Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061882

ARTICLE

TSMS-InceptionNeXt: A Framework for Image-Based Combustion State Recognition in Counterflow Burners via Feature Extraction Optimization

Huiling Yu[1], **Xibei Jia**[2], **Yongfeng Niu**[1] **and Yizhuo Zhang**[1,*]

[1]Software Engineering, Department of Computer Science, Changzhou University, Changzhou, 213146, China
[2]Electrical Engineering, Department of Computer Science, Changzhou University, Changzhou, 213146, China

*Corresponding Author: Yizhuo Zhang. Email: yzzhang@cczu.edu.cn

Received: 05 December 2024; Accepted: 24 February 2025; Published: 19 May 2025

ABSTRACT: The counterflow burner is a combustion device used for research on combustion. By utilizing deep convolutional models to identify the combustion state of a counterflow burner through visible flame images, it facilitates the optimization of the combustion process and enhances combustion efficiency. Among existing deep convolutional models, InceptionNeXt is a deep learning architecture that integrates the ideas of the Inception series and ConvNeXt. It has garnered significant attention for its computational efficiency, remarkable model accuracy, and exceptional feature extraction capabilities. However, since this model still has limitations in the combustion state recognition task, we propose a Triple-Scale Multi-Stage InceptionNeXt (TSMS-InceptionNeXt) combustion state recognition method based on feature extraction optimization. First, to address the InceptionNeXt model's limited ability to capture dynamic features in flame images, we introduce Triplet Attention, which applies attention to the width, height, and Red Green Blue (RGB) dimensions of the flame images to enhance its ability to model dynamic features. Secondly, to address the issue of key information loss in the Inception deep convolution layers, we propose a Similarity-based Feature Concentration (SimC) mechanism to enhance the model's capability to concentrate on critical features. Next, to address the insufficient receptive field of the model, we propose a Multi-Scale Dilated Channel Parallel Integration (MDCPI) mechanism to enhance the model's ability to extract multi-scale contextual information. Finally, to address the issue of the model's Multi-Layer Perceptron Head (MlpHead)neglecting channel interactions, we propose a Channel Shuffle-Guided Channel-Spatial Attention (ShuffleCS) mechanism, which integrates information from different channels to further enhance the representational power of the input features. To validate the effectiveness of the method, experiments are conducted on the counterflow burner flame visible light image dataset. The experimental results show that the TSMS-InceptionNeXt model achieved an accuracy of 85.71% on the dataset, improving by 2.38% over the baseline model and outperforming the baseline model's performance. It achieved accuracy improvements of 10.47%, 4.76%, 11.19%, and 9.28% compared to the Reparameterized Visual Geometry Group (RepVGG), Squeeze-erunhanced Axial Transoformer (SeaFormer), Simplified Graph Transformers (SGFormer), and VanillaNet models, respectively, effectively enhancing the recognition performance for combustion states in counterflow burners.

KEYWORDS: Counterflow burner; combustion state recognition; InceptionNeXt; dilated convolution; channel shuffling

1 Overview

A counterflow burner is a combustion device commonly used in the industrial and energy sectors. It achieves the mixing of combustion air and fuel through a counterflow arrangement, thereby enhancing combustion efficiency, reducing the emission of harmful substances, and enabling efficient thermal energy

conversion. Counterflow burners are widely applied in gas turbines, boilers, combustion laboratories, and other energy conversion devices, particularly in scenarios that require precise control of the combustion state. Accurately identifying the combustion state within a counterflow burner can optimize the combustion process, enhance combustion efficiency, reduce the emission of harmful gases, and decrease energy waste. However, identifying the combustion state poses significant challenges due to the complex and stochastic nature of flames, making accurate identification difficult. With the advancement of monitoring technologies, flame imaging combined with algorithms has emerged as a promising method and has been widely adopted. Furthermore, with the development of artificial intelligence and deep learning technologies, recent years have seen a transition in combustion state recognition methods within burner flames from traditional feature extraction models to deep convolutional models. Among deep convolutional models, the InceptionNeXt model inherits the traditional advantages of the Inception series, enabling it to process information at multiple scales simultaneously. In flame state recognition, the features of flames are significant at various scales. The overall shape of the flame can be perceived at larger scales, while the detailed information of the flame edges needs to be captured at smaller scales. InceptionNeXt can utilize convolutional kernels of different sizes across various branches to extract multi-scale features, thereby comprehensively describing the flame state. Compared to traditional single-scale convolutional neural networks, it does not lose key information at certain scales due to fixed convolutional kernel sizes. InceptionNeXt has demonstrated outstanding performance in other image classification tasks, which share similarities with flame state recognition, as both require feature extraction and classification of targets within images. In image classification tasks, it can accurately identify objects of different categories, indicating that its feature extraction capabilities and classifier performance are reliable. Flame state recognition can be viewed as a specialized image classification task, where flame images are categorized into different state classes. Based on InceptionNeXt's strong performance in other similar tasks, it is reasonable to hypothesize that it can also achieve good results in flame state recognition tasks. However, despite its favorable performance, InceptionNeXt still faces challenges in combustion state recognition, such as difficulty in adapting to the dynamic changes of flames, susceptibility to losing key information in flame images, and having a relatively small receptive field. Therefore, we propose the TSMS-InceptionNeXt model, which enhances combustion state recognition capabilities by optimizing the model's ability to extract flame features.

The main contributions of our study can be summarized as follows:

(1) We analyze the defects in the feature extraction part, Stages, of the InceptionNeXt recognition framework and introduce the Triplet Attention mechanism in the Stages componentto enhance the model's ability to capture dynamic flame features.
(2) We analyze the deficiencies in the MetaNeXt module of the InceptionNeXt recognition framework and introduce the SimC mechanism to improve its ability to recognize key information.
(3) We propose a MDCPI mechanism, which combines multi-scale dilated convolutions with parallel channel and spatial attention mechanisms to enlarge the model's receptive field.
(4) We propose a ShuffleCS mechanism that improves the model's MlpHead by facilitating the interaction of information across different channels.
(5) We introduce the Focal Loss function to replace SoftTargetCrossEntropy, thereby enhancing the model's focus on hard-to-recognize states.

The remainder of this paper is organized as follows: Section 2 introduces related work; Section 3 describes the feature extraction-optimized TSMS-InceptionNeXt; Section 4 outlines the experimental setup and analyzes the results; Section 5 concludes the paper.

2 Related Work

Traditional methods for recognizing combustion states inside burners rely on visual observation and sensor-based detection. However, these methods are highly subjective, slow to respond, and lack precision. With advancements in detection technology, using algorithms to process flame images has emerged as a promising approach and is widely employed for combustion state recognition. Unlike traditional detection methods, machine learning approaches based on flame images identify combustion states by extracting image features. Malpica et al. [1] proposed a gradient-free combustion state recognition method based on the thermochemical state features of flame images. However, due to a lack of prior knowledge, this method is limited to specific combustion scenarios. Sitaraman et al. [2] introduced a machine learning method to recognize combustion states in Reactivity Controlled Compression Ignition (RCCI) engine combustion systems using heat release rate features from flame images. However, its accuracy is low due to insufficient prior knowledge. Bhattacharya et al. [3] developed a combustion state recognition method based on a probabilistic finite-state automaton, which improved recognition accuracy. However, the approach requires a large amount of data and is computationally inefficient. Compais et al. [4] applied Analysis of Variance (ANOVA) F-variance analysis to select features from flame images and proposed a machine learning method for identifying combustion states in laboratory devices. While it achieves high accuracy with small datasets, it requires significant computational resources, making it unsuitable for real-time applications. Bukkarapu et al. [5] extracted multiple features from flame images and proposed a decision tree model for recognizing burner combustion states. This method achieved high accuracy but incurred substantial computational costs.

From the above methods, it can be observed that machine learning approaches for extracting flame image features and recognizing combustion states face the following issues: heavy reliance on domain experts' prior knowledge, a large amount of required data, low processing efficiency, and high computational cost. Therefore, a more precise and efficient method for extracting flame image features is needed. In recent years, deep learning methods have garnered significant attention in the field of combustion state recognition. By employing multi-layer nonlinear transformations to autonomously extract discriminative features from large amounts of flame data for recognition, deep learning overcomes the reliance on prior knowledge seen in machine learning and other shallow models, while also avoiding the need for complex feature extraction processes. Choi et al. [6] proposed a fusion layer combined with the Residual Network (ResNet) model for recognizing the combustion states of gas turbine burners. This model does not require prior knowledge and captures flame features by extracting image frames. However, the features of each channel are learned independently, with no feature exchange between channels. Roncancio et al. [7] proposed a Convolutional Neural Network (CNN) combined with data augmentation to recognize turbulent flame states using turbulent flame images. The method achieves high accuracy by extracting image features based on heat release rates, but it does not consider dependencies between flame image channels and spatial dimensions, making it unable to handle dynamically changing flames. Omiotek et al. [8] proposed a flame image segmentation method based on the red component of the flame image and employed the Visual Geometry Group 16 (VGG16) model for combustion state recognition. This approach improved the segmentation quality of flame regions and enhanced recognition accuracy, but the model is limited by its small receptive field due to using only 3×3 convolutions. Pereira et al. [9] proposed an improved convolutional neural network for recognizing the states of combustion devices in production lines. The method maintained high accuracy with short processing times and expanded the receptive field. However, the downsampling process led to the loss of local critical information. Natsui et al. [10] combined class activation mapping with convolutional neural networks to recognize different combustion states by analyzing temporal sequence images extracted from videos. This approach overcame disturbances caused by dynamic flame changes, but the model's receptive field was small. Pan et al. [11] proposed a combustion state recognition method based on the Vision-Transformer-Improved Deep Forest Classification (ViT-IDFC) algorithm, improving the feature

extraction component. The model achieved a large receptive field and high accuracy but lacked information exchange between channels. Wang et al. [12] proposed an improved Artificial neural network (ANN) model for recognizing combustion states in Moderate or Intense Low-Oxygen Dilution (MILD) burners. However, the model overlooked spatial relationships between pixels, resulting in an inability to capture critical information in images. Wu et al. [13] proposed a convolutional neural network combining Long Short-Term Memory (LSTM) and Proximal Policy Optimization (PPO) to detect and recognize combustion states in pulverized coal boilers. While this method focused on temporal dependencies in image sequences, it lacked attention to spatial relationships within the images. Lv et al. [14] proposed an unsupervised learning method combining the Convolutional Block Attention Module (CBAM) and Stacked Convolutional Autoencoders (SCAE) to extract features and recognize combustion states from flame images of burners. This approach established dependencies between channel and spatial dimensions and enhanced channel-wise information interaction, but it suffered from a small receptive field.

