Expert Systems With Applications 233 (2023) 120921

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Air pollution measurement based on hybrid convolutional neural network with spatial-and-channel attention mechanism

Zhenyu Wang *, Fucheng Wu, Yingdong Yang

The School of Control and Computer Engineering, North China Electric Power University, Beijing, 102206, China

ARTICLE INFO

Keywords:
Attention
Air quality detection
CNN
Transformer
Fine-grained image recognition

ABSTRACT

Air quality is tightly correlated with human health, and long-term exposure to air pollution can pose a serious health risk to humans. In recent years, image-based air quality detection methods have been proposed and have achieved good accuracy in specific scenarios. However, most of the methods are still based on pure CNNs with fast inference speed but limited accuracy. Some also invoke a single channel or spatial attention mechanism, with improved accuracy but much slower inference speed. To have both advantages we propose the Spatial and Channel Calibration Network (SCCNet). The network combines spatial and channel attention to improve the detection efficiency and accuracy of the model by better extracting global information to focus computational resources on regions that are more important to the task. Our proposed channel averaging pooling (CAP) module significantly reduces the number of parameters in the model while extracting global information, improving the detection speed of the model. We also introduce a discrete cosine transform (DCT) method to transform images from the spatial domain to the frequency domain, which enhances the extraction of fine-grained features and improves the model's classification ability for air quality detection tasks. Our experimental results show that SCCNet achieves an accuracy of 92.17% with about 30 million parameters in an air quality detection task, which is 1.65% and 1.71% more accurate than Swin Transformer (based on spatial attention) and SENet (based on channel attention) for a similar number of parameters. Our code and models will are publicly available at https://github.com/Fucheng-Wu/SCCNet.

1. Introduction

Air quality detection, an important and challenging task in fine-grained image recognition, has received a lot of attention in recent years and has been addressed in fields such as conventional image processing and computer vision. Theoretically, light is diffusely reflected in the air. The poorer the air quality, the denser the fine particles in the air, resulting in a stronger diffuse reflection of light. So the environment images that we capture through the devices in environments with different air quality will differ. Therefore, it is feasible to detect the air quality at the time of filming by processing the ambient images.

Existing image-based air quality detection methods can be divided into two categories: traditional image processing algorithms and deep learning algorithms. Traditional image processing algorithms rely on manual feature extraction methods to process images. In contrast to traditional image processing algorithms, deep neural networks will use designed loss functions and optimization methods for parameter updating, and self-learning through backpropagation. The use of deep learning algorithms to solve air quality detection problems is also a major area of research for us.

As the task of air quality detection has been explored, CNNs (LeCun, Bottou, Bengio, & Haffner, 1998), a classical type of network for deep learning, have also been applied to the task of detecting air quality. CNNs have the advantages of fast convergence, high operational efficiency, and fast inference. But it also makes ordinary CNN networks quickly reach an accuracy bottleneck on air quality tasks, making it difficult to improve further. Inspired by the fact that humans can find important regions in complex scenes naturally and efficiently, attentional mechanisms have been introduced into computer vision.

Since the attention mechanism can extract global information, it can focus on information that is more important to the task with high weights and ignore irrelevant information with low weights, and it can also continuously adjust the weights so that high-value information can also be selected in different situations. Although the introduction of the attention mechanism leads to an improvement in the accuracy of air quality detection, the computational overhead it entails cannot be ignored.

It is worth noting that air quality detection is a special fine-grained image recognition task. In other words, it relies on the extraction of

* Corresponding author.
E-mail addresses: zywang@ncepu.edu.cn (Z. Wang), fcwu@ncepu.edu.cn (F. Wu), ydyang@ncepu.edu.cn (Y. Yang).

https://doi.org/10.1016/j.eswa.2023.120921
Received 31 December 2022; Received in revised form 18 March 2023; Accepted 27 June 2023
Available online 6 July 2023

global information to determine air quality and also requires attention to subtle differences in different air quality images. This is an area that many people have overlooked in their research on air quality.

In this context, how to make the network at the same time have high detection accuracy, fast detection, and a special enhancement of the air quality detection task is a major difficulty and a problem that we want to solve. Based on the above motivation, we propose a hybrid convolutional neural network, SCCNet. To improve the classification accuracy of the network, we propose a combination of spatial and channel attention to focus computational resources on regions that are more important for air quality detection. To solve the problem of high number of parameters and computational effort caused by the introduction of spatial and channel attention, and to improve the detection speed of the network, we propose the CAP method. To perform the air quality detection task better, we introduce a DCT method to convert the feature map from the spatial domain to the frequency domain to improve the network's recognition of fine-grained features. We have also made many improvements to the structure of the model, which will be described in detail in Section 3. Our experimental results show that our proposed network achieves state-of-the-art performance for the same number of parameters and computational cost in the air quality detection task.

We summarize the contributions of this paper as follows :

- We construct an SCC block capable of extracting spatial and channel global information. The spatial calibration branch consists of our proposed channel averaging pooling (CAP) module and the spatial-attention module; the channel calibration branch is obtained by improving the channel-attention module. The SCC module obtains better air quality features and does not increase the deduction time.
- We construct a hybrid convolutional neural network combining spatial and channel attention, called SCCNet. In order to simplify the network structure and fully utilize the SCC blocks, the SCCNet backbone consists of multiple SCC blocks and corresponding downsampling modules. SCCNet significantly improves the classification performance of air quality detection tasks.
- We introduce the DCT module to enhance the network's ability to extract information in the frequency domain. The DCT module can help the air quality detection model to perform special fine-grained image classification and improve the network classification performance.
- We established a new high-quality environmental image dataset Get-AQI in One shot-4 (GAOs-4). Through careful screening and inspection, we not only solved the long-tail distribution problem that appeared in the previous dataset but also tripled the scale of the dataset.

The rest of the paper is organized as follows. Section 2 describes the related work, Section 3 details the proposed method, Section 4 describes the experimental results, Section 5 discusses the role of some structures, and Section 6 concludes the paper.

2. Related works

Due to the rapid development of deep learning in recent years, many efficient algorithms have emerged among them. The underlying networks range from CNNs in the beginning to Transformer-based attention networks now. The attention mechanism has also changed from the channel attention mechanism at the beginning to the spatial attention mechanism now. With the development of deep learning, the method of detecting air quality by image-based deep learning algorithms has received more and more attention.

