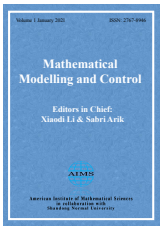
Mathematical Modelling and Control

https://www.aimspress.com/journal/mmc



MMC, 5(4): 355–367.
DOI: 10.3934/mmc.2025024
Received: 10 April 2024
Revised: 19 May 2024
Accepted: 24 June 2024
Published: 22 October 2025

Research article

Multi-modal fusion traffic information classification algorithm based on distributed optical fiber sensing

Chun Shan*, Dongping Liu, Yewen Huang, Xiaoyan Huang, Tongyi Zou, Jiayi Li and Shaoming Liu

School of Electronics and Information, Guangdong Polytechnic Normal University, Guangzhou 510665, China

* **Correspondence:** Email: shanchun@gpnu.edu.cn.

Abstract: The realization of intelligent transportation is inseparable from the perception and processing of traffic information. In the field of intelligent transportation, video surveillance, infrared, magnetic induction sensors, other methods have been applied, to a certain extent to meet the needs. However, the above method has some limitations, such as complex installation, high cost, blind spot detection, weather influence. Therefore, this paper proposes a traffic information classification algorithm based on distributed optical fiber sensing, which uses the distributed perception ability of optical fiber to obtain traffic vibration data, combined with video data, and realizes high-precision classification of vibration signals through neural network. This method uses multi-modal fusion technology to extract key features of fiber distributed sensing data and video data respectively through ResNet and video understanding network, and then extracts multi-scale features corresponding to the two modes from different levels of the feature extraction network. Finally, multi-modal cross-attention feature enhancement fusion module is used to achieve multi-modal feature fusion. The information complementation between different modes is realized effectively, and the classification of traffic information is completed. In this study, we conducted tests on the traffic data set we collected, and the framework showed excellent performance in various types of identification, which can provide a reference for the construction of intelligent transportation.

Keywords: intelligent transportation; multimodal fusion; distributed optical fiber sensing; neural network

1. Introduction

As economy rapidly expands, the number of automobiles and the overall length of roads are constantly growing, leading to an increase in road traffic flow. Therefore, it is very important to monitor and manage traffic information. Relying on the construction of basic traffic facilities, it is difficult for traffic management departments to implement comprehensive, systematic, real-time supervision and maintenance of road traffic. Consequently, the utilization of artificial intelligence algorithms for the recognition of traffic target information has become a prevalent research focus in the realm of intelligent transportation.

At present, in the road traffic scene, the target recognition mainly relies on video surveillance, infrared, magnetic induction sensors, other means, different types of sensors and recognition technology have different working principles, and have their own different recognition effects. The video image recognition technology is flexible in installation, has a high recognition rate, and can obtain a variety of traffic information, but its real-time performance is poor, it is more sensitive to weather conditions, and it cannot work normally in the environment such as haze and heavy rain. Moreover, for long-distance traffic monitoring, a large amount of data will be generated, which puts forward high requirements for network transmission performance and device cache performance. Infrared and laser detection methods are easy to cause damage to the human body, but also more sensitive to extreme weather. Magnetic induction sensor detection technology has high precision, portability and easy placement, but the price is high, and it will cost a considerable amount to form a distributed sensing system with its help [1]. Not only that, many detection devices will produce a large number of

multi-modal data, which is easy to violate people's data privacy [2–4], and they also face a host of security issues during the deployment phase [5].Therefore, how to achieve high accuracy classification perception of traffic information under the premise of cost control has important research and application value.

To address this issue, this paper uses distributed acoustic sensing (DAS) to collect data. The system can directly use the existing road side communication fiber, eliminating the installation cost. In addition, DAS system can realize distributed measurement. When vehicles, pedestrians and other targets move near the optical fiber, continuous distributed information in time and space of the targets in the measured area can be established in real time [6–8], the amount of data is relatively small, and the detection blind spots can be reduced [9, 10]. Because the DAS system collects vibration signals, it is not easy to violate people's privacy.

On this basis, this paper proposes a distributed fiber optic traffic information classification sensing algorithm based on multi-modal fusion. The algorithm mainly includes DAS data processing module, video data processing module and multi-modal feature fusion module. The DAS data processing module then extracts features from the DAS data using a fine-tuned ResNet50 network that was pre-trained on the ImageNet dataset. The video data processing module extracts video data features through an efficient video understanding network. The multi-modal feature fusion module can compress, enhance and fuse the feature information of the two modes, realize the complementary information of the two modes, and complete the final classification and recognition. Through the fusion of the two modes, the model can comprehensively consider the effective information of the two kinds of data, reduce the influence of interference information, and finally improve the detection accuracy. In this paper, the algorithm is evaluated on a dataset we constructed and the results demonstrate that it performs favorably when compared with different input combinations, showcasing its effectiveness in classifying various categories.

2. Related works

Both theoretical research on distributed optical fiber sensing technology and applied research on vibration signal recognition are undergoing rapid development.

2.1. Distributed optical fiber sensing technology theory

In 1976, the optical time domain reflectometer (OTDR) was successfully developed and rapidly emerged as a standard tool for fault diagnosis and localization in optical fiber links [11]. However, In long-distance optical communication systems, cascaded repeaters are often used to compensate for the signal light's transmission loss, so the accompanying spontaneous radiation noise is amplified and accumulated step by step. The wide spectrum light source and direct detection method used by OTDR are difficult to suppress the wide spectrum noise by optical filtering method, and its performance will deteriorate rapidly in the cascaded fiber link. Therefore, P. Healey et al. employed the self-heterodyne detection approach and developed the coherent optical time domain reflectometer (COTDR) [12]. Later, researchers further enhanced the system's anti-noise performance by incorporating narrow-width lasers with high coherence and heterodyne coherent detection optical path structures [13–15], as well as various methods [16–18]. However, due to the use of narrow linewidth light source, the loss curve obtained by COTDR will produce random fluctuations due to the interference of coherent fading noise, and it is difficult to obtain the loss measurement accuracy comparable to that of OTDR [19–21]. Therefore, by studying the statistical characteristics of coherent fading noise [22–24], researchers propose methods such as frequency shift average noise reduction, which effectively realizes the suppression of fading noise [25–27]. Later, the concept of the phase-sensitive optical time domain reflectometer (ϕ-OTDR) was introduced, which further reduced the noise power and increased the frequency response range [28–30].