Yu et al. [15] proposed dilated convolution to efficiently expand the receptive field, addressing the issue of reduced resolution in traditional convolutions when increasing the receptive field. This method also captures multi-scale contextual information. Jaderberg et al. [16] proposed the spatial attention mechanism based on spatial invariance, solving the problem of traditional CNN failing to recognize images effectively after transformations. Hu et al. [17] introduced the channel attention mechanism based on the significance of different channel features, addressing the issue of traditional CNN neglecting channel features. Zhang et al. [18] proposed the channel shuffle strategy to enhance channel information utilization, addressing the problem of traditional convolutions losing channel information. Wang et al. [19] proposed the Efficient Channel Attention (ECA) mechanism, which avoids the information loss caused by dimensionality reduction in traditional channel attention and has low computational cost. Yang et al. [20] introduced the Simplified Attention Mechanism (SimAM) inspired by human brain attention, addressing the issue of traditional spatial and channel attention mechanisms requiring additional parameters, which increase model complexity. Misra et al. [21] proposed the Triplet Attention mechanism based on cross-dimensional interaction, which enables better understanding of complex input feature structures through channel and spatial information interaction. Ding et al. [22] proposed the RepVGG model based on re-parameterization. This model features a simple structure and fast inference speed but relies primarily on local convolution operations, lacking interaction modeling between channel and spatial dimensions, and has a small receptive field. Wan et al. [23] introduced the Seaformer model based on self-attention, enabling interaction modeling between spatial and channel dimensions in lightweight models. However, it tends to overlook critical information in detail-rich or long-distance dependency images, and self-attention limits feature exchange across channels. Wu et al. [24] proposed the SGFormer model based on the Transformer architecture, utilizing a multi-scale design to extract features at different scales with a large receptive field. However, its lightweight structure limits its ability to fully capture long-distance detail information and features despite maintaining a large receptive field. Chen et al. [25] proposed the VanillaNet model based on basic convolutional modules. By stacking convolutional layers, the receptive field was expanded. However, the lack of inter-channel feature interaction in its design resulted in insufficient fusion of channel information, reducing recognition performance.

From the current state of research in deep learning, it is evident that deep learning methods based on flame image feature extraction have been widely applied to combustion state recognition tasks. These methods not only enable combustion state recognition but also extract potential flame features through feature maps, leading to better performance. However, current deep learning methods for combustion state recognition based on flame image feature extraction suffer from issues such as small receptive fields, lack of channel interaction, and low recognition accuracy. With the deepening of research, Yu et al. [26] inspired by models such as Inception [27], ConvNeXt [28], and Vision Transformer (ViT) [29], decomposed large kernel depth convolutions into four parallel branches and proposed an InceptionNeXt model based on the

CNN architecture. Experimental results show that InceptionNeXt outperforms traditional CNN models on both the ImageNet-1K and ADE20K datasets, demonstrating its high potential for image recognition tasks. Subsequent research has also confirmed the effectiveness of the InceptionNeXt model in various image recognition tasks. Li et al. [30] addressed the issue of insufficient receptive field in the InceptionNeXt model by proposing an InceptionNeXt Attention Differentiable Binarization Network (INA-DBNet) scene text detection method based on semantic segmentation. This method integrates InceptionNeXt and a multi-scale attention mechanism to resolve the receptive field limitations of existing text detection methods. However, the introduction of new modules decreased the computational efficiency of the model. Lau et al. [31] proposed an AudioRepInceptionNeXt audio recognition model to address the lack of inter-channel information interaction in InceptionNeXt. This model was improved by using an inverted bottleneck design, where parallel multi-scale depth separable convolution kernels were placed before the 1×1 expansion layer of InceptionNeXt to enable inter-channel information exchange. However, the model relies on large convolution kernels to capture global information, leading to a neglect of local critical details. Although some improvements have been made in these studies, there are still some deficiencies. In summary, to address the common issues in the InceptionNeXt model and other models in combustion state recognition, and to more accurately identify the combustion states of counterflow burners, this paper proposes an improved InceptionNeXt model optimized for feature extraction. The proposed model was tested on a visible light flame image dataset of a counterflow burner, which includes six common combustion states of the burner. This dataset covers a wide range of combustion conditions used in the laboratory and provides a foundation for experimental design.

In recent years, large-kernel depthwise convolution has been widely studied and adopted due to its large receptive field. However, its high memory access cost results in low model efficiency. Reducing its memory access cost could improve detection speed. To address this, InceptionNeXt proposed an Inception Depthwise Convolution, which significantly improves model efficiency while maintaining the performance of large-kernel depthwise convolution.

Let the input be X, and the computation of traditional large-kernel depthwise convolution is expressed as Eq. (1).

$$X' = DWConv_{k\times k}^{C\to C}(X) \tag{1}$$

In the equation, X' represents the result after the large-kernel depthwise convolution operation; $DWConv_{k\times k}^{C\to C}()$ denotes a depthwise convolution with a kernel size of $k \times k$, where C represents the input and output channels.

The computation of the Inception Depthwise Convolution is expressed as Eq. (2).

$$\begin{aligned} X_{hW}, X_W, X_h, X_{id} &= Split(X) \\ X'_{hW} &= DWConv_{k_s\times k_s}^{g\to g}(X_{hW}) \\ X'_W &= DWConv_{1\times k_b}^{g\to g}(X_W) \\ X'_h &= DWConv_{k_b\times 1}^{g\to g}(X_h) \\ X'_{id} &= X_{id} \\ X' &= Concat\left(X'_{hW}, X'_W, X'_h, X'_{id}\right) \end{aligned} \tag{2}$$

In the equation, X_{hW}, X_W, X_h, X_{id} are the four decomposed convolution branches; k_s and k_b represent the sizes of the decomposed convolution kernels; and $Concat()$ denotes the operation of concatenating the outputs of each branch.

By comparing Eqs. (1) and (2), it is evident that traditional large-kernel depthwise convolutions require $k \times k$ operations each time, whereas Inception Depthwise Convolution decomposes the traditional large-kernel depthwise convolution into four parallel branches: a small square convolution (kernel size $k_s \times k_s$), two strip convolutions (kernel sizes $1 \times k_b$ and $k_b \times 1$), and an identity mapping. Each parallel branch requires far less computation than $k \times k$, thus significantly reducing computational cost. The small square convolution preserves the ability of traditional convolutions to capture small-scale features, the two orthogonal strip convolutions extend the receptive field and reduce computational cost, and the identity mapping keeps certain channels unchanged. This design addresses the high memory access cost and reduced computational speed of traditional large-kernel depthwise convolutions. By decomposing large-kernel depthwise convolutions into multiple small-kernel convolutions and one-dimensional strip convolutions, it retains the advantage of a large receptive field while reducing computational complexity. Using InceptionNeXt as the baseline model provides a solid foundation for subsequent combustion state recognition tasks.

The InceptionNeXt image recognition model comprises three main components: Input for image preprocessing, Stages for feature extraction, and MlpHead for classification. The Input component preprocesses the counterflow burner flame images by resizing, performing convolution operations, and applying batch normalization. The Stages component uses the MetaNeXtBlock module to extract feature information from the flame images and outputs feature maps. The MlpHead component applies global average pooling to the feature maps, followed by two fully connected layers for feature transformation, ultimately producing recognition results of size Num_classes. The structure of the baseline model is shown in Fig. 1. The black dashed boxes on the left side of Fig. 1 indicate the parts of the model, and the text in the corners of the black dashed boxes indicates the name of the part, while the two black dashed boxes on the right side illustrate the internal structure of the MetaNeXtBlock and the Inception deep convolution in the original model of InceptionNeXt, respectively. To address the common issues of the InceptionNeXt baseline model and other models in combustion state recognition, strategies for improving the baseline model are proposed in the following sections.

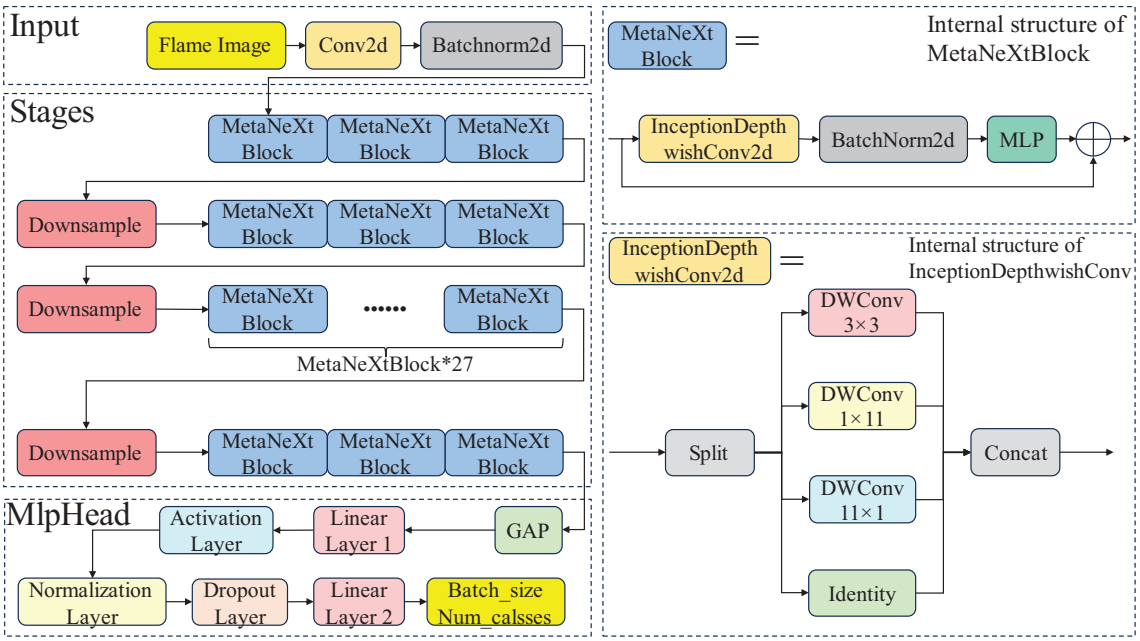


Figure 1: Schematic diagram of the InceptionNeXt baseline model structure

3 Our Methods

The structure of the proposed TSMS-InceptionNeXt model is shown in Fig. 2, with the improvements highlighted by red dashed lines. Firstly, the Triplet Attention mechanism is introduced before the Stages component, applying attention across the height, width, and Red Green Blue (RGB) dimensions to enhance the model's ability to capture dynamic flame features. Second, the MetaNeXtBlock in the feature extraction component is improved by integrating the SimAM and ECA mechanisms to increase the module's focus on critical image information. Next, a MDCPI mechanism is proposed after the Stages component to expand the model's receptive field. Finally, a ShuffleCS mechanism is proposed and embedded into the MlpHead to enhance information interaction between channels, further improving the model's recognition performance. The dashed box on the right in Fig. 2 represents the structure of the improved MetaNeXtBlock.

