

Article

Audio and Video Scene Classification with Cross-Modal Attention Mechanism

Mingze Xia *, Guisheng Yin and Yuxin Dong

College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China

* Correspondence: xiamingze@hrbeu.edu.cn

Abstract

Scene classification aims to identify scene categories by analyzing environmental information. In real-world scenes, scene features are primarily captured through acoustic and visual modalities. However, environmental complexity and information diversity pose significant challenges to classification performance. To address the issues of insufficient information interaction and inadequate feature fusion in traditional segmented fusion methods for audio–visual data, this paper proposes an audio–visual scene classification method based on a cross-modal attention mechanism. The fusion mechanism of multi-modal features is investigated, and scene classification performance is enhanced by optimizing the feature fusion strategy. The proposed method consists of three components: a cross-modal attention module, a gating unit, and a residual connection. The cross-modal attention module achieves adaptive feature alignment by establishing dynamic correlations between audio and visual features. The multi-modal gating unit employs an adaptive gating mechanism to dynamically adjust the contribution weight of each modality, thereby alleviating the information loss problem commonly encountered in traditional methods. The residual connection module preserves the original modality features to prevent information degradation. The model's performance is evaluated through testing and validation on real-scene audio–visual datasets. Multiple sets of experimental results on these datasets demonstrate that the proposed cross-modal attention method achieves a significant improvement in classification accuracy.

Keywords: cross-modal; scene classification; multi-modal fusion; attention mechanism

1. Introduction

In real life, people's perception and understanding of the surrounding environment rely on the synergy of the auditory and visual perception systems, which provide us with rich and diverse information and serve as an important basis [1] for interacting with the environment and comprehending the things around us. Inspired by this, scene classification [2] can improve model performance by jointly leveraging auditory and visual information from audio and video data. In the field of data management and retrieval, audio–visual scene classification enables the orderly categorization of large-scale audio–visual data and enhances the efficiency with which users search for specific scene content. In the field of intelligent security, abnormal scenes in surveillance videos can be identified in real time through audio–visual scene classification [3] technology, and early warning mechanisms can be triggered promptly to build a strong technical defense [4] line for protecting life and property safety. Furthermore, for autonomous driving technology,



Academic Editor: Giuseppe Ciaburro

Received: 24 May 2026

Revised: 25 June 2026

Accepted: 8 July 2026

Published: 15 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

audio–visual scene recognition provides a reliable perceptual basis [5] for vehicle decision-making systems by analyzing sound and image information from the road environment in real time.

In the field of scene classification research, single-modal and multi-modal methods represent the two mainstream directions. At present, single-modal scene classification research has achieved significant progress in both the visual and auditory domains, laying the foundation for multi-modal studies. For example, in the acoustic modality, researchers have employed audio feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCCs), combined with deep learning models such as convolutional neural networks, to effectively analyze key information including the frequency, timbre, and rhythm of sound, thereby enabling the classification of different scenes. In the visual modality, researchers utilize image recognition algorithms and convolutional neural network techniques to accurately identify features such as object distribution, spatial layout, color, and texture in images, and subsequently accomplish the scene classification task.

However, the information provided by a single modality has inherent limitations in complex scene classification. Since a single modality can only reflect partial characteristics of a scene, the extracted features are one-sided and incomplete, making it difficult to accurately identify subtle differences [6] between scenes. Taking the acoustic modality as an example, the problem of background noise is particularly prominent in audio scene classification tasks. For instance, on bustling and noisy city streets, the sounds of vehicle engines and horns, crowd bustle, and various environmental noises superimpose on one another, forming complex acoustic reverberation that seriously interferes with the quality of audio signals. This interference reduces the signal-to-noise ratio of the audio signal, and key acoustic features become submerged by noise, which not only diminishes the classification accuracy of the model but also causes the model to misclassify across scenes in its classification decisions. The visual modality also has limitations. When lighting conditions are dim or objects are occluded, image feature extraction is adversely affected. For example, in nighttime surveillance images, insufficient illumination renders image details blurred, and it becomes difficult for the model to accurately extract key features for scene recognition from the limited and fuzzy visual information, resulting in the model's inability to recognize the scene correctly.

