential attributes of multi-level features for adaptive hierarchical reweighting during aggregation, and cross-layer connections are made in a biased way. In this section, the overall architecture of the CLSANet model is presented first. Then we introduce the details of three novel modules, SPM, TFM, and FSCHead, respectively. In particular, SPM is applied to the backbone, TFM is embedded at the end of feature encoding, and FSCHead directly is the head of our whole network. These novel cross-layer connection structures effectively prevent the dilution of meaningful information and pursue accurate smoke detection. Finally, we briefly describe the loss function for training.

3.1. Whole Network Architecture

We propose a lightweight CLSANet for multi-scale convergence of smoke features with different preferences, and Figure 2 shows its overall architecture. Specifically, besides the usual convolutional operation, C2F block [41], and path aggregation network (PAN) [42], it also contains three novel modules including SPM, TFM, and FSCHead. As we can see, in the backbone, the three features F_{SPM}^{out0} , F_{SPM}^{out1} , and F_{TFM}^{out} are all calculated by the cross-layer fusion in the spatial perception SPM before being fed into the PAN. In SPM, the feature maps are combined with texture details mined from lower layers four times larger than themselves to enhance and integrate the selective information. Indirect feature exchange between distant layers ensures that low-level context can be dynamically preserved until the final output layer. The highest feature, F_{TFM}^{out} , also undergoes the texture federation TFM. This is because the deep layers experience more convolutional computation and correspondingly need to be supplemented with more underlying details. TFM aggregates the low-level spatial texture attention STA with the high-level fully connected attention FCA after spatial pyramid pooling-fast (SPPF) [43] processing to gain access to rich contexts. Subsequently, FSCHead performs a self-cooperation mechanism between neighboring layers on the three outputs of the PAN, F_{FSC}^{in0} , F_{FSC}^{in1} , and F_{FSC}^{in2} . The localization and classification tasks perform adjacent layer cooperation via deep decoupling.

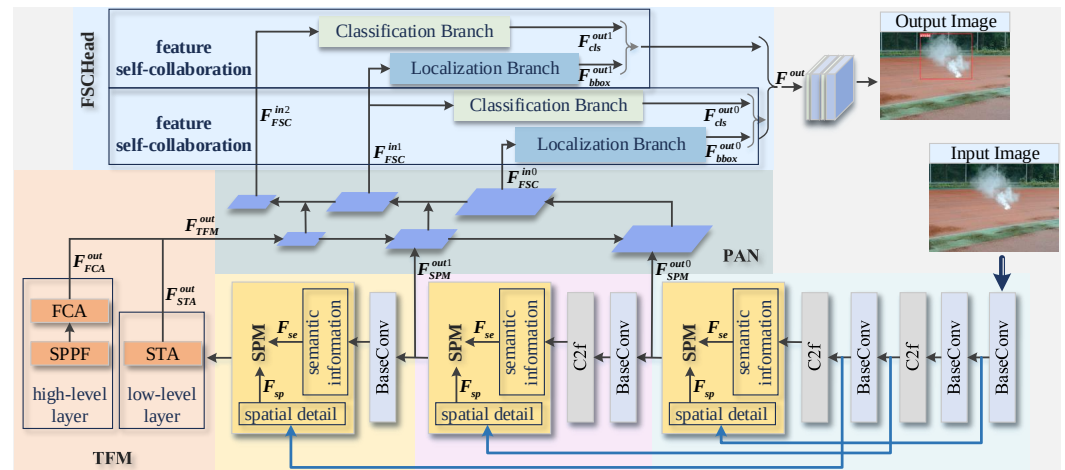


Figure 2. The whole network architecture of CLSANet. For catering to the different functions of high- and low-level features, cross-layer connections have divergent preferences during progressive feature fusion, allowing precise smoke detection. The backbone features are subjected to spatial restoration across four scales by SPM before being output into the pyramid PAN. The deepest coding layer also suffers from the TFM module to further complement its texture details. On the three outputs from the PAN, FSCHead carries out self-collaboration between adjacent layers for the localization and classification tasks.

3.2. Spatial Perception Module

As the network deepens, the receptive field is gradually getting larger and the semantic expression capability is also enhanced. But it also reduces the resolution of feature maps

and lots of spatial details become blurred [44]. Therefore, we design the SPM to perform cross-layer feature supplementation, as shown in Figure 3. It is employed to the input feature maps of pyramidal networks to enhance their spatial perception across four scales.

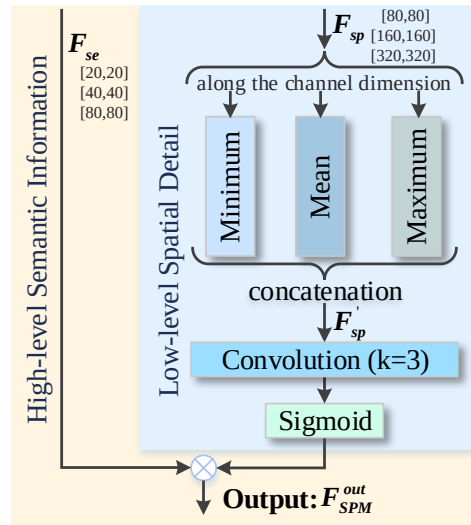


Figure 3. The detailed illustration of the proposed SPM. It reinforces deep feature F_{se} by spatial information F_{sp} from low layers across four scales, mitigating the dissipation of underlying attributes during encoding.

It is well known that the common smoke usually has two colors, black and white, and thus SPM simultaneously extracts both the maximum and minimum values on the shallow layer. In addition, because the pixel values of blurred boundaries and smoke regions are susceptible to the influence of ambient colors, SPM also takes the average operation to preserve the contextual details of smoke. Then after minimum, mean, and maximum operations, the three texture features are refined by a convolution with kernel 3. At last, the low-level spatial guidance maps are formed by a sigmoid function, which are loaded on the deep layers to bridge the loss of texture details. It is interesting to note that there is a fourfold difference in scale between the feature maps of the semantic and spatial layers in SPM. Such cross-layer connections allow texture information to be dynamically retained in each output layer of the network backbone, effectively addressing the issue of dissipation of contextual details during the feature encoding process.

As previously mentioned, when the low-level spatial feature is denoted as F_{sp} , the high-level semantic input is recorded as F_{se} , and the acquired intermediate feature containing textures by minimum, mean, and maximum operations is noted as F'_{sp} , the output of SPM F_{SPM}^{out} can be formulated as:

$$\begin{cases} F'_{sp} = Cat[\min(F_{sp}), \text{mean}(F_{sp}), \max(F_{sp})] \\ F_{SPM}^{out} = F_{se} \otimes \sigma(\text{Conv}(F'_{sp})) \end{cases}, \quad (1)$$

where minimum, mean, and maximum operations are along the channel dimension, Cat denotes the concatenation, $Conv$ means the convolution with a kernel size of 3, σ is the sigmoid function, and $\otimes$ denotes the element-wise multiplication.

3.3. Texture Federation Module

Many detailed features gradually dissipate after multi-layer convolutional operations. For this reason, we design the spatial perception SPM in the backbone. But at the deepest layer of feature encoding, this issue progressively accumulates, and it is insufficient to compensate for spatial details by only relying on SPM. Therefore, we devise the texture federation TFM after SPM modification. It is arranged in the last layer of the backbone to reinforce the semantic features and further supplement the meaningful spatial details

faded in the deep network maps. The elaborate structure of TFM is illustrated in Figure 4. Following the cross-layer design, STA is applied to preserve valuable low-level texture details. As for the high-level features, after the adaptive dimension of the SPPF structure, they are input into the FCA to strengthen the deep semantic information through fully connected attention. The final low- and high-level features are integrated and exported as the deepest feature encoding.

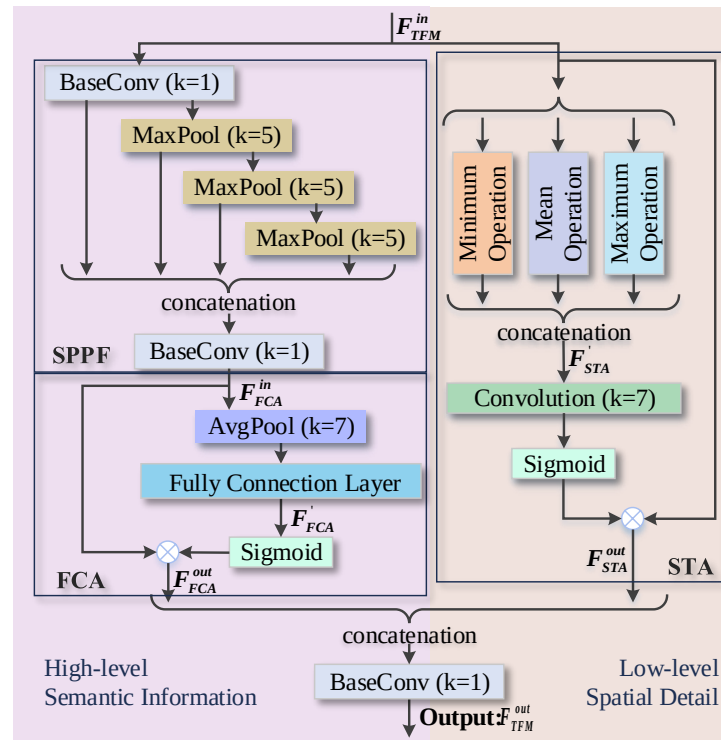


Figure 4. The elaborated structure of the TFM module. The STA is to preserve valuable texture details on the low-level path and the fully connected attention in FCA can reinforce the meaningful semantics of high-level layers.

