M^{3}RTNet：Combustion State Recognition Model of MSWI Process Based on Res-Transformer and Three Feature Enhancement Strategies

Jian Zhang *, Rongcheng Sun , Jian Tang , Haoran Pei

Posted Date: 18 March 2025

doi: 10.20944/preprints202503.1326.v1

Keywords: municipal solid waste incineration; feature enhancement; Resnet; Transformer; EMA; context feature fusion

Article

M³RTNet: Combustion State Recognition Model of MSWI Process Based on Res-Transformer and Three Feature Enhancement Strategies

Jian Zhang [1,*], Rongcheng Sun [1], Jian Tang [2] and Haoran Pei [3]

[1] School of Computer Science, Nanjing University of Information Science & Technology, Nanjing 210044, China; jianzhang@nuist.edu.cn (J.Z.); 202312490650@nuist.edu.cn (R.S.)

[2] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China; freeflytang@bjut.edu.cn

[3] Department of Computer and Information Science, The University of Mississippi, Oxford, USA; hpei@go.olemiss.edu

* Correspondence: jianzhang@nuist.edu.cn

Abstract: Aiming at the problem of low accuracy of flame combustion state recognition in incinerator during municipal solid waste incineration, this paper proposes a Res-Transformer flame combustion state recognition model based on three feature enhancement strategies. In this paper, Res-Transformer is used as the backbone network of the model to effectively integrate local flame combustion features and global features. Firstly, we introduce an efficient multi-scale attention module into Resnet, which uses a multi-scale parallel sub-network to establish long and short dependencies; Then, a deformable multi-head attention module is designed in the Transformer layer, and the deformable self-attention is used to extract long-term feature dependencies. Finally, we design a context feature fusion module to efficiently aggregate the spatial information of the shallow network and the channel information of the deep network, and enhance the cross-layer features extracted by the network. Experiments demonstrate the effectiveness and robustness of this method.

Keywords: municipal solid waste incineration; feature enhancement; Resnet; Transformer; EMA; context feature fusion

1. Introduction

With the global industrialization process and social and economic development, the amount of municipal solid waste(MSW) is increasing significantly [1,2]. This rapid growth has brought serious challenges to municipal solid waste treatment [3]. In this context, Municipal solid waste incineration (MSWI) technology stands out as a widely adopted treatment method [4]. MSWI is an efficient waste treatment method. The core of the technology is to convert municipal solid waste into ash, flue gas and recoverable heat energy through a high-temperature combustion process [5]. This technology can not only greatly reduce the volume of waste, but also effectively control the generation of secondary pollution, and can also realize the reuse of resources [6]. The MSWI process consists of six stages: solid waste fermentation, solid waste incineration, waste heat exchange, steam generation, flue gas purification, and flue gas emission [7]. Among them, the solid waste incineration stage occupies a core position in the MSWI process, which is not only an effective way to reduce and recycle waste, but also its technical control directly affects the flue gas purification and emission quality [8]. Reasonable regulation of the incineration process can reduce the generation of harmful substances, ensure that environmental standards are met, and protect the environment and public health [9]. The essence of solid waste combustion is a thermochemical treatment process, its core lies in the oxidation reaction under high temperature conditions, the combustible substances in solid waste into gaseous products (such as carbon dioxide, water vapor, nitrogen, etc.) and a small amount of solid residues

 doi:10.20944/preprints202503.1326.v1

(such as ash). This process not only realizes the volume reduction and harmless treatment of waste, but also is accompanied by the release and conversion of energy, which provides the basis for subsequent energy recovery and utilization [10].

Therefore, the reasonable regulation of solid waste incineration process is particularly important. However, this process still faces multiple challenges. Among them, the instability of combustion state is a particularly intractable problem, which directly leads to the difficulty of pollutant discharge to meet the standards [11]. At the same time, it also exacerbates the problems such as slag, ash accumulation and equipment corrosion in the furnace, and may even cause safety accidents such as furnace explosion in serious cases. In view of this, it is particularly important to maintain the combustion stability during the incineration process [12], which is directly related to the efficiency and environmental effectiveness of the entire treatment process. It is worth noting that the operation of many current MSWI facilities still relies on manual intuitive judgment of the flame burning state during solid waste incineration to adjust the control strategy. Although this method is practical to some extent, it is easily limited by personal experience, subjective judgment and insufficient intelligence level, so it is difficult to meet the urgent needs of MSWI process optimization operation. Therefore, exploring and constructing a combustion state recognition model that can not only adapt to the complex and variable environment of MSWI, but also have a high degree of robustness has become an important direction of current research. The construction of this model not only requires accurate identification of the combustion state, but also should be able to effectively guide the adjustment of the control strategy to ensure the efficient and stable operation of the MSWI process and meet the strict environmental emission standards at the same time.

In recent years, artificial intelligence technology has been more and more widely used in the field of MSWI combustion state recognition [13]. For example, Duan et al. [14] proposed a model to identify burning conditions in MSWI process based on multi-scale color moment features and Random Forest (RF). Firstly, the image is preprocessed by dehazing and denoising. Then, based on the pre-set scale, the color moment features of flame images at different scales are extracted by using sliding Windows. Finally, taking the classification accuracy as the evaluation criterion, the RF algorithm based on feature selection was used to realize the accurate identification of the burning state. However, due to the scarcity of abnormal burning state images and the high labeling cost, it is difficult to obtain enough image burning state anomalies. Therefore, Guo et al. [15] proposed a deep convolutional generation based burning State adversarial Network (DCGAN) for abnormal image generation. Moreover, Ding et al. [16] introduced the typical MSWI process and summarized the main control requirements. Then, the related control methods were summarized and the applicability and development status were analyzed. Finally, the main problems and difficulties in the current MSWI process control were summarized. At the same time, another method by Zhang et al. [17] proposed a combustion state recognition model for MSWI process based on convolutional neural network. Guo et al. [18] constructed a flame image dataset classified according to the position of the burning line of the flame in the MSWI burning image, and used a variety of deep learning models to verify the feasibility of flame burning state recognition, which provides a certain reference for subsequent research work. Tian et al. [19] firstly enhanced the original flame image by data enhancement methods such as rotation and adding noise to expand the scale of labeled samples and overcome the problem of high cost of manual labeling. Then, the VGG19 model pre-trained on ImageNet is used as the base model, and the output of the last layer of the middle layer is used as the model output. By enhancing the flame image data set and fine-tuning the model parameters, feature transfer learning is realized. Pan et al. [20] chose Lenet-5 as the recognition model for flame burning state recognition. In order to obtain a suitable CNN model to automatically extract the flame image features, the authors tested the influence of multiple CNN networks on the flame image features and obtained the optimal structure setting of the CNN network. Moreover, Pan et al. [21] combined the experience of industry experts and the research results in related fields to study the construction of the classification standard and benchmark database of the flame combustion state of waste disposal. The combustion classification criteria based on normal burning, partial burning, channeling burning