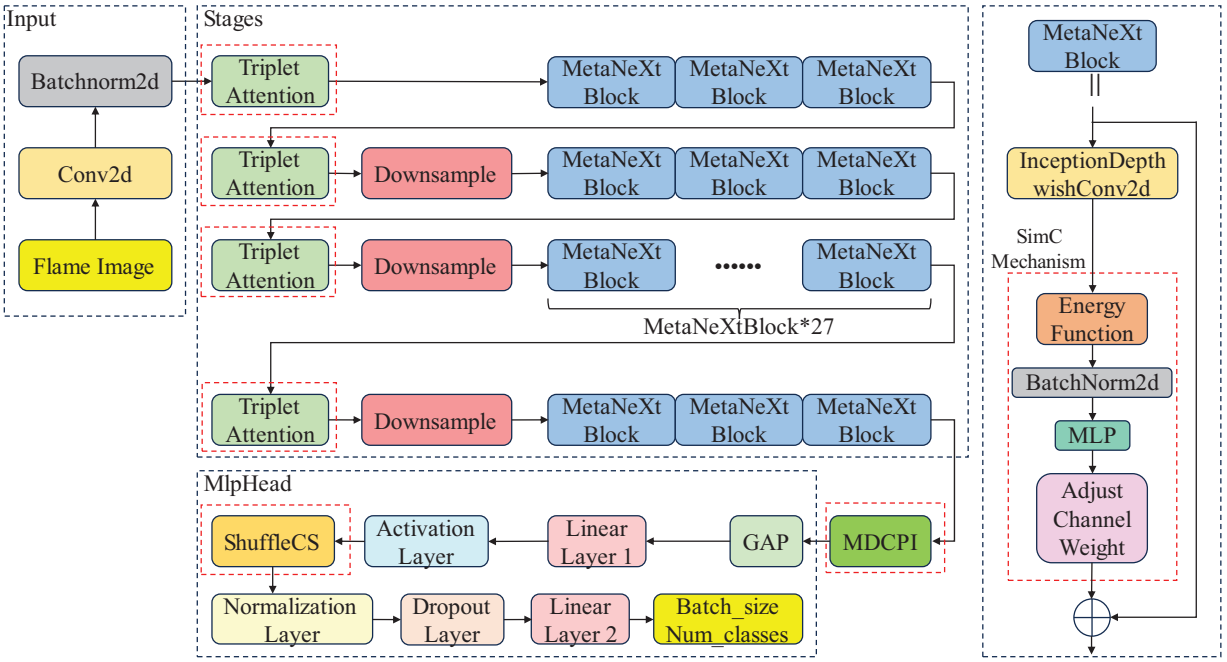


Figure 2: Schematic of the TSMS-InceptionNeXt model structure optimized for feature extraction

3.1 Triplet Attention Mechanism

The InceptionNeXt baseline model struggles to capture the dynamic features of flames. To address this issue, the Triplet Attention mechanism is introduced in the model's Stages component, with the aim of enhancing the model's ability to capture dynamic features of flame images. The structure of the Triplet Attention mechanism is shown in Fig. 3.

Existing attention mechanisms predominantly rely on the aggregation of global or local information. In contrast, Triplet Attention integrates the height, width, and channel information of images, enabling it to more effectively extract subtle feature variations in flame images compared to traditional self-attention mechanisms. This enhancement improves the model's classification capability under complex flame conditions.

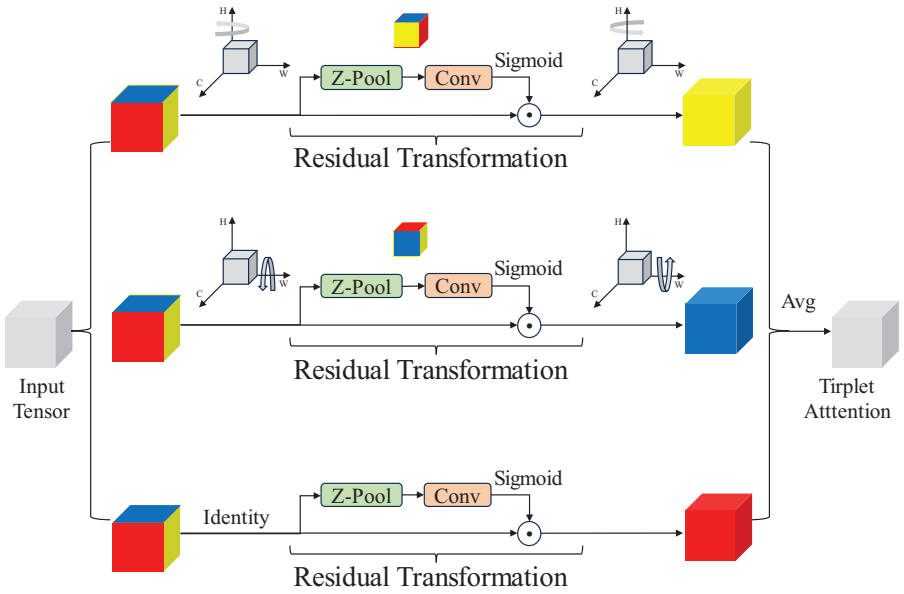


Figure 3: Structure of the triplet attention mechanism

Triplet Attention focuses on the input features across multiple dimensions by dividing the feature map into three dimensions: height, width, and channel. It then applies attention separately to each dimension using a three-branch structure. For an input flame image $M \in R^{C\times H\times W}$, the computations are performed through three branches, and the final weights are obtained by averaging, as described by Eqs. (3)–(6).

$$M^*_{H^-} = \left(M_{H^-}\sigma\left(\omega_1 * \left(Z - pool\left(M_{H^-}\right)\right)\right)\right)_{H^+} \tag{3}$$

$$M^*_{W^-} = \left(M_{W^-}\sigma\left(\omega_2 * \left(Z - pool\left(M_{W^-}\right)\right)\right)\right)_{W^+} \tag{4}$$

$$M^* = M\sigma\left(\omega_3 * \left(Z - pool\left(M\right)\right)\right) \tag{5}$$

$$M' = \frac{1}{3}\left(M^*_H + M^*_W + M^*\right) \tag{6}$$

In the equations, H^- represents a 90° counterclockwise rotation along the H-axis; W^- represents a 90° counterclockwise rotation along the W-axis; σ denotes the activation function; ω_1 and ω_2 are convolution kernels; $*$ indicates the convolution operation; H^+ represents a 90° clockwise rotation along the H-axis; W^+ represents a 90° clockwise rotation along the W-axis; and $Z - pool$ preserves feature representations.

The three-branch design of Triplet Attention enables it to apply attention separately from different spatial dimensions (height, width, and channel). Over different times and combustion states, flame shapes and colors undergo changes. In terms of shape, flame features vary vertically and horizontally. Triplet Attention extracts attention information from the height and width of flame images, dynamically adjusting the model's sensitivity to these changing features, thereby better capturing variations in flame shapes; Regarding color, changes in flame colors under different combustion states essentially involve alterations in Red Green Blue (RGB) components, which are reflected through the channel dimension of the image. Triplet Attention extracts attention information from the channel dimension, allowing the model to adapt more flexibly to changes in flame colors. In summary, these two improvements enhance the model's ability to capture the dynamic features of flames.

3.2 SimC Mechanism Based on Energy Function

The Inception deep convolution used in the InceptionNeXt baseline model decomposes large-kernel deep convolutions, which leads to the neglect of key features in the image during the decomposition process. To address the issue of Inception deep convolutions overlooking key information and thereby improve the model's recognition accuracy, this paper designs a SimC mechanism based on an energy function, which is embedded after the Inception deep convolution. The fundamental principle of the SimC is to enhance pixels and channels containing critical information. Its structure is shown in Fig. 4.

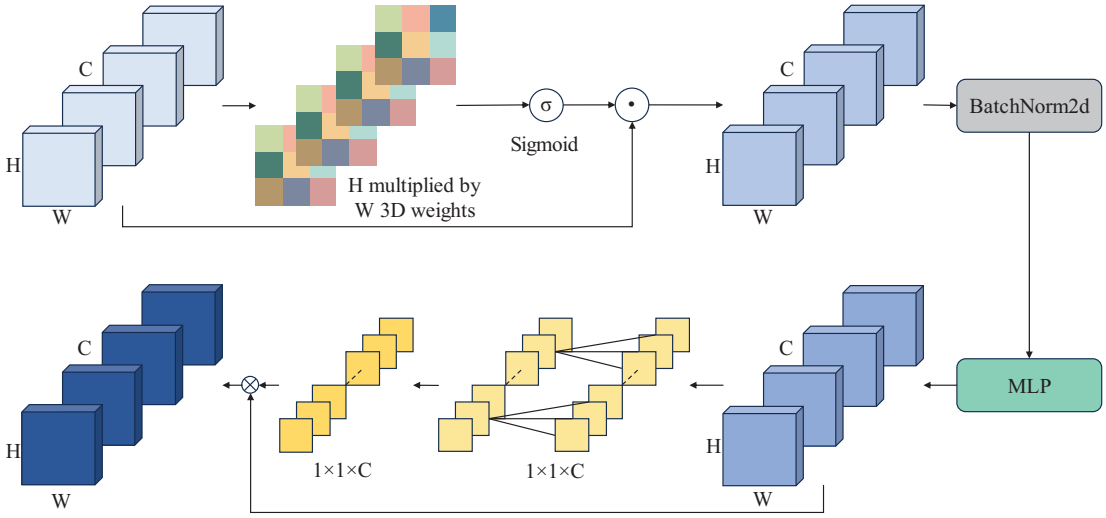


Figure 4: Structure of the SimC mechanism

The SimC mechanism optimizes information extraction by selectively focusing on features, thereby reducing the impact of redundant features. Compared to traditional convolutional networks, it places greater emphasis on the similarities and differences between features, enabling the model to more accurately identify key features in variable flame images and thereby enhancing recognition accuracy.

