Received: 15 June 2024 | Revised: 13 August 2024 | Accepted: 14 September 2024 | Published online: 30 September 2024

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2025, Vol. 4(1) 89–96
DOI: 10.47852/bonviewJCCE42023628

BON VIEW PUBLISHING

The Fire-ViT Model for Tunnel Fire Detection with Vision Transformer Improvement

Xiaobing Liu[1,2,*] and Vladimir Y. Mariano[1]

[1]*College of Computing and Information Technologies, National University, Philippines*

[2]*College of Information Engineering, Jiangxi Communications Vocational and Technical College, China*

Abstract: Smoke detection in tunnels presents unique challenges, including low light conditions, visual obstructions like smoke and headlights, and the need to analyze ultrahigh-resolution images. To address these challenges, this study introduces an innovative model named Fire-ViT, leveraging the vision transformer (ViT) architecture. Unlike traditional convolutional neural networks that often struggle with false positives under these complex conditions, Fire-ViT incorporates an attention mechanism and multiperceptron layer, significantly enhancing its capability to discern details in high-resolution images specific to tunnel environments. The model's performance is outstanding, achieving an accuracy rate of 99.87% on a high-resolution tunnel image dataset, markedly surpassing that of conventional models. Notably, Fire-ViT not only elevates detection accuracy and robustness but also cuts training time in half. This efficiency, coupled with its adaptability to the intricate tunnel environment, makes Fire-ViT an ideal solution for early warning systems against fires in tunnels, fulfilling the demand for high-standard, fine-grained fire detection.

Keywords: Fire-ViT, tunnel fire dataset, tunnel fire detection, fire alarm, visual transformer

1. Introduction

With the rapid development of China's social economy and the significant enhancement of its scientific and technological strength, road tunnels, as an important part of the transportation network, have grown rapidly to 24,850 by the end of 2022, with a total length of 26,784.3 km. Tunnels play a central role in modern transportation systems, but the potential risk of tunnel fires has increased, posing a challenge to safe operations. The narrow and closed characteristics of the tunnel make the evacuation and rescue operations extremely difficult once the fire occurs, often resulting in heavy losses. Therefore, the construction of early fire monitoring and warning system is very important to improve tunnel safety [1].

At present, the tunnel fire detection technology is mainly composed of the Internet of Things (IoT) sensor alarm system and machine vision fire identification technology, the two technologies complement each other to form an intelligent protection network to ensure that fire can be quickly detected and responded to. However, early IoT detectors are vulnerable to environmental factors, resulting in insufficient detection accuracy and frequent false alarms. In addition, infrared and optical sensors require manual verification, increasing operation and maintenance costs [2].

In recent years, the development of AI technology has provided a new method for the early monitoring and warning of highway tunnel fire. Combining machine vision and IoT smoke sensing systems, video-based fire detection enables efficient fire identification and response by analyzing characteristics such as the color, shape, and rate of change of flames and smoke [3]. Researchers are developing new algorithmic models, exploring ways to reduce data processing costs, and building larger fire datasets to enhance the model's generalization.

Although the application of machine vision technology in tunnel fire warning has broad prospects, it still faces challenges such as accuracy, processing large-size images, real-time performance, cost, multimode recognition, hazard degree recognition, and anomaly recognition [4]. Future research will focus on the development of machine vision algorithms that can achieve higher accuracy, real-time performance and robustness in low-illumination, large-size images, and small-scale datasets. Multimodal data fusion will be the key to reduce cost, improve recognition accuracy, and reduce false positives.

This study effectively improves the application effect of machine vision technology in tunnel fire early warning system, with the specific objectives as follows:

1) Construct a tunnel fire dataset containing more than 5000 high-resolution images, covering 20 types of tunnel fire simulation images under different scenarios, to ensure the universality and practicality of the dataset.
2) Using the core algorithm framework vision transformer (ViT), a model called Fire-ViT was developed to automatically identify key fire characteristics such as flame, smoke, and temperature changes in tunnels.
3) The performance of Fire-ViT model was compared with other convolutional neural network (CNN) related algorithms, and the model performance was optimized by adjusting model

***Corresponding author**: Xiaobing Liu, College of Computing and Information Technologies, National University, Philippines and College of Information Engineering, Jiangxi Communications Vocational and Technical College, China. Email: liux2@students.national-u.edu.ph

hyperparameters and using data enhancement techniques to reduce the false positive rate and improve the accuracy and response speed of the model.

2. Literature Review

In 2019, Tian et al. [5] proposed a multitype of flame detection method based on Faster R-CNN, which improved the average detector accuracy rate by 3.03% to 8.78%, showing strong anti-interference ability. Then, Li et al. [6] based on the improved Mask R-CNN algorithm in 2020, improved the fire image detection accuracy by more than 5% through the bottom-up fusion of the feature pyramid and optimization of the loss function, effectively enhancing the generalization ability of the model. These research results of fire monitoring based on deep learning not only play an important role in the field of fire prevention and control but also provide valuable experience for the application of deep learning technology in other fields such as image recognition and processing.