 doi:10.20944/preprints202503.1326.v1

and smoldering are expounded, and the flame combustion state image database for machine learning is constructed. Finally, based on various classical algorithms in the field of machine vision, the flame combustion image database is modeled and tested. Motivated by this, paper [22] used deep Forest Classification with Improved ViT (VIT-IDFC) algorithm to identify burning states. Based on the transformer encoding layer of the pre-trained ViT model, multi-layer visual transformation features are extracted from the flame image. The experience of domain experts is used to select deep features. By taking the selected ViT visual transformation features and the original flame image as the input of the cascade forest, an IDFC model is constructed to identify the flame burning state. Yang et al. [23] proposed a YOLOv5-based method for MSWI process burning state recognition, which uses a backbone network for feature extraction and a head layer for state recognition. Guo et al. [24] performed data augmentation through DCGAN, then expanded the sample again through non-generative data augmentation, and finally constructed a convolutional neural network to identify the burning state. Hu et al. [25] used artificial multi-exposure image fusion dehazing algorithm, feature normalization, trap filter, median filter and other preprocessing means to dehaze and represent the flame image. Then, multiple features such as brightness, flame, color and principal component were extracted from the flame image to represent the multi-view image, and the multi-view features were reduced based on mutual information. Finally, a burning state recognition model was constructed based on image features and deep forest model.

The above studies show that artificial intelligence techniques, especially deep learning-based methods, are able to show good performance in the existing MSWI burning state recognition due to their strong feature learning and expression capabilities. However, there are still some problems in the existing recognition methods. For example, the burning flame of MSWI image is diverse, the shape is complex, the size is also very different, and the boundary with the background image is fuzzy, which leads to the model cannot fully extract effective features for recognition. Aiming at the above problems, we propose M^{3}RTNet model, and use three different feature enhancement strategies to enhance the feature extraction ability of the model, so as to better extract flame burning features. The main contributions are as follows: (1) We use the Res-Transformer structure as the backbone network, and combine the local feature extraction ability of Resnet and the global feature extraction advantage of Transformer to extract local flame combustion features and global information. (2) An efficient multi-scale attention(EMA) module is introduced into the Resnet network, and a multi-scale parallel sub-network is used to establish the long and short dependence relationship to strengthen the recognition ability of the flame burning area, so as to improve the classification performance of the residual neural network. (3) A deformable multi-head attention module(DMAM) is designed in the Transformer layer, which uses deformable self-attention to extract long-term feature dependencies and enhance its global feature extraction ability. (4) A context feature fusion module(CFFM) is designed to efficiently aggregate the spatial information of the shallow network and the channel information of the deep network, and enhance the cross-layer features extracted by the network. Experimental results show that the proposed model effectively improves the recognition accuracy of flame burning state in MSWI process.

2. Materials and Methods

2.1. Introduction of Flame Combustion Image

In the solid waste combustion stage, the combustion process can be subdivided into three stages, namely drying, burning and embers. After solid waste enters the incinerator from the feed port, it first enters the drying stage. At this stage, the water in the waste is rapidly evaporated by the action of the high temperature furnace gas. With the removal of moisture, the waste gradually enters the combustion phase. At this stage, the combustible components in the waste react violently with the oxygen in the furnace, releasing a large amount of heat energy and generating flue gas and ash. This process is the key link of energy conversion in solid waste incineration, and it is also the stage with the highest temperature and the most intense reaction. Subsequently, the gas produced by

 doi:10.20944/preprints202503.1326.v1

combustion and some incompletely burned solid particles enter the embers stage. At this stage, the remaining combustible material continues to react with oxygen until it is completely burned out, forming the final flue gas. At the same time, ash and slag accumulate at the bottom of the furnace and are eventually discharged out of the furnace through the slag discharge system. In this study, combined with Pan et al. 's work in paper [21], the MSWI flame burning states are divided into four states: normal burning, partial burning, channeling burning and smoldering.

As shown in Figure 1, it shows four typical flame burning states. Figure 1 (a) shows normal burning. It can be seen that the burning line is distributed in a straight line and concentrated in the pixel band corresponding to the burning section, and the burning is bright and smooth. Figure 1 (b) shows partial burning. It can be seen that the burning line is curvilinear and runs through the burning area. The flame heights are scattered, bright but scattered. Figure 1 (c) shows the channeling combustion, the combustion line is scattered distribution, and the flame is locally channeling. Figure 1 (d) shows the braising. It can be seen that there is a large area of black block area caused by lack of fire inside the furnace.

Figure 1. Typical flame burning state of the MSWI process:(a)normal burning, (b) partial burning, (c) channeling burning, (d)smoldering.

2.2. Methods

In order to effectively use the global and local features in MSWI flame burning images and improve the recognition ability of the model for different types of MSWI flame burning states, we propose an MSWI flame burning state recognition model based on three feature enhancement strategies: M^{3}RTNet. The overall structure of this model is shown in Figure 2.