The specific calculation formulas for SimC are provided in Eqs. (7)–(10), where Eq. (7) defines the energy function.

$$e_i^* = -\frac{4\left(\hat{\sigma}^2 + \lambda\right)}{\left(t_i - \hat{\mu}\right)^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{7}$$

$$\tilde{X} = sigmiod\left(\frac{1}{E}\right) \odot X \tag{8}$$

$$k = \psi(C) = \left|\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}\right|_{odd} \tag{9}$$

$$out = Conv1D_k\left(\tilde{X}\right) \tag{10}$$

In the equations, e_i^* is used to determine pixel similarity, where i represents the pixel index; $\hat{\sigma}^2$ denotes the variance of all neurons in a single channel; λ is the regularization term; t_i represents the i-th pixel in the input feature map of a single channel; $\hat{\mu}$ is the mean of all pixels in a single channel; $\tilde{X}$ is the enhanced tensor; E is the sum of e_i^*; $\odot$ denotes the Hadamard product; C represents the number of channels; b and γ

adjust the ratio of channels to kernel size; *odd* refers to odd numbers; and $Conv1D_k()$ is a one-dimensional convolution operation with a kernel size of k.

For the input feature tensor, the SimC mechanism uses an energy function based on local image similarity to identify highly similar pixels and enhance them in subsequent operations, thereby emphasizing key features critical for accurate recognition. After obtaining the feature tensor with enhanced pixels, to maintain the lightweight advantage of Inception Depthwise Convolution, fast one-dimensional convolution operations are used to learn the importance of each channel relative to others, thereby increasing the weights of channels related to critical information. This achieves the goal of focusing on important information in the feature map. The SimC mechanism adaptively weights the input feature maps and dynamically adjusts channel weights, effectively improving the model's sensitivity to critical image information and addressing the issue of Inception Depthwise Convolution overlooking important details.

3.3 Multi-Scale Dilated Channel Parallel Integration (MDCPI) Mechanism Based on Dilated Convolutions

To expand the model's receptive field and enhance its focus on the contextual information of flame images, this paper designs a Multi-Scale Dilated Channel Parallel Integration (MDCPI) mechanism based on dilated convolutions. This mechanism uses multiple dilation rates and integrates channel and spatial attention mechanisms to improve the model's perception of multi-scale contextual information. The structure of the MDCPI mechanism is shown in Fig. 5.

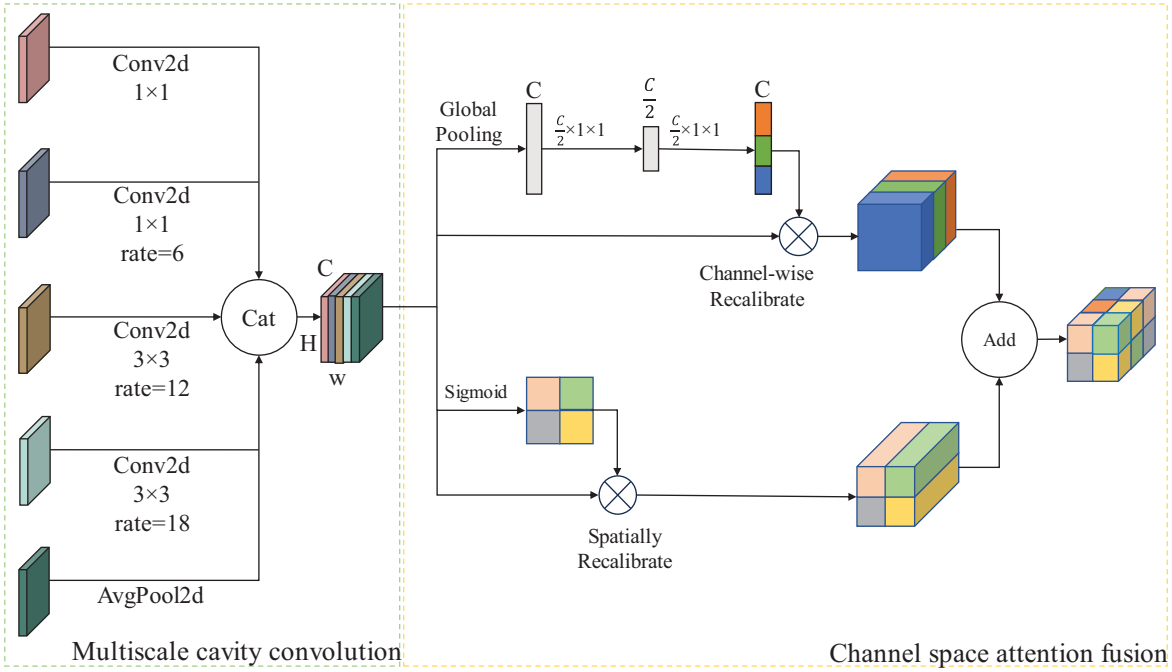


Figure 5: Structure of the MDCPI mechanism

When dealing with tasks involving flame images characterized by multi-scale features, existing convolutional layers often constrain the size of the receptive field. The MDCPI mechanism introduces multi-scale dilated convolutions, effectively expanding the receptive field and enabling the model to capture flame features at multiple levels. Compared to single-scale convolution methods, the MDCPI mechanism not only

enhances the model's expressive capacity but also significantly improves the accuracy of identifying different combustion states.

The workflow of the MDCPI mechanism consists of two steps. The first step is multi-scale dilated convolution, which utilizes five parallel convolution branches with different dilation rates to extract features at various scales. These features are concatenated along the channel dimension, producing a composite feature map as the output. Dilated convolutions increase the spacing between convolutional kernels by introducing holes, thereby expanding the receptive field without altering the kernel size. Unlike traditional convolutions, dilated convolutions perform convolution operations at intervals of pixels, allowing the convolution to cover a broader area of the input image. One of the key features of the MDCPI mechanism is its multi-scale dilated convolutions, which utilize multiple dilation rates to perform convolution operations at various scales. This approach enables the fusion of information across multiple levels and effectively captures features at different scales, thereby expanding the receptive field. Consequently, the model can extract more useful information from larger regions of the image, enhancing its ability to understand the image comprehensively. The expressions for multi-scale dilated convolution are shown in Eqs. (11)–(16):

$$F_1(x) = ReLU(BN(Conv2D(x, 1\times 1, dilation = 1))) \tag{11}$$

$$F_2(x) = ReLU(BN(Conv2D(x, 3\times 3, dilation = 6))) \tag{12}$$

$$F_3(x) = ReLU(BN(Conv2D(x, 3\times 3, dilation = 12))) \tag{13}$$

$$F_4(x) = ReLU(BN(Conv2D(x, 3\times 3, dilation = 18))) \tag{14}$$

$$F_5(x) = ReLU(BN(Conv2D(GlobalAvgPool(x), 1\times 1))) \tag{15}$$

$$F_{concat}(x) = Concat(F_1(x), F_2(x), F_3(x), F_4(x), F_5(x)) \tag{16}$$

In Eqs. (10)–(15), $ReLU()$ is the nonlinear activation function; $BN()$ represents batch normalization; $Conv2D()$ is a two-dimensional convolution; $dilation$ indicates the dilation rate; $GlobalAvgPool()$ denotes global average pooling; and $Concat()$ refers to concatenation of five tensors along a specific dimension.

However, experiments revealed that the feature maps output by dilated convolutions lacked certain pixels. This is due to the grid effect inherent in dilated convolutions: the dilation rate causes the convolution kernel to skip pixels during sampling, resulting in a grid-like distribution of sampled points rather than a continuous pixel region, leading to the loss of local information. To address the issue from the first step, the second step employs parallel channel and spatial attention mechanisms. The feature maps output from the first step are processed in parallel through channel and spatial attention mechanisms, and the outputs of the two parallel branches are fused via element-wise addition, producing the final enhanced output feature map.

The MDCPI mechanism significantly expands the model's receptive field through multi-scale dilated convolution, enabling it to capture a broader range of contextual information in flame images. Due to the grid effect of dilated convolution, the feature maps processed by multi-scale dilated convolution are calibrated using parallel channel and spatial attention mechanisms. Channel attention enhances global contextual information, mitigating the issue of local information loss caused by dilated convolutions and making features more comprehensive along the channel dimension; Spatial attention corrects the discontinuities in the spatial dimension of features, ensuring spatial consistency of the feature map and counteracting the local irregularities caused by the grid effect. Finally, the features from the two parallel branches are integrated to ensure that the model's receptive field is expanded, enabling it to effectively capture the multi-scale contextual information of flame feature maps.

3.4 Channel Shuffle-Guided Channel-Spatial Attention (ShuffleCS) Mechanism

In the InceptionNeXt baseline model, the MlpHead relies solely on Global Average Pooling (GAP) to extract global features, which isolates all computations and prevents information exchange between groups, thereby reducing the representational power of the input features. To address this issue, a channel-shuffling based feature expression mechanism, ShuffleCS, is designed. The structure of this mechanismis shown in Fig. 6.