We first describe the specific process of the low-level network path. Smoke, as a salient target with fuzziness, we similarly introduce the minimum, mean, and maximum values to perceive texture. Specifically, when the input feature of the TFM is notated as F_{TFM}^{in} , the preliminary texture-aware feature F_{STA}' is obtained by concatenating the results of the minimum, mean, and maximum computations. Then the F_{STA}' is subjected to a convolution operation with kernel 7 and a sigmoid function to obtain the final smoke spatial filter, which will be used to filter out the invalid noise and selectively enhance the meaningful texture details in input F_{TFM}^{in} . The output of STA module F_{STA}^{out} can thus be expressed as:

$$\begin{cases} F_{STA}' = Cat[\min(F_{TFM}^{in}), \text{mean}(F_{TFM}^{in}), \max(F_{TFM}^{in})] \\ F_{STA}^{out} = F_{TFM}^{in} \otimes \sigma(\text{Conv}(F_{STA}')) \end{cases}, \quad (2)$$

where minimum, mean, and maximum operations are along the channel dimension, Cat denotes the concatenation, $Conv$ means the convolution with a kernel size of 7, σ is the sigmoid function, and $\otimes$ denotes the element-wise multiplication. Note that the convolutional kernel size here is 7, which is because the TFM is applied to the highest layer of the backbone, and a larger receptive field can drum up the development of global contextual dependencies.

As for the high-level branch, the characteristic flow passes through SPPF and FCA sequentially. The SPPF structure [43] enables adaptive dimension generation via multi-scale

spatial containers with little increase in computational effort. When the input is denoted as F_{TFM}^{in} , the output F_{SPPF}^{out} of SPPF can be simply obtained by:

$$F_{SPPF}^{out} = BConv(Cat[BConv(F_{TFM}^{in}), \max(BConv(F_{TFM}^{in}))], \max(\max(BConv(F_{TFM}^{in})), \max(\max(\max(BConv(F_{TFM}^{in})))))), \quad (3)$$

where $BConv$ means the base convolution including convolution, batch normalization, and SiLU activation function, Cat represents the concatenation, and $\max$ means the maximum computation with a kernel of 5×5 .

Next, the F_{SPPF}^{out} serves as the input for our FCA structure and is denoted as F_{FCA}^{in} . One distinct difference between FCA and existing image attention approaches is that it breaks through the spatial location constraints caused by 2D convolution and instead takes advantage of only one fully connected layer. In FCA, the F_{FCA}^{in} first undergoes the average pooling to refine its spatial information, and then a simple one-layer full connection, where the operation objects of the neurons are the individual channels in the original image feature maps, reweights the global channels to introduce more possible feature representation. The obtained channel mask at this point is denoted as F'_{FCA} , and immediately after it is imported into the activation function for final channel scores. The output F_{FCA}^{out} of the FCA is derived by multiplying the initial input F_{FCA}^{in} with the channel scores, and it is formulated as:

$$\begin{cases} F'_{FCA} = FC(avg(F_{FCA}^{in})) \\ F_{FCA}^{out} = F_{FCA}^{in} \otimes \sigma(F'_{FCA}) \end{cases}, \quad (4)$$

where avg denotes the adaptive average pooling, FC means the linear fully connection, σ is the sigmoid function, and $\otimes$ denotes the element-wise multiplication.

Finally, to recover the spatial context details in the top-level encoding, we directly concatenate the F_{STA}^{out} after spatial texture attention on the low-level pathway and the F_{FCA}^{out} after fully connected attention on the high-level pathway, and then modify the channel numbers via a convolution operation so that the size of the input and output feature maps of the whole texture federation TFM module can be consistent. Therefore, the output feature F_{TFM}^{out} of TFM can capture richer smoke context and the process can be defined as:

$$\begin{cases} F_{STA}^{out} = STA(F_{TFM}^{in}) \\ F_{FCA}^{in} = SPPF(F_{TFM}^{in}) \\ F_{FCA}^{out} = FCA(F_{FCA}^{in}) \\ F_{TFM}^{out} = BConv(Cat[F_{STA}^{out}, F_{FCA}^{out}]) \end{cases}, \quad (5)$$

where Cat denotes the concatenation operator and $BConv$ means the base convolution.

3.4. Feature Self-Collaboration Head

After going through the feature pyramid, the network holds three branches at distinct scales. The conventional detection head, either anchor-based [45] or anchor-free [46], carries out the localization and classification tasks simultaneously, resulting in little communication between the different paths. Several studies [47–49] indicate that high-level features in deep layers encode the semantic information and acquire an abstract description of smoke, while low-level features in shallow layers retain spatial details for rebuilding the smoke boundaries. We hence propose the feature self-collaboration FSCHead, as presented in Figure 5. With cross-layer cooperation, the high-level paths are used only for the classification task, and the low-level layer is employed merely for smoke localization.

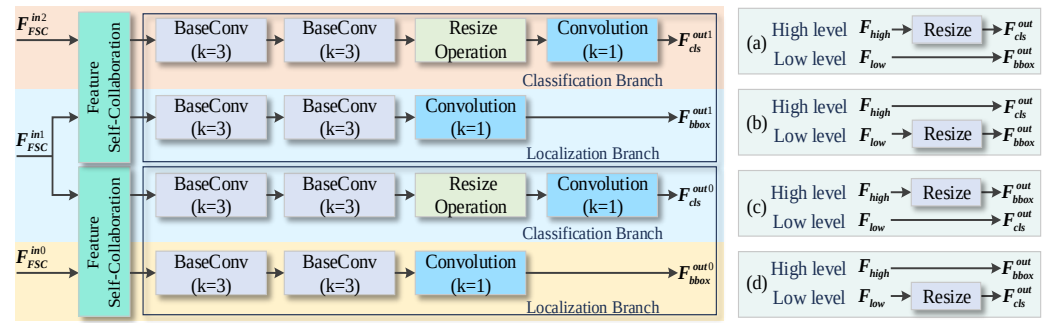


Figure 5. The specific design of FSCHead. Considering the decoupling of detection tasks on different branches, there are four strategies (a–d) available, and strategy (a) attains the best results consistent in both theory and experiment. As a result, FSCHead is designed to use high-level semantic features solely for smoke classification, while the low-level spatial details are only adopted for smoke localization. This cross-layer connection effectively removes redundant noise and refines the smoke localization and classification information.

The essential idea of our FSCHead module is to adapt the feature computation to fit the appropriate detection task based on the attribute preferences of different layers, instead of indiscriminately conducting localization and classification. Here are four exclusive strategies, as displayed in Figure 5a–d. Figure 5a,b classify smoke on the deep layers which are rich in semantics, and localization is performed on the low layers which contain spatial details. Due to the scale disparity on the respective branches, Figure 5a adjusts the classification branch to match the scale of the localization branch via up-sampling, and Figure 5b tunes the scale of the localization branch to be consistent with that of the classification branch by down-sampling. In contrast, Figure 5c,d implement smoke localization on the high level and the classification task is deployed on the low level. The former produces high-resolution feature maps after up-sampling the localization branch, while the latter down-samples the classification information and thus exports low-resolution detection maps.

In addition, referring to the anchor-free mechanisms [50] and taking into account the real-time requirements for smoke detection, each branch is concretely explored. The specific classification branch and localization branch are composed of two layers of base convolution and one 2D convolution. The base convolution with a kernel of 3 is designed to reinforce the comprehension of the features and recover discriminative smoke information. The mere 2D convolution, whose kernel size is 1, adaptively modifies the channel number of the features according to the assigned task. Figure 5a achieves the best results both theoretically and practically, and such feature self-collaboration mechanism with high layers for classification and low layers for localization can simply and directly eliminate the redundancy and preserve the meaningful smoke features. Denote the three inputs of different scales of FSCHead as F_{FSC}^{in0} , F_{FSC}^{in1} , and F_{FSC}^{in2} , respectively. The classification and localization outputs of the low-level layers are designated as F_{cls}^{out0} and F_{bbox}^{out0} , and F_{cls}^{out1} and F_{bbox}^{out1} separately represent the corresponding outputs on the high-level branches. The total output is noted as F^{out} , and then the transfer process of features in FSCHead can be described by:

$$\begin{cases} F_{cls}^{out0} = \text{Conv}(\text{Rs}(\text{BConv}(\text{BConv}(F_{FSC}^{in1})))) \\ F_{bbox}^{out0} = \text{Conv}(\text{BConv}(\text{BConv}(F_{FSC}^{in0}))) \\ F_{cls}^{out1} = \text{Conv}(\text{Rs}(\text{BConv}(\text{BConv}(F_{FSC}^{in2})))) \\ F_{bbox}^{out1} = \text{Conv}(\text{BConv}(\text{BConv}(F_{FSC}^{in1}))) \\ F^{out} = \text{Cat}[F_{cls}^{out0}, F_{bbox}^{out0}, F_{cls}^{out1}, F_{bbox}^{out1}] \end{cases} \quad (6)$$

where BConv means the base convolution, Conv means the convolution with a kernel of 1, Rs denotes the up-sampling, and Cat denotes the concatenation operator.