 doi:10.20944/preprints202503.1326.v1

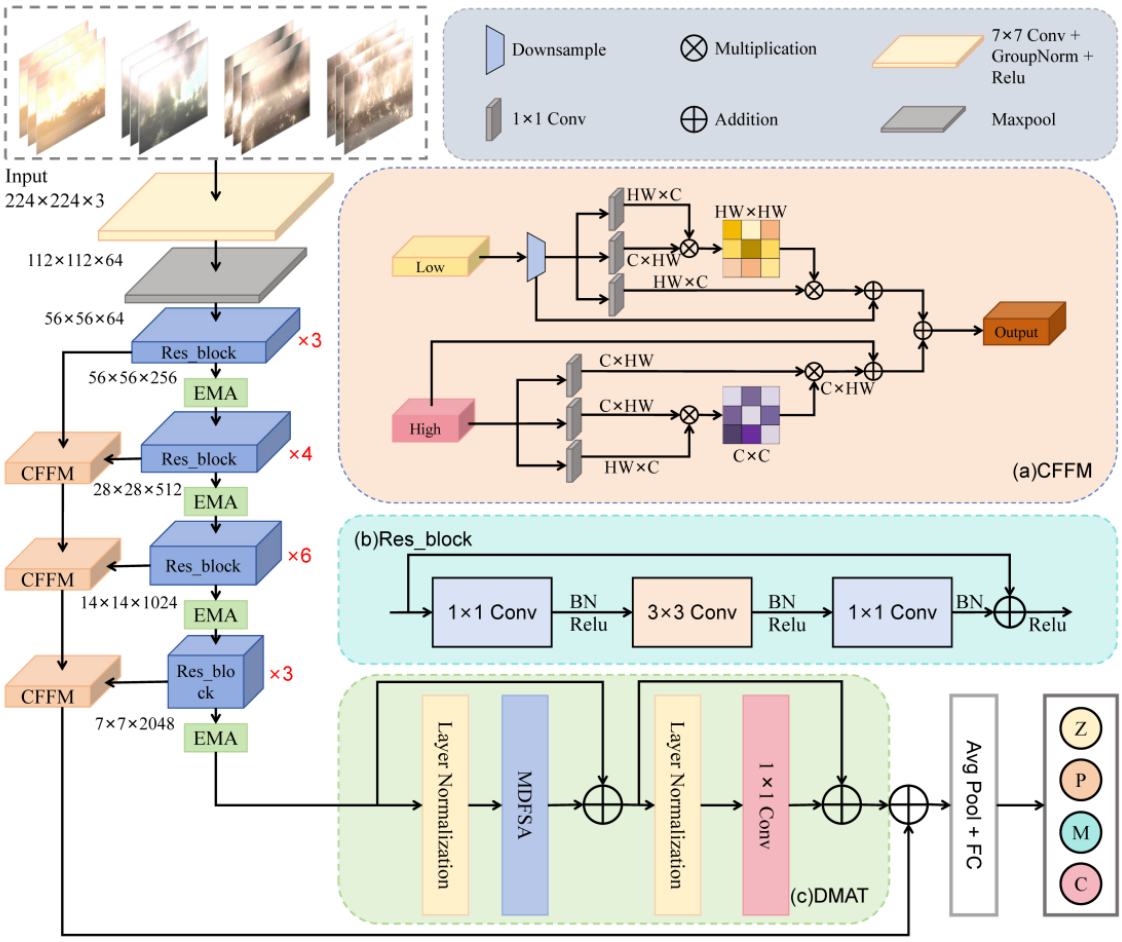

Figure 2. The overall structure of the M³RTNet model.

M³RTNet is an end-to-end architecture that consists of four key components: a multi-scale attention residual network, a deformable multi-head attention Transformer (DMAT), three context feature fusion modules(CFFM), and a classifier.

Firstly, the flame burning image is input into the multi-scale attention residual network to initially extract the flame features. Subsequently, these feature maps are fed into the proposed deformable multi-head attention Transformer to establish a wide range of feature dependencies. Thus, the multi-head attention feature map is obtained. At the same time, we extract the features of each layer of Resnet to fuse the context information, so as to improve the classification performance of the network. Finally, the features obtained by the deformable multi-head attention Transformer and the context feature fusion modules are fused, and then input into the classification layer to realize the recognition of MSWI images. Each submodule of M³RTNet will be discussed in detail in the following subsections.

2.2.1. The Multi-Scale Attention Residual Network

The multi-scale attention residual feature extraction network is improved on the basis of Resnet50 [26], which is mainly composed of Stem module, residual block (Res_block) and Efficient Multi-scale Attention (EMA) module. Firstly, the flame burning image is input into the Stem module, which includes 7×7 convolutional layer, group normalization layer, ReLU activation function and Max pooling layer. Then, three residual blocks are stacked repeatedly in Stage1, four residual blocks in Stage2, nine residual blocks in Stage3 and six residual blocks in Stage4. All residual blocks adopt the basic bottleneck residual block, as shown in Figure 2 (b). The residual block consists of 1×1 convolution, 3×3 convolution, batch normalization and Relu activation function, and the residual connection is used to effectively alleviate the gradient disappearance problem. On this basis, we

 doi:10.20944/preprints202503.1326.v1

introduce an efficient multi-scale attention module [27] after each Stage, which makes the model focus on the burning area to extract more useful features in the flame burning image.

The overall structure of the EMA module is shown in Figure 3. It uses the feature grouping strategy to process the input feature data in parallel, so as to accelerate the model training. At the same time, it integrates multi-scale parallel subnetworks and cross-space learning methods to capture both short-term and long-term dependencies.

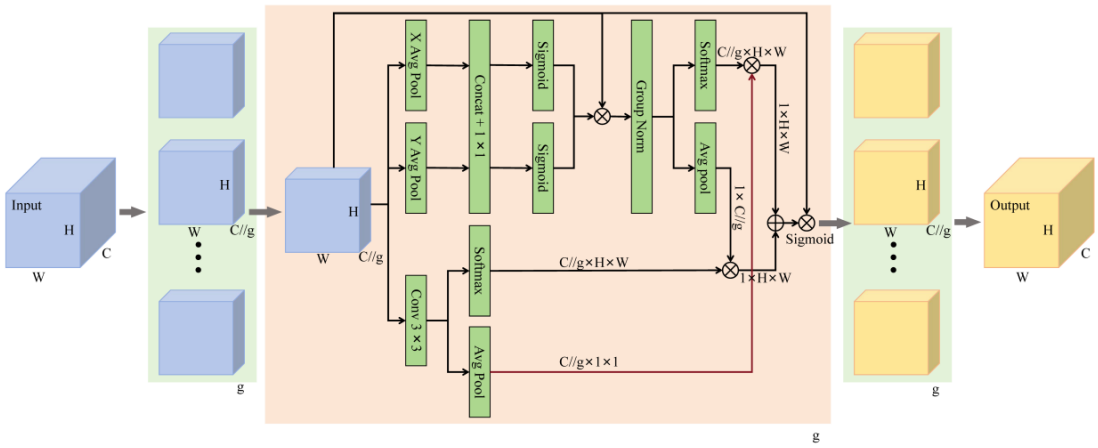

Figure 3. The overall structure of the EMA module.

As shown in the figure, for a given input feature map, EMA module first divides it into g sub-feature groups, thus letting each sub-feature group learn different semantics. The feature grouping method enables the model to allocate and process the model on more GPU resources. This grouping method not only strengthens the feature learning of semantic regions, but also compresses the noise. Then, the EMA module adopts three parallel paths to extract the attention weight descriptors of the grouped feature maps. Two paths are 1x1 branches and the third path is 3x3 branches. A one-dimensional global average pooling operation is used in the 1x1 branch to encode channel information in two spatial directions, respectively. The 3x3 branch captures the multi-scale feature representation by 3x3 convolution. In this way, EMA is able to not only encode information across channels to adjust the importance of different channels, but also retain precise spatial structure information into the channels. EMA enriches feature aggregation by providing cross-spatial information aggregation methods in different spatial dimension directions. Specifically, the output of the 1x1 branch encodes global spatial information via 2D global average pooling, while the output of the 3x3 branch is directly transformed into the corresponding dimensional shape. These outputs are then aggregated by matrix dot product operations to generate the first spatial attention map. Finally, the output feature maps within each group are aggregated by a Sigmoid function of the two generated spatial attention weight values, capturing pixel-level pairing relationships and highlighting the global context of all pixels [28].

In summary, the EMA module is a parallel attention mechanism that is mainly used in computer vision tasks. Its main goal is to help the model capture the interaction between features at different scales and thus improve the performance of the model.

