Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/120/2025.19482

Research on Forest Fire Detection Method Based on Multi-Modal Data

Jiabo Liu[1,a,*]

[1]*School of Life Sciences, Sichuan University, Sichuan, China*
a. 2022141240135@stu.scu.edu.cn
**corresponding author*

Abstract: Forest fire is one of the major disasters that threaten the ecological environment and the sustainable development of social economy. Aiming at the limitations of traditional fire monitoring methods in terms of real-time, accuracy and environmental adaptability, this paper proposes a multimodal data fusion forest fire detection method based on Transformer architecture. This method realizes fire classification and precise positioning by fusing the feature information of visible light images and infrared images. The model design includes multimodal input encoding, cross-modal attention mechanism and multi-head attention mechanism, aiming to fully explore the complementary characteristics of the two modal data, and enhance the stability and convergence of the model through residual connection and layer normalization technology. The experiment is based on the public FLAME dataset. The results show that the proposed method is superior to the traditional method in fire classification accuracy and positioning accuracy, with a classification accuracy of 94.5% and an IoU of 0.86, which significantly improves the fire recognition ability in complex scenes. This study provides effective technical support for the intelligent monitoring of forest fires.

Keywords: Multimodal fusion, Forest fire detection,Transformer, Cross-modal attention.

1. Introduction

As one of the most important ecological disasters in the world, forest fires not only cause irreversible damage to forest resources and the ecological environment, but also threaten human life and property safety and the sustainable development of social economy[1]. In recent years, with the intensification of climate change and the frequent occurrence of extreme weather events, the frequency and intensity of forest fires have increased significantly, which has put forward higher requirements for early monitoring and precise positioning of fires[2]. However, due to the random occurrence, complex regional distribution and rapid development of forest fires, traditional fire monitoring methods have significant limitations in real-time, accuracy and environmental adaptability[3][4]. The rise of multi-sensor data fusion technology has provided a new solution for forest fire monitoring. By fusing visible light images and infrared images, sensors of different modalities can complement each other and make up for the shortcomings of a single sensor limited by ambient light, weather conditions, etc. Visible light images can capture the visual characteristics of fires[5], such as smoke and open flames, while infrared images are highly sensitive to heat sources and can effectively detect fire areas under low light or smoke obstruction conditions. Multimodal data fusion has become a research hotspot in the field of forest fire monitoring.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/120/2025.19482

In recent years, the development of deep learning technology has provided a powerful tool for multimodal data fusion[6]. Among them, the Transformer architecture has made breakthrough progress in natural language processing, computer vision and other fields due to its excellent modeling capabilities and parallelization characteristics. Compared with traditional convolutional neural networks (CNNs), Transformers can adaptively capture global and local features through the attention mechanism and achieve efficient fusion of cross-modal features[7]. However, in the task of forest fire detection, how to effectively combine the characteristics of visible light and infrared images and design an adaptable and efficient Transformer model is still a challenging research problem. This paper focuses on forest fire detection based on multimodal data fusion and proposes a fire detection method based on the Transformer architecture. This method extracts deep features from visible light and infrared images and cross-modal fusion, and uses a multi-head attention mechanism to efficiently model fire features. At the same time, in order to further improve the stability and robustness of the model, this paper introduces residual connection and layer normalization technology. The proposed method significantly outperforms traditional methods in both classification and localization tasks, and can provide important support for intelligent monitoring of forest fires.

2. Method

2.1. Problem Definition

In this study, we propose a Transformer-based multimodal data fusion method to improve the accuracy of forest fire detection. By fusing the multimodal features of visual images (RGB images) with infrared images, we can fully use the complementary information of the two modalities and enhance the model's ability to identify and locate fires. Specifically, given a pair of visual images I_v and infrared images I_{ir}, our goal is to design a model F that can process and fuse the features of these two modalities to output the fire's classification results and the fire area's location information.

$$F: (I_v, I_{ir}) \rightarrow (Y_c, Y_l) \tag{1}$$

Among them, Y_c represents the classification results of fire and non-fire, and Y_l represents the spatial coordinates of the fire area.

2.2. Transformer Architecture Design

The core goal of this paper is to achieve fire classification and accurate positioning of fire areas by multimodal data fusion based on the Transformer architecture. The overall architecture is shown in Figure 1.