3.5. Hybrid Loss Function

In our network, a hybrid loss function is introduced to access the gap between the predicted results and the ground truths and direct the subsequent training. CLSNet has two computational outputs for each smoke target, the classification branch and the localization branch, and the total loss L_{Loss} is made up of three parts. The classification loss L_{cls} is computed on the classification branch, and the regression loss L_{reg} and confidence loss L_{dfl} are derived from the localization branch. The total loss L_{Loss} is expressed as:

$$L_{Loss} = \sum_{i=1}^n L_{loss}^i = \sum_{i=1}^n (L_{cls}^i + \alpha L_{reg}^i + \beta L_{dfl}^i), \quad (7)$$

where α and β are the weight parameters, $i \in [0, 1]$ denotes the two different forward paths in the head, and L_{loss}^i signifies the sum of losses on the forward path i .

Specifically, classification in smoke detection is a binary task, and a binary cross-entropy loss (BCE) is adopted to guide its optimization. BCE is easy to deploy and has a high computational efficiency. As for smoke localization, it is essentially a regression task, and a complete intersection over union loss (CIOU) is employed to penalize the inconsistent results on it. CIOU takes into account the intersection over union, centroid distance, and relative proportions between the predicted and true values. It is suitable for smoke detection tasks that have various shapes. Furthermore, since smoke detection is binarised and its distribution is highly steep, a distributional focal loss (DFL) is thus implemented in the regression branch to further refine the coordinates of the detection boxes after decoding their integrals. As a result, we can acquire the L_{cls}^i , L_{reg}^i , and L_{dfl}^i losses by:

$$\begin{cases} L_{cls}^i = BCE(cls^i, cls_{gt}^i) \\ L_{reg}^i = CIOU(bbox^i, bbox_{gt}^i) \\ L_{dfl}^i = DFL(bbox^i, bbox_{gt}^i) \end{cases}, \quad (8)$$

where $i \in [0, 1]$ means the two different forward paths in the head, cls and $bbox$ denote the predicted classification probability and $bbox$ coordinate, respectively, and cls_{gt} and $bbox_{gt}$ represent the corresponding ground truths, respectively.

4. Experimental Evaluation

In this section, we first present the two databases used for the experiments, and immediately after, the specific training details are described. Then, we compare CLSNet with state-of-the-art object detection methods and several available smoke detection algorithms. Thirdly, an ablation analysis for each component is organized. In addition, we present the model size and time complexity to validate that the CLSNet is lightweight and efficient. Furthermore, to verify that our algorithm can accurately focus on the smoke regions, we carry out a feature visualization discussion. Finally, the detection results of our CLSNet in real scenarios are given, and we analyze the success and failure cases, respectively.

4.1. Database

There are two available databases used in this paper, which are the annotated real smoke database of Xi'an Jiaotong University (XJTU-RS) [51] and the real smoke and forest background database of University of Science and Technology of China (USTC-RF) [52], respectively. Their statistics are shown in Table 1.

Table 1. Statistical data of the smoke databases.

Database	Total	Training	Validation	Testing
XJTU-RS [51]	6845	4791	1369	685
USTC-RF [52]	12,620	3155	3155	6310

XJTU-RS contains a total of 6845 images, all of which are taken in real scenes and annotated manually. There are 4791 images in the training set, 1369 images in the validation set, and 685 images in the test set. To make the trained model meet the requirements for practical application, XJTU-RS acquires the original images from two real but unlabeled benchmark databases: Keimyung University (KMU) [53] and University of Science and Technology of China (USTC) [52]. Their smoke images are taken in various natural scenarios, e.g., indoor smoke, forest smoke, playground smoke, and farmland smoke. To enhance the robustness of XJTU-RS, stratified random sampling is arranged in almost every scene. Finally, the smoke regions are demarcated with rectangular boxes and corresponding labels supported by a published annotation tool. Some sample images of XJTU-RS are displayed in Figure 6a, and they are manually annotated and quite realistic.

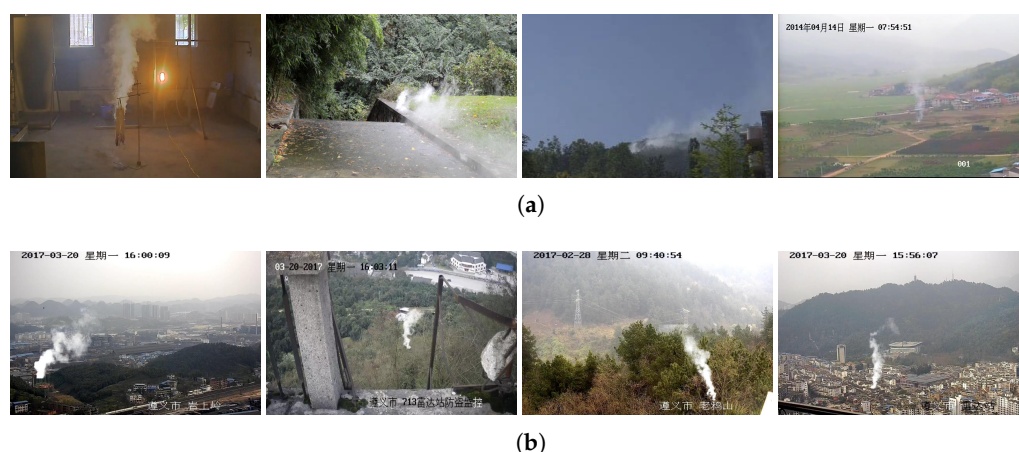


Figure 6. Sample images of (a) the XJTU-RS [51] and (b) the USTC-RF [52] databases. The XJTU-RS database is manually labeled from realistic images, while the USTC-RF database is synthesized from forest images and simulated smoke frames jointly.

USTC-RF comprises 12,620 synthesized smoke images, out of which 3155 are selected for training, 3155 for validation, and the remaining 6310 for testing. In this dataset, in order to precisely extract the smoke, the real smoke image is photographed against a green background, and based on it, the pure smoke regions are derived. There are 2800 smoke frames selected from 10 smoke videos with green backgrounds. To further increase the smoke diversity, rendering software is employed to generate the smoke plumes prepared for the synthesized images, where the initial flow rate, airflow, illumination, and viewing angle are all set randomly. A total of 1000 synthetic smoke plumes are obtained from such simulations. At last, the smoke plumes are freely changed in shape and inserted into 12,620 forest background images at random locations. Each synthetic image includes only one wisp of smoke and the location of the smoke can be provided automatically during the insertion process. Examples of images from USTC-RF are presented in Figure 6b.

4.2. Implementation Details

The experiments are implemented on a PC with an Intel CPU i7-8700K @ 3.2 GHz, 16 GB of RAM, and an NVIDIA RTX2080Ti with 11 GB. Our proposed CLSNet is trained using the Pytorch framework, and it is lightweight with only 2.38 M, allowing real-time detection on such a hardware platform. In addition to the regular convolution and C2F operations, the spatial perception SPM and texture federation TFM modules are introduced in the backbone for biased cross-layer feature communication. PAN pyramid is closely followed for multi-scale feature fusion, and finally, the FSCHead deploys the feature self-cooperation mechanism on the high-level and low-level branching channels. It also applies the anchor-free mechanism to alleviate the inherent conflict between classification and localization. During the training period, CLSNet is optimized by the stochastic gradient descent (SGD) with a weight decay of 0.0005, a momentum of 0.937, and an initial

learning rate of 0.01. Mosaic strategy [54] which splices four random images is exploited to augment the data and complicate the detection background. As for the sample assignment strategy, CLSNet employs the task-aligned assigner which selects positive data based on the weighted scores of classification and regression. The threshold of intersection over union for non-maximum suppression is set to 0.7 and the coefficients of classification loss, regression loss, and distributional focal loss are 0.5, 7.5, and 1.5, respectively.

To avoid overfitting, CLSNet is trained on the training database and validated on the validation database. The images for training and validation are different. Moreover, the training is automatically terminated when the network has no improvement in the last 50 epochs in training.

In this paper, we adopt widely used metrics to evaluate the performance of CLSNet, including the average precision (AP) and the average recall (AR). In addition, the average precision of detecting small (AP_S), medium (AP_M), and large (AP_L) areas are also explored to further delve into the superiority of our model. Note that the smoke regions in the USTC-RF database are all larger than the scale of the defined large area, i.e., 96×96 . It means that the AP_S and AP_M are null and the values of AP_L and AP are the same. So in the experimental results, we no longer give the AP_S , AP_M , and AP_L on USTC-RF database.