2.2.2. The Deformable Multi-Head Attention Transformer

The feature map extracted by the multi-scale attention residual network learns the local context details from the input image, but it does not contain the global context information, which is crucial in MSWI image analysis. Therefore, we introduce a deformable multi-head attention Transformer to capture local and global feature relationships in feature maps. This ultimately helps the model to learn salient patterns from burning regions. The framework of this module is shown in Figure 2(c), which consists of several main components, namely layer normalization, deformable multi-head

 doi:10.20944/preprints202503.1326.v1

attention module, 1×1 convolution, and residual connection. The purpose of introducing deformable multi-head attention module in it is to use the self-attention of deformable convolutions to extract long-term feature dependencies, so as to obtain a refined feature map.

The deformable multi-head attention Transformer takes as input the feature maps generated in the backbone network and then utilizes layer normalization to normalize each feature activated, which helps to stabilize and speed up the learning process. Then, the normalized feature map is input into the deformable multi-head attention module to obtain the attention feature map. To enhance feature propagation throughout the module, residual connections are introduced in this study. We take advantage of the residual connection, add and fuse the input and output, then further perform layer normalization and a 1×1 convolution to obtain a better feature representation, and finally obtain the final feature map using the residual connection.

Figure 4 illustrates the structure of the deformable multi-head attention module used in this study. In the solid waste incineration flame image, the burning area has a unique pattern of random spatial distribution. Therefore, extracting these features from different burning regions is extremely important for better recognition. Self-attention is a core component of the Version Transformer, which helps model a wide range of global feature relationships and allows the model to focus on key domains of the input. However, self-attention utilizes standard convolutions with a fixed kernel size of 1×1, thus extracting features from a fixed receptive field. This procedure does not take into account the need for different receptive fields. Therefore, we apply a dynamic filtering method called deformable convolution [29] on the spatial dimension of SA to generate feature tensors. The deformable convolution provides the benefit of an adjustable receptive field that varies according to the scale of the burning region and thus has the ability to adapt to geometric changes in the burning region. The deformable convolution gives the convolution window I={(-1, -1), (-1,0),... ,(1,1)} on the feature map adds an offset $\{\Delta d_n | n=1,\dots, D\}$, where $D=|I|$. Therefore, for each position d_0 in the output feature map, *DFConv* is calculated as follows:

$$DFConv(d_0) = \sum w(d_n)x(d_0 + d_n + \triangle d_n) \quad (1)$$

For the input feature map F_n in Figure 4, we first apply a convolution operation to it with a kernel size of 1×1, and then calculate its query, key and value by deformable convolution as follows:

$$Q(F_n) = DFConv_{3\times3}(Conv_{1\times1}(F_n)) \quad (2)$$

$$K(F_n) = DFConv_{3\times3}(Conv_{1\times1}(F_n)) \quad (3)$$

$$V(F_n) = DFConv_{3\times3}(Conv_{1\times1}(F_n)) \quad (4)$$

Then we perform spatial reshaping, and obtain the spatial similarity of the feature vectors by matrix multiplication. Finally, we generate the spatial attention feature map by the activation function:

$$A_{sp} = \sigma(Q(F_n) \otimes K(F_n)^T) \quad (5)$$

Furthermore, the spatial attention feature map is multiplied with $V(F_n)$, and then the reshaping operation is performed again to obtain the attention feature map based on deformable convolution. To explore rich spatial contextual information from the input feature maps, we use four empirically determined heads. Finally, we perform element addition on the four deformable attention feature maps to obtain the final attention feature map. Through the introduction of deformable multi-head attention, the model is able to extract sufficient spatial feature relationships from the input feature maps. In short, the use of multi-head self-attention mechanism and deformable convolutions in the proposed deformable multi-head attention Transformer helps to extract global and local feature dependencies.

 doi:10.20944/preprints202503.1326.v1

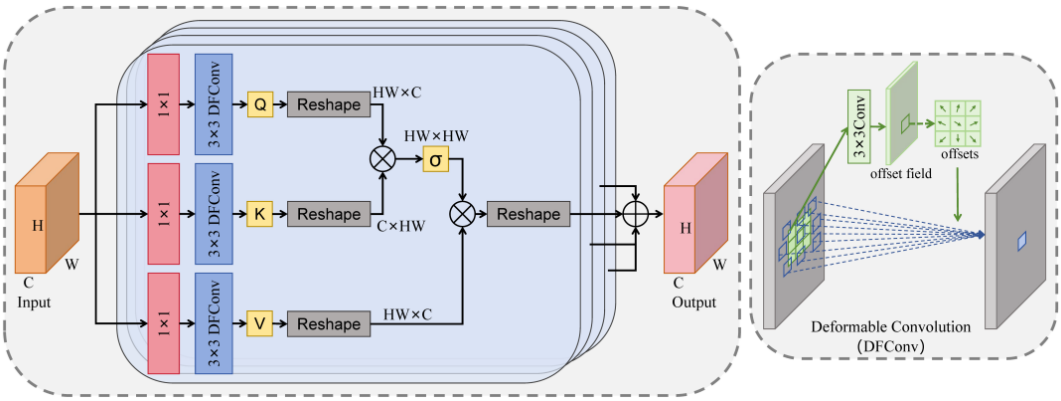

Figure 4. The overall structure of the deformable multi-head attention module.

2.2.3. The Context Feature Fusion Module

The convolution operation in the residual network loses the underlying texture details while extracting features, so that the high-level features and low-level features are distributed at both ends of the network. High-level features have stronger semantic information, but their resolution is low and their perception of details is poor. Shallow features have high resolution and contain more information such as location details, edges and textures, but they are low semantic and noisy due to inadequate feature extraction. In addition, the information concerned by different layers in the feature extraction network is also different, and using the features of different layers to fuse context information can improve the classification performance of the network. However, simple addition is easy to cause information redundancy and cannot make full use of the advantages of both. Therefore, this paper designs a context information fusion module to make up for the lack of deep semantic information with shallow semantic information, as shown in Figure 1 (a). The low-semantic information such as texture and shape of the shallow network is enhanced by spatial attention, and the high-semantic information of the deep network is enhanced by channel. The filtered channel and spatial information were added to efficiently fuse the context information of the shallow and deep layers of the image, so as to retain more useful information and improve the classification performance of the model.

For the low-level feature map of the original input, it is first downsampled to adjust its shape to be consistent with the high-level feature map. Then it goes through three identical 1×1 convolution branches for feature mapping, whose original size is C×H×W, and then transpose them to map the 3D features into 2D features, and the shapes become HW×C, C×HW, and HW×C, respectively. The first two branches are multiplied to obtain the spatial attention value, and then the third branch is multiplied to obtain the feature map filtered in the spatial dimension. Finally, the shape of the feature map is transposed to C×H×W.

For the high semantic feature map E, the same three same 1×1 convolution branches were used for feature mapping, and then the transpose operation was performed to make their shapes into C×HW, C×HW and HW×C. The feature maps of the last two branches were multiplied to obtain the channel attention value, and then the filtered feature map in the channel dimension was obtained by multiplying it with the first branch. Finally, the feature map shape is transposed to C×H×W.

