

Multi-Scale Hybrid CNN-Transformer for Smoke Detection in Satellite Images

Tony Zhang
University of Michigan
1301 Beal Ave, Ann Arbor, MI 48109
ttzhan@umich.edu

Robert Dick
University of Michigan
1301 Beal Ave, Ann Arbor, MI 48109
dickrp@umich.edu

Abstract

This paper addresses the challenge of smoke detection from satellite images, which is crucial for identifying and mitigating wildfires. Smoke can vary in size, shape, and texture, making it difficult to classify using remote sensing images. While several CNN architectures have been proposed for smoke detection, they have limitations in modeling long-range context in images due to convolution's bias towards learning local relationships. To address this limitation, the paper describes a hybrid network that combines CNN and transformer architectures. The transformer architecture leverages multi-head attention to learn long-range, global relationships among different image regions.

Initially, multi-scale features are extracted by adding the transformer architecture after each CNN layer. Additionally, another transformer layer is appended to capture relationships among features in different receptive fields, significantly improving model accuracy. The proposed approach is evaluated on the USTC.SmokeRS dataset of remote sensing images for smoke detection and outperforms prior methods.

1. Introduction

Wildfires can have a significant negative impact on human health and the environment. The flames can cause property damage, and the fire can create smoke that can cause respiratory problems and aggravate existing medical conditions. Therefore, timely and accurate detection of smoke and thus fire is crucial for protecting human health and the environment. Image-based smoke detection, particularly from aerial images, can help identify the location and severity of wildfires.

Smoke detection in remote sensing images presents several challenges. Smoke can have a low contrast with its surrounding environment, making it difficult to detect. Smoke can occur at various scales, from small fires to large wildfires or industrial accidents. In addition, smoke can be obscured by atmospheric interference, such as haze or clouds,

which can scatter or absorb the light, making the smoke less visible.

While traditional CNNs can extract fine-grained spatial details, they have difficulty capturing global relationships among regions of an image due to their small receptive fields. CNNs typically suffer from using a limited receptive field for modeling scenes in images. While increasing CNN depth can increase the size of the receptive field, the benefit adding more layers diminish once a certain number of layers are incorporated [8]. Thus, the accuracy benefits of increasing CNN depth saturate before approaches such as our CNN-Transformer hybrid method.

By using transformers, the model can learn the relationships between all image pixels instead of only learning those within local neighborhoods, resulting in a more comprehensive analysis of the image. Additionally, transformers do not have any inductive biases related to locality, making them more suitable for capturing global information. Therefore, to achieve more accurate smoke detection models for aerial images, the use of advanced techniques such as hybrid CNN-Transformer architectures can improve accuracy.

The proposed method consists of three steps. (1) The CNN backbone network is used to extract multi-scale features, taking advantage of the locality feature of convolution. (2) A transformer architecture is integrated after each layer of the CNN to capture long-range relationships within the multi-scale features. (3) To enhance feature fusion and facilitate inter-layer information exchange, we employ an additional transformer layer that is stacked on top of the existing model.

We summarize our contributions as follows:

1. We describe a CNN-transformer model to extract multi-scale feature representations and long-range contextual relationships for smoke classification.
2. We incorporate an additional transformer layer on top of the existing model to perform feature fusion, allowing for the exchange of information with the other layers of the CNN.
3. We incorporate the patch attention mechanism that mod-

els the relative importance of each patch. This allows the Transformer to focus on the most important patches of the image when modeling long-range context of the pixels of the feature map.

2. Related Work

This section provides an overview of related work on smoke detection in satellite images.

2.1. Smoke Detection

Image-based smoke detection in computer vision has been an active research area for many years. One of the earliest approaches to smoke detection in images was based on manually-constructed features [10, 19]. For instance, Yuan [19] used dynamic texture analysis for video smoke detection. This approach, however, suffered from several limitations, including sensitivity to lighting conditions and the inability to distinguish between smoke and other similar objects, such as clouds.