2.1. Channel attention

The channel attention mechanism was first proposed by Hu, Shen, and Sun (2018). via Squeeze-and-Excitation network (SENet). The core idea of SENet is a SE module. The Squeeze module captures the global spatial information and the Excitation module captures the relationship between channels to recalibrate the channel weights. Gao, Xie, Wang, and Li (2019) propose the use of global second-order pooling blocks (GSoP) to improve the Squeeze module based on SENet. GSoP obtains the correlation between channels by calculating the covariance matrix of different channels, collecting global information while also modeling higher-order feature data. Lee, Kim, and Nam (2019) proposed a lightweight style-based recalibration module (SRM), which uses the mean and variance of the input feature maps to improve the ability of the compression module to capture global information. SRM uses CFC (Channel-wise fully-connected layer) at the fully-connected point to reduce the computational effort, which improves the accuracy and reduces the computational effort at the same time. Wang, et al. (2019) proposed the efficient channel attention block (ECA), which improves the quality of the results by directly modeling the correspondence between weight vectors and inputs by one-dimensional convolution to obtain the relationship between channels. Yang, Zhu, Wu, and Yang (2020) proposed gated channel transform (GCT), which not only reduces the computational effort in the calibration module but also establishes an explicit channel relationship. Qin, Zhang, Wu, and Li (2020) improved the compression module to obtain a more powerful representation. They proved that global average pooling is only a special case of the DCT and further proposed multispectral channel attention to improve the network performance. Due to the effectiveness of the SE attention module, the lightweight network MobilenetV3 was proposed by Howard et al. (2019) and EfficientNetV2 was proposed by Tan and Le (2021). They both added SE modules to the network, which significantly improved the performance of the network.

2.2. Spatial attention

Since deep neural networks have a huge computational cost, Mnih, Heess, Graves, and Kavukcuoglu (2014) proposed the recurrent attention model (RAM) to focus all the limited computational resources on the important regions. RAM uses a recurrent neural network (Zaremba, Sutskever, & Vinyals, 2014) and reinforcement learning to make the network learn where to pay attention, reducing the number of computations in the network and improving the classification results. Jaderberg, Simonyan, Zisserman, et al. (2015) proposed the spatial transformation network (STN), which enables CNNs to focus on important regions while implementing features such as translation, rotation, and scaling. STN is the first attention mechanism that explicitly predicts important regions for deep neural networks. In 2017, Vaswani et al. (2017) proposed self-attention, which has been a great success in the field of natural language processing and shows great potential for application to computer vision. Wang, Girshick, Gupta, and He (2017) were the first to introduce self-attention into computer vision. The self-attention mechanism increases the perceptual field and enhances the network's ability to understand global information. Dosovitskiy et al. (2020) proposed the Vision Transformer, which is the first pure Transformer structure for image processing and can obtain results comparable to convolutional networks. Liu et al. (2021) borrowed many design concepts and prior knowledge of CNNs to propose Swin Transformer, which introduced a window-shifting operation of convolution to improve the network performance. Mehta and Rastegari (2021) proposed MobileViT, which uses a hybrid architecture of CNN and Transformer, fusing the advantages of convolution and Transformer as a lightweight, general-purpose, low-latency network model.

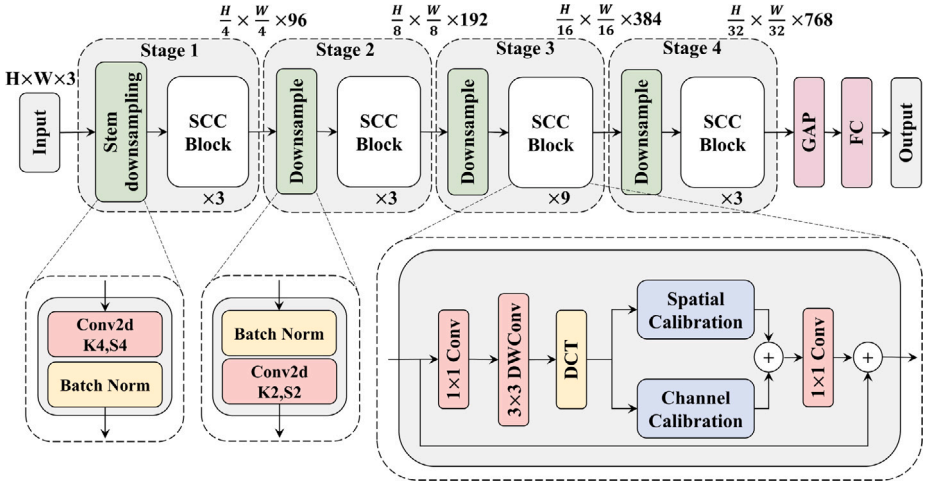


Fig. 1. The network structure of SCCNet. The process of extracting features is done in four stages, each of which contains a downsampling module and an SCC block. The width and height of the output feature map and the number of channels for each stage are marked on the top of the figure. GAP denotes the global average pooling layer and FC denotes the fully connected layer. For simplicity, the normalization layer and activation layer are omitted in the figure.

2.3. Deep learning algorithm

2.3.1. Convolutional neural network based methods

With the great popularity of CNNs, more people are using CNNs to extract image features and perform classification, which eventually achieves good results. Zhang et al. (2016) proposed an effective CNN-based neural network model for air pollution level estimation, which was designed with a negative logarithmic ordinal classifier in the last layer of the network and was able to improve the ordinal recognition capability of the model. Pan, Yu, Miao, and Leung (2017) used a deep learning approach for the accurate estimation of air quality using transmittance maps of environmental images as input. Chakma, Vizena, Cao, Lin, and Zhang (2017) used a CNN-based approach to estimate PM2.5 concentrations in environmental images, and it used two transfer learning methods to improve the performance of the network. Ma, Li, Han, and Yang (2018) proposed an algorithm for estimating air pollution using a hybrid deep neural network. The dark channel images of the environmental images were obtained by the dark channel prior proposed by He, Sun, and Tang (2010). The hybrid convolutional neural network was trained using the original and dark channel images as inputs, and good results were achieved on both synthetic and real datasets. Rijal et al. (2018) integrated multiple deep-learning networks to estimate PM2.5 concentrations from environmental images. The results predicted by three CNN models VGG-16, Inception-v3, and Resnet50 were combined in a nonlinear manner to obtain the final predicted PM2.5 concentrations. Wang, et al. (2019) proposed a two-channel weighted convolutional network integrated learning algorithm to measure air quality, which uses a two-channel convolutional neural network to extract features and perform weighted connections on different parts of the environmental image. The algorithm proved to be effective through experiments. Zhang, Fu, and Tian (2020) proposed a deep learning-based air quality evaluation model, AQC-Net. The model introduces a self-supervised module to capture the interdependence of air quality information in environmental images, enhancing local information important to the task, and improve the representation capability of the model. Wang, Yue, and Song (2021) proposed a video-based two-channel 3D convolutional network in which the semantic channels guide the network to learn region-level features and features from both channels are combined for predicting air quality.