Finally, the feature maps filtered by high and low layers were added to obtain the feature result map.

2.2.3. The Classifier

In the classification layer, we additively fuse the feature maps obtained from the deformable multi-head attention layer and the context feature fusion attention layer, then use the global average pooling layer, and finally use the fully connected layer with two nodes to identify the burning state of the MSWI image.

 doi:10.20944/preprints202503.1326.v1

3. Experiments and Results

3.1. Introduction of Flame Combustion Image

The experiments in this study used Intel(R) Core (TM) i5-13400F processor 4.60 GHz CPU, Windows 11(64-bit) operating system, pytorch2.1.0 framework and NVIDIA CUDA interface model for acceleration.

The network input image size is 224×224×3 pixels, and the selected optimization strategy is Adam algorithm. We used 16 batch sizes of the dataset to train the model for 100 epochs. No pre-trained models were used in any of the experiments, and each model was trained starting from an initial state.

3.2. Evaluation Metrics

Through the quantitative comparison of the experimental results of the classification model, the advantages and disadvantages of the classification model can be judged. We mainly use accuracy (Acc), precision (Pre), recall (Rec) and F1 score as evaluation indicators to analyze the recognition effect of the network model proposed in this study on the burning state of MSWI. The mathematical expression of the evaluation index is as follows:

$$\mathrm{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$\mathrm{Pre} = \frac{TP}{TP + FP} \tag{7}$$

$$\mathrm{Rec} = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = \frac{2 \times \mathrm{Pre} \times \mathrm{Rec}}{\mathrm{Pre} + \mathrm{Rec}} \tag{9}$$

In the equation: TP is the number of model predictions correctly labeled as positive, FP is the number of model predictions incorrectly labeled as positive, TN is the number of model predictions correctly labeled as negative, and FN is the number of model predictions incorrectly labeled as negative.

3.3. Flame Burning Images Dataset

The data set used in this experiment is the flame burning image data set created by Pan et al in literature [21], which is from a MSWI factory in Beijing. Inside the incinerator, the left and right sides are equipped with high temperature resistant cameras for capturing flame video. After collecting the flame video from the left and right cameras in the field, the first step was to remove the segments that did not describe the burning state clearly. Next, the remaining video clips are classified according to the burning state classification criteria shown in Figure 1. These classified video clips were subsequently sampled using a MATLAB program at a consistent rate of 1 frame per minute, resulting in the extraction of flame image frames. Finally, the total number of typical burning state images obtained from the left and right stoves is 3289 and 2685, respectively. Due to the symmetrical distribution of the left and right grate images, in order to improve the generalization ability of the model in this paper, the left and right grate images are merged into a data set for experiments, and the ratio of training set, validation set and test set is 0.7:0.15:0.15. The amount of data corresponding to each typical burning state is shown in Table 1.

Table 1. The distribution of the flame burning image dataset.

Grate	Amount	Normal	Partial	Channeling	Smoldering	Size
Left	3289	655	1176	1044	414	720×576
Right	2685	564	1002	534	585	720×576

 doi:10.20944/preprints202503.1326.v1

3.4. Model Experimental Results

Figure 5 represents the loss and accuracy during model training. The left figure shows the loss plotted against epochs for the training and validation sets. It can be observed that both the training loss and the validation loss gradually decrease with the increase of epochs, which indicates that the performance of the model on both the training and validation sets is constantly improving. In particular, the loss decreases relatively fast in the first 20 epochs, after which the loss gradually levels off and the validation loss fluctuates up and down around the training loss.

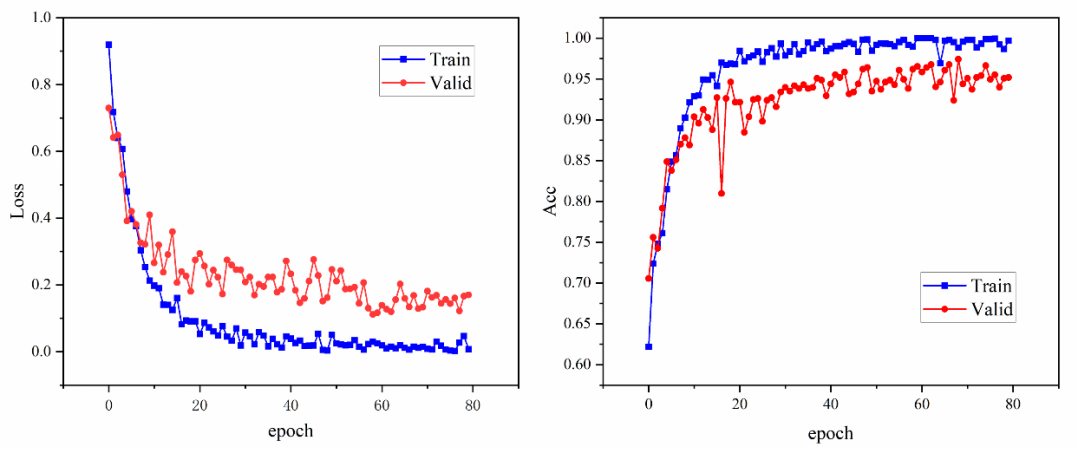

Figure 5. The training process of M³RTNet.

The right figure shows the training and validation set accuracy plotted against epochs. It can be seen from the figure that both the training accuracy and the validation accuracy gradually increase with the increase of epochs and become stable after about 20 epochs. The training accuracy is always higher than the validation accuracy, which reflects that the model performs better on the training set than on the validation set, but the gap between the two is not large, indicating that the model does not significantly overfit.

The two figures show that the loss of the model gradually decreases and the accuracy gradually improves during the training process, and it tends to be stable after a certain number of epochs, which achieves a good training effect, indicating that the model has effectively learned the features of the flame burning image.

3.4. Results of Ablation Experiments

In order to evaluate the effectiveness of the modules, each module is tested through different network models, as shown in Table 2, and eight experiments are conducted in turn based on the residual network in this experiment.