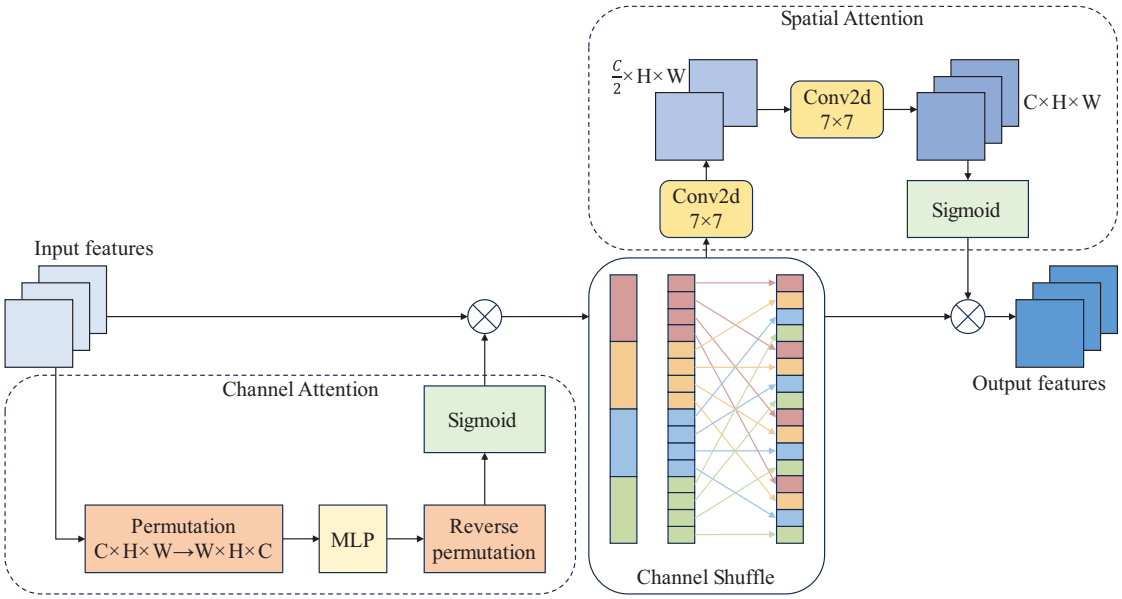


Figure 6: Structure of the ShuffleCS mechanism

Traditional convolutional neural networks often lack effective utilization of interactions between channels. In contrast, the ShuffleCS mechanismenhances inter-channel interactions through channel shuffle techniques, thereby strengthening the expressive capability of features.

The ShuffleCS mechanismuses a channel attention submodule combined with a multilayer perceptron (MLP) to enhance inter-channel dependencies, achieving an initial mixing of channel information. To prevent insufficient mixing, channel shuffle operations are introduced. The aforementioned feature fusion may overlook spatial dependencies, so a spatial attention submodule combined with a 7 × 7 convolution is used to enhance the effectiveness of feature fusion.

The specific calculation formulas for the ShuffleCS mechanismare provided in Eqs. (17)–(19):

$$T_c = Sigmoid\left(\sigma\left(W_{avg}\cdot F_{avg}+b_{avg}\right)+\sigma\left(W_{max}\cdot F_{max}+b_{max}\right)\right)\odot T \tag{17}$$

$$T_s = Shuffle\left(T_c\right) \tag{18}$$

$$T_s' = Sigmoid\left(W_a\cdot\left[AvgPool\left(T_s\right);MaxPool\left(T_s\right)\right]+b_a\right)\odot T_s \tag{19}$$

In this mechanism, T_c is the feature map weighted by channel attention, $Sigmoid()$ and σ are activation functions, W_{avg} and W_{max} are the weights of the fully connected layers, F_{avg} is the global average pooling, and F_{max} is the global maximum pooling, b_{avg} and b_{max} are the bias terms, $\odot$ denotes element-wise multiplication, T_s is the shuffled feature map, $Shuffle()$ is the operation that shuffles the grouped channels, T_c is the feature map after channel attention weighting, and T_s' is the feature map after spatial attention

processing, W_a is the convolution kernel, $AvgPool\ (T_s)$ is the global average pooling operation on the feature map T_s along the channel dimension, and $MaxPool(T_s)$ is the global maximum pooling operation on the feature map T_s along the channel dimension, b_a is the bias term, T_s is the input feature map.

3.5 Loss Function Optimization

The loss function in the InceptionNeXt baseline model is SoftTargetCrossEntropy, whose fundamental principle is to calculate the discrepancy between the predicted probability distribution and the target probability distribution, using this discrepancy to update model parameters. The formula is provided in Eq. (20).

$$L = -\sum_{i=1}^{C} y_i \log(p_i) \tag{20}$$

Here, L is the loss value; C is the number of categories; y_i represents the target probability distribution, which is the probability of a sample belonging to the i-th category; and p_i represents the predicted probability distribution, which is the model's predicted probability of the sample belonging to the i-th category.

In the combustion state recognition task of this study, the six classes have varying levels of recognition difficulty. However, SoftTargetCrossEntropy cannot effectively emphasize attention to the hard-to-recognize states. To address the model's insufficient focus on hard-to-recognize samples, this paper adopts the Focal Loss [32] function as a replacement for SoftTargetCrossEntropy. Focal Loss dynamically adjusts the loss function weights to make the model pay more attention to hard-to-recognize states. The formula is provided in Eq. (21).

$$FL(p_t) = -\sum_{c=1}^{C} \alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{21}$$

Here, p_t represents the probability obtained by applying the softmax function to the logit output; C is the number of categories; α_t is a parameter to adjust the weight of positive and negative samples; and γ is a parameter to adjust the weight of easy and hard samples.

4 Experimental Results and Analysis

4.1 Experimental Environment and Dataset Description

To validate the effectiveness of the improved InceptionNeXt model for combustion state recognition, experiments were conducted on a visible light flame image dataset of a counterflow burner. The experiments were performed on a 64-bit Ubuntu operating system using an NVIDIA GeForce RTX 4060 Ti GPU, the PyTorch deep learning framework, and Python as the programming language. The experimental environment configuration is shown in Table 1.

The dataset used in this study originates from the work of Kang et al. [33] (2022), which includes visible flame images of a counterflow burner in six different combustion states, captured by the FASTEC TS-5 high-speed CMOS camera. The FASTEC TS-5 high-speed CMOS camera is sensitive to wavelengths ranging from 350 to 950 nm, covering the visible light spectrum, and is equipped with a high sampling rate of 100 frames per second. The authors simulated six different combustion states by adjusting the gas ratios entering the counterflow burner and separately collected flame videos under each of these six combustion states. After collecting the video data, the authors performed frame extraction on the videos to select visible flame images corresponding to the six combustion states, totaling 2640 images with a resolution of 640 × 480 pixels each. The working principle of the counterflow burner is illustrated in Fig. 7.

Table 1: Experimental environment configuration

Experimental environment configuration	Version information
CPU	13th Gen Intel(R) Core (TM) i5-13490F
GPU	NVIDIA GeForce RTX 4060 Ti
Operating system	Ubuntu 18.04
GPU memory	16 G
Programming language	Python 3.8
CUDA	11.8
Deep learning framework	PyTorch 2.3.1

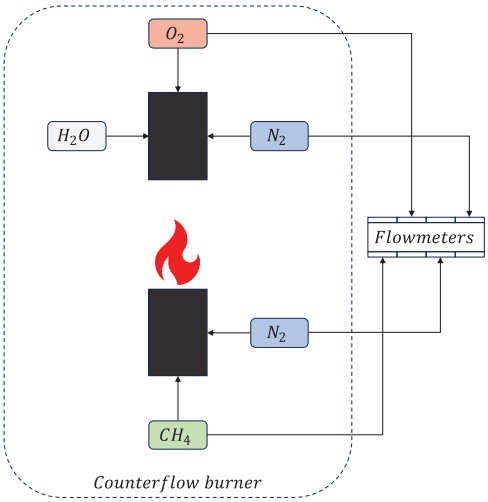
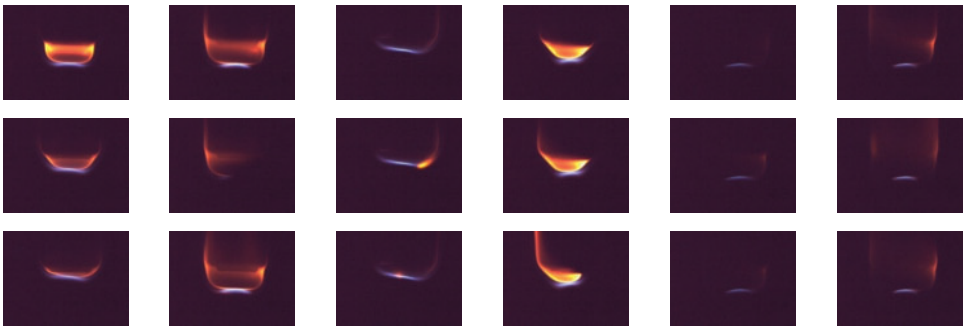


Figure 7: The working principle of counterflow burner

Based on the gas ratios entering the counterflow burner, the six combustion states in the dataset are named as follows: Oxygen-Enriched Stable Combustion (OESC), High Oxygen Combustion (HOC), Nitrogen-Diluted High Oxygen Combustion (ND-HOC), Low Oxygen and Low Methane Combustion (LO-LMC), High Nitrogen-Diluted Moderate Combustion (HNMC), and Balanced Combustion (BC). This dataset essentially covers various combustion conditions of the counterflow burner. Utilizing this dataset for experiments facilitates the validation of the effectiveness of the combustion state recognition model proposed in this paper. Visible flame images of the six combustion states are shown in Fig. 8. To simplify subsequent discussions, the corresponding English abbreviations will be used to refer to each combustion state in the following sections.

In this study, to expand the dataset, each flame image undergoes occlusion processing, where random regions of the image are selected and obscured with black rectangles. Subsequently, noise is applied to the images. These types of noise include rotation, flipping, blurring, noise addition, brightness and contrast adjustments, and perspective transformations, which enhance the diversity and robustness of the images through these operations. By repeatedly using existing images and applying various noise combinations multiple times, we successfully expanded the dataset to 6000 images. After eliminating lower-quality images, we ultimately selected 5820 images covering the six combustion states, thereby providing sufficient data for subsequent model training. These augmentation methods aim to simulate the various changes that

flames in the counterflow burner might undergo, including different lighting conditions, angles, and image interferences. Through these techniques, the model can better adapt to complex real-world combustion environments, enhance its robustness to variations in flame images, and reduce the risk of overfitting. This study employs a stringent dataset partitioning strategy to ensure the fairness and reliability of the deep learning model during training and evaluation processes. The complete dataset comprises 5820 images, covering six different categories. For each category, the dataset is divided into three parts: training set, validation set, and test set, with the number of images in the three parts in the ratio of 70:20:7. This partitioning method ensures that the model can learn from ample training data and evaluate its generalization ability and actual performance through the validation and test sets.