Therefore, in complex and changeable real-world scenarios, traditional single-modal scene classification can no longer meet practical demands, and there is an urgent need to develop scene classification technologies based on multi-modal data fusion. However, this technology faces multiple challenges. First, at the data level, audio is presented in the form of time series, with its characteristics mainly reflected in the frequency domain, time domain, and semantic level, whereas video data are organized as image sequences, with features including visual content, spatial layout, and temporal dynamic information. The essential differences between modalities give rise to problems in cross-modal feature alignment and integration. Second, model-level challenges include the difficulty of rationally designing fusion strategies. Existing approaches such as early fusion, late fusion, or hybrid fusion strategies require further exploration regarding how to select the optimal fusion strategy based on the characteristics of specific scenes. Finally, it is necessary to consider how to balance the contribution weights of different modalities during the fusion process, so as to avoid situations where a certain modality dominates model decisions while other modalities are ignored. To address the above issues, this paper systematically explores feature extraction, fusion strategies, and weight balancing mechanisms in multi-modal scene classification, with the aim of proposing a more effective scene classification method.

In this paper, we deeply explore the potential correlation between audio and video modalities and give full play to their complementary advantages. The main innovations are as follows:

1. The cross-modal attention mechanism is introduced to enable information exchange between audio and visual features prior to feature fusion.
2. The multi-modal gated fusion module employs an adaptive gating mechanism to dynamically adjust the contribution weight of each modality during feature fusion, thereby effectively alleviating the information loss problem commonly encountered in traditional fusion methods.
3. The residual connection module links the original modality features with the fused features, thereby not only preserving the unique characteristics of each modality but also preventing excessive loss of the original features during the fusion process.

The main contributions of this paper are threefold. First, we propose a unified multi-modal fusion framework that synergistically combines cross-modal attention, adaptive gating, and residual connections for audio–visual scene classification. While these components have been individually explored in other domains, their integration for urban scene understanding is both novel and non-trivial. Second, we provide a key design insight based on extensive experiments: adding a pre-fused feature stream does not necessarily improve performance, suggesting that direct cross-modal interaction between the original modalities is more effective. Third, we contribute a comprehensive empirical evaluation on the TAU Urban Audio-Visual Scenes 2021 dataset, including ablation studies on each component and sensitivity analyses of key hyperparameters, thereby establishing a strong baseline for future research.

When multiple modalities are fused at early or intermediate stages, the original modality-specific information can be diluted or lost. Our residual connection explicitly preserves the original modality features throughout the fusion process. Most existing approaches use fixed fusion strategies (e.g., concatenation or addition) that treat all modalities equally regardless of the input content. However, the importance of audio vs. visual information varies significantly across different scenes—for example, audio is more informative in a “subway station” (announcements, door sounds), while visual is more informative in a “park” (landscape, greenery). Our adaptive gating mechanism dynamically adjusts the contribution of each modality based on the input content.

2. Related Works

Acoustic scene classification (ASC) [7] mainly identifies different indoor and outdoor environments through acoustic signals. ASC has been widely applied in security monitoring, environmental perception, wearable devices, and smart homes. Influenced by the Detection and Classification of Acoustic Scenes and Events (DCASE) [8] challenges, audio scene classification technology has achieved breakthroughs in model accuracy, environmental adaptability, and computational efficiency. In the development of audio scene classification technology, feature extraction is a critical link. Currently, there are two main technical paths: one is the feature engineering approach based on signal processing, which constructs a robust acoustic feature representation by mining effective time-frequency features such as Mel-Frequency Cepstral Coefficients [9] and chroma features. The other is the automatic learning of features based on deep learning. At present, research exhibits a trend of multi-technology integration. By improving model robustness through data augmentation and feature transformation, optimizing computational efficiency using lightweight network architectures, and exploring cross-modal learning to enhance classification performance, these technologies are driving the advancement of audio scene classification technology.