More recently, deep learning-based approaches have gained popularity for smoke detection. A CNN-based method called the DCNN incorporates a dual-channel neural network for smoke detection [6]. Lin et al. proposed a method for video smoke detection using faster R-CNN and 3D CNN to extract spatio-temporal information [12].

Ding et al. developed a video-based smoke detection algorithm consisting of two parts: one part extracts spatial features from individual frames, while another extracts temporal features from consecutive frames [4]. Some researchers have proposed using CNN-based object detection models for smoke detection, such as Faster R-CNN model for video frames [7].

Moreover, remote sensing-based smoke detection has been an active area of research in recent years. For instance, SmokeNet uses spatial and channel-wise attention modules for satellite smoke scene detection [1], where a feature map in a CNN contains a spatial dimension and a channel dimension. Spatial attention emphasizes the important regions in the feature maps, whereas channel attention emphasizes the most informative channels. In addition, Chen et al. proposed a method for satellite smoke detection via a self-adaptive feature aggregation technique that combines both global and salient features. It uses a two-branch network to capture spatial and spectral features, respectively [2].

2.2. Image Classification

Image classification is a long-standing problem in computer vision. State-of-the-art image classification methods have undergone significant improvements in recent years, including VGGNet [15], ResNet [9], and DenseNet [11]. These models use deeper architectures and various techniques such as skip connections to improve accuracy.

ResNeXt is an extension of the ResNet architecture that focuses on improving the network's scalability and performance [18]. It achieves this by introducing cardinality that controls the number of parallel paths in a block, allowing for greater diversity in feature representation. Another widely known model is EfficientNet with a smaller model size than previous architectures [16]. EfficientNet uses a compound scaling method that balances model depth, width, and resolution.

CoAtNet combines convolution and transformers through a vertical layout design with a focus on efficient feature reuse and parameter sharing across different resolutions [3]. The convolutional layers are applied to the spatial dimensions and the attention layers are applied to the channel dimensions. Moreover, ConvNeXt [14] extends the ResNeXt model by using grouped convolutions, inverted bottlenecks, larger kernel sizes.

Another approach to image classification is based on attention mechanisms from transformers [17]. For example, the Vision Transformer (ViT) model uses a transformer architecture with self-attention to capture long-range relationships in the image. It processes an image as a sequence of patches, which are then fed into a transformer encoder. In particular, ViT demonstrates the effectiveness of attention mechanisms for image classification without the use of convolutional operations [5]. Furthermore, the Swin Transformer introduces a hierarchical architecture, each stage consisting of patch merging followed by a series of transformer blocks [13]. Patch merging aggregates information from the previous stage and produces a new representation with a higher resolution.

3. Overview of the Proposed Model

This section provides an overview of the proposed model and describes each of its key components in detail.

3.1. Overview

While CNNs have been able to obtain a higher accuracy on classification tasks, they can have an inability to obtain contextual information from the image at larger scales due to a lower receptive field. On the other hand, the transformer network has been shown to be effective in capturing long-range relationships among different image regions. Specifically, it is important to model the relationship of both nearby and distant objects to classify smoke, especially in complex backgrounds. By combining the strengths of CNNs and transformers, the proposed architecture can capture both local and global information from the image, resulting in more accurate smoke detection.