2.2. Vibration signal recognition applications

Vibration signal recognition technology is divided into active detection and passive detection, and distributed

optical fiber sensing technology belongs to passive detection. Compared with active detection, passive detection has the advantages of long effective distance, small size, low power consumption, strong anti-interference ability, good concealment, outstanding anti-reconnaissance ability, etc., and a large number of researches have been carried out at home and abroad [31]. In 2012, OptaSense established nearly 40 km of pilot projects on highways to evaluate the ability of DAS systems to sense ground vibration signals generated by targets such as vehicles [32]. In 2020, the experimental results of NEC Company showed that the accuracy of the DAS system to detect vehicle speed reached 90%. In the same year, Huiyong Liu et al. used DAS technology to accurately locate and identify vibration sources such as vehicles passing through optical fibers [33]. The related research in this field has developed rapidly, and has achieved many excellent results, and has a wide application prospect in the future.

3. Data acquisition

The DAS equipment used for data collection in this paper was developed by Guangzhou Agizhi Photosensitive Electronics Technology Co., LTD. The model is DAS 1550-40-FA. The device provides real-time acquisition, analysis and display of vibration information every 0.4 m across tens of kilometers of fiber deployed along the infrastructure. DAS 1550-40-FA can also be deployed directly on the laid cable, only the cable must be single-mode fiber. The main technical parameters are shown in Table 1.

Table 1. The main technical parameters of DAS equipment.

Maximum distance	Resolve space	Sample interval	Maximum frequency	Minimum frequency
40 km	20 m	0.4 m	50 KHz	0.1 Hz

The equipment has the characteristics of low coherent fading noise and high detection sensitivity, which is suitable for harsh detection environment and single-end measurement. It can be used to monitor external event disturbances and record data, such as vehicle detection, vessel tracking, geological analysis, seismic surveys, intrusion detection, and more. The device is shown in Figure 1.

The traffic vibration signals collected in this paper are divided into five categories: vehicle, pedestrian, electric bicycle, trolley, background noise. The analysis of these signals can help monitor vehicle traffic flow, pedestrian flow, and road use. Based on the distributed fiber sensing principle above and the introduction of DAS equipment, the vehicle vibration signal acquisition scheme adopted is shown in Figure 2.

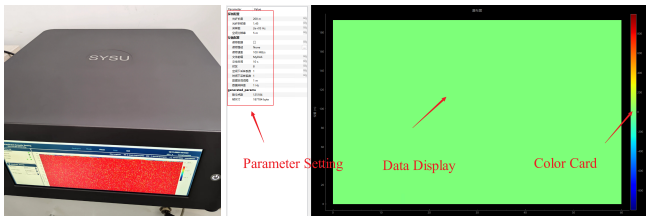


Figure 1. Physical picture of DAS device. On the left is the DAS device diagram, and on the right is the data acquisition interface diagram.

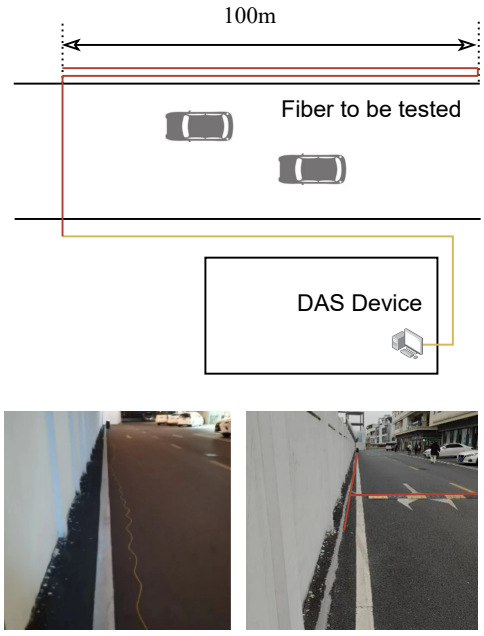


Figure 2. Data acquisition scheme of Traffic vibration signal. On the left is a diagram of the data acquisition scheme, and on the right is the actual fiber laying circuit diagram.

When collecting data, the DAS device is first installed in the laboratory and provided with uninterrupted power supply

from the indoor power supply. Then, a section of single-mode optical fiber is fixed on the edge of the road with cement and connected to the DAS equipment in the nearby laboratory across the road, ensuring good coupling of the optical fiber and the ground.The length of the optical fiber to be tested is 200 m. Due to the limitation of the experimental site and the need to enhance the signal, the 200 m optical fiber is laid on the edge of the road once and again.

According to Nyquist sampling theorem, the frequency of traffic information is concentrated in the middle and low frequency region, so the sampling frequency of DAS device is set to 2000 Hz, and the interval between sampling points is set to 2 m [34]. After careful comparison and analysis, we find that the optical fiber area to be measured is about 200 m to 400 m of the total length of the optical fiber, so the spatial axis index of HDF5 files in the optical fiber area to be measured is about 100 to 200. The result of a HDF5 file‘s waterfall visualization is shown in Figure 3.

In this paper, when collecting DAS data, video is taken at the same time for data processing. By aligning the timeline of DAS data and video data, the target vibration event signals in DAS data are accurately located and screened. The shooting video data is shown in Figure 3.