Rambabu et al. [10] proposed the RSZero-CSAT framework, which enables interaction between visual and semantic features through a bidirectional Transformer. The SAVT branch uses semantic attributes to guide visual feature extraction, while the VSAT branch optimizes semantic representation through visual features. A feature extension encoder and a multi-level semantic collaboration loss are introduced to promote learning and collaboration within the model. Sun et al. [11] proposed a multi-modal aerial scene classification method to address the problem of sensor data heterogeneity. In this method, an adaptive zero-crossing rate module converts time-series sensor data into image representations to achieve modal alignment at the original data level, a dedicated alignment fusion module is designed to accomplish information fusion at the feature level, and a multi-modal alignment loss function is optimized for training. Experiments demonstrate that the average classification accuracy of the proposed method surpasses that of existing methods on multiple benchmark datasets.

In recent years, the fusion of visual and auditory signals has driven the rise [12] of multi-modal analysis. Traditionally, the field of auditory analysis has focused on extracting various types of information from audio signals, including frequency, intensity, timbre, sound source, and temporal variation trends. The visual field, on the other hand, concentrates on extracting rich spatial information from images and videos, such as the shape, color, and location of objects, as well as the overall layout of scenes. With the continuous advancement of deep learning technologies, recurrent neural networks (RNNs) [13] and their variants, such as Long Short-Term Memory (LSTM) [14] networks and Gated Recurrent Units (GRUs) [15], have been widely adopted in acoustic scene classification tasks due to their capability in processing temporal information in audio sequences. In the visual domain, convolutional neural networks (CNNs) can efficiently extract hierarchical features from images, progressively advancing from low-level features such as edges and textures to abstract representations of objects and scenes. Recently, Yin et al. [16] proposed a Multi-Modal Multi-Scale Temporal Enhancement Network that captures temporal dependencies at multiple scales for audio–visual scene classification. This work highlights the importance of temporal modeling in multi-modal fusion, complementing our focus on cross-modal feature interaction. In the domain of driving scene understanding, Cantarini et al. [17] introduced the A3CarScene dataset, which provides synchronized audio–visual recordings for autonomous driving scenarios. While their dataset is tailored to driving contexts, our work targets a broader range of urban scenes, encompassing both indoor and outdoor environments.

The attention mechanism has made significant contributions to the field of multi-modal fusion. The Multi-modal Attention Fusion Network (MAFnet) proposed by Brousmiche et al. [18] dynamically fuses audio and visual features through modal and temporal attention mechanisms, employs FiLM layers to achieve cross-modal interaction, and addresses the problem of divergent modal learning speeds via a random discarding strategy. The proposed network effectively improves the performance of video event recognition. Building upon this, Wang et al. [19] proposed a multi-modal audio–visual scene classification method that extracts audio and visual features using CNNs and ResNet, respectively, and employs a bidirectional attention mechanism to achieve cross-modal fusion. By introducing hierarchical feature fusion and an adaptive optimization strategy, the proposed method outperforms the single-modal approach by more than 15% in accuracy on mainstream datasets and exhibits strong robustness, particularly in noisy environments. Meanwhile, important progress has also been made in lightweight attention design. Duan et al. [20] proposed the DG-SCT attention mechanism, which dynamically adjusts the parameters of a pre-trained model through mutual audio–visual prompting. This method keeps the pre-trained model frozen and introduces only a small number of additional parameters

to achieve cross-modal feature optimization, thereby providing an efficient solution for multi-modal learning.