(**a**) OESC (**b**) HOC (**c**) ND-HOC (**d**) LO-LMC (**e**) HNMC (**f**) BC

Figure 8: Visible light flame images of six combustion states

4.2 Evaluation Metrics

To further validate the effectiveness and accuracy of the model, the performance of combustion state recognition was evaluated using Precision (P), Recall (R), F1-score (F1), overall Accuracy (Acc), TOP-3 Accuracy (TOP-3), Log Loss, number of Parameters (Params), and Frames Per Second (FPS).

Precision indicates the proportion of images predicted by the model to belong to a certain combustion state that actually belong to that state. A high precision indicates a low error rate. Recall represents the proportion of images that are correctly identified by the model out of all images that actually belongs to a certain combustion state. A high recall indicates the model is capable of identifying most of the true positive cases with minimal omissions. The F1-score represents the harmonic mean of precision and recall, providing a comprehensive measure of the model's performance in recognizing the six combustion states. Accuracy indicates the proportion of correctly identified images out of all flame images. TOP-3 Accuracy denotes the proportion of cases where the correct category is included among the top three most likely categories predicted by the model. Log Loss measures the difference between the predicted states and the true states. A higher Log Loss indicates lower confidence in the model's predictions. The specific formulas for the above evaluation metrics are provided in Eqs. (22)–(26):

$$P = \frac{TP}{TP + FP} \tag{22}$$

$$R = \frac{TP}{TP + FN} \tag{23}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{24}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{25}$$

$$LogLoss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,C}\log\left(p_{i,C}\right) \tag{26}$$

In the equations, TP (True Positive) represents the number of samples belonging to a certain combustion state and correctly identified as such; FP (False Positive) denotes the number of samples that do not belong to a certain combustion state but are incorrectly classified as such; TN (True Negative) indicates the number of samples that do not belong to a certain combustion state and are correctly classified as other states; FN (False Negative) represents the number of samples that belong to a certain combustion state but are incorrectly classified as other states; N is the total number of samples; C is the number of categories; $y_{i,C}$ denotes the true label of the i-th sample for category C; and $p_{i,C}$ is the predicted probability of category C by the model.

Since the dataset includes six types of flame images, the macro average and weighted average of precision, recall, and F1-score for each category must be calculated. The macro average is the simple average of each metric across all categories without considering the number of samples in each category, reflecting the model's balanced performance across categories; The weighted average, on the other hand, is weighted based on the number of samples in each category, providing a closer approximation of the model's performance in real-world scenarios.

4.3 Training Process

When training the model on the dataset, the learning rate (model_lr) was set to 0.0001, the batch size to 16, the number of epochs to 150, and the gradient clipping parameter to 5.0. For data preprocessing and augmentation, random rotation by 10 degrees, Gaussian blur, adjustments to color saturation and brightness, and the Mixup method were used, followed by resizing and normalization. During training, the AdamW optimizer with L2 regularization was used, and the learning rate was adjusted using a cosine annealing learning rate scheduler. The Exponential Moving Average (EMA) method was employed to smooth time-series data, with the EMA decay coefficient set to 0.9998, making the model more consistent and stable. The Loss and Accuracy (Acc) curves for the InceptionNeXt baseline model and the TSMS-InceptionNeXt are shown in Fig. 9.

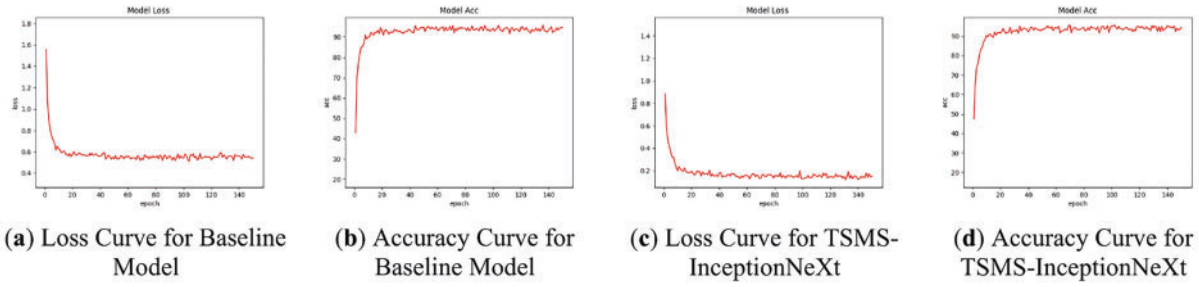


Figure 9: Loss and accuracy curves of the baseline and TSMS-InceptionNeXt during training

From the Loss curve in Fig. 9, it can be observed that the TSMS-InceptionNeXt converges after 80 epochs, indicating that the improved InceptionNeXt model successfully achieves convergence on the dataset, with parameters adapting to the flame image features and effectively performing the recognition task. Additionally, the converged loss value of the TSMS-InceptionNeXt stabilizes around 0.22, significantly lower than that of the original model, indicating better fitting capability, more accurate capture of flame

image features, and reduced errors. Furthermore, the TSMS-InceptionNeXt exhibits faster convergence, demonstrating enhanced optimization efficiency. From the Acc curve in Fig. 9, it is evident that the overall accuracy of the TSMS-InceptionNeXt surpasses that of the original model, indicating improved feature extraction capability.

4.4 Comparative Experiments

To validate the superiority of the improved InceptionNeXt model in the combustion state recognition task, comparative experiments were conducted using other classical models, including RepVGG, SeaFormer, SGFormer, and VanillaNet, on the same dataset. The improved InceptionNeXt model and other models were trained and tested on the same devices and datasets. In the comparative experiments, when selecting other models, the application scenarios, performance metrics, and data types of the other models were similar to those of the model proposed in this study. To ensure fairness in parameter settings, all models used identical parameters: a learning rate (model_lr) of 0.0001, a batch size of 16, 150 epochs, a gradient clipping parameter set to 5.0, and the AdamW optimizer combined with L2 regularization during training. A Cosine Annealing Learning Rate Scheduler was employed to adjust the learning rate, and the EMA method decay coefficient (model_ema_decay) was set to 0.9998. The performance of each model was evaluated based on the macro and weighted averages of Precision, Recall, and F1-score for each category, as well as overall Accuracy, TOP-3 Accuracy, Log Loss, and the number of Parameters. The comparative experiment results are shown in Table 2.

Table 2: Comparative experiment results

Model	Macro Avg			Weighted Avg			Acc/%	TOP-3 Acc/%	Log Loss	Params (MB)	FPS
	Precision	Recall	F1-Score	Precision	Recall	F1-Score					
RepVGG	80.12	74.84	75.80	80.02	75.24	75.89	75.24	91.43	0.6892	29.89	419.33
SeaFormer	83.04	80.67	81.07	82.96	80.95	81.13	80.95	93.81	0.5674	6.18	217.14
SGFormer	77.57	74.35	74.74	77.60	74.52	74.82	74.52	90.48	0.7436	84.06	48.64
VanillaNet	80.85	76.07	76.60	80.82	76.43	76.73	76.43	92.62	0.8407	294.15	255.85
TSMS-InceptionNeXt	87.14	85.45	85.79	86.97	85.71	85.82	85.71	95.95	0.4901	256.59	114.02

From the overall results of the comparative experiments, the TSMS-InceptionNeXt outperformed the comparison models across all metrics. For macro-average F1-scores, the TSMS-InceptionNeXt showed significant improvements of 9.99%, 4.72%, 11.05%, and 9.19% compared to RepVGG, SeaFormer, SGFormer, and VanillaNet, respectively, reflecting more balanced classification performance across all categories. Weighted average F1-scores improved by 9.93%, 4.69%, 11.00%, and 9.09%, respectively, demonstrating that the TSMS-InceptionNeXt achieved higher recognition performance even when considering category sample sizes. This validates the effectiveness of expanding the receptive field in the model. In terms of overall classification accuracy, the TSMS-InceptionNeXt achieved an outstanding performance of 85.71%, surpassing the next best VanillaNet (80.95%) by 4.76%. This indicates the TSMS-InceptionNeXt's superior accuracy in the combustion state recognition task, validating the effectiveness of channel shuffling in enhancing performance. For TOP-3 accuracy, the TSMS-InceptionNeXt reached 95.95%, outperforming the next best SeaFormer (93.81%) by 2.14%. This demonstrates that even when the TSMS-InceptionNeXt's TOP-1 prediction is incorrect, it still includes the correct combustion state in its TOP-3 predictions, enhancing reliability and validating the focus on critical image information. Regarding Log Loss, the TSMS-InceptionNeXt achieved a score of 0.4901, reducing Log Loss by 0.1991, 0.0773, 0.2535, and 0.3506 compared to RepVGG, SeaFormer, SGFormer, and

VanillaNet, respectively. This indicates higher confidence in the combustion state recognition of the TSMS-InceptionNeXt. Although the TSMS-InceptionNeXt has a relatively large parameter size (256.59 MB), it demonstrated higher accuracy and stability across multiple performance metrics. In contrast, SeaFormer has a parameter size of only 6.18 MB but falls short of the TSMS-InceptionNeXt in terms of precision and recall.

In the comparative experiments, the confusion matrices of the proposed TSMS-InceptionNeXt and the comparison models on the test set are shown in Fig. 10. Observing the accurate recognition counts for each state (OESC, HOC, ND-HOC, LO-LMC, HNMC, BC) across models, it is evident that the TSMS-InceptionNeXt achieves a higher total of correctly identified samples (i.e., the sum of diagonal elements for all states) compared to all comparison models. This reflects a consistent improvement across all state categories. The misclassification counts of the TSMS-InceptionNeXt are more evenly distributed, with no concentration in specific states, indicating that the model performs relatively balanced across all combustion states rather than relying on a single class. This validates the effectiveness of incorporating Focal Loss into the model. Specifically, in terms of state categories, the TSMS-InceptionNeXt achieved significant advantages in recognizing easily confusable categories (OESC and LO-LMC). For the OESC state, the number of misclassified samples by the TSMS-InceptionNeXt was reduced by 12, 6, 14, and 10 compared to other models (RepVGG, SeaFormer, SGFormer, and VanillaNet, respectively). For the LO-LMC state, the number of misclassified samples was reduced by 12, 5, 15, and 16 compared to the same models. This indicates that the TSMS-InceptionNeXt is more accurate in identifying subtle differences in easily confusable state categories, thereby validating the importance of focusing on key local detail features in flame images for performance enhancement.