2.2.1. Input Encoding

We design a two-stream input encoding to convert the original visual image (RGB image) and infrared image (IR image) into feature representations suitable for Transformer model processing. In this study, we use a combination of patch embedding and position encoding to convert image data into a vector sequence for further input into Transformer for processing. For the input visual image $I_v \in R^{H_v \times W_v \times C}$, where H_v and W_v are the height and width of the image, respectively, and C is the number of channels. We first divide the image into several non-overlapping image patches of size $P_v \times P_v$, each of which has a size of $P_v \times P_v \times C$, and flatten each image patch into a one-dimensional vector, each image block x_i^v can be expressed as:

$$x_i^v = Flatten(I_v[i]) \in R^{P_v^2 C}, i = 1,2, \dots, N_v \tag{2}$$

DOI: 10.54254/2755-2721/120/2025.19482

Next, in order to preserve the position information of the image block in the original image, we add a learnable position encoding P_i^v, where d is the dimension of the embedding space. The final embedding representation of each image block is:

$$z_i^v = x_i^v + p_i^v, i = 1,2,\dots,N_v \tag{3}$$

The same operation is also applicable to infrared images $I_{ir} \in R^{H_{ir} \times W_{ir} \times C_{ir}}$, by dividing the infrared image into image blocks of size $P_{ir} \times P_{ir} \times C_{ir}$ and flattening each image block, the final embedding representation of each infrared image block is: $z_i^{ir} = x_i^{ir} + p_i^{ir}, i = 1,2,\dots,N_{ir}$. In this way, we successfully transformed the image data of the two modalities into feature sequences with spatial information, providing a basis for multimodal fusion.

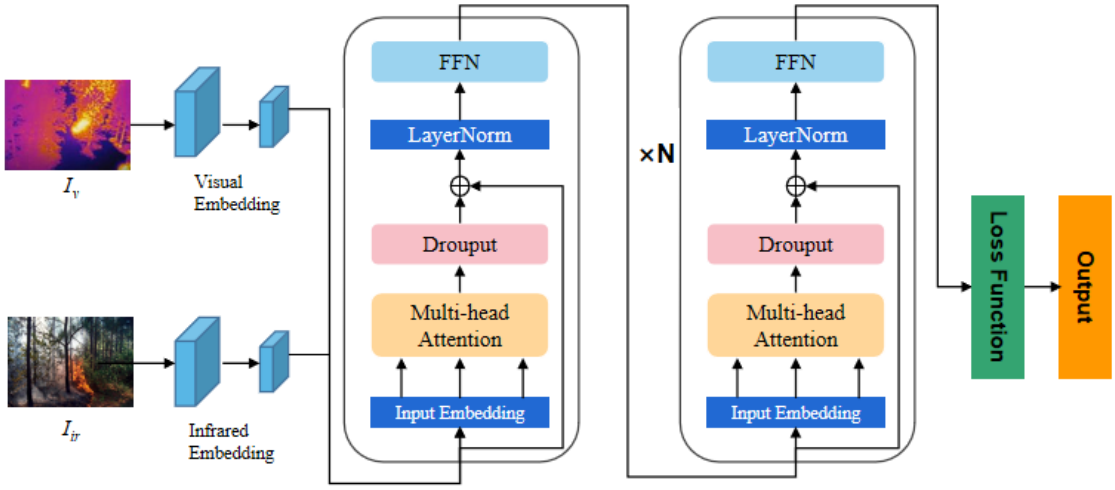

Figure 1: Fire detection network based on multimodal data fusion of Transformer

2.2.2. Visual Embedding and Multimodal Fusion

In order to make full use of the complementarity of these two modalities, we adopt the Transformer architecture for feature extraction and multimodal fusion. First, we input the visual image and the infrared image for encoding. The specific processing method has been described in detail in Section 2.2.1. Through the patch embedding technology, the two images are converted into their respective vector sequences Z_v and Z_{ir}. These vector sequences will be input into the feature extraction modules of visual images and infrared images respectively. For visual images, the following self-attention calculation formula is used to extract features from the input sequence:

$$\text{Attention}(Q_v, K_v, V_v) = \text{softmax}\left(\frac{Q_v K_v^{\cdot}}{\sqrt{d_k}}\right) V_v \tag{4}$$

Among them, $(Q_v, K_v, V_v)$are the query, key and value matrices obtained by linearly transforming the visual input Z_v, and d_k is the dimension of each attention head. In this way, the model can learn the dependencies between regions in the visual image and extract global information. Similarly, the infrared image Z_{ir} is extracted through a similar self-attention mechanism to obtain infrared feature representation.