4.3. Comparison with Regular Objects Detection Methods

In this section, some popular and state-of-the-art object detection methods, which are Faster R-CNN [55], SSD [12], RetinaNet-50 [13], RetinaNet-101 [13], EfficientDet [14], YOLOv5 [15], YOLOv7 [56], YOLOX [16], EdgeYOLO-Tiny [17], and EdgeYOLO-S [17], are compared with the proposed CLSNet on the same test environment. We also conducted experiments on two transformer-based networks, an improved YOLOv5 based on transformer prediction head (TPH-YOLOv5) [57] and a real-time detection transformer (RT-DETR) [58], respectively. The AP_S , AP_M , AP_L , AP, and AR of different networks on XJTU-RS database are presented in Table 2, but since the AP_S and AP_M are null and the values of AP_L and AP are the same on the USTC-RF database, there are only the corresponding AP and AR for the different algorithms given in Table 3. The best results in Tables 2 and 3 are shown in bold font.

Table 2. Experimental results of the proposed and ten state-of-the-art methods on XJTU-RS.

Method	$AP_S/\%$	$AP_M/\%$	$AP_L/\%$	AP/%	AR/%
Faster R-CNN [55]	38.5	60.2	68.0	65.5	71.2
SSD [12]	49.6	61.2	67.9	65.9	68.7
RetinaNet-50 [13]	52.9	63.2	71.8	68.9	70.0
RetinaNet-101 [13]	48.6	63.2	72.7	69.1	70.1
EfficientDet [14]	46.2	63.3	67.6	65.4	69.6
YOLOv5 [15]	51.3	63.1	69.0	66.1	70.3
YOLOv7 [56]	43.8	63.2	74.1	70.9	70.0
YOLOX [16]	48.1	63.2	74.6	71.3	70.1
EdgeYOLO-Tiny [17]	45.5	63.0	74.3	71.1	70.2
EdgeYOLO-S [17]	45.2	63.5	74.2	71.1	70.5
TPH-YOLOv5 [57]	51.8	63.7	74.4	71.3	71.9
RT-DETR [58]	47.0	65.1	74.3	71.4	71.0
CLSNet (ours)	50.7	66.1	76.2	73.3	72.1

Table 3. Experimental results of the proposed and ten state-of-the-art methods on USTC-RF.

Method	AP/%	AR/%
Faster R-CNN [55]	78.4	82.5
SSD [12]	81.0	85.2
RetinaNet-50 [13]	82.9	86.0
RetinaNet-101 [13]	86.8	90.7
EfficientDet [14]	60.1	63.7
YOLOv5 [15]	88.5	90.7
YOLOv7 [56]	90.5	92.4
YOLOX [16]	90.7	92.3
EdgeYOLO-Tiny [17]	91.2	92.8
EdgeYOLO-S [17]	91.1	92.8
TPH-YOLOv5 [57]	91.0	92.9
RT-DETR [58]	90.7	93.1
CLSANet (ours)	94.4	95.3

From the experimental statistics, it is quite obvious that our algorithm achieves the best performance on both the XJTU-RS and USTC-RF databases. Specifically, benefiting from cross-layer connections, our CLSANet attains the best results on all metrics except the AP_S on the XJTU-RS which is only 2.2% less than the optimal result of RetinaNet-50 in Table 2. The feature extraction of Faster R-CNN and SSD is just a simple stack of several convolutional layers without feature modification and supplementation among distinct levels. As a result, the AP of Faster R-CNN and SSD is 65.5% and 65.9% on XJTU-RS and 78.4% and 81.0% on USTC-RF, respectively. RetinaNet-50 and RetinaNet-101 have much better comprehensive capabilities than them. It is attributed to the residual unit in RetinaNet, which also belongs to a kind of simple cross-layer connection. What's more, RetinaNet-101 has more residual units than RetinaNet-50. It results in a more sufficient exchange of indirect information between the distant layers, and thus RetinaNet-101 is slightly more effective than RetinaNet-50. In contrast, EfficientDet produces the lowest AP on both databases, 65.4% and 60.1%, and its AR is the second worst on XJTU-RS and the worst on USTC-RF, 69.6% and 63.7%, respectively. The bi-directional feature pyramid network (BiFPN) [59] in EfficientDet allows cross-layer fusion on feature encoding maps of the backbone output at three scales. Evidently, EfficientDet trades information only on depth-coded feature maps and it is not sufficient for the detection of fickle smoke.

As for the YOLO series, which has gained enormous popularity in recent years, it can be found that their entire performance is relatively superior. To be specific, the bottleneck cross-stage partial network (CSPNet) of YOLOv5 [60] enables cross-layer fusion of internal information through feature concatenation. Referring to VoVNet's [61] strategy, YOLOv7 modifies the CSPNet module as an efficient layer aggregation network (ELAN) [56]. It adopts a stack structure to deepen the feature transmission path. Consequently, YOLOv7 which has more cross-layer communication convincingly behaves better than YOLOv5. Furthermore, YOLOX applies the anchor-free mechanism in its head, and such decoupling between the classification and localization is also employed in our method. As shown in Table 2, the AP of YOLOX on the XJTU-RS is relatively acceptable with 71.3%. In addition, EdgeYOLO is designed specifically for edge devices. From Table 3, the EdgeYOLO-Tiny has the sub-optimal results on the USTC-RF in terms of AP with 91.2%. It is because the EdgeYOLO is designed to further enrich the feature diversity of input images through a flexible data augmentation. The detection performance of both EdgeYOLO-Tiny and EdgeYOLO-S on the XJTU-RS and USTC-RF databases is favorable and indistinguishable.

In recent years, many teams have developed various object detection models using the transformer architecture. We compare CLSANet with two transformer-based models TPH-YOLOv5 and RT-DETR. The results are shown in Tables 2 and 3. Since the transformer's self-attention mechanism allows the model to simultaneously consider all positions of the input sequence and learn the dependencies between different positions, the results achieved by TPH-YOLOv5 and RT-DETR on both databases are significant. In particular, on

XJTU-RS database, RT-DETR achieves the AP only second to our algorithm, which is 71.4%. The global attention in transformer allows the model to capture global dependencies, which increases the complexity while delivering fine detection performance. The number of parameters for TPH-YOLOv5 and RT-DETR are 41.49 M and 31.99 M, respectively. Compared to our CLSNet which is only 2.38 M, it is clear that transformer-based algorithms are not suitable for resource-limited edge devices.

The comparison among these object detection algorithms indicates that the smoke detection performance can be improved to varying degrees by reinforcing cross-layer feature exchange. Based on the results in Tables 2 and 3, we can conclude that smoke detection, as a simple binary task, is not boosted as the model deepens, but relies more on the specific network design. Our CLSNet specifically devises three strategies, SPM, TFM, and FSCHead, to realize dynamic cross-layer communication with different biases based on smoke attributes. As a result, it obtains the best smoke detection results on both XJTU-RS and USTC-RF databases.

4.4. Comparison with Special Smoke Detection Models

To further validate the performance of the CLSNet, we select several smoke detection networks in recent years and conduct experiments both on XJTU-RS and USTC-RF. As we can see, they are DCNN [24], W-Net [36], SASC-YOLOX [51], Deep CNN [62], STCNet [63], and MVMNet [64], respectively. Since most available warning networks are only intended to classify smoke, for the sake of a fair comparison, we uniformly substitute a conventional anchor-free head for the fully connected layer at the end of the smoke classification network. The corresponding results on XJTU-RS and USTC-RF are listed in Tables 4 and 5, respectively, with the best results in bold font.

Table 4. Comparison results of different smoke detection algorithms on XJTU-RS.

Method	$AP_S/\%$	$AP_M/\%$	$AP_L/\%$	$AP/\%$	$AR/\%$
DCNN [24]	46.0	62.5	67.7	66.0	70.6
W-Net [36]	44.1	59.8	67.1	65.1	70.4
SASC-YOLOX [51]	53.5	64.8	75.5	72.7	71.4
Deep CNN [62]	36.8	56.2	65.1	62.7	67.8
STCNet [63]	37.0	49.8	60.3	57.2	63.5
MVMNet [64]	53.5	65.2	73.9	71.4	70.7
CLSNet (ours)	50.7	66.1	76.2	73.3	72.1

Table 5. Comparison results of different smoke detection algorithms on USTC-RF.

Method	$AP/\%$	$AR/\%$
DCNN [24]	85.3	87.6
W-Net [36]	77.0	80.7
SASC-YOLOX [51]	92.3	94.0
Deep CNN [62]	87.6	89.6
STCNet [63]	70.9	75.5
MVMNet [64]	88.8	90.7
CLSNet (ours)	94.4	95.3

Specifically, Cao et al. [63] devised a novel spatiotemporal cross network (STCNet). It involves a spatial pathway to capture the smoke texture and a temporal route to extract motion information. It can be seen that STCNet has the most poor detection performance among these methods on the two databases, with AP of 57.2% and 70.9%, and AR of 63.5% and 75.5%, respectively. This may be because the STCNet, as a residual frame structure for video smoke detection, is not appropriate for training sets which are labeled image by image in XJTU-RS and USTC-RF.

An energy-efficient deep CNN was proposed by Khan et al. [62]. It slightly outperforms STCNet in experiments and is the second worst on XJTU-RS. However, Deep

CNN has only simple convolution and fully connected layers but without any cross-layer communication, which results in the loss of much critical smoke information as the network deepens, especially small smoke features. Consequently, Deep CNN is the least in terms of AP_S with only 36.8%.