With the rapid development of machine vision technology, the field of fire detection has also ushered in significant technological progress. From different angles, researchers have proposed a variety of effective fire detection solutions, which have significantly improved the accuracy and efficiency of fire detection. For example, the real-time fire warning and monitoring method based on the human visual attention mechanism utilizes the FMF algorithm and the dynamic frame difference method to effectively identify the flame region in the video frame [7]. These innovations not only provide strong technical support for fire prevention and control but also bring new research directions to the field of tunnel fire monitoring.

In recent years, the transformer architecture has made great progress in the computer vision space. It was originally used for natural language processing tasks, such as machine translation and language modeling, but has since been applied to image classification and intensive prediction tasks and has shown excellent performance [8]. ViT demonstrates its advantages over multiple image recognition benchmarks by applying it directly to image patch sequences, reducing the computational resources required compared to traditional convolutional networks. Detection transformer (DETR) and faster-RCNN (ViT-FRCNN) models based on vision converters improve target detection from a new perspective, respectively. DETR directly outputs prediction in parallel through transformer architecture. However, ViT-FRCNN combines adaptive clustering converters to achieve end-to-end target detection [9]. Fax-RCNN model (ViT-FRCNN) based on vision converters (Josh Beal,2021) is one of the most widely used target detection methods [10]. Swin transformer improves efficiency and is compatible with a wide range of visual tasks by introducing shift window calculations and layered design [11]. Pyramid vision transformer (PVT) model shows its versatility and high efficiency in different visual tasks, especially in object detection and segmentation tasks [12]. The positioning and focusing vision transformer (LF-ViT) effectively reduces the computing and memory requirements by reducing significant spatial redundancy, while the optimization method proposed by Ataiefard et al. [13] can improve the training throughput by reducing unnecessary token interactions.

In the field of fire monitoring, research based on ViT is relatively limited, but the proposal of fire former model marks a new exploration in this field [14]. This model optimizes the forest fire monitoring algorithm by using self-attention mechanism and significantly improves the stability and reliability of tunnel fire monitoring and early warning system by constructing the spatial feature relationship between smoke and surrounding natural environment elements. The successful application of fire former model opens a new direction for fire monitoring research based on ViT. Future research can further explore the optimization strategy of ViT architecture to improve the accuracy and real-time performance of fire monitoring.

Although deep learning and machine vision technologies have made significant progress in the field of fire monitoring, especially for image classification tasks, relatively little research has been done on tunnel fire monitoring and ViT-based research. The particularity of tunnel environment, such as narrow space, light variation, and smoke interference, poses higher challenges for fire monitoring algorithms. At present, the number of studies on the application of machine vision in tunnel fire monitoring in the literature is limited, indicating that this field has not received widespread attention. Although ViT has achieved great success in areas such as image classification, it does not mean that it can be directly applied in the field of fire monitoring, which requires in-depth research on its adaptability and performance. In general, VIT-based tunnel fire monitoring research is currently in its infancy, which provides a broad space for future research, and it is necessary to develop new algorithms to adapt to the special challenges of the tunnel environment and explore the practical application potential of ViT in fire monitoring.

3. Theoretical Framework

In this study, a new tunnel fire detection model Fire-ViT is developed by improving transformer model to efficiently identify fire location, scale, and burning degree. This model integrates HD and thermal imaging technology and realizes accurate identification and prediction of fire characteristics through four main steps: image acquisition, data preprocessing, feature extraction, and classification recognition, as shown in Figure 1. In the feature extraction phase, the model utilizes multihead self-attention mechanisms and multiperceptron (MLP) blocks, combined with LN standardization and residual connectivity, to optimize the network to extract key flame and smoke features. Part of the classifier uses MLP Head to further distinguish smoke, flame and other interfering factors, and realize the classification detection of fireworks.