Figure 6 shows the ablation experimental results of different network configurations in the flame burning state recognition task of MSWI process, and the effectiveness of the proposed method is verified by comparing the training and validation accuracy. As can be seen from the left figure, the proposed method shows the characteristics of rapid convergence at the early stage of training, and the training accuracy is significantly higher than other configurations. It is also seen from the right figure that the accuracy of the proposed model in the validation set is always ahead and significantly higher than that of other configurations, which proves the key role of the three feature enhancement strategies proposed in this paper.

 doi:10.20944/preprints202503.1326.v1

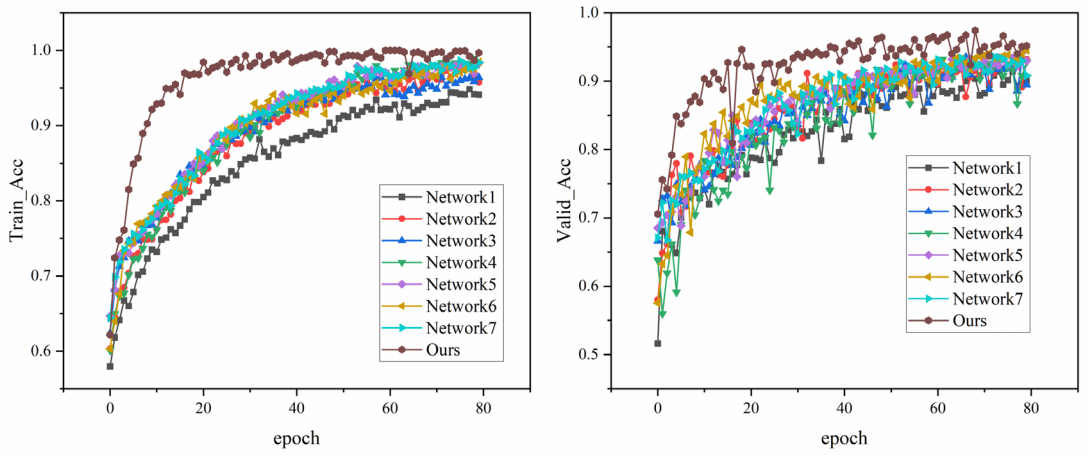

Figure 6. Training and validation accuracy during ablation experiments.

Table 2 provides detailed experimental results for each model. Compared with Network1, the performance parameters of Network2 are improved after adding EMA module, and the accuracy, precision, recall and F1 score of Network2 are increased by 0.93%,1.2%,0.55% and 0.88% respectively, which proves the effectiveness of EMA module. After adding DMAT module, the accuracy, precision, recall and F1 score of Network3 are increased by 0.4%,1.14%,0.2 % and 0.67%, respectively, which proves that DMAT module can make the network have better extracted features. After adding the CFFM module to Network4, the accuracy rate is increased by 0.27%, the precision rate is increased by 0.57%, the recall rate is increased by 0.05%, and the F1 score is increased by 0.31%. It is verified that the CFFM module can enhance the feature fusion of different stages and enhance the feature extraction ability of the model.

Table 2. The results of ablation experiments.

Name	Details	EMA	DMAT	CFFM	Acc	Pre	Rec	F1
Network1	Resnet50				0.9192	0.9145	0.9250	0.9197
Network2	Resnet50+EMA	√			0.9285	0.9265	0.9305	0.9285
Network3	Resnet50+DMAT		√		0.9232	0.9259	0.9270	0.9264
Network4	Resnet50+CFFM			√	0.9219	0.9202	0.9255	0.9228
Network5	Resnet50+EMA+DMAT	√	√		0.9483	0.9508	0.9453	0.9480
Network6	Resnet50+EMA+CFFM	√		√	0.9523	0.9518	0.9513	0.9515
Network7	Resnet50+DMAT+CFFM		√	√	0.9364	0.9409	0.9363	0.9386
Network8	M³RTNet	√	√	√	0.9616	0.9615	0.9607	0.9611

The evaluation indexes of Network5, 6 and 7 with two modules are higher than those of Network2,3 and 4 with only one module. The model with three modules has the best performance, and compared with the initial Networkl model, the accuracy of pneumonia classification is increased from 91.92% to 96.16%, the precision is increased from 91.45% to 96.15%, the recall is increased from 92.5% to 96.07%, and the F1 score is increased from 91.97% to 96.11%. It can be concluded that the proposed model has the best performance and the best performance in MSWI combustion state recognition.

In addition, in order to investigate the difference between the labels predicted by different models for the classification of four types of samples and the real situation, this paper uses a confusion matrix to visualize the test results of ablation experiments, as shown in Figure 7. Through the comparison of confusion matrix, it can be seen that the proposed model has better classification effect and can realize the accurate identification of MSWI burning state.

 doi:10.20944/preprints202503.1326.v1

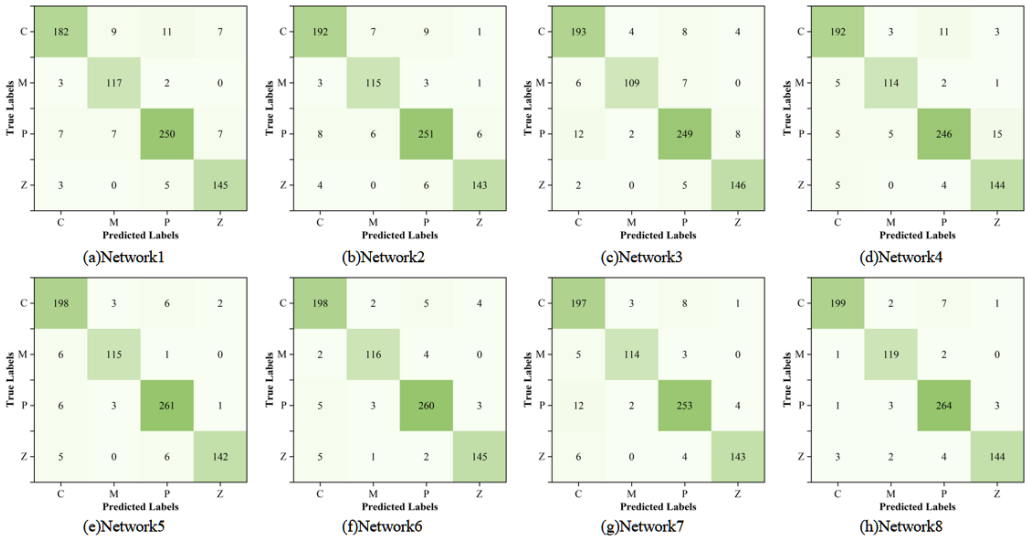

Figure 7. The confusion matrix for ablation experiments.

3.3. Results of Comparative Experiment

In order to verify the recognition ability of the proposed model for MSWI burning state, it is compared with the classical deep learning method in Table 3 on the same data set. The experimental results are as follows:

Figure 8 shows the performance of different models in the training process, the left figure shows the accuracy of each model on the training set, and the right figure shows the accuracy on the validation set. In conclusion, the proposed model not only performs well on the training set, but also has strong generalization ability on the validation set. The specific experimental results are shown in Table 3.