Through the above method, we finally get a total of 3888 DAS data events, each of which corresponds to the occurrence of a vibration event. The data included a total of 866 vehicle vibration events, 954 electric bicycle vibration events, 505 trolley vibration events, 946 pedestrian vibration events, and 617 background events, as shown in Table 2.

Figure 3. Data acquisition scheme of Traffic vibration signal. On the left is a diagram of the data acquisition scheme, and on the right is the actual fiber laying circuit diagram.

Table 2. Statistical diagram of vibration event categories. A total of 3888 vibration event data were collected.

Layout	Cement burial				
Event	Car	Bicycle	Handcart	Step	Background
Number	866	954	505	946	617

In the video data, when the vehicle passes and other events occur, on the basis of one-to-one alignment with the DAS data, the video footage of five seconds before and after the vehicle passes is captured from the original video as a training data, also named according to the category label. The video training data came from the filmed video, and each event corresponded equally to a five-second video training data and a DAS training data, so the total number of video clips was also 3888. Finally, DAS data and video footage data are integrated into the data set used in this paper.

4. Method

In the present study, we propose a novel multimodal fusion framework that can efficiently integrate two different types of data sources from DAS and video. As shown in Figure 4, with this fusion, we can capture the characteristics of traffic vibration events from different angles, enabling a more accurate classification of vehicles or other vibration sources. In this framework, we use the distributed sensing data collected by DAS devices and the synchronized surveillance video as the training basis to solve the problem of insufficient accuracy of single mode recognition. In addition, this chapter will also explore how to effectively integrate the information of these two modes, and the specific contribution of modules, such as multimodal fusion to improve classification performance.

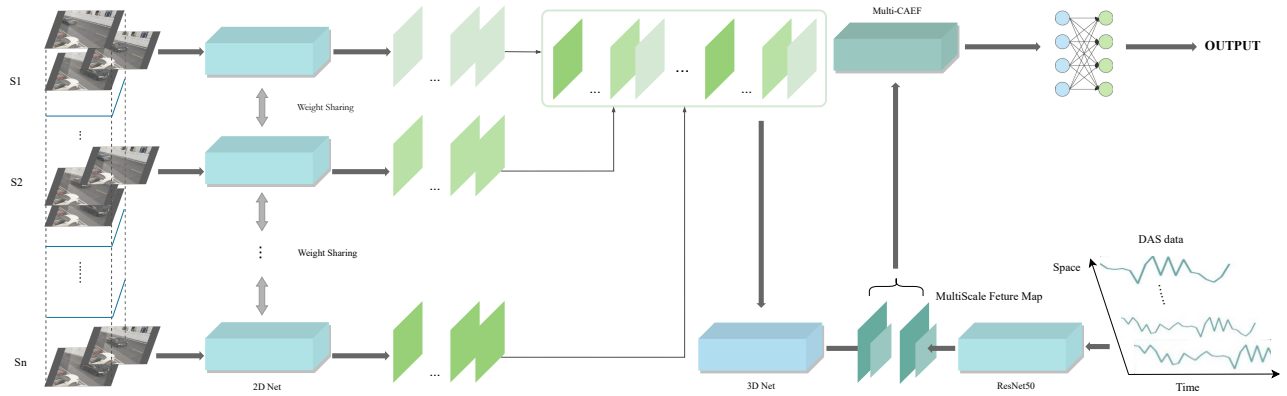


Figure 4. The overall flowchart of the framework. It consists of three main parts: DAS data processing module, video data processing module and Multimodal cross attention Feature enhancement fusion module. In the DAS data processing module, in the DAS data processing module, we use the ResNet50 to extract the multi-scale feature information of the DAS data. Similarly, in the video data processing module, we use an efficient video understanding network to extract the multi-scale feature information of the video data; finally, in the multi-modal feature fusion module, effective fusion of multi-scale features is realized, and classification results are obtained through the classifier.

4.1. DAS data processing module

DAS data is stored as a two-dimensional array of time-space, where each element represents the vibration intensity at a specific point in time and location. Therefore, DAS data can be processed through 2D convolutional networks to achieve feature extraction. In this study, ResNet50 [35] pre-trained on ImageNet1K [36] was used as the backbone network for processing DAS data. Before DAS data is processed, it must be preprocessed first to improve the effectiveness of network learning. The pre-processing steps include wavelet threshold denoising to remove some of the interference noise. Specifically, each channel data of DAS data is taken as a one-dimensional data signal, and after wavelet transformation, the effective signal generates larger wavelet coefficients than noise, so we can choose an appropriate threshold based on the wavelet coefficients. The wavelet coefficients that are greater than the threshold are considered to be generated by the effective signal, while the wavelet coefficients that are lower than the threshold are considered to be noise and are set to 0 to achieve the purpose of denoising.

Loading the pre-trained ResNet50 model is a key step in the DAS data processing process. ResNet50 is a kind of deep residual network, its core principle is to solve the degradation problem in deep network by introducing residual learning. In ResNet, the input is not only passed through the weighted layer, but also directly to the subsequent layer through the jump connection, which permits the network to acquire the residual function connecting the input and output, efficiently educating the deep network.

ResNet is comprised of five distinct stages, with Stage 0 exhibiting a straightforward design that can be viewed as a preparatory Bottleneck for input processing. The succeeding four stages, meanwhile, are constructed using bottleneck structures that share similarities. In Stage0, enter a DAS two-dimensional array of data shapes (C, W, H). Then the input data goes through three successive operations. First, a convolutional layer of 77 convolutional nuclei and 64 channels with step size of 2 is passed. Subsequently, the output of the convolutional layer undergoes normalization through a batch normalization layer. Lastly, the ReLU activation function is employed on the normalized output, and a MaxPooling operation is implemented to further downscale the feature map's dimension, thereby decreasing the computational load.