With the widespread adoption of the Transformer architecture, its advantages in temporal sequence modeling have become increasingly prominent. Sahu et al. [21] proposed a novel self-attention module to address the insufficient application of Transformers in video understanding and the difficulty of exploiting strong temporal relationships between adjacent video frames to obtain powerful frame-level representations. This module leverages both local and global temporal relationships between video frames to capture more comprehensive contextual information for individual frames, enabling the model to understand video content at multiple granularities. Meanwhile, multi-scale feature fusion has also achieved breakthroughs. Zhou et al. [22] proposed a Feature Pyramid Attention Network (FPANet) to tackle the problem of multi-scale feature fusion in audio–visual scene classification tasks. This method constructs a Feature Pyramid Attention Module (FPAM), which unifies feature scales across different levels through dimensional alignment and employs pyramid spatial attention and channel attention to focus on key spatial regions and temporal frames, thereby verifying its capability to effectively attend to semantically relevant regions.

In terms of data augmentation, Wang et al. [23] proposed an audio–visual scene classification method that combines joint modeling with data augmentation. A pre-trained image network is used to extract video features, while the audio network is trained from scratch to achieve modality-specific representations. Meanwhile, the effectiveness of various image enhancement operations in RandAugment is systematically compared. On the TAU 2021 dataset [24], the proposed method achieves the best result of the DCASE competition with an accuracy of 94.2%, thereby verifying the effectiveness of cross-modal collaborative learning in scene classification tasks. In the field of decision fusion, Pham et al. [25] constructed an audio–visual dataset containing five types of crowded scenes and proposed a multi-modal deep learning framework. By fusing Mel, CQT, and Gammatone spectral features and fine-tuning the visual network, combined with the PROD decision fusion strategy, a classification accuracy of 95.7% is achieved on the self-built dataset, which verifies the effectiveness of multi-modal fusion for complex scene understanding.

In terms of self-supervised learning, Suzuki et al. [26] proposed an audio–visual category association method based on two-stage self-supervised contrastive learning. In the first stage, one-to-one, one-to-many, and many-to-many correspondences are learned through self-supervised contrastive learning. In the second stage, for cases without correspondence, a dataset is constructed by analyzing the variation trends of category embeddings for contrastive learning. Experiments show that the F1 score of category embedding reaches 76.6%, which is 1.86 percentage points higher than that of the first stage, and the method can effectively predict audio–visual correspondences in real-world environments. This approach overcomes the limitations of traditional one-to-one associations and offers a new perspective for multi-modal scene understanding in complex environments. In the field of generative methods, Sung-Bin et al. [27] proposed a novel method for sound-to-visual scene generation through audio–visual latent alignment. To bridge the information gap between audio and visual modalities, a staged learning strategy is introduced. First, audio features are mapped to the visual latent space via cross-modal alignment, and then scene images are synthesized using a pre-trained generator. To improve generation quality, the method incorporates sound source localization techniques to select strongly correlated audio–visual pairs as training data. Experiments on the VEGAS and VGGSound datasets demonstrate that the generation quality of the proposed method significantly outperforms that of existing approaches. The study further shows that controllable editing of the gener-

ated results can be achieved through simple manipulations of the input waveform or the latent space, offering new insights for cross-modal generation tasks.

Gadzicki et al. [28] conducted systematic studies on modality integration that laid an important foundation for the field. They compared early and late fusion strategies in multi-modal scene classification and found that early fusion, which integrates data directly at the feature level, is suitable for scenes where modalities are well aligned and can effectively capture inter-modal correlations. In contrast, late fusion, which first models each modality independently and then fuses the results, offers greater flexibility.

Overall, compared with single-modal scene classification, multi-modal scene classification benefits from a more robust data foundation. Multi-modal fusion enables the model to perceive a scene from multiple dimensions, and information from different modalities can complement each other, thereby providing a more complete and accurate description of the scene. However, the optimization of multi-modal feature extraction methods and the selection of more suitable fusion models still require further exploration. In terms of feature extraction, although numerous feature extraction methods have been developed for different modalities, designing extraction methods that are well suited to the characteristics of each modality remains a challenge. In addition, how to design adaptive fusion models that can automatically adjust the fusion strategy according to the characteristics of different modal data and task requirements is also a current research focus and difficulty.