Therefore, this paper details a hybrid CNN-Transformer model. The CNN backbone network is responsible for capturing low-level and high-level features, which are then passed through the transformer network to capture global

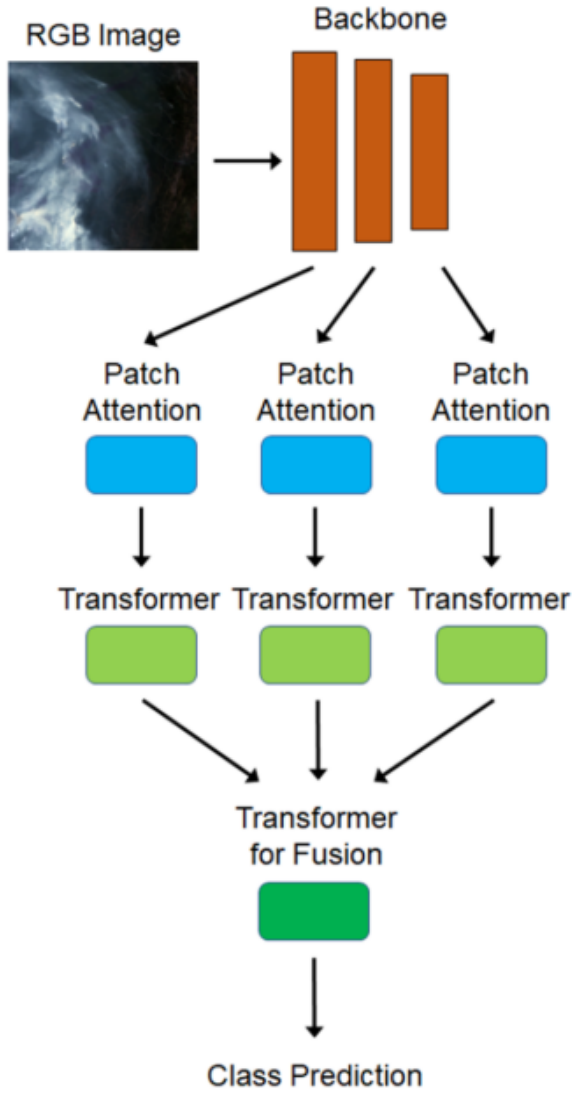


Figure 1. Framework of the proposed method.

relationships between different regions of the image. In the last stage, we perform adaptive feature refinement in order to consider relationships at multiple length scales by stacking another transformer on top of the existing architecture.

Overall, this multi-scale approach allows the model to detect smoke regardless of its size or location in the image. Additionally, the use of multi-scale features enhances the model’s robustness to noise and other forms of interference that may obscure smoke in remote sensing images. The next subsection will provide details about the transformer architecture before going further into the proposed model’s architecture.

3.2. Transformer Architecture

The design of our model uses the encoder from the original Transformer [17]. The input to the Transformer encoder is a 1D vector of feature embeddings. Initially, we split an image into a sequence of 2D patches and flatten each into a 1D vector. Specifically, an image $x \in \mathbb{R}^{H \times W \times C}$ is reshaped to form a sequence of N patches $x = [x_1, \dots, x_N] \in \mathbb{R}^{N \times P^2 \times C}$, where the patch resolution is P , the number of patches is $N = HW/P^2$, and the number of channels from the original image is C .

After each patch is flattened, it is linearly projected to produce a sequence of patch embeddings. Then, it captures positional information via learnable position embeddings, which are added to the input sequence of tokens and then fed as an input to the transformer encoder. The transformer encoder is a set of L layers in order to generate a sequence of long-range contextual features. Each layer is composed of a multi-head self-attention (MSA) block and a multi-layer perceptron (MLP) block. Additionally, the layer norm (LN) is applied before the MSA and the MLP, and residual connections are added after the MSA and the MLP.

The multi-head self-attention encodes the input sequence by linearly projecting it into different feature spaces through the key, query, and value [17]. They are represented as $K \in \mathbb{R}^{n \times d_m}$, $Q \in \mathbb{R}^{n \times d_m}$, and $V \in \mathbb{R}^{n \times d_m}$, respectively. In this case, n is the length of the input sequence and d_m is the dimension of each feature. The MSA linearly projects the input differently h times in order to form h different heads. The attention weight for each head is computed based on its corresponding key, query, and value, shown in the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_m}} \right) V \quad (1)$$

First, we compute a dot product between Q and K and scale it by $1/\sqrt{d_m}$. Then, we use a softmax function on the output and multiply it by V .