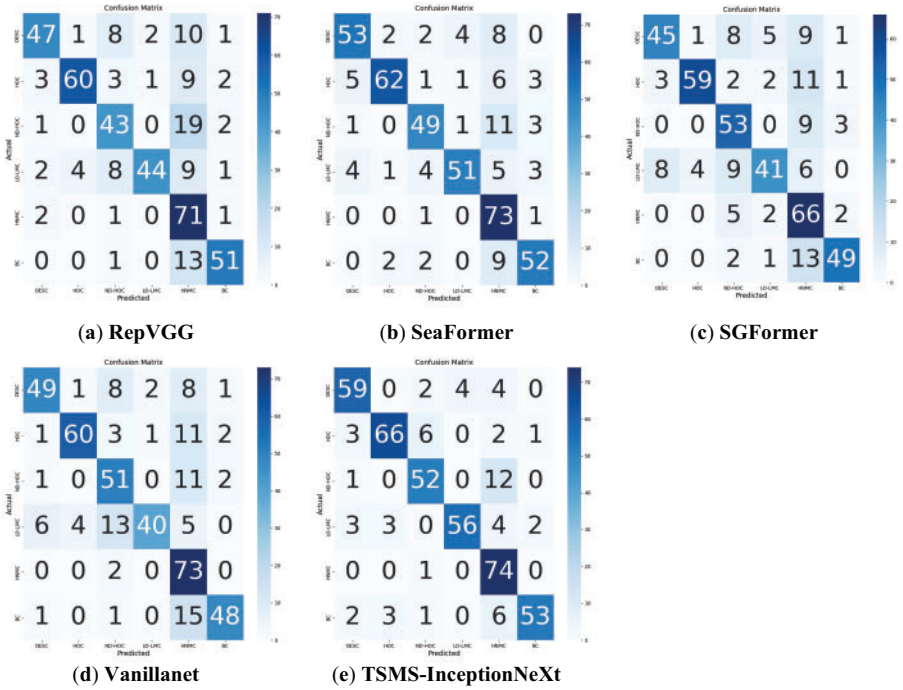


Figure 10: Confusion matrices of the improved and comparison models on the test set

Table 2 shows that the TSMS-InceptionNeXt has slower inference speed than other models, including the traditional InceptionNeXt, due to more complex attention mechanisms and multi-scale convolution modules requiring greater computational resources. However, it achieves significantly higher accuracy,

which is essential for precisely identifying flame combustion states. In tasks involving counterflow burners, where combustion states change slowly, high precision is more critical than speed. Therefore, the current trade-off between precision and computational efficiency is justified. To further enhance computational efficiency, we plan to implement network lightweighting techniques, optimize the model with more efficient architectures, reduce computational load, and explore deployment on edge devices in future work.

4.5 Ablation Study

To validate the effectiveness of the proposed improvements to the InceptionNeXt model for combustion state recognition, InceptionNeXt was used as the baseline model. Under the same dataset conditions, an ablation study was conducted to analyze the contributions of five improvement strategies: Triplet Attention, SimC, MDCPI, ShuffleCS, and Focal Loss. In Table 3, "Groups" represents the experiment indices, "√" indicates the use of a mechanism, and "×" indicates its absence. The results of the ablation study are shown in Table 3.

Table 3: Ablation study results

Groups	Triplet attention	SimC	MDCPI	ShuffleCS	Focal Loss	Macro Avg F1-Score	Weighted Avg F1-Score	Acc/%	TOP-3 Acc/%	Log Loss	Params (MB)
1	×	×	×	×	×	83.61	83.63	83.33	95.48	0.5741	179.61
2	√	×	×	×	×	83.61	83.64	83.33	93.57	0.5658	179.61
3	√	√	×	×	×	83.24	83.22	83.10	94.76	0.5495	179.61
4	√	√	√	×	√	84.72	84.52	84.52	94.76	0.5117	256.19
5	√	√	√	√	√	85.79	85.82	85.71	95.95	0.4901	256.59

Using the baseline InceptionNeXt model's performance across various evaluation metrics as the foundation for the ablation study analysis. Firstly, by introducing Triplet Attention, the weighted average F1-score saw a slight improvement. This is because adding Triplet Attention to the Stages componentenables the model to capture flame feature variations from the height, width, and channel dimensions, thereby enhancing the model's adaptability to dynamic flame changes. Additionally, the overall accuracy remained at 83.33%, and the number of model parameters remained unchanged, validating the effectiveness of incorporating this mechanism in enhancing model performance. Secondly, with the addition of the SimC mechanism, the Log Loss decreased to 0.5495. This is because incorporating the SimC mechanism into the Stages component allows the model to focus on key features important for accurate recognition through the energy function. Experimental results indicate an improvement in the model's stability. Subsequently, by adding the MDCPI mechanism and Focal Loss, the macro and weighted average F1-scores reached 84.72% and 84.52%, respectively, representing increases of 1.48% and 1.3%. The overall accuracy rose to 84.52%, and the Log Loss significantly decreased to 0.5117. This indicates that the combination of MDCPI and Focal Loss plays a significant role in improving recall rates and overall recognition performance, while also being more effective in reducing the confidence of misclassifications. It validates the contribution of expanding the receptive field to enhancing model performance. The calculation of Log Loss is based on the difference between the predicted probabilities and the true labels. The MDCPI mechanism enhances the model's ability to extract boundary, detail, and global information through multi-scale dilated convolutions, thereby improving the model's capability to recognize these hard-to-classify samples. Meanwhile, Focal Loss helps the model focus more on hard-to-classify samples, enhancing the learning of these challenging instances. Since these samples typically have higher losses, Focal Loss can reduce the prediction errors associated with them. Through the synergistic effect of the MDCPI mechanism and Focal Loss, the model is better able

to understand these difficult samples during training, achieve greater accuracy in handling hard-to-classify instances, reduce erroneous predictions, and ultimately attain a significant decrease in Log Loss. Finally, after incorporating the ShuffleCS mechanism, the macro and weighted average F1-scores reached 85.79% and 85.82%, respectively, representing increases of 1.07% and 1.30%. The overall accuracy improved to 85.71%, and the TOP-3 Accuracy rose to 95.95%. The Log Loss further decreased to 0.4901, enhancing the model's prediction stability and recognition performance. This improvement is attributed to the modification of the MlpHead component with the ShuffleCS mechanism, which allows the model to enhance its ability to recognize different states through feature mixing. Although the number of parameters slightly increased to 256.59 MB, the significant performance gains validate the effectiveness of inter-channel information fusion in enhancing model performance.

In the experiments, a synergistic interaction between the MDCPI mechanism and Focal Loss was discovered. The experimental data for different mechanism combinations are presented in Table 4. As shown in Table 4, introducing the MDCPI mechanism or Focal Loss individually results in a decrease in the model's recognition accuracy. However, when used in combination, their performance is enhanced. The MDCPI mechanism enhances both the global and local representations of flame features but simultaneously introduces redundant information, leading to increased noise and altered feature distributions, which negatively impact the model's performance. Focal Loss assigns excessively high weights to hard-to-classify samples in this dataset, thereby diminishing the model's focus on easily identifiable samples. Therefore, when combined, Focal Loss helps the model better utilize the features provided by the MDCPI mechanism and focus on easily confusable categories, thereby improving recognition performance.

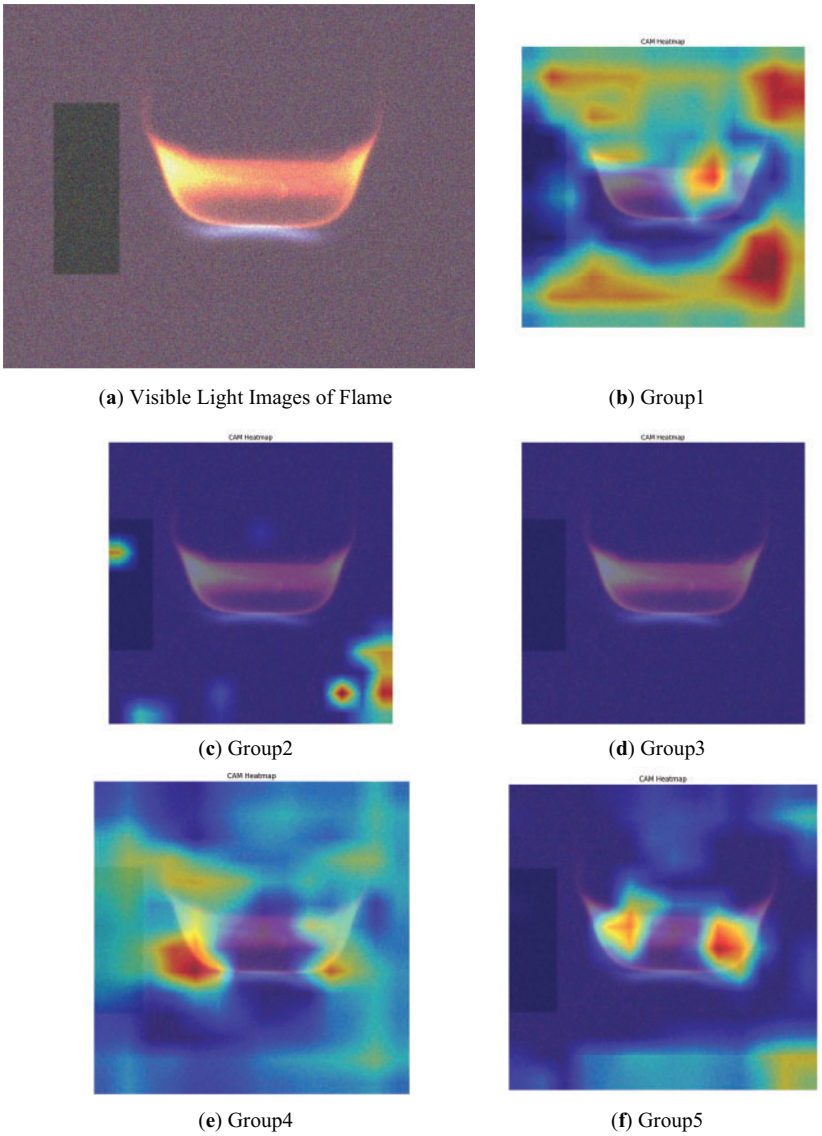
Table 4: Experimental data for different mechanism combinations

Groups	Triplet attention	SimC	MDCPI	ShuffleCS	Focal Loss	Macro Avg F1-Score	Weighted Avg F1-Score	Acc/%	TOP-3 Acc/%	Log Loss	Params (MB)
1	×	×	√	×	×	80.14	80.31	80.00	94.29	0.6068	256.19
2	×	×	×	×	√	82.50	82.49	82.14	93.57	0.5645	179.61
3	×	×	√	×	√	84.07	84.20	84.05	93.81	0.5382	256.19

A heatmap is a visualization tool that represents data intuitively through varying color intensities. Heatmaps can illustrate the key regions that a model focuses on when processing images. Warmer colors indicate higher attention to those areas, while cooler colors denote lower attention levels. In the context of flame combustion state recognition, heatmaps can reveal the flame regions that the model deems critical for determining the combustion state, such as high-temperature areas and intensely burning sections. This aids in understanding the model's decision-making process, allowing us to assess whether the model accurately focuses on task-relevant image features rather than being distracted by irrelevant backgrounds or noise.