In order to effectively fuse the features of visual images and infrared images, we adopt a cross-modal attention mechanism[8]. Different from the single-modal self-attention, the cross-modal attention mechanism uses the features of the visual image as the query (Query), and the features of

DOI: 10.54254/2755-2721/120/2025.19482

the infrared image as the key (Key) and value (Value), and performs cross-modal feature fusion through the attention mechanism.

$$\text{CrossAttention}(Q_v, K_{ir}, V_{ir}) = \text{softmax}\left(\frac{Q_v K_{ir}^{\cdot}}{\sqrt{d_k}}\right) V_{ir} \tag{5}$$

Among them, Q_v is the visual image feature, K_{ir} and V_{ir} are the key and value of the infrared image respectively. In this way, the model can associate the information in the visual image with the heat source information in the infrared image and capture the complementary features between the two. This process can effectively improve the accuracy of fire detection, especially in complex scenes, where infrared images can help the model identify the fire source, while visual images provide more scene context information.

$$Z_f = \text{Concat}(\text{CrossAttention}(Q_v, K_{ir}, V_{ir}), Z_v) \tag{6}$$

Among them, Z_f is the fused feature representation, which contains the comprehensive information from the visual image and the infrared image. Next, these features will be input into the subsequent Transformer module to continue multi-layer feature extraction and enhancement.

2.2.3. Multi-head Attention Mechanism

In Transformer, the multi-head attention mechanism is used to learn multiple attention patterns in different subspaces at the same time, thereby enhancing the expressiveness of the model. In the multi-head attention mechanism, multiple independent attention heads work in parallel, and each head learns different feature patterns through different transformations

$$\text{head}_i = \text{Attention}\left(QW_q^i, KW_k^i, VW_v^i\right) \tag{7}$$

Among them, W_q^i, W_k^i, W_v^i is the learnable weight matrix of the i th head, and Q, K, V are the input query, key and value respectively. Finally, the outputs of multiple attention heads are concatenated and the final output is obtained through linear transformation:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W_o \tag{8}$$

In order to enhance the learning ability and stability of the model, the Transformer model also uses residual connections and layer normalization.

$$Z' = \text{LayerNorm}\left(Z + \text{MultiHead}(Q, K, V)\right) \tag{9}$$

In addition, the feed-forward network (FFN) in the Transformer model further enhances the representation ability of features through non-linear activation functions such as ReLU.

$$Z'' = \text{LayerNorm}\left(Z' + \text{FFN}(Z')\right) \tag{10}$$

To improve the expressiveness of the model, multiple Transformer blocks are stacked into a deep network structure. In this study, we used a 6-layer Transformer stack. The model improves the ability to recognize fire information in visual images and infrared images by continuously adjusting the feature representation of each layer.

2.2.4. Loss Function

To train the model, we design a loss function that combines classification and localization. The total loss consists of two parts: classification loss and localization loss:

$$\text{L} = \text{L}_{cls} + \lambda \text{L}_{loc} \tag{11}$$

DOI: 10.54254/2755-2721/120/2025.19482

Among them, L_{cls} is the classification loss, using cross entropy loss, and L_{loc} is the positioning loss, using L1 loss:

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log P_c^i \tag{12}$$

$$L_{loc} = \frac{1}{N}\sum_{i=1}^{N} \left|Y_l^i - \hat{Y}_l^i\right| \tag{13}$$

By integrating classification and location losses, the model can simultaneously optimize the classification accuracy and location accuracy of fires.

3. Experimental analysis

3.1. Dataset and Experiment settings

The FLAME (Forest Fire Localization and Assessment through Multi-sensor Environmental Data) dataset is an important public dataset specifically used for forest fire detection and location research. The dataset integrates multimodal sensor data, including visible light images, thermal infrared images, and other auxiliary information (such as ambient temperature, humidity, and wind speed).

In the model setting, the number of Transformer model layers is 6, the number of attention heads is set to 8, the embedding dimension is 512. The Adam optimizer is used as the optimization method, the initial learning rate is 10-4, and the batch size is 64. In order to comprehensively evaluate the performance of the model, the experimental evaluation indicators used accuracy, precision, recall, and F1 score. IOU was used to measure the overlap between the predicted bounding box and the true bounding box.