2.3.2. Attention network based methods

With the Vision Transformer proposed by Vaswani et al. refreshing various lists, the attention network has attracted more attention. Wang, Yang, and Yue (2022) proposed a dual-output Vision Transformer (DOViT) algorithm to predict local air quality levels and AQI levels, which uses a multiple self-attention (MSA) mechanism to process environmental images and achieve higher classification accuracy. Wang, et al. (2022) proposed a hybrid AQI prediction model based on a CNN and attention gate unit (AGU), which not only solved the vanishing gradient and extended gradient problems of RNN but also improved the prediction performance of AQI.

With the development of deep learning, image-based air quality detection algorithms are performing better and better, but there are still some problems. Although pure convolutional networks are easy to optimize and have fast inference, they are deficient in classification accuracy. Some people have successfully improved the classification performance of the network by introducing spatial attention or channel attention mechanism. But it also leads to the problem of too many parameters that are difficult to optimize and slow down the inference speed, which is difficult to implement in practical applications. Therefore, we intend to improve the detection efficiency and accuracy of the network by combining spatial and channel attention to extract global information and focus computational resources on regions that are more important for air quality detection. To improve the detection speed of the model through a subtle network structure design that significantly reduces the number of high references and computational effort caused by the attention mechanism. To enhance the classification performance of the model for air quality detection tasks by transforming the environmental images from the spatial domain to the frequency domain to display the fine-grained features of the images.

3. Method

Inspired by the spatial attention and channel attention approaches, we construct a module called the SCC block that fuses spatial and channel attention. The inverted residual structure and DCT (discrete cosine transform) module are used to improve the representation of the feature maps, and the Spatial Calibration and Channel Calibration branches are used to correct the input feature maps to obtain more useful features for air quality classification. Our network outperforms previous methods in terms of performance, and the inference speed remains largely consistent.

3.1. The overall structure of SCCNet

The total structure of SCCNet is shown in Fig. 1, we designed a simple and effective architecture to explore the effectiveness of the SCC block. We used the common convolutional network layered backbone.

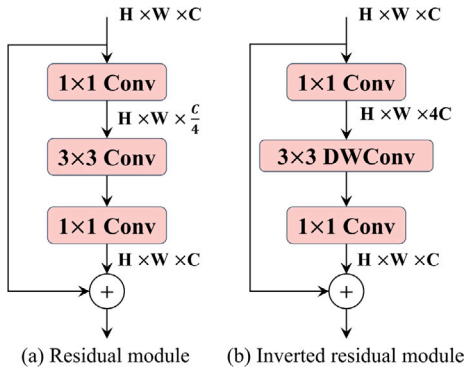


Fig. 2. Comparison of the residual module and inverted residual module, where DWConv represents Depthwise convolution. The width, height, and number of channels of the output feature map of each module are marked.

The main body of the network consists of four stages (S1-S4), each of which comprises a downsampling module and an SCC block. The use of a separate downsampling layer is for the SCC block to be able to focus more on feature extraction. In particular, to speed up the network feature extraction, a convolutional layer with a convolutional kernel size of 4 and a step size of 4 is used in the first downsampling layer of the network to implement. The resolution of each stage is half that of the previous stage, and the number of channels is doubled. In our network, the same SCC block is used for the entire backbone. The final classification head is implemented by a global average pooling layer and a fully connected layer.

3.2. Spatial-and-channel calibration block

The SCC block shown in Fig. 1 aims to model the input feature map's global space and channel information with fewer parameters to extract features that are more conducive to air quality classification.

Formally, for the input feature map $X_{in} \in \mathbb{R}^{H\times W\times C}$, the SCC block is processed by applying the inverted residual structure. The inverted residual structure is proposed in Mobilenetv2 (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018) and is similar to the residual block proposed by ResNet (He, Zhang, Ren, & Sun, 2016). Residual blocks are first downscaled to extract features and then upscaled to output feature maps reduces the number of parameters and speeds up inference, but compresses the information in the channels, thus reducing the quality of the results. The inverted residual structure first projects the tensor to a higher-dimensional space through a 1 × 1 convolution layer to obtain more object representations, then a 3 × 3 Depth-wise convolution to process the feature maps of each channel to produce $X_L \in \mathbb{R}^{H\times W\times 4C}$, and finally a 1 × 1 convolution layer to reduce the number of channels to the original number of channels, as shown in Fig. 2(b). The comparison between the inverted residual structure and the residual structure is shown in Fig. 2. The inverted residual structure can extract features better without increasing the computational effort.

To improve the representation of air quality in environmental images, we introduce the DCT to convert the feature map from the spatial domain to the frequency domain for processing.

To achieve our goal of combining spatial and channel attention to improve network performance and obtain the most favorable feature maps for air quality classification, we calibrate the pixel and channel information of the feature maps by using the Spatial Calibration and Channel Calibration modules, respectively. Finally, the output feature map $X_{out} \in \mathbb{R}^{H\times W\times C}$ can be obtained by summing the shortcut with the input, as shown in Fig. 1. The SCC block can guide the network to respond differently to different regions and channels of the environment image, shifting the attention to the most important parts of an image and ignoring the less important parts, improving the network classification performance.

3.2.1. Discrete cosine transform

Because the air quality classification task is a fine-grained classification with small gaps between each category, when air quality changes, the changes are often not visible enough in the environmental image. Therefore, we introduce the DCT to convert the feature map from the spatial domain to the frequency domain to enhance the representation of air quality variations. Typically, the two-dimensional DCT is formulated as follows:

$$f^{2d}_{h,w} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x^{2d}_{i,j}\cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right), \quad s.t. \quad h\epsilon\{0,1,\ldots,H-1\},\ w\epsilon\{0,1,\ldots,W-1\}. \tag{1}$$

where the spectrum of the two-dimensional DCT is $f^{2d} \in \mathbb{R}^{H\times W}$, the input feature map is $x^{2d} \in \mathbb{R}^{H\times W}$, H and W are the height and width of the input feature map, respectively. Corresponding, the inverse two-dimensional DCT can be written as:

$$x^{2d}_{i,j} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} f^{2d}_{h,w}\cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right), \quad s.t. \quad i\epsilon\{0,1,\ldots,H-1\},\ j\epsilon\{0,1,\ldots,W-1\}. \tag{2}$$

We call the common term of the two as the basis function:

$$B^{i,j}_{h,w} = \cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right) \tag{3}$$

The environment image after DCT is shown in Fig. 3. The highlights preserve the low-frequency information of the environment image, reflecting the information of the flat region in the spatial domain image. The other regions preserve the high-frequency information, reflecting the information of the abrupt change region (edge contour detail information) in the spatial domain image. From Fig. 3, we can see that two environmental images with different air quality taken at the same angle have distinct differences in the frequency domain images after DCT. The frequency domain image after DCT contains more information about the edge contour details because the ambient image with better air quality is sharper. Environmental images with poor air quality have low resolution, and the frequency domain images after DCT contain limited information on edge contour details. Subsequent experiments demonstrate that the frequency domain images after DCT are more sensitive to changes in air quality, making it easier for the network to extract useful features.