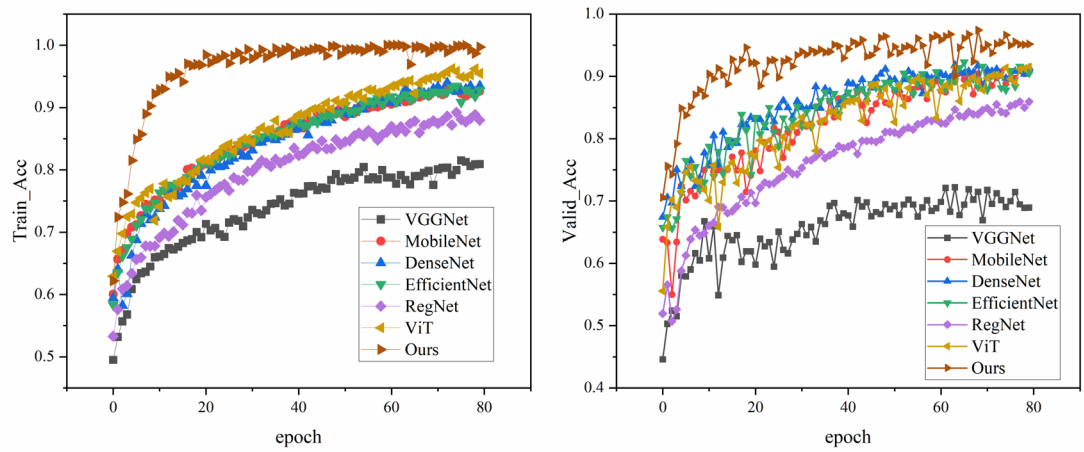

Figure 8. Training and validation accuracy during comparative experiment.

Table 3. The results of comparative experiments.

Name	Acc	Pre	Rec	F1
VGGNet [30]	0.9192	0.9145	0.9250	0.9197
MobileNet [31]	0.9285	0.9265	0.9305	0.9285
DenseNet [32]	0.9232	0.9259	0.9270	0.9264
EfficientNet [33]	0.9219	0.9202	0.9255	0.9228
RegNet [34]	0.9483	0.9508	0.9453	0.9480
ViT [35]	0.9523	0.9518	0.9513	0.9515

 doi:10.20944/preprints202503.1326.v1

M³RTNet	0.9364	0.9409	0.9363	0.9386

The experimental results show that the accuracy rate, precision rate, recall rate and F1 score of the proposed model are 96.16%, 96.15%, 96.07% and 96.11%, which are better than other networks and have better classification performance. In this study, the confusion matrix is used to visualize the results of the test set of each model, and the results are shown in Figure 8. From the comparison of the confusion matrix, it can be seen that the recognition ability of the model proposed in this paper for MSWI states is more balanced and more effective than other classification networks.

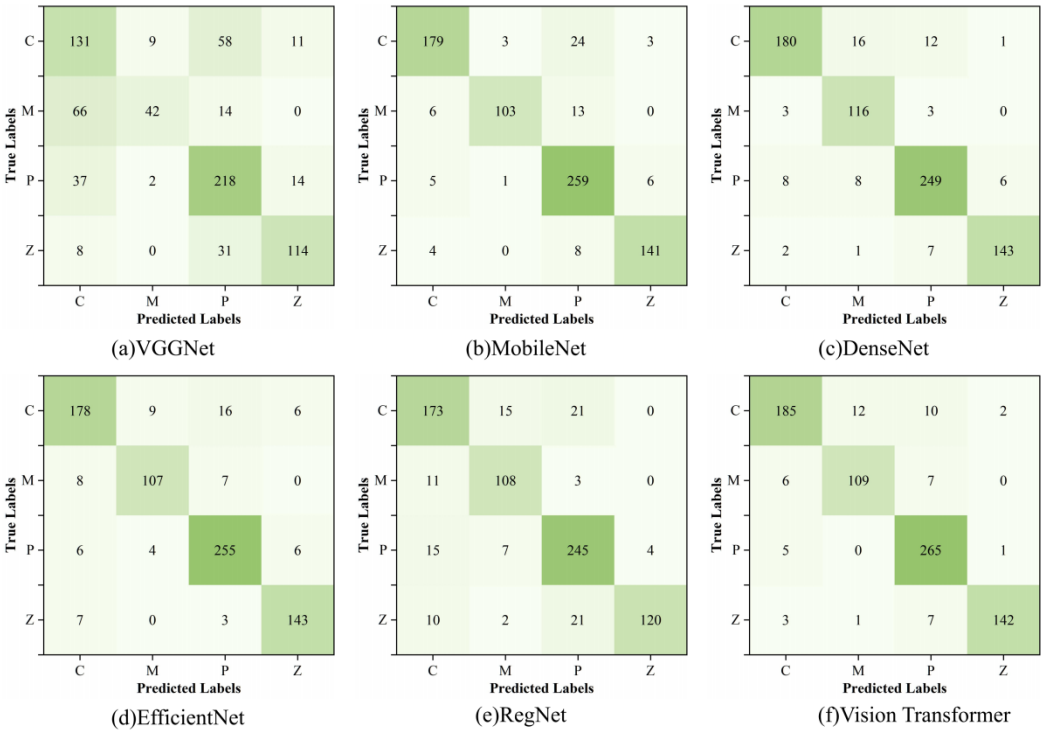

Figure 8. The confusion matrix for comparative experiments.

4. Conclusions

In this study, we propose an MSWI flame burning state recognition model based on three feature enhancement strategies, which integrates the EMA module, greatly enhances the model's ability in multi-scale feature extraction and efficient attention allocation, and effectively alleviates the limitations of traditional attention mechanisms in computational complexity and feature fusion. The DMAT module is used at the end of the network. The advantages of CNN and Transformer were combined to make the network fully extract the global features and local features of MSWI images, and the high-level semantic information was globally modeled to obtain the global features of high-level semantic information. The CFFM module is designed to fuse the spatial information such as texture and edge of the shallow network and the channel information of the deep network to further enhance the feature extraction ability of the network. The experimental results on the MSWI dataset show that the accuracy, precision, recall and F1 score of the proposed model are 96.16%, 96.15%, 96.07% and 96.11%, respectively, which proves that the model proposed in this study can accurately identify the flame burning state in the MSWI process. Assist the field staff to adjust the control strategy in time.