The transformer architecture produces an embedding for each patch, and we take the average of all patch embeddings, which serves as the outputs. In our model, the MSA contains 8 heads, each transformer module has 6 layers, the embedding dimension is 128, and the size of the hidden layer is 256.

3.3. Network Architecture

Figure 1 shows the architecture of the proposed model. Initially, we use a ResNet-50 backbone, pretrained on ImageNet, to extract features, where the $(i + 1)$ -th layer is denoted as f_i ($i = 0, 1, 2$). The CNN feature map of the 1st layer’s output f_0 contains a high resolution and maintains more image details. On the other hand, the CNN feature

map of the 3rd layer’s output f_2 contains more semantic information about the objects in the image. The CNN is effective at mapping the image to feature representations using the inductive bias of locality. It, however, cannot capture the long-range context of the scene well.

Next, we utilize multiple transformer encoders T_i^0 ($i = 0, 1, 2$) after the $(i + 1)$ -th layer of the CNN, each drawing out multi-scale features by operating on different receptive fields initially. In particular, we incorporate the patch attention mechanism P_i ($i = 0, 1, 2$) before each transformer, which models the relative importance of each patch. We split each feature map $f_i \in \mathbb{R}^{H \times W \times C}$ into a sequence of 2D patches $\tilde{f}_i \in \mathbb{R}^{N \times P^2 \times C}$, where i is the output of the $(i + 1)$ -th layer of the CNN. In our model, each patch has a spatial resolution of 4×4 .

From the output of the $(i + 1)$ -th layer of the CNN f_i , the operations described before are depicted below:

$$\tilde{f}_i = \text{Patch}(f_i), \quad (2)$$

$$\hat{f}_i = \text{Position_Embeddings}(\text{Linear}(P_i(\tilde{f}_i))), \quad (3)$$

$$\tilde{\tilde{f}}_i = T_i^0(\hat{f}_i), \quad (i = 0, 1, 2) \quad (4)$$

Before feeding the output of the patch attention $P_i(\tilde{f}_i)$ into the transformer T_i^0 ($i = 0, 1, 2$), each patch is flattened into a 1D vector and linearly projected to produce a sequence of patch embeddings. Then, the model explicitly learns the value of position embeddings for each position in the sequence and adds them to the 1D vector. The resulting features \hat{f}_i are then fed to the transformer encoder T_i^0 .

To achieve better feature fusion and inter-layer information exchange, we add an extra transformer encoder T^1 on top of the current model, thereby combining low-level and high-level features into a fused representation f . This fusion mechanism ensures that the model captures both fine-grained spatial details and global relationships between different regions of the image, resulting in more accurate smoke detection.

$$\tilde{f} = \text{Concat}(\tilde{f}_0, \tilde{f}_1, \tilde{f}_2), \quad (5)$$

$$f = T^1(\tilde{f}) \quad (6)$$

Finally, a fully-connected layer FC_f is performed on f to output a vector v corresponding to the values for each individual class:

$$v = \text{FC}_f(f) \quad (7)$$

3.4. Patch Attention

Figure 2 shows the architecture of Patch Attention. The input to the transformer is a flattened 1D vector representing a sequence of 2D patches of the feature map. We preprocess each patch through our Patch Attention mechanism which models the relative importance of each patch. This allows

the Transformer to focus on the most important patches of the image when modeling long-range context of the pixels of the feature map.

We split each feature map $f_i \in \mathbb{R}^{H \times W \times C}$ from the CNN backbone into a sequence of 2D patches $\tilde{f}_i \in \mathbb{R}^{N \times P^2 \times C}$, where i is the output of the $(i + 1)$ -th layer of the CNN. The patch resolution is P , the number of patches is $N = HW/P^2$, and the number of channels from the original image is C . In our model, each patch has a spatial resolution of 4×4 .