3.2.2. Spatial calibration branch

We design the spatial calibration branch to use global spatial information to correct the weights of each pixel in the feature map and to guide the network to focus on important regions of the environmental image. The spatial calibration branch consists of a total of three modules and is structured as shown in Fig. 4. The squeeze channel information module is used to compress the channel information of the feature map $X_{in} \in \mathbb{R}^{H\times W\times C}$ by our proposed CAP (channel averaging pooling) layer to obtain the tensor $X_c \in \mathbb{R}^{H\times W\times 1}$. The mathematical expression for channel averaging pooling is as follows:

$$y_{i,j} = \frac{1}{N}\sum_{n=0}^{N-1} x^n_{i,j}, \quad s.t. \quad i\epsilon\{0,1,\ldots,H-1\},\ j\epsilon\{0,1,\ldots,W-1\}. \tag{4}$$

where $x^n_{i,j}$ is denoted as the pixel value on channel n of the feature map, N is the total number of channels. The CAP process is shown in Fig. 5(b), which not only extracts information from all channels simply and efficiently but also drastically reduces the amount of subsequent computation. In the global spatial representation module we need to model the global information association. In the CNN class of models, calculating the association between two pixels that are far apart can only be achieved by dilated convolution or expanding the convolution kernel. Dilated convolution loses much information, making the extracted information incoherent. Expanding the convolution kernel increases the computational effort of the model significantly and reduces

(a) Good air quality (b) Bad air quality

Fig. 3. Original environment image and DCT image.

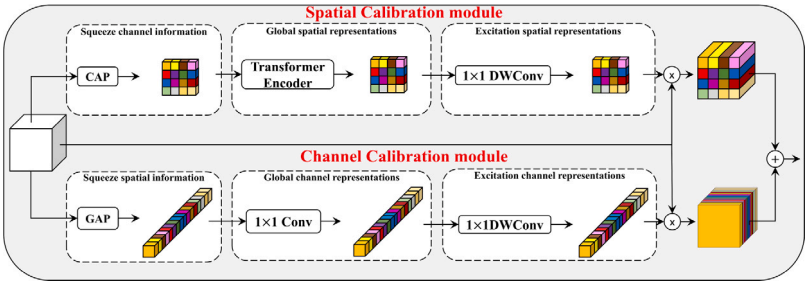


Fig. 4. Structure diagram of the spatial calibration module (upper half) and the channel calibration module (lower half), where CAP represents channel averaging pooling and GAP represents global average pooling.

the speed of model inference. In contrast, the self-attention mechanism proposed by Transformer is a distance-independent method to calculate the association between two pixels. So here we intend to use Transformer to model global information association. The global spatial representation module inputs the compressed tensor into Transformer Encoder to calculate the self-attentiveness using global spatial information to obtain the global representation. The equation for self-attention is:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

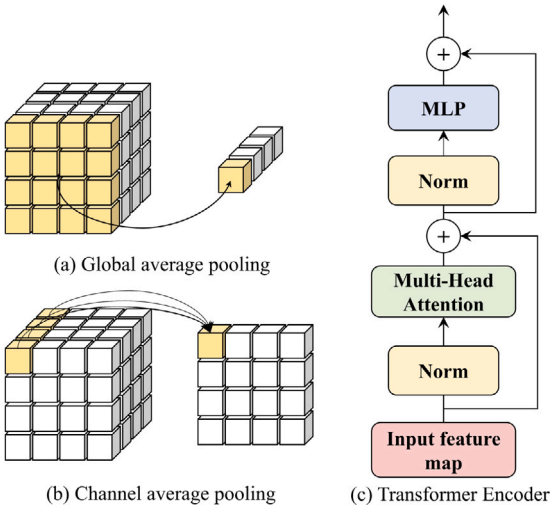
where Q,K,V are the vectors obtained by multiplying the input tensor with the parameter matrix, d_k is the number of dimensions of the tensor. The spatial excitation module first reshapes the tensor as $X_r \in \mathbb{R}^{1\times HW}$. And then, use the Depth-wise module with 1×1 convolution kernel to change the weight of each pixel and reshaped to its original shape. Finally, the spatially calibrated feature map is obtained by multiplying the input feature map with the spatial calibration sensor.

Transformer module has a great advantage over convolution in extracting global information, but it is usually accompanied by a large amount of computation, making training difficult. Our proposed processing of channel averaging pooling allows this step to be implemented with a small amount of computation.

The spatial calibration module recalibrates the weights of each pixel by extracting the global spatial information and all the channel information of the input feature map, shifting the attention to the important parts of the environment picture and ignoring some unimportant parts.

3.2.3. Channel calibration branch

To use all channel information to correct the weights of each channel in the feature map and guide the network to focus on the important objects of the environment image, we design the channel calibration branch. Like the spatial calibration branch, the channel calibration branch also consists of three corresponding modules, the structure of which is shown in the lower part of Fig. 4. The squeeze spatial information module compresses the global spatial information of the input feature map $X_{in} \in \mathbb{R}^{H\times W\times C}$ by global average pooling to obtain the tensor $X_G \in \mathbb{R}^{1\times 1\times C}$. The mathematical expression for global averaging pooling is as follows:



(a) Global average pooling

(b) Channel average pooling

(c) Transformer Encoder

Fig. 5. Comparison of the visualization of GAP and our proposed CAP (left) and the structure of the Transformer Encoder (right).