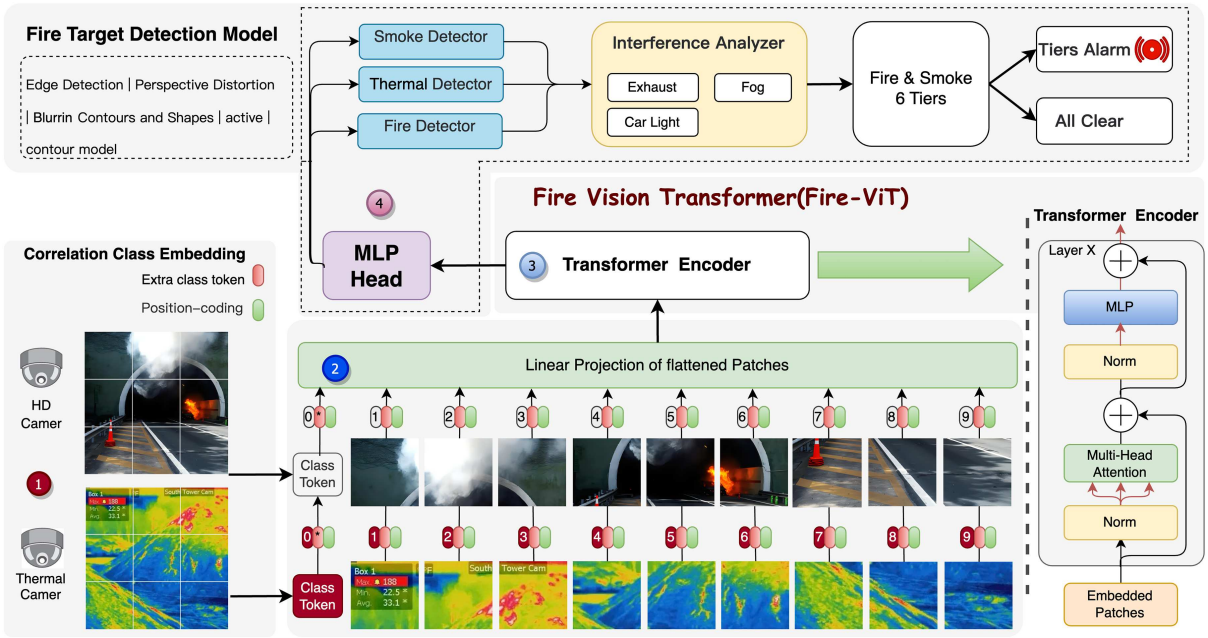
Key innovations in the Fire-ViT model include the combination of high-definition and thermal images to extract richer pyrotechnic features, parallel computing to improve classification prediction speed, and the introduction of technical improvements such as coding loss, data enhancement, supervised learning, and denoizing training to meet the specific needs of tunnel fire detection. These improvements not only improve the accuracy and robustness of the model but also speed up the calculation speed and provide earlier warning detection capability for tunnel fires. Through this technical scheme, Fire-ViT model provides an effective solution strategy for tunnel fire warning and management, demonstrating its potential and value in dealing with complex fire scenarios.

4. Research Methodology

4.1. Research design

This research has successfully developed a Fire-ViT fire detection algorithm based on PyTorch and completed efficient training and evaluation on Kaggle platform. With CUDA accelerated Tesla T4 GPU and the latest PyTorch 2.0.0 release, combined with the Adam optimizer and refined training strategy, the model training efficiency is ensured.

Figure 1
Fire-ViT conceptual framework



As shown in Figure 2, the Fire-ViT model makes use of transformer architecture, especially the design of multiple multihead attention layers and MLP layers, to realize in-depth learning of global and local image features.

Traditional convolutional neural networks (CNNs) are often difficult to achieve ideal fitting results when faced with the challenge of insufficient dataset. In addition, although PVT network is suitable for intensive partition training of images to achieve high output resolution, it is easy to affect the calculation speed due to too large dataset when processing large-size tunnel images with more noise [15]. Therefore, the transformer model, which has both high resolution and low compute and memory costs, was chosen as the solution in this study. With its multihead self-attention mechanism, transformer model can segment large-size images into small blocks without relying on many recursion and convolution operations and use linear embedded sequences of these image blocks as inputs [16]. This processing is like labeling words in natural language processing (NLP) and training an image classification model through supervised learning.

Feature extraction in this study aims to segment the candidate flame region by the color of flame and smoke, and then extract the key features of the fire region in the image, such as the flame formation position, shape, temperature, burning speed, smoke density, and the change rate of smoke diffusion area. The encoder of the transformer model consists of alternating multihead self-attention layer (MSA) and multilayer perceptron (MLP) blocks. Before each block, linear transformations are performed by application layer normalization (LN), and then residual connections are used to help optimize the deepening network structure, thereby delivering critical feature information and continuously extracting global and local features of the image. Although batch normalization (BN) is generally more efficient than LN in the field of image processing, LN is chosen to process images in this study because BN is not suitable for cases where the time series length is not fixed. Each step of the transformer model is automatically converted to parallel matrix operations, an advantage that makes it superior to CNN in handling large-size images [17].