Existing methods typically project audio and video features into a common space through independent linear projections, but they do not explicitly model the cross-modal correlations during this projection. This results in aligned features that may still be sub-optimal for cross-modal interaction. Our cross-modal attention mechanism addresses this by computing attention weights that explicitly measure the relevance between each audio feature and each video feature.

3. Proposed Method

3.1. Cross-Modal Attention Mechanism Scene Classification Method

Audio and Video Scene Classification Network

In terms of feature extraction, as shown in Figure 1: (1) For the audio modality, this paper proposes an audio scene classification method based on improved scSE network and multi-attention mechanism. This method adopts a progressive feature optimization strategy to enhance the discriminative capability of the model for complex acoustic scenes. (2) For the video modality, this paper proposes a video scene classification method based on Residual Gated Recurrent Unit (Res-GRU). This method leverages the hierarchical structure of the convolutional neural network (ResNet50) to progressively extract features, and constructs a hybrid deep learning model by combining the advantages of recurrent neural networks in temporal data processing, to collaboratively improve the capability of spatial feature extraction and temporal modeling. For detailed technical specifications, please refer to Table 1.

Table 1. Architectural specifications.

Configuration	Value
Audio backbone	VGGish (pre-trained on AudioSet)
Audio feature dimension	128
Video backbone	ResNet-50 (pre-trained on ImageNet)
Video feature dimension	2048

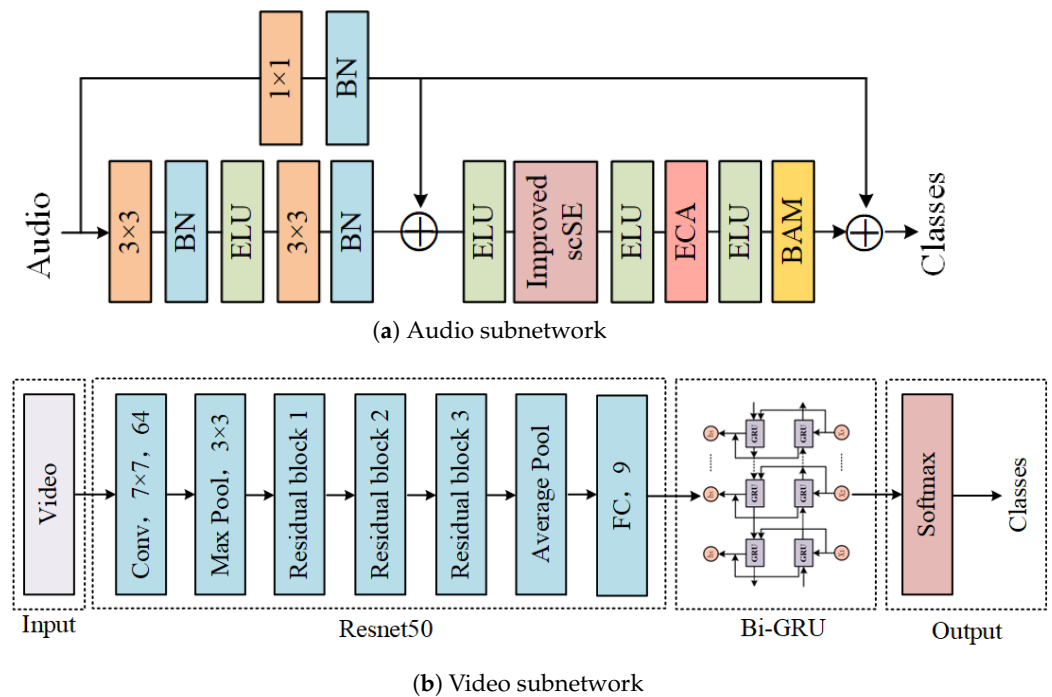


Figure 1. Audio–video network for feature extraction.