Next, we generate a feature for each patch in the feature map through average pooling $\tilde{f}_i^{\text{avg}} \in \mathbb{R}^{N \times C}$, where the size of the pooling window is 4×4 . The features are concatenated to form a 1D vector \hat{f}_i^{avg} . After that, we will learn an attention map A_i for each patch in the feature map f_i via fully connected (FC) layers:

$$A_i = \sigma(W_2 \delta(W_1 \hat{f}_i^{\text{avg}})), \quad (8)$$

where σ is ReLU, δ is sigmoid, r represents the parameter for the gating operation, $W_1 \in \mathbb{R}^{(N \times C) \times \frac{N \times C}{r}}$ represents the first FC layer, and $W_2 \in \mathbb{R}^{\frac{N \times C}{r} \times (N \times C)}$ represents the second FC layer. The FC layers reduce the dimension of the vector before expanding it back to the original dimension. The parameter r is set to 2.

In the end, the attention map for each image region is generated $A_i \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times C}$. The attention map is then upsampled to the dimension of the feature map $A_i \in \mathbb{R}^{H \times W \times C}$. After that, an element-wise multiplication is performed with A_i and the corresponding feature map f_i to result in f_i^{attn} . Finally, an element-wise summation is performed with f_i^{attn} and f_i through a residual connection directly with the input.

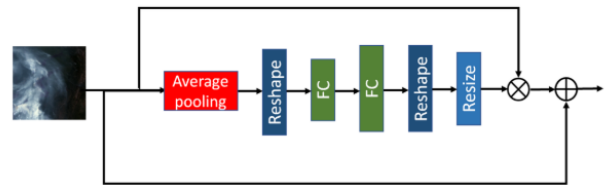


Figure 2. Architecture of Patch Attention.

3.5. Fusion of Multi-Scale Features

After the low-level and high-level feature maps are individually refined by a transformer model, another transformer module is used to represent the relationship of the fine-grained spatial details and global relationships between different regions of the image.

One method of performing fusion is simply concatenating low-level and high-level features. However, low-level

Table 1. Confusion matrix of the proposed method.

Predicted / Actual	Cloud	Dust	Haze	Land	Seaside	Smoke
Cloud	455	1	0	5	7	3
Dust	4	363	16	14	4	13
Haze	1	24	371	5	3	29
Land	1	7	4	386	2	6
Seaside	0	4	2	0	386	3
Smoke	5	5	8	1	1	352

features contain more spatial information about the input image but less semantic information. The converse is true for high-level features. Low-level features and high-level features are complementary. As both features have very different representations, a more sophisticated architecture can better model the relationship between different feature sets.

The refined feature maps $\tilde{f}_n \in \mathbb{R}^{H \times W \times C}$ (where $n = 0, 1, 2$) are concatenated to form a single vector $\tilde{f} = [\tilde{f}_0, \tilde{f}_1, \tilde{f}_2]$. It is then fed as an input to the transformer model T_1 , which outputs a refined vector f . Finally, the result goes through a fully connected layer FC_f . The output is a vector with elements corresponding to the relative probabilities of each class in the dataset.

The transformer can implicitly embed semantic information from high-level features into the low-level features and vice versa. Low-level features have a higher spatial resolution and contain fine-grained spatial information about the input image such as color and shape, but without semantic information about the relationship of both nearby and distant objects. High-level features are the opposite, containing rich semantic information about the relationship of both nearby and distant objects.

By this approach, low-level features are embedded with semantic information. This technique allows for the extraction of information beyond the boundaries of the receptive fields, while preserving the spatial details of the image. For high-level features, this technique can also effectively embed fine-grained details to bridge the gap between different feature representations.

4. Experimental Results

We first introduce the dataset and the experimental protocol. Next, we evaluate our proposed method on images containing smoke and other scenes and compare it with other methods.