This study refers to the research specifically solving the deep transformer training problem [18–20], through parallel computing to improve ViT, several serial transformer blocks are processed in parallel, which is equivalent to reducing the depth of the model and increasing the width of the model, which can accelerate the reasoning speed of the model. See area 2 in Figure 2. We replace the serial computation with two parallel computation operations, formula (1) and formula (2):

$$x_{l+1} = x_l + mhsa_{l,1}(x_l) + mhsa_{l,2}(x_l) \tag{1}$$

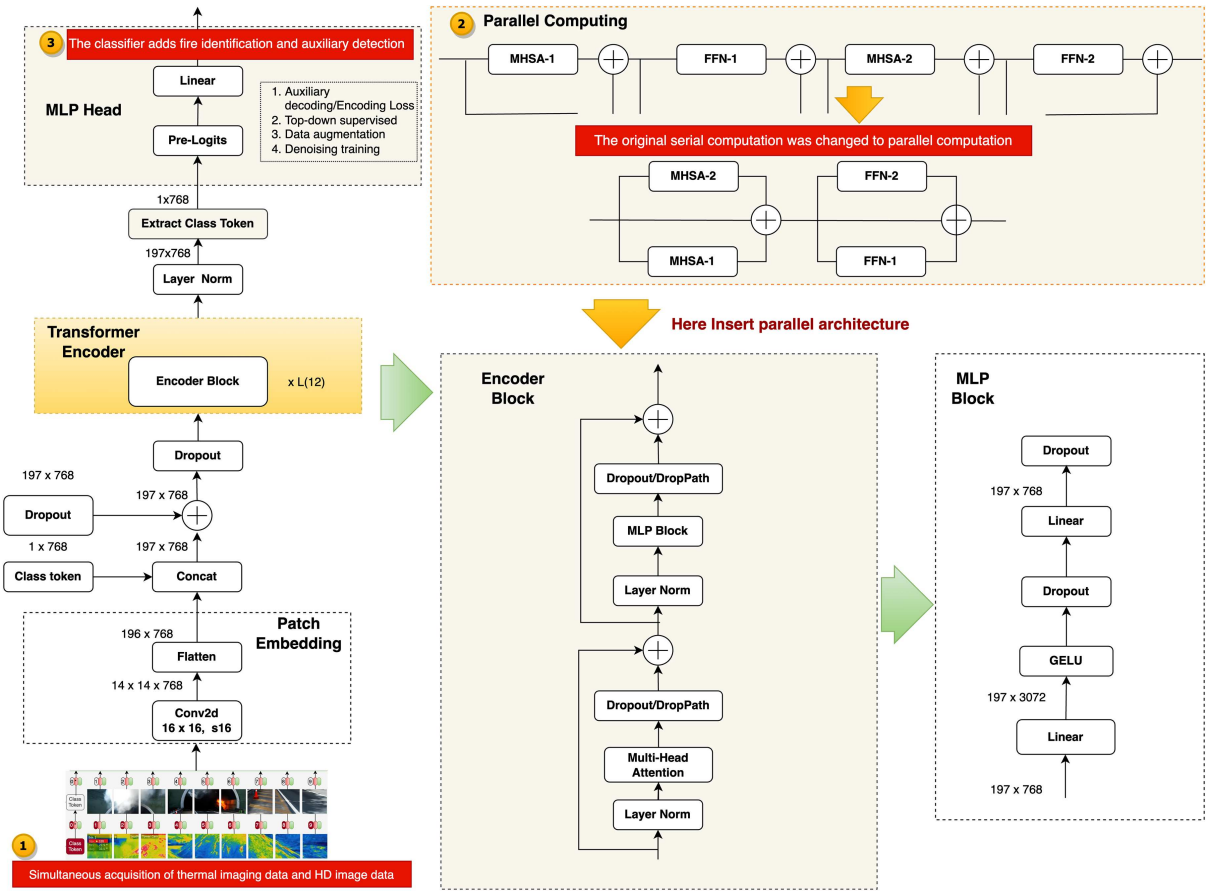
$$x_{l+2} = x_{l+1} + ffn_{l,1}(x_{l+1}) + ffn_{l,2}(x_{l+1}) \tag{2}$$

In terms of model optimization, several innovative strategies such as coding loss optimization, data enhancement, supervised learning, and denoizing training are adopted to comprehensively improve the classification accuracy, stability, and generalization ability of the model.

4.2. TunnelFire dataset

In this study, TunnalFire2024, a practical tunnel pyrotechnic monitoring dataset, was constructed and filmed inside a tunnel under construction in Jiangxi Province, China, in April 2024, addressing the need of tunnel fire monitoring. We simulated and recorded six different fire scenarios, including no fire, fog, exhaust, smoke, small fire, and large fire, with the aim of comprehensively capturing the key characteristics of the fire scenario. Capturing data from both perspectives with HD and thermal imaging cameras increases the diversity and richness of the dataset. Each scene was filmed separately into a video, totaling 20, each about 3 minutes long, ensuring high-quality data and detail capture.