In Stage1, the main change compared to Stage0 is that

two bottlenecks are used. Bottleneck1 is used for simplified cases where the quantity of input channels is equivalent to the number of output channels and does not require an additional 11 convolution layer for dimensional matching. Let the input of shape (C, W, H) be x, the output after 3 convolutional blocks and associated BN and RELU be $F(x)$, add the two together to $F(x) + x$, and then pass through a ReLU activation function to obtain Bottleneck1 output, which has the shape of (C, W, H). The number of channels in the input x matches that of the output $F(x)$. The formula can be expressed as:

$$O_1 = x + \text{Conv}_{1\times1}(\text{Conv}_{3\times3}(\text{Conv}_{1\times1}(x, C))). \tag{4.1}$$

Bottleneck2 is employed in scenarios where the number of input channels differs from the number of output channels. The main difference between Bottleneck2 and Bottleneck1 is that the former includes an additional 11 convolution layer on the residual connection to adjust for inconsistencies in the number of input and output channels so that they can perform element summation operations. Let the input of shape (C, W, H) be x, the output after 3 convolutional blocks and related BN and RELU be $F(x)$, the output after connecting a convolutional block with residuals to adjust the number of channels be $G(x)$, and the sum of the two is $F(x) + G(x)$. The formula can be expressed as:

$$O_2 = \text{Conv}_{1\times1}(x, C) + \text{Conv}_{1\times1}(\text{Conv}_{3\times3}(\text{Conv}_{1\times1}(x, C'))). \tag{4.2}$$

The remaining stages of ResNet50 are similar in structure to Stage1 and will not be detailed here. After passing through the ResNet50 backbone network, the output after these 4 stages are respectively taken as multi-scale features for fusion with video features.

4.2. Video processing module

The video data processing module uses the efficient convolutional network for online video understanding (ECO) [37] to achieve feature extraction of video clips. Due to the large amount of information redundancy in adjacent frames of video, ECO uses a pre-trained 2D convolutional network to process only one frame of video at regular intervals. In order to enable the convolutional network itself to learn long-term semantic information, 3D convolutional layer is employed to extract features from the collection of frames captured at a specific time interval. In the specific training of the model, the video is segmented into N equal-length sub-sequences. In the video data processing module, each video training data sample is divided into N pieces during training. In this paper, our video frame rate is 60 frames/sec, so the total frame number of a single video data sample is 300 frames. In the experiment, we set N to 10, which can achieve a good balance between experimental accuracy and computational complexity. Each time a video is trained, a frame of video is randomly sampled in each time period and input into the convolutional network. In terms of long training sessions, every frame in the video is fully utilized. In terms of single forward and backpropagation, the model only needs to process N frames for a video (N is fixed), so the model runs very fast.

To be precise, consider the input video V represented as a four-dimensional tensor $\in \mathbb{R}^{T\times H\times W\times C}$, where T represents the temporal dimension, encompassing the total number of frames, while H and W correspond to the vertical and horizontal dimensions of each frame, respectively. Additionally, C denotes the number of channels present in the video. To minimize computational complexity and redundancy, the model initially selects frames from the video at regular intervals, and selects frames with intervals of s to form the sampled video V_s. The sampled video can be expressed as $V_s \in \mathbb{R}^{T'\times H\times W\times C}$, where $T' = \lceil T/s \rceil$. For each sampling frame, a pre-trained 2D convolutional network is applied for feature extraction. For frame i, $I_i \in \mathbb{R}^{H\times W\times C}$, the 2D convolution operation can be expressed as:

$$F_{2D}^i = f_{2D}(I_i). \tag{4.3}$$

Where f_{2D} represents the operation of the 2D convolutional network, and F_{2D}^i is the feature representation of frame i.

The 2D feature F_{2D}^i of all sampled frames is stacked along the time dimension to form $F_{2D}^{stack} \in \mathbb{R}^{T'\times H'\times W'\times C'}$. Here, H, W and C represent the height, width, and channel count of the output from the 2D convolution layer, respectively. Subsequently, this stacked feature tensor is subjected to a 3D convolution layer for further processing.

$$F_{3D} = f_{3D}(F_{2D}^{stack}). \tag{4.4}$$

f_{3D} represents the operation of the 3D convolutional network, and f_{3D} is the feature representation after processing the 3D convolutional layer.

As shown in Figure 4, in the first half of the model, The pre-trained image classification model was employed to extract distinct features for each individual frame of the video. Then the features obtained from each frame are stacked, and the 3D convolutional network is used for cross-frame feature extraction in the second half. Finally, feature maps extracted from different network layers of 3D convolutional network are used as video multi-scale features for fusion with DAS data multi-scale feature maps.

4.3. Multimodal cross attention feature enhancement fusion

The above two sections extract the multi-scale feature maps of DAS data and video data respectively, and the multi-modal cross attention feature enhancement fusion method described in this section realizes the fusion of the two modal features. Figure 5 shows the network architecture of the multi-scale and multi-modal feature fusion module outlined in this section comprises three components: a feature compression module (FCM), cross-modal cross feature enhancement module (CFE) and multi-modal feature fusion module (MFF).

To minimize the parameter count and memory consumption of the model, the FCM module is applied before the CFE module to compress the size of the feature map. This module uses pooling operations to compress features. Specifically, the pooling operation uses average pooling and maximum pooling, which are frequently employed to decrease the spatial dimensions of the feature map without introducing additional parameters.

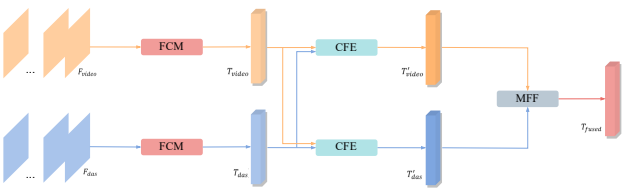


Figure 5. Multimodal cross attention feature enhancement fusion framework. It is divided into three parts: FCM, CFE, MFF.