$$y_n = \frac{1}{H\times W}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}^n, \quad s.t. \quad n\epsilon\{0,1,\ldots,C-1\}, \tag{6}$$

where $x_{i,j}^n$ is denoted as the pixel value on channel n of the feature map. The global average pooling is obtained by summing and averaging all pixel values for each channel, and the whole process is shown in Fig. 5(a). The global channel representation module captures the relationships between all channels using a 1×1 convolutional layer to obtain a global representation. The channel excitation module is implemented by a 1×1 Depth-wise convolutional layer. Because each convolution kernel acts on only one channel, it is easier to correct the

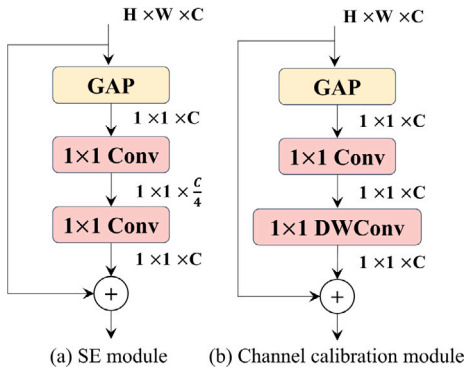


Fig. 6. Structure comparison of the SE (Squeeze-and-Excitation) module (left) and our proposed CC module (right), where GAP represents global average pooling.

Table 1
Comparison of the SE module with the CC module in terms of arithmetic volume, parameters, and accuracy.

Model	FLOPs	#params.	Accuracy
SCCNet(+SE)	0.23 G	5.48 M	91.91
SCCNet(+CC)	**0.23 G**	**6.99 M**	**92.17**

weights of each channel. Finally, the channel-corrected feature map is obtained by multiplying the input feature map with the channel calibration vector.

Our channel calibration branch is improved by the SE module (Hu et al., 2018). The SE module reduces the number of channels to avoid high model complexity. However, this strategy cannot directly model the correspondence between the weight vectors and the inputs, which reduces the quality of the results. To overcome this problem, we do not reduce the number of channels in the feature map and replace the 1×1 convolutional layer with a 1×1 Depth-wise convolutional layer. After the change, each convolution kernel only performs the convolution calculation for one channel, which makes it easier to correct the weights of each channel while reducing the parameters significantly. Both avoid high model complexity and improve the quality of the results. A comparison of the structures is shown in Fig. 6. We also designed experiments to compare, and the experimental results in Table 1 show that the improved channel calibration module works better.

The channel calibration module recalibrates the weights of each channel by extracting the global spatial information and all the channel information of the input feature map, shifting the attention to the feature objects that are important for the environment picture.

4. Experiment

4.1. Environmental image dataset

To make the experimental results more accurate and the model more generalizable, we propose a new dataset GAOs-4 (Get-AQI in One shot-4) based on GAOs-2. We used cell phone photography to manually collect 5700 environmental images at various times of the day in the Beijing area as our air quality dataset.

When collecting images, we try to avoid the interference of external factors on the images. For example, avoid capturing data in bad weather and low light, avoid direct sunlight to affect the representation of air quality in images, etc. The time, location, and AQI (air quality index) value of the acquisition are recorded at the same time as the image is acquired. The images were then divided into six categories based on the correspondence between AQI and air quality classes in Table 2. The sample images for each category are shown in Fig. 7.

Table 2
The AQI range corresponds to the air quality level.

AQI	0–50	51–100	101–150	151–200	201–300	300
Grade	1	2	3	4	5	6

To improve the dataset's quality and enhance the network's classification performance, we carefully filter the collected dataset. We removed images that were blurred due to improper photography, were dim due to backlighting, and images where the sky and buildings were covered by nearby objects. The sample of deleted environmental images is shown in Fig. 8. Because there is a certain long-tail distribution in GAOs-2, in which there are far more pictures with good air quality than those with poor air quality, there is some discrepancy in the results of the experiment. Therefore, we continuously filter and adjust the captured images to remove the redundant images with good air quality so that the images in each category are evenly distributed. We ended up with a high-quality environmental image dataset containing 3700 images, which is more than three times the data volume of GAOs-2. We divided them into 2960 training images and 740 validation images. The number of images in each category is shown in Fig. 9. The GAOs-4 dataset is compared with the GAOs-2 dataset as shown in Fig. 10. The GAOs-4 dataset is now open access (Wang & Wu, 2022).

4.2. Implementation details

The input network image is RGB three-channel, and the data is enhanced by random cropping to 224×224 size and random horizontal flipping. To prevent the network from overfitting during the training process, we add a Droppath (Huang, Sun, Liu, Sedra, & Weinberger, 2016) module to the main branch of each SCC block. The Droppath module has the Stochastic Depth feature, which skips that SCC block with a 10% probability.

The network is designed based on the PyTorch framework, using the AdamW optimizer (Loshchilov & Hutter, 2017). The base learning rate of the network is set to 5e-4 and the weight decay is set to 5e-2. To slow down the early overfitting of the model to the mini-batch in the initial stage during training and to maintain the stability of the model in the deeper layers, we use the Cosine Warmup (He et al., 2019) method. The learning first increases linearly from a very small value to a preset learning rate and then decays according to a cos function trend. We set the epoch of warmup to 5. Our model is trained on the NVIDIA GeForce RTX 3090 Ti platform, and the training loss is stable at around 0.08 after training 70 epochs.

4.3. Experimental results

With the evaluation of the GAOs-4 validation dataset, we compare the proposed SCCNet with spatial-channel attention with other representative networks. We divide the networks into three categories according to model size and compare them in the case of similar model sizes. Including pure convolutional networks (ResNet He et al., 2016, MobileNetv2 Sandler et al., 2018, ConvNeXt Liu et al., 2022), convolutional networks with channel attention (SENet Hu et al., 2018, EfficientNetv2 Tan & Le, 2021, MobileNetv3 Howard et al., 2019), and neural networks with global spatial attention (Swin Transformer Liu et al., 2021, MobileViT Mehta & Rastegari, 2021). These models are compared in three aspects: several parameters, inference speed, and classification accuracy, and the experimental results are shown in Table 3. We can find that in the network model with a similar number of parameters, the accuracy of the network with an attention mechanism is significantly improved compared to the ordinary pure convolutional network. And spatial attention is a bigger boost to the network.