Figure 2
Fire-ViT implementation architecture



We further refined the scene into 20 subclasses and constructed a detailed labeling system based on four core dimensions, such as image type, lighting conditions, flame/smoke size and concentration, and main elements in the image, which provided a solid foundation for model training and verification. To ensure a balanced dataset, at least 200 images were collected from each category, for a total of 5034 images, which were divided into the training set and the test set in a ratio of 7:3.

The TunnalFire2024 dataset, which covers the full range of tunnel fire scenarios, aims to promote global scientific exchange and cooperation, advance tunnel fire monitoring technology, and contribute to public safety. We open this dataset to the Kaggle website to inspire more innovative thinking and solutions.

4.3. Evaluation

To fully evaluate the performance of the Fire-ViT model, we used multidimensional metrics such as accuracy, FPS, recall, accuracy, and F1 scores. The formula is as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \tag{3}$$

$$FPS = \frac{1000ms}{\text{preprocess}+\text{inference}+\text{postprocess}} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{6}$$

$$F_1\ Score = \frac{2\times(\text{Precision}\times\text{Recall})}{\text{Precision}+\text{Recall}} \tag{7}$$

These measures take into account the model's predictive accuracy, speed, and overall performance. By comparing with the actual labels in the test set, the concepts of TP (true positive example), FP (false-positive example), TN (true-negative example), and FN (false-negative example) are used to calculate the above indexes, and the model performance is further analyzed through the confusion matrix. During the evaluation process, we also made hyperparameter adjustments and data enhancements to the model to improve its performance and generalization ability. This comprehensive evaluation method ensures the efficiency and accuracy of Fire-ViT model in tunnel Fire image prediction and classification.

4.4. Experimental Config

This study utilized the TunnalFire2024 dataset to conduct a detailed comparison and analysis of models such as Fire-ViT, ViT, CNN, ImageNet, ResNet, and VGG-16 based on two key metrics: FPS and Accuracy. CNNs are pivotal in computer vision, enabling feature extraction and learning from images for tasks like classification and recognition [21]. ImageNet, developed by Stanford University, is a crucial CNN-based image classification algorithm that uses a vast database of tagged images to advance research in this

area [22]. ResNet, or Residual Neural Network, introduces residual blocks to enhance training of deep networks, significantly improving model performance and accuracy [23]. VGG-16, devised by the University of Oxford, is a classic CNN model known for its depth and efficiency. It utilizes small convolution kernels and pooling layers to capture intricate image details, setting a benchmark in image classification [24, 25].

4.4.1. Parameters setting

Multiple parameters are covered: the learning rate is set to 0.0001 to balance the update speed and convergence stability; The batch size is chosen as 16 to reduce memory consumption and accelerate model convergence. The training period is set as 50 to improve the learning ability and avoid overfitting. Dropout was used as a regularization technique to reduce the risk of overfitting. The activation function selected RELU to accelerate the convergence speed and improve the sparse expression ability. The optimizer uses Adam to speed up training and improve stability due to its advantages of momentum and adaptive learning rate; The loss function was cross-entropy loss, which was suitable for multi-classification problems, and the SoftMax layer was used to optimize the model performance. The model adopted a pretraining strategy and used the model pretrained on ImageNet to accelerate convergence and improve the performance of new tasks. Finally, the model output layer is set to 20 categories to adapt to the specific image recognition task. The setting of these parameters aims to fully tap the potential of the Fire-ViT model and improve the accuracy and efficiency of tunnel fire detection.

4.4.2. Quantitative results

As can be seen from the performance report Table 1 of Fire-ViT, Fire-ViT, and ViT achieve 99.87% and 99.34% accuracy values, respectively, which proves their accuracy in detecting and segmenting flame pixels. Due to the use of both global and local features, with transformer as the Backbone, they segment tunnel fire pixels well and provide finer fire details. However, in terms of robustness, the loss value of Fire-ViT (0.0204) is significantly lower than the loss value of ViT (0.0609) and is significantly lower than other classical CNN and deep CNN models, and there is little difference between Fire-ViT and actual tunnel fireworks image classification. The parallel processing module of Fire-ViT has significantly improved the accuracy and robustness of tunnel fireworks image classification.

Fire-ViT parallel computing processing ensures that the number of model channels matches, and the input data includes high-definition images with a resolution of 2160×3840 and thermal images with a resolution of 720 × 544 to unify into images with a resolution of 384×384. During the test, 1511 images are divided into 20 categories. Considering that there are many categories of image classification, the accuracy of the model may reduce the recognition accuracy due to interference. Therefore, using the pretrained model for classification and the transfer method can not only extract high-level and finer features faster but also reduce the computing power demand and the segmentation flame pixel positioning time in the training process.

Table 2 shows the comparison and analysis of Fire-ViT, ViT, CNN, ImageNet, ResNet, and VGG-16 models in inference speed (FPS) and classification accuracy (accuracy) based on Tunnal-Fire2024 dataset. The results show that Fire-ViT outperforms other networks in terms of performance metrics.