Let F be the feature graph taken out in the first place, and its size is CWH. First, global average pooling and maximum pooling operations are carried out on it. The formula is expressed as follows:

$$\begin{aligned} F_a &= \text{AvgPooling}(F, S), \\ F_m &= \text{MaxPooling}(F, S). \end{aligned} \tag{4.5}$$

Then, a learnable momentum factor λ is used to aggregate the results obtained by the two methods of average pooling and maximum pooling. The formula is expressed as follows:

$$F_o = \lambda \cdot F_a + (1 - \lambda) \cdot F_m. \tag{4.6}$$

Where The input feature map is denoted as F, while S represents the pooled kernel's size. Respectively, F_a and F_m represent the feature mappings obtained after performing average pooling and maximum pooling and F_o is the feature after final compression. Finally, compared with the original feature, the dimension of the compressed feature map is reduced from CWH to $C(WH)/S$, thus reducing the dimensions of the query (Q), key (K), value (Y) in the CFE module, and reducing the overall computational complexity.

Instead of capturing the local features of different modes, CFE enables each single mode to learn more complementary information from the other mode. First, the input feature map is flattened into a set of sequences, and a learnable location embed is added, which is a trainable parameter that encodes spatial information between different sequences. The video and DAS features are then enhanced using two CFE modules, which do not share parameters. The CFE module uses a multi-head cross-attention mechanism [38] to enable the model to understand the correlation between the two modal features from different perspectives. Specifically, it works by projecting the sequence of one of the modes onto two independent matrices (values V and keys K) and projecting the sequence of the other mode onto the other matrix (query Q). After the correlation matrix established by dot product operation is regularized by softmax function, the resulting vector is projected back into the original space by nonlinear transformation and residual connection. The specific structure is shown in Figure 6. The whole process is expressed by the formula:

$$T'_{video} = F_1(\{T_{video}, T_{das}\}). \tag{4.7}$$

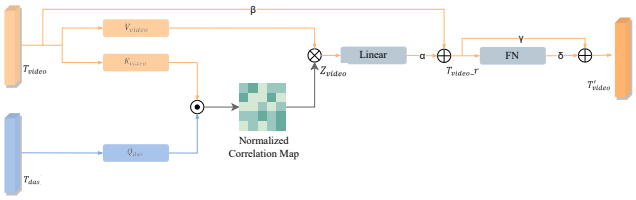


Figure 6. Cross-modal feature enhancement module. The multi-head attention mechanism is adopted.

T_{video} and T_{das} represent the feature sequence input into the CFE module, T'_{video} and T'_{das} represent the enhanced video feature and DAS data feature of the CFE module, respectively. F_1 represents a CFE module.

Specifically, for the video feature sequence enhanced CFE module, the video feature is first projected onto two independent matrices T_{video} and T_{das}, a set of values (V_{video}) and keys (K_{cideo}) are calculated, and then the DAS data feature is projected onto another independent matrix T_{das}. A set of queries (Q_{das}) is calculated, which can be expressed as follows:

$$\begin{aligned} V_{video} &= T_{video} W^V, \\ K_{video} &= T_{video} W^K, \\ Q_{das} &= T_{das} W^Q. \end{aligned} \tag{4.8}$$

Where W^V, W^K and W^Q are weight matrices.

Next, the correlation matrix is constructed through the dot product operation, and then the correlation function is normalized with $softmax$ function, which is used to represent the similarity between the two modal features, and Z_{video} is obtained, which is expressed as follows:

$$Z_{video} = softmax\left((Q_{das} K^T_{video})/{D_k}^{-1/2}\right) \cdot V_{video}. \tag{4.9}$$

Reproject Z_{video} back to the original space and add the residual connection:

$$T_{video_r} = \alpha \cdot Z_{video} W^o + \beta \cdot T_{video}. \tag{4.10}$$

Here, W^o denotes the weight matrix preceding the final two fully connected layers, with α and β being adjustable parameters. Subsequently, two fully connected layers are employed to further enhance the global information, ultimately yielding the improved video feature representation as output.The formula is expressed as follows:

$$T'_{video} = \gamma \cdot T_{video_r} + \delta \cdot FN(T_{video_r}). \tag{4.11}$$

Where, γ and δ are learnable parameters, and FN represents the fully connected layer.

Similarly, another CFE module is also used to enhance DAS data features, expressed by the formula:

$$T'_{das} = F_2(\{T_{das}, T_{video}\}). \tag{4.12}$$

Where F_2 represents another CFE module.

After the cross-attention-enhanced feature representations of video and DAS data are obtained, they are input into the MFF module to achieve the final feature fusion. Since the features of the two modes have been very finely cross-fused in the CFE module mentioned above, we adopt a direct additive fusion method in this module. The formula is expressed as follows:

$$T_{fused} = T'_{video} + T'_{das}. \tag{4.13}$$

In the above formula, T'_{video} and T'_{das} are the cross-attention enhanced feature representation of video and DAS data respectively, and T_{fused} is the multi-modal feature representation after fusion. After the fusion is complete, the fused feature T_{fused} can be categorized using the fully connected network and the $softmax$ classification layer.

5. Experiments

The experimental part of this chapter is carried out on the multimodal data set mentioned above, which is divided into two parts, namely video data part and DAS data part. The data set has five categories: Car, Bicycle, Handcart, Step and Background, which are divided according to the ratio of training set: verification set: test set 8:1:1. In the course of the experiment, the multi-CAEF module was respectively ablation experiments. Finally, the effects of different modal data combinations on the experimental results are tested respectively.