In order to enhance the cross-layer feature integration, Yuan et al. [36] designed a wave-shaped neural network (W-Net). It copies and resizes the encoding outputs, and then concatenates them between peaks and troughs. As shown in Tables 4 and 5, the smoke detection performance of the W-Net is further improved. However, its cross-layer communication is simply splicing and does not individually utilize the advantages of different layers.

Gu et al. [24] gave a deep dual-channel neural network (DCNN). The texture details and base information of the smoke are derived from the two channels separately, which can alleviate the dissipation of smoke details. It is more capable than W-Net and has AP of 66.0% and 85.3% on XJTU-RS and USTC-RF, respectively.

In addition, a multi-orientated detection based on a value conversion-attention mechanism module and mixed-NMS (MVMNet) was proposed by Hu et al. [64]. As shown in Tables 4 and 5, the AP of MVMNet is only 1.9% and 5.6% less than that of our CLSNet on the two databases, respectively. It is because MVMNet deploys a conversion-attention mechanism to reinforce the spatial and textural information of smoke.

Wang et al. [51] presented a self-attention and self-cooperation YOLOX (SASC-YOLOX). It obtains attention weights from the low layers and realizes feature reinforcement between the high-level semantics and the low-level spatial information. Such cross-layer connection is endowed with preliminary feature complementary, and accordingly, it is second only to the proposed CLSNet in terms of comprehensive capability in experiments.

Lastly, CLSNet attains the best detection results on both XJTU-RS and USTC-RF, with AP of 73.3% and 94.4%, and AR of 72.1% and 95.3%, respectively. It justifiably validates the effectiveness of our optimization strategies. Cross-layer communication can enhance the network's ability to extract smoke features, and this individualized transfer between high and low layers with bias can further reduce noise and refine valuable smoke features.

4.5. Ablation Study

We first analyze the effectiveness of each module, that is, SPM, TFM, and FSCHead. The ablation experiments are shown in Table 6. Then, a clear comparison between the convolutional block attention module (CBAM) and our proposed SPM and TFM modules is provided in Table 7. Finally, four strategies of feature self-collaboration mechanism are compared to demonstrate the FSCHead design and the corresponding results are provided in Table 8. All ablation experiments are conducted on both XJTU-RS and USTC-RF databases, and the best results are denoted in bold font.

Table 6. Ablation analysis of each component in CLSNet on XJTU-RS and USTC-RF.

No.	[SPM]	[TFM]	[FSCHead]	XJTU-RS [51]				USTC-RF [52]		
				$AP_S/\%$	$AP_M/\%$	$AP_L/\%$	$AP/\%$	$AR/\%$	$AP/\%$	$AR/\%$
(1)				43.9	59.3	71.5	68.1	67.8	90.1	91.8
(2)	✓			51.6	62.9	75.0	71.7	70.7	92.6	93.8
(3)		✓		45.2	63.9	74.5	71.5	70.6	91.8	93.1
(4)			✓	50.0	63.2	74.9	71.6	71.0	91.5	93.0
(5)	✓	✓		45.2	65.5	75.6	72.6	71.5	93.1	94.1
(6)	✓		✓	46.1	65.1	75.9	72.9	72.0	93.6	94.5
(7)		✓	✓	49.0	64.4	75.1	72.2	70.8	93.0	94.2
(8)	✓	✓	✓	50.7	66.1	76.2	73.3	72.1	94.4	95.3

Table 7. Comparison of CBAM with our proposed SPM and TFM modules on XJTU-RS and USTC-RF.

[Module 1]	[Module 2]	XJTU-RS [51]					USTC-RF [52]	
		$AP_S/\%$	$AP_M/\%$	$AP_L/\%$	$AP/\%$	$AR/\%$	$AP/\%$	$AR/\%$
CBAM [65]	CBAM [65]	53.7	65.1	74.7	72.0	71.6	93.3	94.6
CBAM [65]	TFM	46.4	64.9	75.7	72.7	71.7	93.7	94.7
SPM	CBAM [65]	48.1	64.6	73.9	71.3	70.9	93.4	94.7
SPM	TFM	50.7	66.1	76.2	73.3	72.1	94.4	95.3

Table 8. Performance of four feature self-collaboration strategies for FSCHead on XJTU-RS and USTC-RF.

Method	XJTU-RS [51]					USTC-RF [52]	
	$AP_S/\%$	$AP_M/\%$	$AP_L/\%$	$AP/\%$	$AR/\%$	$AP/\%$	$AR/\%$
FSCHead (a)	50.0	63.2	74.9	71.6	71.0	91.5	93.0
FSCHead (b)	45.1	63.0	74.9	71.4	70.3	89.1	90.4
FSCHead (c)	42.9	59.6	72.1	68.6	68.2	87.9	89.8
FSCHead (d)	45.4	60.7	72.1	68.8	68.3	88.1	89.8

4.5.1. Effectiveness of SPM

As shown in Table 6, the addition of SPM obviously brings a boost in terms of all indicator scores. When the model is only equipped with SPM, the AP increases by 5.29% and 2.77% from baseline on both databases and the AR is improved by 4.28% and 2.14%, respectively. When using SPM as a variable, by comparing (1) and (2), (3) and (5), (4) and (6), and (7) and (8), the AP scores are improved with a margin of 5.29%, 1.54%, 1.82%, and 1.52% on XJTU-RS, respectively, and 2.77%, 1.42%, 2.30%, and 1.51% on USTC-RF, respectively. In particular, there is an increase of 17.54% in AP_S on XJTU-RS database, which demonstrates that the small-scale smoke features are effectively preserved. The SPM can retain the low-level spatial details through three cross-layer connections until the highly encoded semantic information.

4.5.2. Effectiveness of TFM

In the baseline of the experiment (1), the highest layer of the backbone is a spatial pyramid pooling (SPP) module. Through comparing (1) and (3), a gain of 4.9% and 1.89% in AP is obtained, respectively, with 4.13% and 1.42% of the improvement occurring in AR . Through comparing (1) and (3), the gains of AP are obtained as 4.9% and 1.89%, and 4.13% and 1.42% improvement occurs in AR , respectively. When comparing (2) and (5), (4) and (7), and (6) and (8), the network applied with the TFM structure convincingly has better metrics on XJTU-RS and USTC-RF. It verifies that TFM exploits FCA and STA attention to perform cross-layer feature filtering along two different dimensions at the deep encoding layer and then achieve the perception of meaningful information.

4.5.3. Effectiveness of FSCHead

Experiment (4) is optimized with FSCHead only. When compared to (1), it increases AP_S , AP_M , AP_L , AP , and AR by 13.90%, 6.58%, 4.76%, 5.14%, and 4.72% on XJTU-RS, respectively, and AP and AR are improved by 1.55% and 1.31% on USTC-RF, respectively. Taking FSCHead as a variable, by comparing (1) and (4), (2) and (6), (3) and (7), and (5) and (8), AP obtains an absolute gain of 1.55%, 1.08%, 1.31%, and 1.40% on the synthetic USTC-RF, respectively. And it mitigates the dissipation of small-scale smoke details during transmission. As shown in Table 6, AP_S intuitively increases by 13.90%, 8.41%, and 12.17% on XJTU-RS by the comparison between (1) and (4), (3) and (7), and (5) and (8), respectively. It sufficiently proves that the feature self-collaboration mechanism of FSCHead can preserve the important features and remove redundancy through cross-layer communication.

4.5.4. Comparison with CBAM

CBAM is a simple but effective attention module for feed-forward convolutional neural networks [65]. It optimizes the network in both channel and spatial dimensions, allowing the improved model to acquire meaningful features from both channel and spatial perspectives. Like cross-layer connection, it is a common way to enhance feature extraction. To further validate the effectiveness of our cross-layer connection strategy in smoke feature extraction and detection, we replace the SPM and TFM modules in CLSNet with CBAM in turn for experiments, and the comparison results are listed in Table 7. The first CBAM-equipped model achieves favorable results on both XJTU-RS and USTC-RF databases and even outperforms some conventional object detection networks, such as RetinaNet-101 [13] and YOLOv5 [15]. However, the overall performance of our proposed CLSNet is still superior to that of CBAM-equipped networks. Except for the AP_5 on XJTU-RS, which is second only to the optimal result of 53.7%, CLSNet outperforms the CBAM-equipped model in all metrics.

4.5.5. Configurations of FSCHead

To determine the appropriate FSCHead design for retaining meaningful smoke features, the experiments are carried out sequentially on four different feature self-collaboration strategies, which are FSCHead (a) to (d), as illustrated in Figure 5a–d. The test results are reported in Table 8, and it can be easily noticed that the detection performance of FSCHead (c) and (d) significantly lags behind FSCHead (a) and (b) on both XJTU-RS and USTC-RF. It may be because the higher layers can accurately extract semantic information after more computation, which is beneficial for classification, whereas the lower layers which have smaller receptive fields can retain spatial details and thus guide the localization. The performance of FSCHead (a) is again superior to that of FSCHead (b), and the best AP on the two databases are 71.6% and 91.5%, respectively. It also suggests that the more information contained in the feature map, the stronger the network is. The variants of the four different configurations confirm that the high-level branch for classification and the low-level branch for localization is an adaptive scheme to the smoke detection task and that the design of our FSCHead plays an important role in CLSNet.