It can be seen from Table 2 that the FPS value of the Fire-ViT model after parallel computing processing (29.37) is significantly higher than that of the ViT model (27.58), and the accuracy is as high as 99.87% at the speed of FPS = 30. In addition, the CNN model and ImageNet model have the best FPS scores of 254.38 and 183.15, respectively, indicating that their inference speed is better than other models, but the accuracy is not ideal. ResNet shows better results in terms of recognition accuracy and speed compared to CNN, ImageNet, and VGG-16. However, it is still difficult to model global information compared to Fire-ViT and ViT, achieving an accuracy of 96.43%, which is 3.44% lower than the accuracy of Fire-ViT (99.87%). The VGG-16 model with 1511 images as a test set is much slower than the CNN, ImageNet, ResNet models with the same test set, and its accuracy (95.83%) is 4.05% lower than the

Table 1
Quantitative results of Fire-ViT and transformer on TunnalFire2024 dataset

Model	Backbone	Input Resolution	Accuracy (%)	Loss
Fire-ViT	ViT	384 × 384	99.87	0.0204
ViT	ViT	384 × 384	99.34	0.0609
CNN	CNN	384 × 384	92.52	0.4040
ImageNet	Mobile Net	384 × 384	96.23	0.2122
ResNet	ResNet	384 × 384	96.43	0.2333
VGG-16	VGG-16	384 × 384	95.83	0.2445

Table 2
384 x 384 image size 43 TunnalFire2024 datasets, in the Fire-ViT, ViT, CNN, ImageNet, ResNet, and VGG-16 model were analyzed

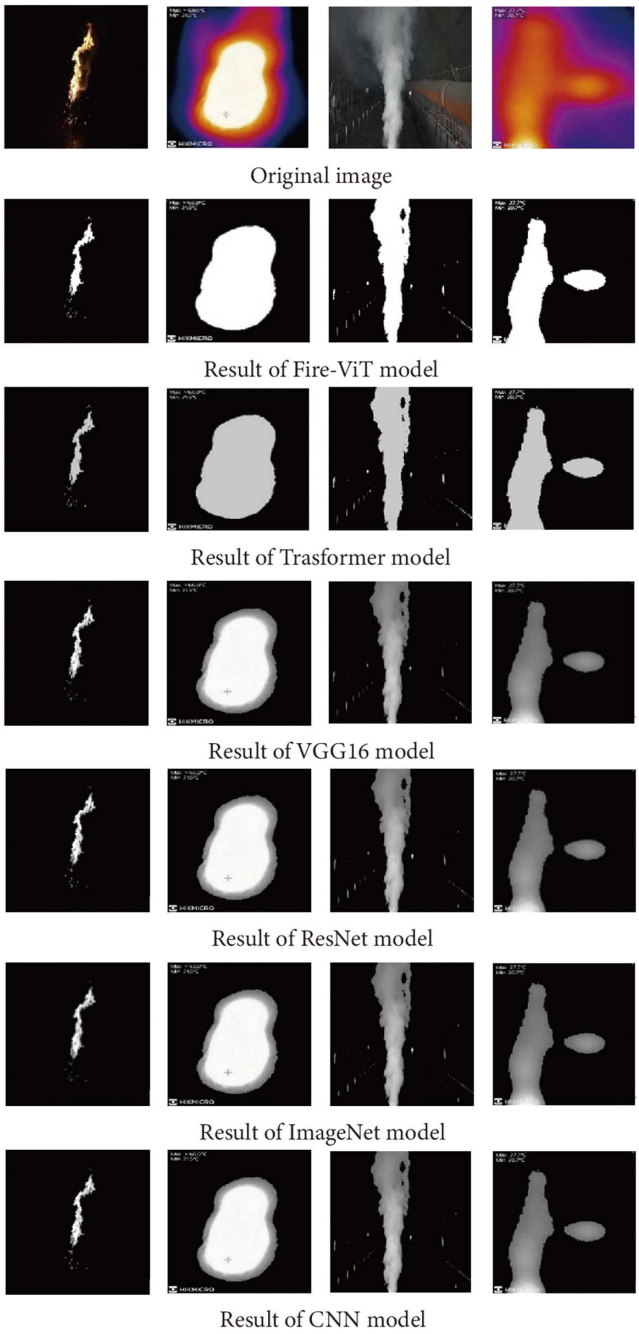
Model	Backbone	Test Data	Accuracy (%)	FPS
Fire-ViT	ViT	1511	99.87	29.37
ViT	ViT	1511	99.34	27.58
CNN	CNN	1511	92.52	254.38
ImageNet	Mobile Net	1511	96.23	183.15
ResNet	ResNet	1511	96.43	92.93
VGG-16	VGG-16	1511	95.83	57.11

accuracy (99.87%) of the Fire-ViT model. The reason is that there are more network layers than other models.