5.1. The influence of different scale feature fusion on the results

In this section, we verify the experimental comparison results of the fusion of multi-scale feature maps at different

levels of the backbone network by different number of multi-CAEF modules to explore their impact on model performance. When the number of multi-CAEF is equal to 0, it means that each mode feature is directly added and fused after being compressed by FCM module without being enhanced by CFE module. Based on the benchmark network, this section gradually increases the number of multi-CAEF modules based on feature graphs of different scales. The changes of the model in four key performance indicators, namely accuracy, precision, recall and F1 score [39], were observed, and the results were shown in Table 3.

Table 3. Comparison of experiments with different input modal combinations. The experimental results of different modal input combinations are shown.

Model	Multi-CAEF	Acc	P	R	F1
Baseline	0	94.3	95.1	94.9	94.9
Baseline	1	96.5	96.7	96.0	96.8
Baseline	2	97.1	97.5	97.4	97.1
Baseline	3	97.7	97.9	98.0	97.8
Baseline	4	97.9	98.2	98.5	98.3

The experimental results show that with the increase of the number of multi-CAEF modules, the performance of the model presents an obvious trend of improvement in general, especially the classification accuracy and reliability of the model can be significantly increased. This phenomenon is attributed to the the multi-modal feature fusion module can effectively integrate information from different modes, and through comprehensive consideration and fusion of multi-level and different size feature maps, the model can understand and represent the input data more comprehensively.

5.2. Comparison of input experimental results of different modal combinations

In this section, we explore the performance of a single modal input (video or DAS) and its combination (video unite video, DAS unite DAS, video unite DAS) in a multi-class classification task. The experiments were all completed on the basic architecture of four multi-CAEF modules. Detailed performance evaluation was carried out to make quantitative analysis, including accuracy, precision, recall and F1-score of various categories. The results are shown in Table 4.

Table 4. Comparison of experiments with different input modal combinations. The experimental results of different modal input combinations are shown.

Input	Class	Acc	P	R	F1
Video	Car	93.7	94.9	93.2	94.0
	Bicycle	93.5	94.5	93.0	93.7
	Handcart	93.8	94.8	93.3	94.0
	Step	93.4	94.6	93.2	93.8
	Background	93.6	94.6	93.0	93.7
	Average	93.6	94.6	93.1	93.8
DAS	Car	92.4	93.9	92.5	93.1
	Bicycle	92.3	93.3	91.8	92.5
	Handcart	92.6	94.7	91.9	93.2
	Step	92.4	94.8	92.9	93.8
	Background	92.0	93.1	90.4	91.7
	Average	92.3	93.9	91.9	92.8
Video+ Video	Car	94.5	96.1	96.6	96.3
	Bicycle	95.9	96.1	96.9	96.4
	Handcart	95.2	95.2	96.0	95.5
	Step	95.8	95.6	95.4	95.4
	Background	95.3	97.3	96.4	96.8
	Average	95.3	96.0	96.2	96.0
DAS+ DAS	Car	94.4	93.8	94.0	93.8
	Bicycle	93.1	94.1	92.2	93.1
	Handcart	94.4	93.2	94.2	93.6
	Step	93.6	93.9	93.6	93.7
	Background	92.1	93.3	91.7	92.4
	Average	93.5	93.6	93.1	93.3
Video+ DAS	Car	97.7	98.3	98.7	98.4
	Bicycle	98.6	97.8	98.4	98.0
	Handcart	98.8	98.9	99.1	98.9
	Step	96.9	98.2	98.5	98.5
	Background	97.8	97.9	97.9	97.9
	Average	97.9	98.2	98.5	98.3

When video and DAS modes were combined, there was a significant improvement in performance across all categories, with an average accuracy of 97.9 percent. This combination significantly improves classification accuracy and other performance indicators, especially in the recognition of bicycle and handcart categories, both accuracy and F1-score are very high, which indicates the great potential of multi-modal fusion in improving classification performance. Combining video and DAS modes allows the model to learn supplementary information from different data sources, making the classifier more robust in the face of diverse data.

From the combination of video unite video and DAS unite DAS, compared with single video or DAS, the average

performance of dual-channel data input is higher than that of each single channel input, which indicates that multi-CAEF module can play a certain role even if it is a single mode input. Multi-CAEF helps the model to notice the useful information that is difficult to utilize in the single mode input, which fully proves the effectiveness of multi-CAEF.

Taken together, the experimental results clearly demonstrate the advantages of multimodal fusion in complex classification tasks. With a well-designed modal fusion strategy, significant performance gains can be achieved, especially in situations where information needs to be extracted and integrated from different data sources. Future work could further explore the complementarity between different modes and how to optimize multi-modal fusion strategies to improve classification accuracy and robustness in complex environments.

5.3. Result visualization

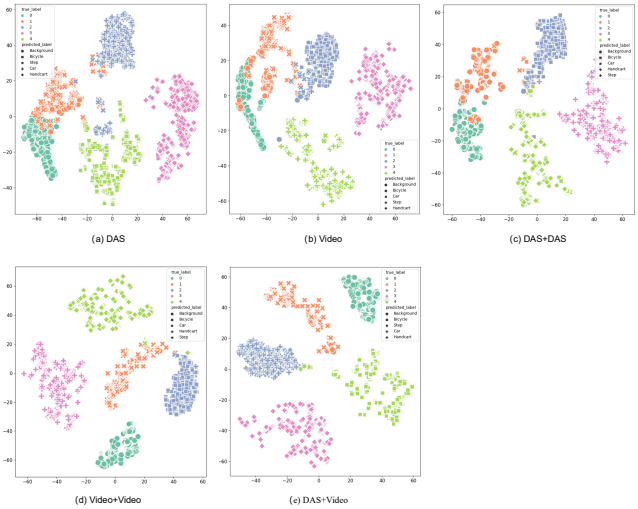


Figure 7. Visualization results of different modal combinations. The visual results of five different modal combinations of inputs are presented.