4.1. Dataset and Implementation Details

This work uses the USTC_SmokeRS dataset for experimental analysis [1]. It consists of 6,225 remote sensing images in the RGB format with a spatial resolution of 256×256 . The images are categorized into six different classes: cloud, dust, haze, land, seaside, and smoke. The dataset represents scenes from different areas around the world.

The training-testing split is approximately 60% to 40%. During the training phase, we set the learning rate to $1e-4$, the batch size to 8, and the number of epochs to 30. We also set the momentum parameter to 0.9 and use the Adam optimizer to update model parameters. The cross-entropy loss (\mathcal{L}_{CE}) is used to compare the predicted class to the ground-truth class, based on the output vector of class probabilities.

Table 2. Results of smoke detection with other methods.

Methods	Accuracy
ResNeXt-101 [18]	0.835
ResNeXt-152 [18]	0.809
EfficientNet [16]	0.758
Vision Transformer [5]	0.745
Swin Transformer [13]	0.855
CoAtNet [3]	0.860
ConvNeXt [14]	0.760
Proposed	0.929

4.2. Model Comparison

We compare our model with previous smoke classification methods, as shown in Table 2. Each method’s accuracy is calculated using the same parameters for a fair comparison. Experiments on the USTC_SmokeRS dataset demonstrate that our model improves accuracy by 6.9% compared to the best existing method. The confusion matrix for the six different classes in the dataset is shown in Table 1.

Transformers rely on patch-based processing, where the image is divided into patches and each patch is treated as a token. This can lead to a loss of spatial information and make it difficult to capture fine-grained details in the image. Moreover, CNNs have a strong inductive bias towards local features, which allows them to capture patterns at different scales and orientations. Vision Transformers and Swin Transformers do not have this bias, which can make them less effective at capturing local features. In particular, Vision Transformers require additional data to compensate for the lack of this locality bias.

ResNeXt, CoAtNet, and ConvNeXt do not include attention mechanisms like those used in Transformers, which limits their ability to perform fine-grained feature extraction

and manipulation. Additionally, these CNN-based models may struggle with modeling long-range dependencies in sequential data due to their reliance on pooling and sub-sampling layers. In contrast, the CNN-Transformer hybrid model, with its use of self-attention mechanisms, is better equipped to model long-term dependencies across sequences.

Table 3. Results of the ablation analysis.

Methods	Accuracy
CNN f_2 + Transformer T_0, T_2	0.861
CNN f_1, f_2 + Transformer T_0, T_1, T_2	0.868
Ours w/o the transformer at the 2nd level T_1 and w/o patch attention P_i	0.859
Ours w/o the transformer at the 2nd level T_1	0.885
Ours using only the 3rd CNN layer f_2	0.872
Ours using only the 2nd and 3rd CNN layers f_1, f_2	0.894
Ours w/o patch attention P_i	0.925
Ours	0.929

4.3. Ablation Analysis

We investigate the contribution of each component in our proposed model. First, we remove the patch attention mechanisms P_i ($i = 0, 1, 2$) after the $(i + 1)$ -th CNN layer while retaining the rest of the model. As shown in Table 3, the model without patch attention achieves an accuracy of 92.5%, underperforming compared to the full model.

Next, we evaluate the importance of multi-scale features. In one variant, we use only the outputs of the second and third CNN layers f_1 and f_2 , appending patch attention modules P_1, P_2 , and individual Transformer encoder layers T_1^0, T_2^0 respectively. An additional Transformer encoder T_1 is added on top of the current model, achieving an accuracy of 89.4%.

In another case, we use only the output from the third CNN layer f_2 , followed by patch attention P_2 , a Transformer encoder T_2^0 , and an additional Transformer encoder T_1 stacked on top. This configuration results in an accuracy of 87.2%.