The experimental results show that our proposed SCCNet achieves the highest accuracy of 92.17% in the air quality detection task by

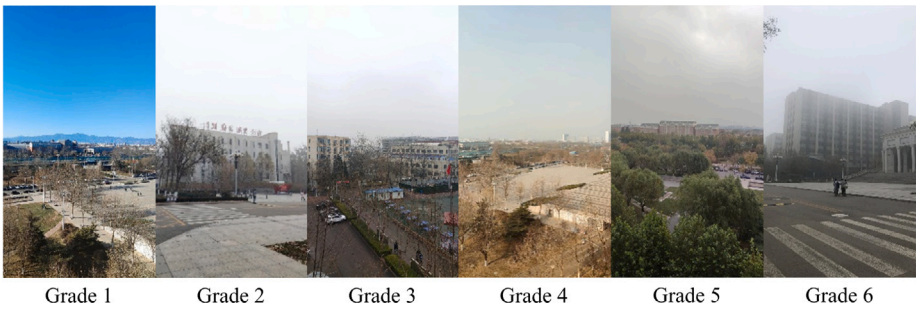


Fig. 7. Samples for each category in the environmental image dataset.

Fig. 8. Sample of deleted environmental images.

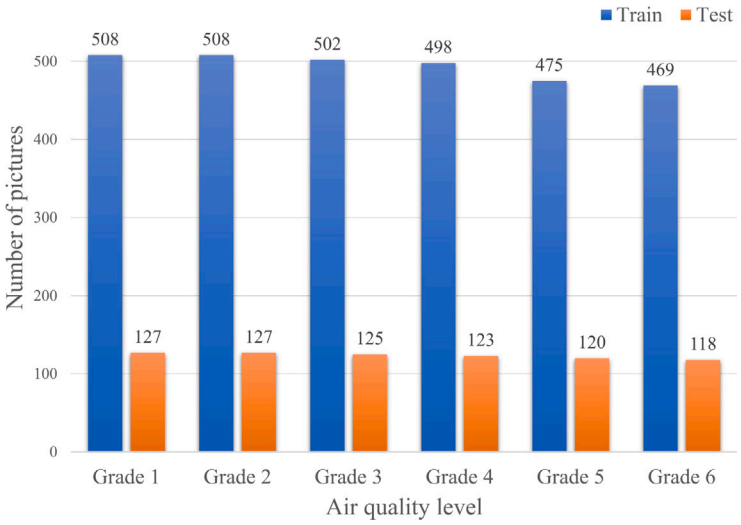


Fig. 9. Number of images in each category in the environmental image dataset.

combining spatial and channel attention and introducing the DCT module, which is 1.45% higher than the Swin Transformer-B model. Our proposed CAP module and a series of improved network structures significantly reduce the number of parameters and computations caused by the attention module, and the number of parameters and computations of SCCNet is only 1/3 of that of the Swin Transformer-B model. The above results illustrate that our proposed SCCNet is easier to deploy to mobile for air quality measurements.

4.4. Comparison of the accuracy of each category

Although our proposed network already performs better than other networks on the air quality dataset, to further explore the classification ability of the network, we observe the classification accuracy of each class through multiple experiments. The result was a bit of a surprise to us. Through Table 4 we can observe that among the six air quality classes, the classification accuracy of the first two categories and the

Table 3
Comparison of various networks on environmental image datasets.

Model	Attention	FLOPs	#params.	Accuracy
MobileNetV2 (Sandler et al., 2018)	None	0.32 G	3.5 M	89.46
MobileNetV3 (Howard et al., 2019)	Channel	0.23 G	5.48 M	90.38
MobileViT-S (Mehta & Rastegari, 2021)	Spatial	1.44 G	5.58 M	90.38
ResNet-50 (He et al., 2016)	None	4.12 G	25.56 M	89.51
SENet (Hu et al., 2018)	Channel	4.13 G	28.07 M	90.46
Swin Transformer-T (Liu et al., 2021)	Spatial	4.36 G	28.29 M	90.52
ConvNeXt-B (Liu et al., 2022)	None	15.38 G	88.59 M	90.01
EfficientNetV2-L (Tan & Le, 2021)	Channel	12.3 G	118.52 M	90.45
Swin Transformer-B (Liu et al., 2021)	Spatial	23.44 G	87.9 M	90.72
SCCNet	**Spatial&channel**	**4.69 G**	**29.28 M**	**92.17**

Table 4
Classification accuracy of our network for each class of the environmental dataset. The first column indicates that five groups of experiments were conducted.

Group	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	AVG
1	97.37	94.21	85.22	86.91	93.31	95.85	91.98
2	96.18	93.86	84.94	88.27	93.14	96.74	92.19
3	96.46	94.37	85.29	86.68	93.54	96.64	92.16
4	96.44	94.38	86.81	87.67	94.26	95.72	92.55
5	96.49	93.30	86.73	85.18	93.61	96.55	91.97
AVG	**96.39**	**94.02**	**85.79**	**86.94**	**93.57**	**96.30**	**92.17**

Table 5
Neighbor accuracy (labels change from grade 1 to grade 1 and grade 2).

Group	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	AVG
1	99.17	99.56	100	99.89	99.43	98.96	99.50
2	99.34	100	99.69	99.78	99.39	99.12	99.56
3	100	99.63	99.74	99.84	99.61	98.85	99.62
4	99.58	99.64	99.62	100	99.58	99.24	99.61
5	99.23	99.26	99.73	99.88	99.16	99.19	99.41
AVG	**99.46**	**99.62**	**99.76**	**99.88**	**99.43**	**99.07**	**99.54**

last two categories is high, but the classification accuracy of the middle two categories is not good enough. The first reason for this may be that when the air quality is in Class 3 and Class 4, the difference reflected in the images is not obvious, resulting in the network not being able to judge easily. The second reason could be that the AQI values obtained during data collection were biased, resulting in wrongly categorizing images to adjacent classes when the AQI values were at the dividing line. This also explains why images with only one adjacent category (grade 1 and grade 6) are classified with significantly higher accuracy. To prove the second point, we designed experiments for verification.

Since our statistical air quality data is obtained from the nearest weather station to the shooting location, there is some error. For example, the distance difference between the location of the weather station and the shooting location, the time difference between the data upload time and the shooting time, and the error of the data itself, etc. Once an image's AQI value is around the air quality class boundary, these deviations will likely cause misclassification. Therefore, we experiment by transforming the labels of an image into labels of itself and adjacent categories (e.g., labels change from grade 1 to grade 1 and grade 2). We use this as another evaluation criterion. Through Table 5, we can find that the classification accuracy of rank three and rank four, which had the lowest classification accuracy before, is now close to 100%. The accuracy of other classes has also been improved, and the overall classification accuracy has reached 99.3%. This indicates that most of the previously misclassified images were classified into adjacent categories due to data bias.