We present GradCAM heatmaps for a typical flame image from the dataset, as shown in Fig. 11. These correspond to the baseline InceptionNeXt model, the progressively TSMS-InceptionNeXts, and the fully improved TSMS-InceptionNeXt model. The group numbers of the model heatmaps in Fig. 11 correspond to the group numbers of the models in Table 3. In Fig. 11b, Group 1 represents the heatmap of the baseline InceptionNeXt model, and in Fig. 11f, Group 5 represents the heatmap of our proposed TSMS-InceptionNeXt model. The heatmap of the original model exhibits widespread and scattered activation regions, indicating that the model's focus areas within the image are not sufficiently concentrated, resulting in imprecise recognition of combustion states. As the model undergoes gradual optimization, the heatmaps of subsequent models display more concentrated and distinct activation regions, primarily focused on the core areas of the

flame and surrounding key regions. This indicates that during the improvement process, the model gradually learned to focus on features closely related to the combustion state, thereby enhancing recognition accuracy and reliability. The heatmap of the fully TSMS-InceptionNeXt ultimately shows highly concentrated and intense activation regions, clearly focusing on the main combustion areas of the flame. This demonstrates a significant enhancement in the model's ability to identify key flame features, validating the effectiveness of the proposed model optimization methods in improving the accuracy of flame combustion state recognition.



(**a**) Visible Light Images of Flame (**b**) Group1

(**c**) Group2 (**d**) Group3

(**e**) Group4 (**f**) Group5

Figure 11: Flame image and corresponding heatmaps from ablation experiments

5 Conclusion

As a combustion device widely utilized in the industrial, energy, and combustion research sectors, accurately identifying the internal combustion states of a counterflow burner is of significant importance. Addressing the issues in traditional and existing deep learning methods, this study selects Inception-NeXt, a state-of-the-art model in image recognition, as the baseline model and proposes an improved

InceptionNeXt-based combustion state recognition model for counterflow burners, optimized for feature extraction. First, the Triplet Attention mechanism was introduced to apply attention across image width, height, and Red Green Blue (RGB) dimensions, enhancing the model's adaptability to capture dynamic flame features. Second, a SimC mechanism was added after the Inception depthwise convolution to address the loss of critical flame information inherent to Inception convolutions. Next, a MDCPI mechanism was embedded after the Stages componentto expand the receptive field, capture broader contextual information in flame images, and synergize with Focal Loss to enhance recognition performance. Finally, a ShuffleCS mechanism was added to the classification head to address the original MlpHead's lack of channel interaction. By fusing features, the MlpHead's ability to recognize different states was improved, further enhancing the model's performance. Experimental results demonstrate that the proposed method is effective for combustion state recognition tasks, achieving an overall accuracy of 85.71%, outperforming the baseline and comparison models. However, the TSMS-InceptionNeXt has a relatively large parameter size, and the proposed improvements were primarily focused on optimizing feature extraction, which presents certain limitations. Future work will focus on finding a balance between overall accuracy and lightweight design to further enhance recognition performance and exploring alternative approaches to optimize model performance.

Acknowledgement: The authors would like to express their gratitude to Ruiyuan Kang, Panos Liatsis, and Dimitrios C. Kyritsis for providing the dataset used in this study.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Huiling Yu, Xibei Jia; data collection: Xibei Jia; analysis and interpretation of results: Xibei Jia; draft manuscript preparation: Xibei Jia, Yongfeng Niu and Yizhuo Zhang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data is not available.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Malpica GR, Ciottoli PP, Valorani M, Im HG. Local combustion regime identification using machine learning. Combust Theory Model. 2022;26(1):135–51. doi:10.1080/13647830.2021.1991595.
2. Sitaraman R, Batool S, Borhan H, Velni JM, Naber JD, Shahbakhti M. Machine learning-based classification of combustion events in an RCCI engine using heat release rate shapes. IFAC-Pap. 2022;55(37):601–7. doi:10.1016/j.ifacol.2022.11.248.
3. Bhattacharya C, Ray A. Thresholdless Classification of chaotic dynamics and combustion instability via probabilistic finite state automata. Mech Syst Signal Process. 2022;164(2):108213. doi:10.1016/j.ymssp.2021.108213.
4. Compais P, Arroyo J, Castán-Lascorz MÁ, Barrio J, Gil A. Detection of slight variations in combustion conditions with machine learning and computer vision. Engineering Appl Artif Intell. 2023;126:106772. doi:10.1016/j.engappai.2023.106772.
5. Bukkarapu KR, Krishnasamy A. Evaluating the feasibility of machine learning algorithms for combustion regime classification in biodiesel-fueled homogeneous charge compression ignition engines. Fuel. 2024;374:132406. doi:10.1016/j.fuel.2024.132406.
6. Choi O, Choi J, Kim N, Lee MC. Combustion instability monitoring through deep-learning-based classification of sequential high-speed flame images. Electronics. 2020;9(5):848. doi:10.3390/electronics9050848.

7. Roncancio R, Kim J, El Gamal A, Gore JP. Data-driven analysis of turbulent flame images. In: AIAA Scitech 2021 Forum; 2021 Jan 11–21. doi:10.2514/6.2021-1787.
8. Omiotek Z, Kotyra A. Flame image processing and classification using a pre-trained VGG16 model in combustion diagnosis. Sensors. 2021;21(2):500. doi:10.3390/s21020500.
9. Pereira R, Rocha E, Pinho D, Santos JP. Convolutional neural networks for identification of moving combustion chambers entering a brazing process. Procedia Comput Sci. 2023;217:1106–16. doi:10.1016/j.procs.2022.12.309.
10. Natsui S, Goto Y, Takahashi J, Nogami H. Pattern analysis of the combustions of various copper concentrate tablets using high-speed microscopy and video-based deep learning. Chem Eng Sci. 2023;276(9-10):118822. doi:10.1016/j.ces.2023.118822.
11. Pan X, Tang J, Xia H, Yu W, Qiao J. Combustion state identification of MSWI processes using ViT-IDFC. Eng Appl Artif Intell. 2023;126(6):106893. doi:10.1016/j.engappai.2023.106893.
12. Wang Y, Liang S, Xu ZQJ, Zhang T, Ji L. Artificial neural network aided unstable combustion state prediction and dominant chemical kinetic analysis. Chem Eng Sci. 2024;300:120567. doi:10.1016/j.ces.2024.120567.
13. Wu X, Zhang H, Chen H, Wang S, Gong L. Combustion optimization study of pulverized coal boiler based on proximal policy optimization algorithm. Appl Therm Eng. 2024;254(6):123857. doi:10.1016/j.applthermaleng.2024.123857.
14. Lv Y, Qi X, Zheng X, Fang F, Liu J. Unsupervised quantitative judgment of furnace combustion state with CBAM-SCAE-based flame feature extraction. J Energy Inst. 2024;116:101733. doi:10.1016/j.joei.2024.101733.
15. Yu F. Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122. 2015.
16. Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks. Adv Neural Inf Process Syst. 2015;28:2017–25.
17. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 7132–41.
18. Zhang X, Zhou X, Lin M, Sun J. Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 6848–56.
19. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 11534–42.
20. Yang L, Zhang RY, Li L, Xie X. Simam: a simple, parameter-free attention module for convolutional neural networks. In: International Conference on Machine Learning. 2021 Jul 18–24. Cambridge, MA, USA: PMLR. p. 11863–74.
21. Misra D, Nalamada T, Arasanipalai AU, Hou Q. Rotate to attend: convolutional triplet attention module. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021 Jan 5–9: Piscataway, NJ, USA: IEEE. p. 3139–48. doi:10.1109/WACV48630.2021.00318.
22. Ding X, Zhang X, Ma N, Han J, Ding G, Sun J. Repvgg: making vgg-style convnets great again. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 13733–42.
23. Wan Q, Huang Z, Lu J, Gang YU, Zhang L. Seaformer: squeeze-enhanced axial transformer for mobile semantic segmentation. In: The Eleventh International Conference on Learning Representations; 2023 May 1–5; Kigali Rwanda.
24. Wu Q, Zhao W, Yang C, Zhang H, Nie F, Jiang H, et al. Simplifying and empowering transformers for large-graph representations. Adv Neural Inf Process Syst. 2024;36:64753–73.
25. Chen H, Wang Y, Guo J, Tao D. Vanillanet: the power of minimalism in deep learning. Adv Neural Inf Process Syst. 2024;36:7050–64.
26. Yu W, Zhou P, Yan S, Wang X. Inceptionnext: when inception meets convnext. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–22; Seattle, WA, USA. p. 5672–83.

27. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 2818–26.
28. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 11976–86.
29. Dosovitskiy A. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
30. Li X, Zhang Y, Liu Y, Yao X, Zhou X. Arbitrary shape text detection fusing InceptionNeXt and multi-scale attention mechanism. J Supercomput. 2024;80(17):25484–509. doi:10.1007/s11227-024-06418-w.
31. Lau KW, Rehman YAU, Po L-M. AudioRepInceptionNeXt: a lightweight single-stream architecture for efficient audio recognition. Neurocomputing. 2024;578(2):127432. doi:10.1016/j.neucom.2024.127432.
32. Lin T. Focal loss for dense object detection. arXiv:1708.02002. 2017.
33. Kang R, Liatsis P, Kyritsis DC. Flame-state monitoring based on very low number of visible or infrared images via few-shot learning. arXiv:2210.07845. 2022.