To address the shortcomings of existing audio–visual scene classification networks in feature fusion, this study proposes a fusion method based on a cross-modal attention mechanism. Currently, the fusion strategies employed by mainstream audio–visual scene classification methods exhibit the following limitations. On the one hand, the majority of networks adopt concatenation-based or addition-based fusion approaches, which merely combine audio and visual features in a simplistic manner. In practice, due to the lack of a deep interaction mechanism, these approaches struggle to effectively associate audio and visual features, making it difficult to mine the potential complementary information between the two modalities. This results in insufficient feature matching accuracy, which ultimately compromises the overall fusion quality. On the other hand, most existing studies on attention-based multi-modal scene classification compute attention weights within the same modality. Even when some methods employ Transformer-based self-attention mechanisms to explore inter-modal feature relationships, they fail to fully consider the intrinsic correlations between audio and visual features when calculating attention weights. As a result, the complementary advantages of the acoustic and visual modalities are difficult to exploit. To address the above problems, this study introduces a cross-modal attention mechanism. Prior to feature fusion, this mechanism facilitates bidirectional information interaction between audio and visual features, enhancing feature complementarity by constructing correlation relationships across modalities. Based on this, two architectures—a two-stream fusion network and a three-stream fusion network—are proposed. During the experiments, it is observed that the performance of the three-stream fusion scheme does not meet expectations and is weaker than that of the two-stream fusion. This phenomenon indicates that although increasing the number of modality streams can theoretically enrich the information dimension, factors such as redundancy among multi-modal features, the complexity of cross-modal attention weight allocation, and information overload during model training constrain the synergy of the three-stream system.

The architecture first extracts high-quality feature representations through dedicated audio subsets and video subsets, respectively, and then introduces a gating mechanism to dynamically adjust the cross-modal information flow. Specifically, the spliced audio and video features undergo weight matrix transformation, bias adjustment, and nonlinear

activation to generate semantic-aware gated values, so as to realize the adaptive interaction of cross-modal information. As an extension of the bi-modal network, the integrated multi modal network takes the original audio and video splicing features as the third modal input, which inherits the core processing flow of the bi-modal network. Both architectures adopt a unified post-processing paradigm to ensure the consistency and comparability of methods. The architecture diagram of the two-stream fusion network is shown in Figure 2, and the architecture diagram of the three-stream fusion network is shown in Figure 3.

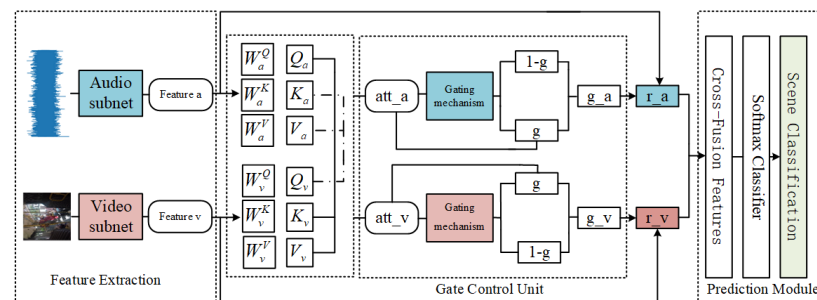


Figure 2. Two-stream fusion network.