4.6. Model Size

The smoke warning is usually used in surveillance systems, so the number of parameters and computational complexity are crucial metrics to evaluate the suitability of the model for resource-constrained edge devices. In this section, we compare CLSNet with some conventional methods in terms of AP , AR , the number of parameters (Params), and frames per second (FPS). The comparative results on XJTU-RS and USTC-RF are shown in Table 9, with the best values in bold.

Table 9. Study of model size and FPS for different detection networks on XJTU-RS and USTC-RF.

Method	XJTU-RS [51]			USTC-RF [52]			Params
	AP/%	AR/%	FPS	AP/%	AR/%	FPS	
Faster R-CNN [55]	65.5	71.2	7.790	78.4	82.5	3.849	41.35 M
SSD [12]	65.9	68.7	39.214	81.0	85.2	24.138	23.75 M
RetinaNet-50 [13]	68.9	70.0	8.421	82.9	86.0	6.917	36.33 M
YOLOv7 [56]	70.9	70.0	25.242	90.5	92.4	28.736	9.14 M
YOLOX [16]	71.3	70.1	232.019	90.7	92.3	265.252	8.94 M
EdgeYOLO-Tiny [17]	71.1	70.2	123.457	91.2	92.8	128.370	5.81 M
EdgeYOLO-S [17]	71.1	70.5	92.851	91.1	92.8	93.545	9.86 M
CLSNet (ours)	73.3	72.1	239.981	94.4	95.3	219.491	2.38 M

It is obvious that the two-stage Faster R-CNN [55] has the most parameters. Accordingly, its computational speed is the slowest among all algorithms, with only 7.790 FPS

and 3.849 FPS on XJTU-RS and USTC-RF, respectively. The bad capability does not allow Faster R-CNN to be applied to smoke detection in the real world. SSD [12] and RetinaNet-50 [13] mainly rely on stacking layers to extract features. This simple deepening of the network increases the number of parameters while increasing the computational complexity. Therefore, SSD and RetinaNet-50 are unsuitable for polymorphic smoke. In particular, the backbone of RetinaNet-50 adopts the residual network (ResNet), which is composed of multiple densely connected convolutional layers. With 36.33 M parameters, its FPS on the two databases are only 8.421 and 6.917.

In YOLO series, such as YOLOv7 [56], YOLOX [16], and EdgeYOLO [17], the network design becomes more exquisite, and cross-layer connections are gradually employed to enhance feature exchange over long distances. Therefore, their accuracy and model size are further developed. The EdgeYOLO-Tiny is only 5.81 M and its *AP* and *AR* on USTC-RF are only second best to our algorithm. The proposed CLSNet achieves consistently optimal results among all the methods, with the best accuracy and recall on both XJTU-RS and USTC-RF. What's more, its model size is only 40.96% of that of the sub-optimal EdgeYOLO-Tiny, i.e., 2.38 M. It verifies the cross-layer connection can ensure the network has superior smoke awareness even with a few parameters. As for the detection speed, it has the optimal FPS on XJTU-RS with 239.981 and the sub-optimal 219.491 on the USTC-RF database, second only to YOLOX. Our algorithm can fully satisfy the real-time requirement of smoke alarm systems. We can declare that the CLSNet is the best candidate to detect smoke in constrained environments and resource-limited equipment.

4.7. Feature Visualization

To validate the accuracy of CLSNet's attention to specific smoke areas, we performed a visualization analysis. The input samples and corresponding results are presented in Figure 7. Because the images in the XJTU-RS database are realistic and cover complex scenes, the weights used for the visualization are obtained on XJTU-RS.

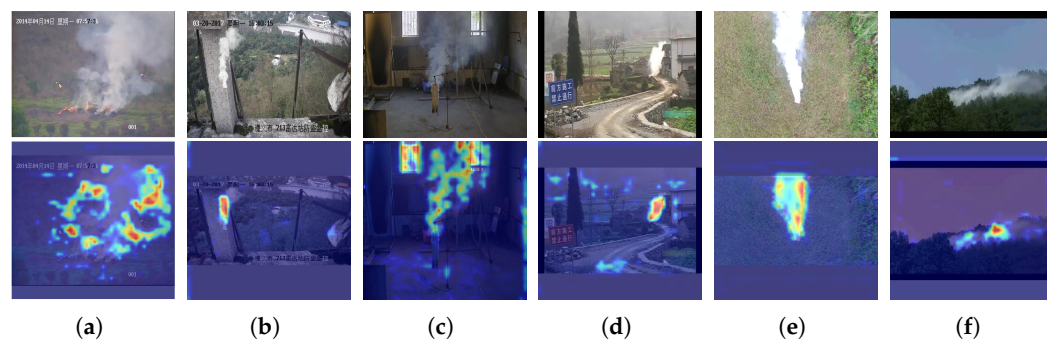


Figure 7. Feature visualization of the proposed CLSNet. (a–f) The first line is the original input and the second line is the corresponding visualization map. The CLSNet can accurately perceive smoke in different scales and forms.

We adopt the classical gradient-weighted class activation mapping (Grad-CAM) method for visualization [66]. It uses the score values corresponding to the target categories to compute the gradient of the feature map, thus generating a heat map to show the important image regions in network prediction. As we can see, the proposed CLSNet model can accurately perceive smoke at various scales, such as large smoke in Figure 7a and small smoke in Figure 7d. Moreover, different concentrations of smoke can also be well focused by CLSNet, such as light smoke in Figure 7c and thick smoke in Figure 7e.

4.8. Practical Application

Smoke detection is a key part of disaster and incident monitoring, so to meet the requirements of real-world applications, CLSNet needs to have favorable reliability in both the trained databases and the untrained scenarios. To validate that the training of the

model is not overfitting and that the CLSNet we obtained has a strong generalization ability, we first randomly select some images from both XJTU-RS and USTC-RF and present their detection results in Figure 8. Then, we download several smoke images from the Internet and the results of CLSNet are visualized in Figure 9.

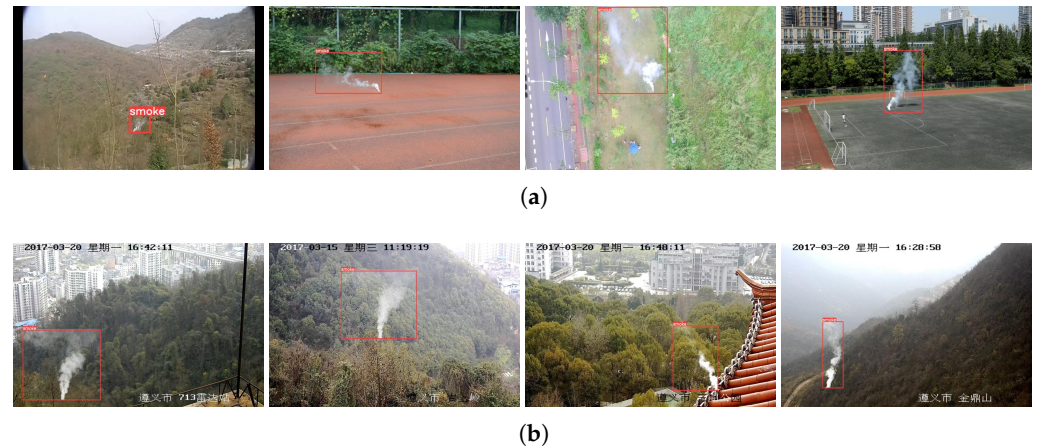


Figure 8. Visual detection for smoke images on (a) XJTU-RS and (b) USTC-RF databases. It can be seen that CLSNet has excellent detection performance on the test images.



Figure 9. Visual detection for smoke images in the real application. These images are randomly downloaded from the Internet and untrained, and our CLSNet can still accurately monitor smoke.

There are four representative smoke images selected from XJTU-RS here and their visualization results of the CLSNet are shown in Figure 8a. As we can see, CLSNet is not affected by the surrounding gray blurred mountains and exhibits remarkable detection of small-scale smoke in the first image. The second and third images revealed that CLSNet can accurately perceive the smoke although it is faint and blended with the background. In addition, as illustrated in the last two, there is no doubt that CLSNet has robust discrimination of obvious white smoke and large regions of smoke blocks. As for USTC-RF, as presented in Figure 8b, we chose the synthetic images from diverse scenarios. It is clear that our CLSNet is able to precisely localize the smoke plume under different drift directions and diffusion conditions.

To further validate the generalization ability of our network, in Figure 9, we download smoke with different morphologies from the Internet, where the images in the first row have high resolution. It can be seen that CLSNet has satisfactory smoke detection performance in these untrained natural environments, including smoke in black and white, smoke in thick and light, smoke at varied scales, smoke with distinct diffusion directions and

morphologies, and images of different capture quality. It is mainly because the cross-layer connections in CLSNet retain as much meaningful smoke information as possible and cut down the interference of noise under variable conditions. All in all, based on these intuitive results, we can claim that our method's training is appropriate. CLSNet has strong robustness and remarkable performance in various smoke environments.

4.9. Failures Analysis

When detecting smoke images, CLSNet inevitably makes some mistakes. We analyze the failure cases from the Internet and the training database, and some samples are displayed in Figure 10.