4.4.3. Qualitative results

This study compares the performance of Fire-ViT with traditional models (including transformer, VGG16, ResNet, ImageNet, and CNN) for fire image segmentation tasks in depth, as shown in Figure 3. The Fire-ViT model performs better than the traditional transformer model in high definition and thermal image processing, accurately identifying and separating flames and backgrounds, and maintaining high efficiency even in complex environmental conditions. It is superior to other models in the accurate detection of small fire areas and flame shapes, showing its application advantages in the field of fire monitoring.

Figure 3
Comparison of image masks generated by Fire-ViT, ViT, ResNet, VGG-16, CNN, and ImageNet



In contrast, although VGG16 performs well in the differentiation of fire pixels, it is insufficient in the sharpness of edges and the capture of small area fire pixels. ResNet is effective in fire shape detection, but the feature extraction accuracy of small fires needs to be improved. The ImageNet model has misjudgment in the identification of small fire areas, while the CNN model can identify the main outline of the fire, but it is not accurate enough in the extraction of fireworks feature information.

In addition, the robustness of the Fire-ViT model was evaluated by using images downloaded from the Internet, and the results showed that it can accurately segment fire pixels and detect the exact shape of the fire under various conditions (such as foggy environments), showing better visual effects and higher accuracy than ViT. These comparative results highlight the advance and practicability of Fire-ViT in fire image processing.

5. Conclusion and Suggestion

In this study, we explore the performance of multiple models in tunnel fire image classification tasks, including Fire-ViT, ViT, CNN, ImageNet, ResNet, and VGG-16. The results show that the Fire-ViT model performs well in both classification accuracy and inference speed, especially in the processing of tunnel fire images. The application of its parallel computing mechanism and transfer learning strategy significantly improves the efficiency and accuracy.

1) Parallel computation and model performance

Through its parallel computing mechanism, Fire-ViT effectively combines the global context information of transformer architecture with the local feature recognition capability like CNN. This design enables Fire-ViT to quickly identify the fine structure and global distribution features in the flame image and achieve a fast and accurate early warning of tunnel fires. On the TunnalFire2024 dataset, Fire-ViT achieved 99.87% accuracy and a speed of inference (FPS) of 29.37, significantly outperforming other traditional models.

2) Model comparison

In contrast, while the ViT model also demonstrated good accuracy (99.34%), its reasoning speed was slightly slower (FPS = 27.58). This may be due to the lack of parallel processing capabilities characteristic of Fire-ViT. In addition, although the traditional CNN model has a fast inference speed, its accuracy is not as good as the Fire-ViT model based on transformer architecture, which shows its limitations in dealing with complex scenes.

3) The importance of image resolution

The study also highlights the importance of evaluating all models at a uniform resolution (384 × 384), which ensures fair comparison between models, and highlights the importance of consistency in preprocessing for the model's ability to generalize.

4) Innovation and advantages of Fire-ViT model

Fire-ViT not only makes significant progress in accuracy and robustness but also optimizes the frame rate of the ViT model to achieve a response speed in the millisecond level, fully meeting the needs of real-time detection. By combining the input of high-definition image and thermal image, using parallel computing technology and optimization of MLP Head classifier, Fire-ViT has

shown its important application value in the field of tunnel fire early warning.

In order to further improve the performance of tunnel fire detection technology, we propose the following suggestions: continuously expand the data set size and explore the strategy of integrating multimodal information to enrich the database of model training; Continuously optimize the Fire-ViT model, including the adjustment of hyperparameters and the application of data enhancement methods, to improve the generalization ability and performance of the model; Consider applying the Fire-ViT model to a real tunnel environment while addressing the challenges of real-time performance and deployment cost; A wide range of scenarios was tested in various tunnel environments to verify the generalization ability and stability of the model under different conditions. Exploring the combination of Fire-ViT model with other existing fire detection technologies, the aim is to build a more comprehensive and efficient fire warning system. By implementing these strategies, we expect to see more research and innovation in the future to further improve the efficiency and accuracy of tunnel fire detection technology to ensure tunnel traffic safety.

Funding Support

This work was sponsored by General Science and Technology Project in Science and Technology Project of Jiangxi Provincial Department of Transportation in 2024: "Fire-ViT Tunnel Fire Warning System Based on Vision Transformer" (2024YB027).

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/xiaobingliu323/tunnelfire2024/data.

Author Contribution Statement

Xiaobing Liu: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Vladimir Y. Mariano:** Methodology, Validation, Resources, Writing – review & editing, Supervision, Project administration.