We conducted experiments on different modal combinations, and the comparison of experimental visual results is shown in Figure 7. The results show that compared with the other four modal combinations, the visualization results of DAS and video combination show clearer boundaries of five types of targets, less overlap between different categories in the feature space, and the data are gathered more closely, and the classification ability of the model has been significantly improved. This shows that the combination of DAS and data can make the model effectively capture the internal relationship between different modes, realize the information complementarity between modes, and use this internal relationship to improve the classification performance of the model.

6. Conclusions

This paper introduces an innovative traffic information perception classification algorithm framework based on multi-modal fusion, which combines video understanding network architecture, ResNet and efficient multi-modal fusion strategy. The experimental results show that the model framework proposed in this chapter significantly improves the classification accuracy of traffic multimodal data sets. In this study, sufficient experiments were conducted on this dataset, including ablation experiments on multi-scale feature fusion and comparison experiments on different mode combinations, to explore in detail the contribution of different strategies to the overall performance of the model. These experimental results not only validate the utility of a single strategy, but also show the importance of their combination to improve the model performance. Therefore, the research in this paper is of great significance, which not only puts forward a new method for traffic information perception and recognition based on distributed optical fiber sensing, but also provides an important reference value for future research in this field. In the future, we plan to further improve the experimental scene, establish a more comprehensive and diversified data set, and carry out traffic information perception and recognition research on this basis to further promote the development of intelligent transportation.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was funded by The National Natural Science Foundation of China under Grant (No.62273108, 62306081), The Youth Project of Guangdong Artificial Intelligence and Digital Economy Laboratory (Guangzhou) (PZL2022KF0006), The National Key Research and Development Program - Research on Key technology of High Frequency broadband mobile communication credit Filter and its Industrialization application - Subproject Circuit Design and Simulation of high frequency broadband Filter (2022YFB3604502), "New Generation Information Technology" Major Science and Technology Project of Guangzhou Key Field RD Plan (202206070001), Special Fund Project of Guangzhou Science and Technology Innovation Development (202201011307), Guangdong Provincial Department of Education Key construction discipline Scientific research ability Improvement Project, Introduction of Talents Project of Guangdong Polytechnic Normal University of China (99166990222), Special Projects in Key Fields of General Colleges and Universities in Guangdong Province (2021ZDZX1016), Natural Science Foundation of Guangdong Province (2024A1515010120), and Special Fund Project of Guangzhou Science and Technology Innovation Development (202201011307).

Conflict of interest

The authors declare that there are no conflicts of interest in this paper.

References

1. S. Taghvaeeyan, R. Rajamani, Portable roadside sensors for vehicle counting, classification, and speed measurement, *IEEE Trans. Intell. Transp. Syst.*, **15** (2014), 73–83. http://doi.org/10.1109/TITS.2013.2273876
2. M. Li, Y. Liu, Z. Tian, C. Shan, Privacy protection method based on multidimensional feature fusion under 6G networks, *IEEE Trans. Network Sci. Eng.*, **10** (2023), 1462–1471. http://doi.org/10.1109/tnse.2022.3186393
3. M. Li, C. Shan, Z. Tian, X. Du, M. Guizani, Adaptive information hiding method based on feature extraction for visible light communication, *IEEE Commun. Mag.*, **61** (2023), 102–106. http://doi.org/10.1109/mcom.001.2200035
4. S. Chun, X. Chen, G. Deng, H. Liu, A trusted NUMFabric algorithm for congestion price calculation at the internet-of-things datacenter, *Comp. Model. Eng.*, **126** (2021), 1203–1216. http://doi.org/10.32604/cmes.2021.012230
5. M. Li, C. Liu, C. Shan, H. Song, Z. Lv, A dual-embedded tamper detection framework based on block truncation coding for intelligent multimedia systems, *Inform. Sciences*, **649** (2023), 119362. http://doi.org/10.1016/j.ins.2023.119362
6. R. I. Crickmore, D. J. Hill, *Traffic sensing and monitoring apparatus*, Patent application number: 20080277568, United States Patent, 2007.
7. T. Parker, S. Shatalin, M. Farhadiroushan, Distributed Acoustic Sensing - a new tool for seismic applications, *Eur. Assoc. Geosci. Eng.*, **32** (2014), 61–69. http://doi.org/10.3997/1365-2397.2013034
8. J. Mestayer, B. Cox, P. Wills, D. Kiyashchenko, J. Lopez, M. Costello, et al., Field trials of distributed acoustic sensing for geophysical monitoring, *Soc. Explor. Geophys.*, 2011, 4253–4257. http://doi.org/10.1190/1.3628095
9. C. Shan, J. Cai, Y. Liu, J. Luo, Node importance to community based caching strategy forinformation centric networking, *Concurr. Comp. Pract. E.*, **31** (2019), e4797. http://doi.org/10.1002/cpe.4797
10. R. O. Duda, P. E. Hart, D. G. Stork, *Pattern classification*, Wiley Interscience, 2001.
11. M. K. Barnoski, S. M. Jensen, Fiber waveguides: a novel technique for investigating attenuation characteristics, *Appl. Optics*, **15** (1976), 2112–2115. http://doi.org/10.1364/AO.15.002112
12. M. Imahama, Y. Koyamada, Hogari, Restorability of Rayleigh backscatter traces measured by coherent OTDR with precisely frequency-controlled light source, *IEICE Trans. Commun.*, **E91.B** (2008), 1243–1246. http://doi.org/10.1093/ietcom/e91-b.4.1243