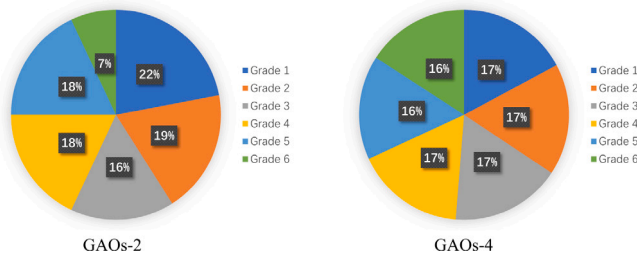


Fig. 10. Comparison of the proportion of GAOs-2 versus GAOs-4 in each category.

5. Discussion

5.1. Function of the DCT module

We experimentally compared the impact of the presence of DCT (discrete cosine transform) on the network performance, and the results are shown in Table 6. We can find that the neural network with the DCT is 0.49% more accurate in classification than the original neural network. This shows that the transformation of the environmental picture from the spatial domain to the frequency domain by the DCT is effective and indeed leads to a significant improvement in the performance of the network in air quality measurement tasks.

Table 6
Impact of DCT on network performance. Ø means no DCT.

Model	FLOPs	#params.	Accuracy
SCCNet(+Ø)	4.69 G	29.28 M	91.68
SCCNet(+DCT)	**4.75 G**	29.62 M	**92.17**

Table 7
The impact of spatial attention and channel attention on network performance. Ø means no attention, CA indicates channel attention, SA indicates spatial attention.

Model	Attention	#params.	FLOPs	Accuracy
SCCNet(+Ø)	None	28.65 M	4.49 G	90.11
SCCNet(+SC)	Spatial	28.93 M	4.58 G	91.21
SCCNet(+CC)	Channel	29.06 M	4.62 G	90.76
SCCNet(+SC+CC)	**Spatial&channel**	**29.28 M**	**4.69 G**	**92.17**

5.2. Comparison of attentional mechanisms

To demonstrate that the spatial and channel calibration modules can indeed recalibrate the weights and improve the feature extraction ability of the network, we also compare the experimental results between networks containing no attention, spatial attention, channel attention, and spatial and channel attention, as shown in Table 7. The experimental results show that both spatial attention and channel attention are improved in terms of accuracy. The former improves more, while the combination of the two makes the greatest improvement in network accuracy. In particular, our proposed attention calibration module has a negligible increase in the number of parameters and computational effort. This shows that our proposed SCC block can not only improve the accuracy of the network substantially but also the inference speed will not be reduced.

The above experiments reflect the importance of spatial and channel calibration models through data that cannot be visualized on the image. Therefore, we want to visualize the role of attentional mechanisms in the image using an attentional heat map. This enables us to observe by which regions and objects the network is classified.

We compared the attention heat map of the network without an attention mechanism and our proposed network, as shown in Fig. 11. We can find a big distinction in the focus of these two networks on environmental images. In the absence of attention, the network only focuses on a small area and thus ignores most of the information in the image. In addition, if the area is split out separately, it is not the best area to judge the air quality with our human visual observation. When the network adds spatial and channel attention, the scope of the network's attention to images becomes large. This makes full use of the information in the image and focuses on areas where the network can make accurate judgments. This could be a big reason for the significant improvement in network performance.

Since the above experimental results reflect the effect of spatial calibration, we designed the experiment to observe the calibration effect of SCCNet on the channels. Our input to the network is usually RGB three-channel color images, but the tensor obtained by convolution in the network is often very large in several channels. The feature maps for each channel represent the different feature objects extracted. After continuous training of the network, the feature objects that are most suitable for air quality classification are selected from all channels. As shown in Fig. 12, SCCNet generates many different feature maps after inputting the environment image. Each feature map is weighted and summed with the channel calibration vectors to obtain the final feature map that facilitates the network's classification.

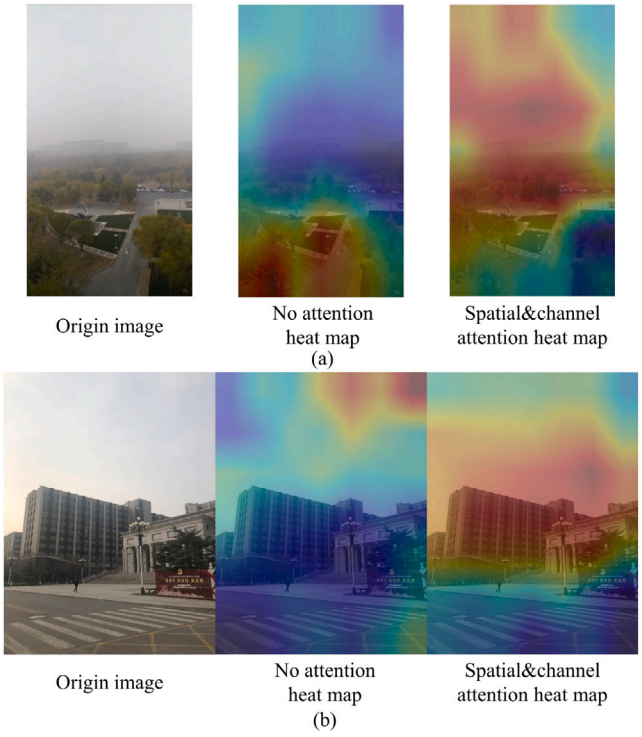


Fig. 11. Attentional heat map of environment images, original environment image(left), an attentional heat map of the no-attentional network(middle), and attentional heat map of SCCNet(right).

5.3. Error sample analysis

Our proposed spatial and channel calibration network achieves 92.17% accuracy on the air quality dataset. If we exclude the deviation of AQI values, the network achieves an amazing 99.3% accuracy rate. Although the accuracy rate is already very high, some images are still difficult to distinguish. The sample images with misclassification are shown in Fig. 13, which are all images with large classification errors. Through careful observation and analysis, we can find that there are two main reasons for misclassified pictures. The first reason is the weather, when the weather is in cloudy weather, the environmental pictures with good air quality are easily classified into the poor air quality class, as shown in Fig. 13(a)(b). In contrast, when the weather is sunny, pictures of poor air quality are easily classified into the good class, as shown in Fig. 13(c). The second reason is the distance from the building when shooting. When the air quality is poor, the faster the light decays in the air. The farther the building is from the person when shooting, the lower the image clarity. If you shoot too close to the object, it is not easy to reflect the quality of air from the image. As shown in Fig. 13(d), the distance affects the classification performance of the network.