 doi:10.20944/preprints202503.1326.v1

References

1. Li, M.; Li, S.; Chen, S.; Meng, Q.; Wang, Y.; Yang, W.; Shi, L.; Ding, F.; Zhu, J.; Ma, R.; et al. Measures for Controlling Gaseous Emissions during Composting: A Review. *International Journal of Environmental Research and Public Health* **2023**, *20*, 3587.
2. Chen, D.M.C.; Bodirsky, B.L.; Krueger, T.; Mishra, A.; Popp, A. The world's growing municipal solid waste: trends and impacts. *Environ. Res. Lett.* **2020**, *15*, 12. https://doi.org/10.1088/1748-9326/ab8659.
3. Funari, V.; Dalconi, M.C.; Farnaud, S.; Nawab, J.; Gupta, N.; Yadav, K.K.; Kremser, K.; Toller, S. Modern management options for solid waste and by-products: sustainable treatment and environmental benefits. *Front. Environ. Sci.* **2024**, *12*, 3. https://doi.org/10.3389/fenvs.2024.1385669.
4. Khan, M.S.; Mubeen, I.; Yu, C.M.; Zhu, G.J.; Khalid, A.; Yan, M. Waste to energy incineration technology: Recent development under climate change scenarios. *Waste Manage. Res.* **2022**, *40*, 1708-1729. https://doi.org/10.1177/0734242x221105411.
5. Kasinski, S.; Debowski, M. Municipal Solid Waste as a Renewable Energy Source: Advances in Thermochemical Conversion Technologies and Environmental Impacts. *Energies* **2024**, *17*, 33. https://doi.org/10.3390/en17184704.
6. Li, K.; Deng, J.; Zhu, Y.; Zhang, W.Y.; Zhang, T.; Tian, C.; Ma, J.W.; Shao, Y.Y.; Yang, Y.F.; Shao, Y.Q. Utilization of municipal solid waste incineration fly ash with different pretreatments with gold tailings and coal fly ash for environmentally friendly geopolymers. *Waste Manage.* **2025**, *194*, 342-352. https://doi.org/10.1016/j.wasman.2025.01.014.
7. Tang, J.; Tian, H.; Xia, H. Interval Type-Ⅱ FNN-based Furnace Temperature Control for MSWI Process. *Journal of Beijing University of Technology* **2025**, *51*, 1-16.
8. Yan, X.; Song, G.W.; Liu, J.Y.; Liu, X.; Wang, H.L.; Hao, Z.P. A comprehensive emission inventory of air pollutants from municipal solid waste incineration in China's megacity, Beijing based on the field measurements. *Sci. Total Environ.* **2024**, *948*, 9. https://doi.org/10.1016/j.scitotenv.2024.174806.
9. Munir, M.T.; Li, B.; Naqvi, M. Revolutionizing municipal solid waste management (MSWM) with machine learning as a clean resource: Opportunities, challenges and solutions. *Fuel* **2023**, *348*, 128548. https://doi.org/10.1016/j.fuel.2023.128548.
10. Gao, C.Q.; Bian, R.X.; Li, P.; Yin, C.Y.; Teng, X.; Zhang, J.R.; Gao, S.D.; Niu, Y.T.; Sun, Y.J.; Wang, Y.A.; et al. Analysis of carbon reduction potential from typical municipal solid waste incineration plants under MSW classification. *J. Environ. Manage.* **2025**, *373*, 9. https://doi.org/10.1016/j.jenvman.2024.123844.
11. Wang, T.Z.; Tang, J.; Aljerf, L.; Qiao, J.F.; Alajlani, M. Emission reduction optimization of multiple flue gas pollutants in Municipal solid waste incineration power plant. *Fuel* **2025**, *381*, 21. https://doi.org/10.1016/j.fuel.2024.133382.
12. Zhou, C.; Cao, Y.; Yang, S. Video Based Combustion State Identification for Municipal Solid Waste Incineration**The work is supported by the National Key Research and Development Plan (2018YFC0214102) of P. R. China. *IFAC-PapersOnLine* **2020**, *53*, 13448-13453. https://doi.org/10.1016/j.ifacol.2020.12.255.
13. Tang, J.; Wang, T.Z.; Xia, H.; Cui, C.L. An Overview of Artificial Intelligence Application for Optimal Control of Municipal Solid Waste Incineration Process. *Sustainability* **2024**, *16*, 41. https://doi.org/10.3390/su16052042.
14. Duan, H.; Tang, J.; Qiao, J. Recognition of Combustion Condition in MSWI Process Based on Multi-scale Color Moment Features and Random Forest. 2019; pp. 2542-2547.
15. Guo, H.; Tang, J.; Zhang, H.; Wang, D. A method for generating images of abnormal combustion state in MSWI process based on DCGAN. 2021; pp. 1-6.

 doi:10.20944/preprints202503.1326.v1

16. Ding, H.; Tang, J.; Qiao, J. Control Methods of Municipal Solid Wastes Incineration Process: A Survey. 2021; pp. 662-667.
17. Zhang, H.; Meng, X.; Tang, J.; Wang, Z.; Duan, H.; Qiao, J. Recognition of Combustion Conditions in MSWI Process Using Convolutional Neural Network. 2021; pp. 6364-6369.
18. Guo, H.; Tang, J.; Heng, X.; Qiao, J. Construction of Combustion Line Quantification Data Set for Municipal Solid Waste Incineration Process. 2022; pp. 1-6.
19. Tian, H.; Tang, J.; Pan, X.; Xia, H.; Wang, T.; Wang, Z. Combustion State Identification of Municipal Solid Waste Incineration Process Based on VGG19 Depth Feature Migration. 2023; pp. 337-342.
20. Pan, X.; Tang, J.; Xia, H. Flame Combustion State Identification Based on CNN in Municipal Solid Waste Incineration Process. 2023; pp. 1-4.
21. Pan, X.; Tang, J.; Xia, H.; Tian, H.; Wang, T.; Xu, W. Construction of flame image classification criteria and reference database for municipal solid waste incineration process. 2023; pp. 343-348.
22. Pan, X.T.; Tang, J.; Xia, H.; Yu, W.; Qiao, J.F. Combustion state identification of MSWI processes using ViT-IDFC. *Eng. Appl. Artif. Intell.* **2023**, *126*, 16. https://doi.org/10.1016/j.engappai.2023.106893.
23. Yang, W.; Tang, J.; Xia, H.; Pang, X.; Cui, C.; Wang, T. Combustion Status Recognition of MSWI process Based on Flame Image by Using YOLOv5. 2024; pp. 2363-2368.
24. Guo, H.T.; Tang, J.; Ding, H.X.; Qiao, J.F. Combustion States Recognition Method of MSWI Process Based on Mixed Data Enhancement. *ACTA AUTOMATICA SINICA* **2024**, *50*, 560-575. https://doi.org/10.16383/j.aas.c210843.
25. Hu, Y.; Tang, J.; Pan, X.; Yang, W.; Cui, C.; Wu, Z. Multi-physical feature extraction and selection method for global representation information of flame images in the MSWI process. 2024; pp. 2374-2379.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27-30 June 2016, 2016; pp. 770-778.
27. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. 2023; pp. 1-5.
28. Wang, K.; Liew, J.H.; Zou, Y.; Zhou, D.; Feng, J. PANet: Few-Shot Image Semantic Segmentation With Prototype Alignment. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 27 Oct.-2 Nov. 2019, 2019; pp. 9196-9205.
29. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), 22-29 Oct. 2017, 2017; pp. 764-773.
30. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
31. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *ArXiv* **2017**, *abs/1704.04861*.
32. Huang, G.; Liu, Z.; Weinberger, K.Q. Densely Connected Convolutional Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* **2016**, 2261-2269.
33. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International conference on machine learning, 2019; pp. 6105-6114.
34. Radosavovic, I.; Kosaraju, R.P.; Girshick, R.B.; He, K.; Dollár, P. Designing Network Design Spaces. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* **2020**, 10425-10433.

 doi:10.20944/preprints202503.1326.v1

35. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv* **2020**, *abs/2010.11929*.