13. P. Healey, Fading in heterodyne OTDR, *Electron. Lett.*, **20** (1984), 30–32. http://doi.org/10.1049/el:19840022
14. P. Healey, Fading rates in coherent OTDR, *Electron. Lett.*, **20** (1984), 443–444. http://doi.org/10.1049/el:19840308
15. J. King, D. Smith, K. Richards, P. Timson, R. Epworth, S. Wright, Development of a coherent OTDR instrument, *J. Lightwave Technol.*, **5** (1987), 616–624. http://doi.org/10.1109/JLT.1987.1075523
16. L. C. Blank, D. M. Spirit, OTDR performance enhancement through erbium fibre amplification, *Electron. Lett.*, **25** (1989), 1693–1694. http://doi.org/10.1049/el:19891132
17. Y. Koyamada, H. Nakamoto, High performance single mode OTDR using coherent detection and fibre amplifiers, *Electron. Lett.*, **26** (1990), 573–575. http://doi.org/10.1049/el:19900375
18. H. Izumita, Y. Koyamada, S. Furukawa, I. Sankawa, The performance limit of coherent OTDR enhanced with optical fiber amplifiers due to optical nonlinear phenomena, *J. Lightwave Technol.*, **12** (1994), 1230–1238. http://doi.org/10.1109/50.301816
19. M. Sumida, OTDR performance enhancement using a quaternary FSK modulated probe and coherent detection, *IEEE Photonics Technol. Lett.*, **7** (1995), 336–338. http://doi.org/10.1109/68.372764
20. C. Shan, X. Wu, Y. Liu, J. Cai, J. Luo, IBP based caching strategy in D2D, *Appl. Sci.*, **9** (2019), 2416. http://doi.org/10.3390/app9122416
21. M. Sumida, Optical time domain reflectometry using an M-ary FSK probe and coherent detection, *J. Lightwave Technol.*, **14** (1996), 2483–2491. http://doi.org/10.1109/50.548145
22. H. Iida, Y. Koshikiya, F. Ito, K. Tanaka, High-sensitivity coherent optical time domain reflectometry employing frequency-division multiplexing, *J. Lightwave Technol.*, **30** (2012), 1121–1126. http://doi.org/10.1109/JLT.2011.2170960
23. K. Shimizu, T. Horiguchi, Y. Koyamada, Characteristics and reduction of coherent fading noise in Rayleigh backscattering measurement for optical fibers and components, *J. Lightwave Technol.*, **10** (1992), 982–987. http://doi.org/10.1109/50.144923
24. H. Izumita, Y. Koyamada, S. Furukawa, I. Sankawa, Stochastic amplitude fluctuation in coherent OTDR and a new technique for its reduction by stimulating synchronous optical frequency hopping, *J. Lightwave Technol.*, **15** (1997), 267–278. http://doi.org/10.1109/50.554377
25. L. Lu, Y. Song, X. Zhang, F. Zhu, Frequency division multiplexing OTDR with fast signal processing, *Opt. Laser Technol.*, **44** (2012), 2206–2209. https://doi.org/10.1016/j.optlastec.2012.02.037
26. X. Zhang, Y. Song, L. Lu, Time division multiplexing optical time domain reflectometry based on dual frequency probe, *IEEE Photonics Technol. Lett.*, **24** (2012), 2005–2008. http://doi.org/10.1109/LPT.2012.2217737
27. M. Chen, Y. Song, X. Zhang, Transient effect to small duty-cycle pulse in cascaded erbium-doped fiber amplifier system, *Opt. Eng.*, **52** (2013), 0325006. http://doi.org/10.1117/1.OE.52.2.025006
28. X. Zhang, X. Chen, L. Liang, S. Zhao, R. He, S. Tong, et al., Enhanced C-OTDRbased online monitoring scheme for long-distance submarine cables, *Acta Opt. Sin.*, **41** (2021), 1306001. http://doi.org/10.3788/AOS202141.1306001
29. H. F. Taylor, C. E. Lee, *Apparatus and method for fiber optic intrusion sensing*, Patent Number: US5194847, United States Patent, 1993. Available from: `http://patents.google.com/patent/US5194847A`.
30. Y. Rao, Z. Ran, K. Xie, *A method for improving the performance of distributed sensing systems using subcarrier technology*, Patent Number: CN101034035A, 2007. Available from: `http://patents.google.com/patent/CN101034035A/zh`.
31. L. Huang, X. Fan, Z. He, Hybrid distributed fiber-optic sensing system by using Rayleigh backscattering lightwave as probe of stimulated Brillouin scattering, *J. Lightwave Technol.*, **41** (2023), 4374–4380. http://doi.org/10.1109/JLT.2022.3213729

32. M. Atluri, M. Chowdhury, N. Kanhere, R. Fries, W. Sarasua, J. Ogle, Development of a sensor system for traffic data collection, *J. Adv. Transport.*, **43** (2009), 1–20. http://doi.org/10.1002/atr.5670430102
33. H. Liu, J. Ma, T. Xu, W. Yan, L. Ma, X. Zhang, Vehicle detection and classification using distributed fiber optic acoustic sensing, *IEEE Trans. Veh. Technol.*, **69** (2020), 1363–1374. http://doi.org/10.1109/TVT.2019.2962334
34. C. Shan, J. Zhou, D. Guo, H. Wang, L. Liu, Q. Wang, et al., Hartley-Domain DD-FTN Algorithm for ACO-SCFDM in Optical-Wireless Communications, *IEEE Photonics J.*, **11** (2019), 1–9. http://doi.org/10.1109/JPHOT.2019.2925007
35. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. http://doi.org/10.1109/CVPR.2016.90
36. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, F. F. Li, ImageNet: a large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, 248–255. http://doi.org/10.1109/CVPR.2009.5206848
37. M. Zolfaghari, K. Singh, T. Brox, ECO: efficient convolutional network for online video understanding, *arXiv*, 2018. http://doi.org/10.48550/arXiv.1804.09066
38. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, *arXiv*, 2017. http://doi.org/10.48550/arXiv.1706.03762
39. D. M. W. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, *Int. J. Mach. Learn. Technol.*, **2** (2011), 37–63.

AIMS Press