6. Conclusion

This paper proposes an air quality detection method that combines channel attention and spatial attention mechanisms. We use channel attention to collect information from all channels to recalibrate the weights of each channel, focusing on objects that benefit the network for classification. We use spatial attention to collect spatial global information, recalibrate the weights of each pixel, and focus on the regions that favor the network for classification. In addition, we also apply the DCT to the air quality detection task. It is found that the network is more sensitive to the environment images after DCT, which further improves the network performance. We also built a new air quality

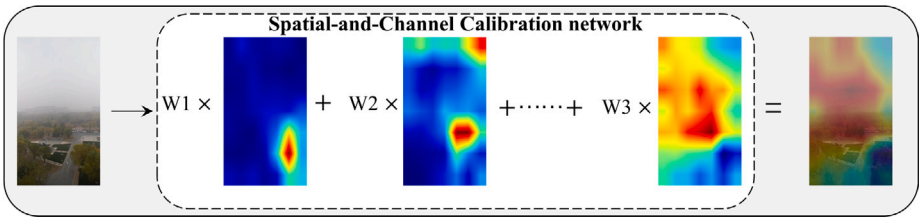


Fig. 12. The process of weighted summation of each channel feature map.

Fig. 13. Sample images of misclassification.

class dataset GAOs-4 and validated our proposed method on GAOs-4. Experimental results show that our method substantially improves the network performance with almost no increase in the number of parameters and computational effort.

Air quality detection is a special fine-grained image classification task, where the environmental images differ very little between each category. The method proposed in this paper will be helpful for researchers with sufficient high-quality environmental image datasets. In a follow-up study, it may be possible to further improve the accuracy of air quality detection if the weather conditions of environmental images can also be utilized.

CRediT authorship contribution statement

Zhenyu Wang: Conceptualization, Methodology, Investigation, Writing – original draft, Supervision, Project administration, Funding acquisition. **Fucheng Wu:** Methodology, Software, Visualization, Investigation, Resources, Writing – original draft, Writing – review & editing. **Yingdong Yang:** Visualization, Formal analysis, Investigation, Data curation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Links to the dataset and code are provided in the article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61976090.

References

Chakma, A., Vizena, B., Cao, T., Lin, J., & Zhang, J. (2017). Image-based air quality analysis using deep convolutional neural network. In *2017 IEEE international conference on image processing* (pp. 3949–3952). IEEE.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv arXiv:2010.11929.

Gao, Z., Xie, J., Wang, Q., & Li, P. (2019). Global second-order pooling convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3024–3033).

He, K., Sun, J., & Tang, X. (2010). Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(12), 2341–2353.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., & Li, M. (2019). Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 558–567).

Howard, A. G., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., et al. (2019). Searching for MobileNetV3. In *2019 IEEE/CVF international conference on computer vision* (pp. 1314–1324).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Huang, G., Sun, Y., Liu, Z., Sedra, D., & Weinberger, K. Q. (2016). Deep networks with stochastic depth. In *European conference on computer vision* (pp. 646–661). Springer.

Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial transformer networks. *Advances in Neural Information Processing Systems*, *28*.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, H., Kim, H.-E., & Nam, H. (2019). Srm: A style-based recalibration module for convolutional neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1854–1862).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision* (pp. 9992–10002).

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11976–11986).

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Ma, J., Li, K., Han, Y., & Yang, J. (2018). Image-based air pollution estimation using hybrid convolutional neural network. In *2018 24th International conference on pattern recognition* (pp. 471–476). IEEE.

Mehta, S., & Rastegari, M. (2021). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. arXiv, arXiv:2110.02178.

Mnih, V., Heess, N. M. O., Graves, A., & Kavukcuoglu, K. (2014). Recurrent models of visual attention. arXiv, arXiv:1406.6247.

Pan, Z., Yu, H., Miao, C., & Leung, C. (2017). Crowdsensing air quality with camera-enabled mobile devices. In *Proceedings of the AAAI conference on artificial intelligence, vol. 31, no. 2* (pp. 4728–4733).

Qin, Z., Zhang, P., Wu, F., & Li, X. (2020). FcaNet: Frequency channel attention networks. In *2021 IEEE/CVF international conference on computer vision* (pp. 763–772).

Rijal, N., Gutta, R. T., Cao, T., Lin, J., Bo, Q., & Zhang, J. (2018). Ensemble of deep neural networks for estimating particulate matter from images. In *2018 IEEE 3rd international conference on image, vision and computing* (pp. 733–738). IEEE.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510–4520).

Tan, M., & Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International conference on machine learning* (pp. 10096–10106). PMLR.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. arXiv, arXiv:1706.03762.

Wang, X., Girshick, R. B., Gupta, A. K., & He, K. (2017). Non-local neural networks. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 7794–7803).

Wang, J., Jin, L., Li, X., He, S., Huang, M., & Wang, H. (2022). A hybrid air quality index prediction model based on CNN and attention gate unit. *IEEE Access*, *10*, 113343–113354.

Wang, Z., & Wu, F. (2022). *Get-AQI in one shot-4 (GAOs-4)*. Mendeley Data, http://dx.doi.org/10.17632/s5hh825ctr.2, URL https://data.mendeley.com/datasets/s5hh825ctr/2.

Wang, Q., Wu, B., Zhu, P. F., Li, P., Zuo, W., & Hu, Q. (2019). ECA-Net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF conference on computer vision and pattern recognition* (pp. 11531–11539).

Wang, Z., Yang, Y., & Yue, S. (2022). Air quality classification and measurement based on double output vision transformer. *IEEE Internet of Things Journal*, *9*, 20975–20984.

Wang, Z., Yue, S., & Song, C. (2021). Video-based air quality measurement with dual-channel 3-D convolutional network. *IEEE Internet of Things Journal*, *8*(18), 14372–14384.

Wang, Z., Zheng, W., Song, C., Zhang, Z., Lian, J., Yue, S., et al. (2019). Air quality measurement based on double-channel convolutional neural network ensemble learning. *IEEE Access*, *7*, 145067–145081.

Yang, Z., Zhu, L., Wu, Y., & Yang, Y. (2020). Gated channel transformation for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11794–11803).

Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. arXiv, arXiv:1409.2329.

Zhang, Q., Fu, F., & Tian, R. (2020). A deep learning and image-based model for air quality estimation. *Science of the Total Environment*, *724*, Article 138178.

Zhang, C., Yan, J., Li, C., Rui, X., Liu, L., & Bie, R. (2016). On estimating air pollution from photos using convolutional neural network. In *Proceedings of the 24th ACM international conference on multimedia* (pp. 297–301).