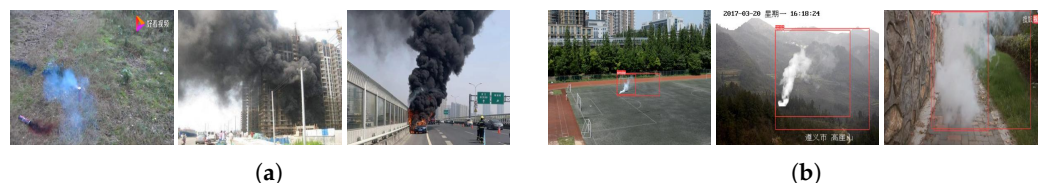


Figure 10. Samples of failed detection by CLSNet. These are the defective results on (a) blue and black smoke and (b) accurate smoke areas respectively.

As we can see, there are two main categories of failure detection. In Figure 10a, CLSNet is unable to correctly identify some smoke with special colors such as blue and black, which is mainly due to the lack of corresponding smoke samples in the training database. It can be addressed by further expanding the database. In Figure 10b, the shortcoming is mainly the inaccurate delineation of the smoke region. CLSNet can successfully detect smoke in these three images, but there are overlaps between the output detection boxes. It may be due to an improper threshold of the non-maximum suppression, resulting in some overlapping frames not being completely suppressed. This can be mitigated by tuning some hyper-parameters according to the specific task. In addition, it inspires us to consider further design and refinement on the decision-making strategy of CLSNet. When the model's calculation of the location and size of the smoke is sufficiently deterministic, it is possible to avoid generating multiple close but slightly different detection boxes.

5. Conclusions

Automatic smoke detection plays an essential role in monitoring fires and safeguarding the ecological environment. In this paper, we propose a lightweight CLSNet with only 2.38 M to detect smoke in real-time. The major difference with the existing algorithms is that CLSNet exploits multiple cross-layer connections with a bias towards high and low-level features. It is based on the fact that the lower layers contain richer spatial details while the higher layers have more precise semantic information. Specifically, we first design the spatial perception SPM in the backbone, which spans four scales and passes shallow features to high layers to compensate for the dissipation of spatial details. Then, the texture federation TFM is proposed and applied to repair the final feature encoding. The STA and FCA in the TFM module assist the cross-layer feature integration along the spatial and channel dimensions, where the full connection layer in the FCA breaks the constraints of spatial location imposed by convolution and facilitates the TFM to acquire more robust semantics. Finally, we propose FSCHead guided by the feature self-collaboration mechanism to preserve the significant information and eliminate redundancy in a concise way. It decouples the detection task and explicitly deploys the task of localization on low-layer branches while smoke classification on high-layer pathways. We conduct extensive experiments on the XJTU-RS and USTC-RF databases, and the results show that our CLSNet achieves state-of-the-art performance compared to both regular object detection algorithms and specialized smoke detection methods. Compared to the baseline, our algorithm improves the precision by 7.64% and 4.77% on XJTU-RS and USTC-RF, respectively.

In future work, our main research direction is firstly to address the existing deficiencies of CLSANet. Improve its detection ability for various colors of smoke, and optimize the size and scale of its anchor boxes to better match the smoke outside the training database. Secondly, it will be deployed to specific edge devices such as unmanned aerial vehicles (UAVs), Internet of Things (IoT) systems, and forest monitoring for field testing. CLSANet is expected to be easily transplanted to portable hardware and maturely applied to fire warnings.

Author Contributions: Conceptualization, J.W.; methodology, J.W. and X.Z.; software, J.W.; validation, J.W. and X.Z.; formal analysis, J.W.; investigation, J.W., X.Z. and C.Z.; resources, X.Z.; data curation, J.W. and X.Z.; writing—original draft preparation, J.W.; writing—review and editing, J.W., X.Z. and C.Z.; visualization, J.W.; supervision, X.Z.; project administration, X.Z.; funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Fund of China (No. 61673316) and in part by the Project Commissioned by the Sichuan Gas Turbine Research Institute of AVIC.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available upon reasonable request from the corresponding author.

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

References

1. Chaturvedi, S.; Khanna, P.; Ojha, A. A survey on vision-based outdoor smoke detection techniques for environmental safety. *ISPRS-J. Photogramm. Remote Sens.* **2022**, *185*, 158–187. [CrossRef]
2. Wang, J.; Zhang, X.; Zhang, C. A lightweight smoke detection network incorporated with the edge cue. *Expert Syst. Appl.* **2024**, *241*, 122583. [CrossRef]
3. Tian, H.; Li, W.; Ogunbona, P.O.; Wang, L. Detection and Separation of Smoke From Single Image Frames. *IEEE Trans. Image Process.* **2018**, *27*, 1164–1177. [CrossRef] [PubMed]
4. Dimitropoulos, K.; Barmoutis, P.; Grammalidis, N. Higher Order Linear Dynamical Systems for Smoke Detection in Video Surveillance Applications. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 1143–1154. [CrossRef]
5. Yar, H.; Khan, Z.A.; Hussain, T.; Baik, S.W. A modified vision transformer architecture with scratch learning capabilities for effective fire detection. *Expert Syst. Appl.* **2024**, *252*, 123935. [CrossRef]
6. Tao, H.; Duan, Q. An adaptive frame selection network with enhanced dilated convolution for video smoke recognition. *Expert Syst. Appl.* **2023**, *215*, 119371. [CrossRef]
7. Cao, Y.; Tang, Q.; Wu, X.; Lu, X. EFNNet: Enhanced Feature Foreground Network for Video Smoke Source Prediction and Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1820–1833. [CrossRef]
8. Guo, F.; Wang, Y.; Qian, Y. Real-time dense traffic detection using lightweight backbone and improved path aggregation feature pyramid network. *J. Ind. Inf. Integr.* **2023**, *31*, 100427. [CrossRef]
9. Li, J.; Zhou, G.; Chen, A.; Lu, C.; Li, L. BCMNet: Cross-Layer Extraction Structure and Multiscale Downsampling Network with Bidirectional Transpose FPN for Fast Detection of Wildfire Smoke. *IEEE Syst. J.* **2023**, *17*, 1235–1246. [CrossRef]
10. Long, J.; Liang, W.; Li, K.; Wei, Y.; Marino, M.D. A Regularized Cross-Layer Ladder Network for Intrusion Detection in Industrial Internet of Things. *IEEE Trans. Ind. Inform.* **2023**, *19*, 1747–1755. [CrossRef]
11. Li, Z.; Lang, C.; Liew, J.H.; Li, Y.; Hou, Q.; Feng, J. Cross-Layer Feature Pyramid Network for Salient Object Detection. *IEEE Trans. Image Process.* **2021**, *30*, 4587–4598. [CrossRef] [PubMed]
12. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Volume 9905, pp. 21–37.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
14. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.
15. Jocher, G. YOLOv5 by Ultralytics. GitHub Repository. 2020. Available online: <https://github.com/ultralytics/yolov5> (accessed on 20 May 2024).
16. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.