References

[1] Zhang, Y., & Huang, X. (2024). A review of tunnel fire evacuation strategies and state-of-the-art research in China. *Fire Technology*, *60*(2), 859–892. https://doi.org/10.1007/s10694-022-01357-5

[2] Zhou, Y., Pang, T., Liu, K., Martin, C. H., Mahoney, M. W., & Yang, Y. (2023). Temperature balancing, layer-wise weight analysis, and neural network training. In *37th Conference on Neural Information Processing Systems,* 1–31.

[3] Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., . . . , & Cao, Y. (2023). EVA: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19358–19369.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . , & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv Preprint:2010.11929*.

[5] Tian, H., Halidan, A., & Du, H. (2019). Multi-type flame detection combined with Faster R-CNN. *Journal of Image and Graphics*, *24*(1), 73–83.

[6] Li, Y., Zhang, Q., Shen, Z., & Zuo, B.. (2019). FMF de huǒ yàn xiǎn zhù xìng jiǎn cè [Flame saliency detection based on FMF]. *China Safety Science Journal*, *29*(5), 56–61. https://doi.org/10.16265/j.cnki.issn1003-3033.2019.05.010

[7] Jia, F. (2023). Fire flame detection in tunnel based on Halcon and ResNet50. In *3rd International Conference on Neural Networks, Information and Communication Engineering,* 256–259. https://doi.org/10.1109/NNICE58320.2023.10105717

[8] Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., . . . , & Han, X. (2021). TransPath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference*, 186–195. https://doi.org/10.1007/978-3-030-87237-3_18

[9] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference,* 213–229. https://doi.org/10.1007/978-3-030-58452-8_13

[10] Beal, J., Kim, E., Tzeng, E., Park, D. H., Zhai, A., & Kislyuk, D. (2020). Toward transformer-based object detection. *arXiv Preprint:2012.09958*.

[11] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., . . . , & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision,*, 10012–10022.

[12] Hu, Y., Cheng, Y., Lu, A., Cao, Z., Wei, D., Liu, J., & Li, Z. (2024). In *Proceedings of the AAAI Conference on Artificial Intelligence,*, *38*(3), 2274–2284. https://doi.org/10.1609/aaai.v38i3.28001

[13] Ataiefard, F., Ahmed, W., Hajimolahoseini, H., Asani, S., Javadi, F., Hassanpour, M., . . . , & Liu, Y. (2024). SkipViT: Speeding up vision transformers with a token-level skip connection. *arXiv Preprint:2401.15293*.

[14] Qiao, Y., Jiang, W., Wang, F., Su, G., Li, X., & Jiang, J. (2023). FireFormer: An efficient transformer to identify forest fire from surveillance cameras. *International Journal of Wildland Fire*, *32*(9), 1364–1380. https://doi.org/10.1071/WF22220

[15] Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., . . . , & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 568–578.

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . , & Polosukhin, I. (2017). Attention is all you need. In *31st Conference on Neural Information Processing Systems,*, 1–15.

[17] Muhammad, K., Ahmad, J., Lv, Z., Bellavista, P., Yang, P., & Baik, S. W. (2019). Efficient deep CNN-based fire detection

and localization in video surveillance applications. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *49*(7), 1419–1434. https://doi.org/10.1109/TSMC.2018.2830099

[18] Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., & Jégou, H. (2021). Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 32–42.

[19] Wang, H., Ma, S., Dong, L., Huang, S., Zhang, D., & Wei, F. (2024). DeepNet: Scaling transformers to 1,000 layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(10), 6761–6774. https://doi.org/10.1109/TPAMI.2024.3386927

[20] Touvron, H., Cord, M., El-Nouby, A., Verbeek, J., & Jégou, H. (2022). Three things everyone should know about vision transformers. In *Computer Vision – ECCV 2022: 17th European Conference,* 497–515. https://doi.org/10.1007/978-3-031-20053-3_29

[21] Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

[22] Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, *8*(2), 20539517211035955. https://doi.org/10.1177/20539517211035955

[23] Hayou, S., Clerico, E., He, B., Deligiannidis, G., Doucet, A., & Rousseau, J. (2021). Stable ResNet. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics,* 1324–1332.

[24] Tao, J., Gu, Y., Sun, J., Bie, Y., & Wang, H. (2021). Research on vgg16 convolutional neural network feature classification algorithm based on transfer learning. In *2nd China International SAR Symposium,* 1–3. https://doi.org/10.23919/CISS51089.2021.9652277

[25] Theckedath, D., & Sedamkar, R. R. (2020). Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks. *SN Computer Science*, *1*(2), 79. https://doi.org/10.1007/s42979-020-0114-9

How to Cite: Liu, X., & Mariano, V. Y. (2025). The Fire-ViT Model for Tunnel Fire Detection with Vision Transformer Improvement. *Journal of Computational and Cognitive Engineering*, *4*(1), 89–96. https://doi.org/10.47852/bonviewJCCE42023628