3.2. Comparative Analysis

In order to comprehensively evaluate the performance of the proposed method, we conducted extensive comparative experiments, as is shown in Table 1. Single-modal CNN (image) and single-modal CNN (infrared) models have limited ability to identify fires, with accuracy rates of 84.2% and 85.6% respectively. This shows that data from a single modality are insufficient to comprehensively capture the characteristic information of fires, especially for complex scenes (such as background light interference or high-temperature non-fire areas). The multi-modal CNN fusion model achieved great performance improvement through simple feature splicing, with the accuracy increased to 88.7%, indicating that the complementarity of multi-modal data to fire features significantly enhances the model's classification ability. This method achieves a classification accuracy of 94.5% through cross-modal and multi-head attention mechanisms, which is better than the multi-modal CNN fusion method and the traditional Transformer model. This result shows that the proposed method can dig deeper into the correlation between visible light and infrared features and improve the ability to identify fire areas. Precision and Recall exceed 92%, indicating that the model’s prediction of fire areas is more accurate, while the missed and false detection rates are significantly reduced.

3.3. Ablation Analysis

Table 2 shows the results of the ablation experiment analysis. After removing the cross-modal attention mechanism, the classification accuracy dropped by 3.5% and the IoU dropped by 5%. This shows that cross-modal attention can significantly enhance the model's ability to extract fire features. The performance after removing the residual connection shows that the residual connection improves

DOI: 10.54254/2755-2721/120/2025.19482

the gradient stability and feature expression ability of the model, and contributes to efficient training and rapid convergence of the model. Single-modal input leads to a significant decrease in classification accuracy and IoU, validating the importance of multi-modal data for the fire detection task.

Table 1: Performance of different methods in fire detection

Method	Accuracy(%)	Precision(%)	Recall(%)	F1
Single-modal(image)	84.2	81.5	80.3	80.9
Single-modal(infrared)	85.6	84.1	82.8	83.4
Multi-modal CNN Fusion	88.7	87.3	86.9	87.1
Multi-modal Transformer(without cross-modal attention)	91.2	89.8	89.3	89.5
Ours	94.5	92.7	93.4	93.0

Table 2: Performance of ablation experiment

Components	Accuracy	IoU
Full model	94.5	0.86
w/o cross-modal attention	91.0	0.81
w/o residual conect	92.0	0.82
Single modal	84.2	0.72

4. Conclusion

In this paper, a Transformer fire detection method based on multimodal data fusion is proposed to address the key challenges in the field of forest fire monitoring. By combining visible light images and infrared images and leveraging the powerful feature modeling capabilities of Transformer, this method successfully solves the problem of limited performance of single modality data in complex scenarios. The cross-modal attention mechanism achieves a deep fusion of the two modal features, significantly improving the accuracy of fire feature extraction.

References

[1] *Flannigan, M. D., Stocks, B. J., & Wotton, B. M. (2000). Climate change and forest fires. Science of the total environment, 262(3), 221-229.*

[2] *Flannigan, M. D., Amiro, B. D., Logan, K. A., Stocks, B. J., & Wotton, B. M. (2006). Forest fires and climate change in the 21 st century. Mitigation and adaptation strategies for global change, 11, 847-859.*

[3] *Barmpoutis, P., Papaioannou, P., Dimitropoulos, K., & Grammalidis, N. (2020). A review on early forest fire detection systems using optical remote sensing. Sensors, 20(22), 6442.*

[4] *Cao, L., Shen, Z., & Xu, S. (2024). Efficient forest fire detection based on an improved YOLO model. Visual Intelligence, 2(1), 20.*

[5] *Lin, J., Lin, H., & Wang, F. (2023). A semi-supervised method for real-time forest fire detection algorithm based on adaptively spatial feature fusion. Forests, 14(2), 361.*

[6] *Xiao, Z., Wan, F., Lei, G., Xiong, Y., Xu, L., Ye, Z., ... & Xu, C. (2023). FL-YOLOv7: a lightweight small object detection algorithm in forest fire detection. Forests, 14(9), 1812.*

[7] *Shahid, M., Chen, S. F., Hsu, Y. L., Chen, Y. Y., Chen, Y. L., & Hua, K. L. (2023). Forest fire segmentation via temporal transformer from aerial images. Forests, 14(3), 563.*

[8] *Lee, S., Park, J., & Park, J. (2024). CrossFormer: Cross-guided attention for multi-modal object detection. Pattern Recognition Letters, 179, 144-150.*