17. Liang, S.; Wu, H.; Zhen, L.; Hua, Q.; Garg, S.; Kaddoum, G.; Hassan, M.M.; Yu, K. Edge YOLO: Real-time intelligent object detection system based on edge-cloud cooperation in autonomous vehicles. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 25345–25360. [CrossRef]
18. Jing, T.; Zeng, M.; Meng, Q.H. SmokePose: End-to-End Smoke Keypoint Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 5778–5789. [CrossRef]
19. Appana, D.K.; Islam, M.R.; Khan, S.A.; Kim, J. A video-based smoke detection using smoke flow pattern and spatial-temporal energy analyses for alarm systems. *Inf. Sci.* **2017**, *418*, 91–101. [CrossRef]
20. Filonenko, A.; Hernández, D.C.; Jo, K. Fast Smoke Detection for Video Surveillance Using CUDA. *IEEE Trans. Ind. Inform.* **2018**, *14*, 725–733. [CrossRef]
21. Prema, C.E.; Suresh, S.; Krishnan, M.N.; Leema, N. A Novel Efficient Video Smoke Detection Algorithm Using Co-occurrence of Local Binary Pattern Variants. *Fire Technol.* **2022**, *58*, 3139–3165. [CrossRef]
22. Hashemzadeh, M.; Farajzadeh, N.; Heydari, M. Smoke detection in video using convolutional neural networks and efficient spatio-temporal features. *Appl. Soft Comput.* **2022**, *128*, 109496. [CrossRef]
23. Tao, H.; Lu, M.; Hu, Z.; Xin, Z.; Wang, J. Attention-aggregated attribute-aware network with redundancy reduction convolution for video-based industrial smoke emission recognition. *IEEE Trans. Ind. Inform.* **2022**, *18*, 7653–7664. [CrossRef]
24. Gu, K.; Xia, Z.; Qiao, J.; Lin, W. Deep Dual-Channel Neural Network for Image-Based Smoke Detection. *Appl. Soft Comput.* **2020**, *22*, 311–323. [CrossRef]
25. Almeida, J.S.; Huang, C.; Nogueira, F.G.; Bhatia, S.; de Albuquerque, V.H.C. EdgeFireSmoke: A Novel Lightweight CNN Model for Real-Time Video Fire-Smoke Detection. *IEEE Trans. Ind. Inform.* **2022**, *18*, 7889–7898. [CrossRef]
26. Mukhiddinov, M.; Abdusalomov, A.B.; Cho, J. A Wildfire Smoke Detection System Using Unmanned Aerial Vehicle Images Based on the Optimized YOLOv5. *Sensors* **2022**, *22*, 9384. [CrossRef] [PubMed]
27. Saydirasulovich, S.N.; Mukhiddinov, M.; Djuraev, O.; Abdusalomov, A.; Cho, Y. An Improved Wildfire Smoke Detection Based on YOLOv8 and UAV Images. *Sensors* **2023**, *23*, 8374. [CrossRef] [PubMed]
28. Munsif, M.; Ullah, M.; Ahmad, B.; Sajjad, M.; Cheikh, F.A. Monitoring Neurological Disorder Patients via Deep Learning Based Facial Expressions Analysis. In Proceedings of the Artificial Intelligence Applications and Innovations, Crete, Greece, 17–20 June 2022; Volume 652, pp. 412–423.
29. Tao, H.; Xie, C.; Wang, J.; Xin, Z. CENet: A Channel-Enhanced Spatiotemporal Network with Sufficient Supervision Information for Recognizing Industrial Smoke Emissions. *IEEE Internet Things J.* **2022**, *9*, 18749–18759. [CrossRef]
30. Chen, W.; Luo, H.; Fang, H.; Chen, L.; Chen, Y.; Ding, J.; Kuo, S. DesmokeNet: A Two-Stage Smoke Removal Pipeline Based on Self-Attentive Feature Consensus and Multi-Level Contrastive Regularization. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 3346–3359. [CrossRef]
31. Li, Y.; Huang, Q.; Pei, X.; Chen, Y.; Jiao, L.; Shang, R. Cross-Layer Attention Network for Small Object Detection in Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2148–2161. [CrossRef]
32. Li, Y.; Zhang, W.; Liu, Y.; Jing, R.; Liu, C. An efficient fire and smoke detection algorithm based on an end-to-end structured network. *Eng. Appl. Artif. Intell.* **2022**, *116*, 105492. [CrossRef]
33. Tao, H.; Duan, Q. Learning Discriminative Feature Representation for Estimating Smoke Density of Smoky Vehicle Rear. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 23136–23147. [CrossRef]
34. Zhang, L.; Lu, C.; Xu, H.; Chen, A.; Li, L.; Zhou, G. MMFNet: Forest Fire Smoke Detection Using Multiscale Convergence Coordinated Pyramid Network with Mixed Attention and Fast-robust NMS. *IEEE Internet Things J.* **2023**, *10*, 18168–18180. [CrossRef]
35. Zhan, J.; Hu, Y.; Zhou, G.; Wang, Y.; Cai, W.; Li, L. A high-precision forest fire smoke detection approach based on ARGNet. *Comput. Electron. Agric.* **2022**, *196*, 106874. [CrossRef]
36. Yuan, F.; Zhang, L.; Xia, X.; Huang, Q.; Li, X. A Wave-Shaped Deep Neural Network for Smoke Density Estimation. *IEEE Trans. Image Process.* **2020**, *29*, 2301–2313. [CrossRef] [PubMed]
37. Jing, T.; Meng, Q.H.; Hou, H.R. SmokeSegger: A Transformer-CNN coupled model for urban scene smoke segmentation. *IEEE Trans. Ind. Inform.* **2024**, *20*, 1385–1396. [CrossRef]
38. Yuan, F.; Zhang, L.; Xia, X.; Huang, Q.; Li, X. A Gated Recurrent Network with Dual Classification Assistance for Smoke Semantic Segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 4409–4422. [CrossRef] [PubMed]
39. Song, K.; Sun, X.; Ma, S.; Yan, Y. Surface Defect Detection of Aeroengine Blades Based on Cross-Layer Semantic Guidance. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–11. [CrossRef]
40. Zhang, Y.; Wu, C.; Guo, W.; Zhang, T.; Li, W. CFANet: Efficient Detection of UAV Image Based on Cross-Layer Feature Aggregation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–11. [CrossRef]
41. Jocher, G.; Chaurasia, A.; Qiu, J. YOLO by Ultralytics. GitHub Repository. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 1 May 2024).
42. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
43. Yang, X.; Bist, R.; Subedi, S.; Wu, Z.; Liu, T.; Chai, L. An automatic classifier for monitoring applied behaviors of cage-free laying hens with deep learning. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106377. [CrossRef]

44. Liu, D.; Zhao, L.; Wang, Y.; Kato, J. Learn from each other to Classify better: Cross-layer mutual attention learning for fine-grained visual classification. *Pattern Recognit.* **2023**, *140*, 109550. [\[CrossRef\]](#)
45. Liu, Y.; Ma, C.; Kira, Z. Unbiased Teacher v2: Semi-supervised Object Detection for Anchor-free and Anchor-based Detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9809–9818.
46. Ding, J.; Li, W.; Pei, L.; Yang, M.; Ye, C.; Yuan, B. Sw-YoloX: An anchor-free detector based transformer for sea surface object detection. *Expert Syst. Appl.* **2023**, *217*, 119560. [\[CrossRef\]](#)
47. Zhang, D.; Zhao, J.; Chen, J.; Zhou, Y.; Shi, B.; Yao, R. Edge-aware and spectral-spatial information aggregation network for multispectral image semantic segmentation. *Eng. Appl. Artif. Intell.* **2022**, *114*, 105070. [\[CrossRef\]](#)
48. Tang, L.; Yuan, J.; Ma, J. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Inf. Fusion* **2022**, *82*, 28–42. [\[CrossRef\]](#)
49. Wang, C.; Zhang, D.; Zhang, L.; Tang, J. Coupling Global Context and Local Contents for Weakly-Supervised Semantic Segmentation. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *earle access*.
50. Yan, Y.; Li, J.; Qin, J.; Zheng, P.; Liao, S.; Yang, X. Efficient Person Search: An Anchor-Free Approach. *Int. J. Comput. Vis.* **2023**, *131*, 1642–1661. [\[CrossRef\]](#)
51. Wang, J.; Zhang, X.; Jing, K.; Zhang, C. Learning precise feature via self-attention and self-cooperation YOLOX for smoke detection. *Expert Syst. Appl.* **2023**, *228*, 120330. [\[CrossRef\]](#)
52. Zhang, Q.; Lin, G.; Zhang, Y.; Xu, G.; Wang, J. Wildland forest fire smoke detection based on faster R-CNN using synthetic smoke images. *Procedia Eng.* **2018**, *211*, 441–446. [\[CrossRef\]](#)
53. Ko, B.; Ham, S.; Nam, J. Modeling and Formalization of Fuzzy Finite Automata for Detection of Irregular Fire Flames. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 1903–1912. [\[CrossRef\]](#)
54. Qi, J.; Liu, X.; Liu, K.; Xu, F.; Guo, H.; Tian, X.; Li, M.; Bao, Z.; Li, Y. An improved YOLOv5 model based on visual attention mechanism: Application to recognition of tomato virus disease. *Comput. Electron. Agric.* **2022**, *194*, 106780. [\[CrossRef\]](#)
55. He, Z.; Zhang, L. Multi-adversarial faster-rcnn for unrestricted object detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6668–6677.
56. 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 IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
57. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
58. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. *arXiv* **2023**, arXiv:2304.08069.
59. Zhu, F.; Wang, Y.; Cui, J.; Liu, G.; Li, H. Target detection for remote sensing based on the enhanced YOLOv4 with improved BiFPN. *Egypt. J. Remote Sens. Space Sci.* **2023**, *26*, 351–360. [\[CrossRef\]](#)
60. Ju, C.; Guan, C. Tensor-cspnet: A novel geometric deep learning framework for motor imagery classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 10955–10969. [\[CrossRef\]](#) [\[PubMed\]](#)
61. Lu, X.; Li, W.; Xiao, J.; Zhu, H.; Yang, D.; Yang, J.; Xu, X.; Lan, Y.; Zhang, Y. Inversion of Leaf Area Index in Citrus Trees Based on Multi-Modal Data Fusion from UAV Platform. *Remote Sens.* **2023**, *15*, 3523. [\[CrossRef\]](#)
62. Khan, S.; Muhammad, K.; Mumtaz, S.; Baik, S.W.; de Albuquerque, V.H.C. Energy-Efficient Deep CNN for Smoke Detection in Foggy IoT Environment. *IEEE Internet Things J.* **2019**, *6*, 9237–9245. [\[CrossRef\]](#)
63. Cao, Y.; Tang, Q.; Lu, X.; Li, F.; Cao, J. STCNet: Spatiotemporal cross network for industrial smoke detection. *Multimed. Tools Appl.* **2022**, *81*, 10261–10277. [\[CrossRef\]](#)
64. Hu, Y.; Zhan, J.; Zhou, G.; Chen, A.; Cai, W.; Guo, K.; Hu, Y.; Li, L. Fast forest fire smoke detection using MVMNet. *Knowl. Based Syst.* **2022**, *241*, 108219. [\[CrossRef\]](#)
65. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; Volume 11211, pp. 3–19.
66. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.