We also construct a variant of the model without the top Transformer layer T_1 , using only an element-wise sum followed by a fully connected layer: $v = \text{FC}_f(\tilde{f}_0 + \tilde{f}_1 + \tilde{f}_2)$ after the outputs from T_i^0 ($i = 0, 1, 2$). Without inter-layer information exchange, the model achieves an accuracy of 88.5%.

In a similar model that also omits patch attention modules, we perform an element-wise sum followed by a fully connected layer: $v = \text{FC}_f(\tilde{f}_0 + \tilde{f}_1 + \tilde{f}_2)$, resulting in a reduced accuracy of 85.9%.

We further construct a method that uses only the third CNN layer f_2 , a Transformer encoder T_2^0 , and a final fully

connected layer FC_f , yielding an accuracy of 86.1%. Another configuration that uses both the second and third CNN layers f_1 and f_2 , Transformer encoders T_1^0 and T_2^0 , and a final FC layer FC_f after an element-wise sum $\tilde{f} = \tilde{f}_1 + \tilde{f}_2$, achieves an accuracy of 86.8%.

5. Conclusion

In this paper, we have presented a CNN-Transformer hybrid network for smoke detection in satellite images. We extract multi-scale features by incorporating the transformer architecture after each layer of the CNN. This serves to model the global relationship between different image regions. After that, we append another transformer layer to effectively represent interactions between features at different receptive fields, thus significantly improving the model performance. Our proposed method improves classification accuracy compared to past work. For future work, we plan to analyze different forms of data for smoke detection such as video and other modalities such as depth and infrared.

References

- [1] Rui Ba, Chen Chen, Jing Yuan, Weiguo Song, and Siuming Lo. Smokenet: Satellite smoke scene detection using convolutional neural network with spatial and channel-wise attention. *Remote Sensing*, 11:1702, 2019. 2, 5
- [2] Shikun Chen, Yichao Cao, Xiaoqiang Feng, and Xiaobo Lu. Global2salient: Self-adaptive feature aggregation for remote sensing smoke detection. *Neurocomputing*, 466:202–220, 2021. 2
- [3] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. In *Advances in Neural Information Processing Systems*, pages 3965–3977. Curran Associates, Inc., 2021. 2, 5
- [4] Zhipeng Ding, Yaqin Zhao, Ao Li, and Zhaoxiang Zheng. Spatial-temporal attention two-stream convolution neural network for smoke region detection. *Fire*, 4(4), 2021. 2
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10687–10696, 2020. 2, 5
- [6] Ke Gu, Zhifang Xia, Junfei Qiao, and Weisi Lin. Deep dual-channel neural network for image-based smoke detection. *IEEE Transactions on Multimedia*, 22(2):311–323, 2019. 2
- [7] Sumayea Benta Hasan, Shakila Rahman, Md. Khaliluzzaman, and Siddique Ahmed. Smoke detection from different environmental conditions using faster r-cnn approach based on deep neural network, 2020. Preprint. 2
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2015. 1

- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. [2](#)
- [10] Chao-Ching Ho. Machine vision-based real-time early flame and smoke detection. *Measurement Science and Technology*, 20:045502, 2009. [2](#)
- [11] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. [2](#)
- [12] Gaohua Lin, Yongming Zhang, Gao Xu, and Qixing Zhang. Smoke detection on video sequences using 3d convolutional neural networks. *Fire Technology*, pages 1–21, 2019. [2](#)
- [13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10257–10266, 2021. [2](#), [5](#)
- [14] Zhuang Liu, Hanzi Mao, Chaozheng Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE Conference on Computer Vision and Pattern Recognition*, pages 11966–11976, 2022. [2](#), [5](#)
- [15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [2](#)
- [16] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114, 2019. [2](#), [5](#)
- [17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. [2](#), [3](#)
- [18] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017. [2](#), [5](#)
- [19] Feiniu Yuan. Video-based smoke detection with histogram sequence of lbp and lbpv pyramids. *Fire Safety Journal*, 46(3):132–139, 2011. [2](#)