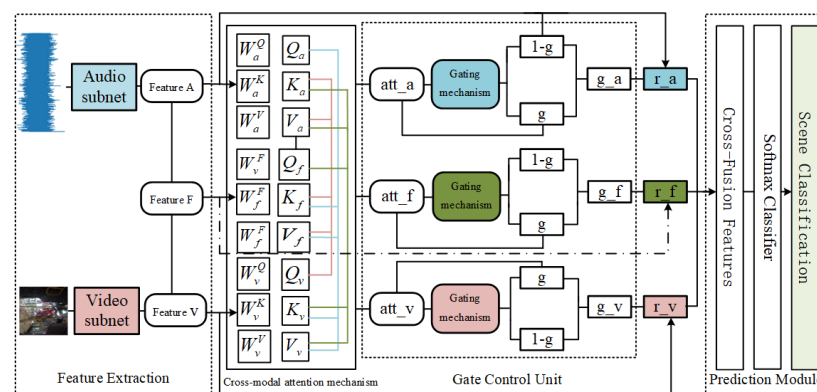


Figure 3. Three-stream fusion network.

3.2. Cross-Modal Attention Mechanism

Cross-modal attention is a multi-modal data processing method, which establishes the interaction between different modalities by dynamically calculating the attention weights between different modalities. The core of this method is to establish a dynamic correlation mechanism between modalities. Given the feature representations of the two modalities, the attention weights are obtained by normalization processing, and these weights are used to perform weighted fusion of features, so as to improve the expression ability of the fused features.

Different from the traditional shallow fusion methods such as feature concatenation or simple addition, the cross-modal attention mechanism can explicitly model the complex correlation between modalities. The core of the cross-modal attention mechanism is bidirectional information complementarity: the features of each modality are adaptively weighted according to the attention weight of the other modality, so as to realize deep interaction between modalities. Compared with the common attention mechanism that only focuses on the internal relationship of a single modality, cross-modal attention enforces the information exchange between modalities, and effectively improves the accuracy of feature matching by dynamically focusing on key cross-modal information. More importantly, the proposed mechanism has the advantage of robustness: when there is noise interference in one modality, the other modality can be effectively compensated by the attention weight to ensure the stable performance of the system. Based on sound, depending on the

two-mode-using cross-modal attention mechanism, inspired by tri-modal approaches, after the introduction of the audio–visual mode as the joined-together third mode, based on three-stream applications across the modal attention mechanism, as shown in Figure 4.

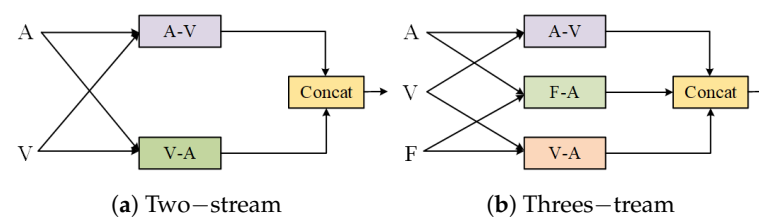


Figure 4. Cross-modal attention mechanism.

In early multi-modal research, feature fusion methods based on early fusion were widely adopted before the introduction of attention mechanisms. This kind of method uses a mathematical expression such as Equation (1):

$$I_{concat} = Concatenate(A \cdot V) \quad (1)$$

where A and V represent the feature vectors of audio and video, respectively. By directly connecting and combining the feature vectors of different modalities, this method retains the original feature information of each modality to the greatest extent. However, this fusion strategy has several significant limitations. First, this approach lacks explicit modeling of the interrelations and interactions between modalities. This lack of modeling ability makes it difficult for the model to fully exploit the potential correlation between different modal features, and thus cannot effectively use the complementary information between modalities to improve the performance of the model. Secondly, this method uses an equal-weighting strategy to deal with all modal features, ignoring that the importance of different modal features for specific tasks may be significantly different. This indiscriminate treatment may lead to the key features not being paid enough attention, while the secondary features receive too much attention, which ultimately affects the discriminative ability of the model.

More importantly, because the method combines the original features directly, when there are noise or outliers in a certain modal feature, these interference factors will directly affect the quality of the whole fusion feature. This effect will significantly reduce the noise propagation model of robustness and reliability prediction. Especially when dealing with complex multi-modal data, these defects become more prominent, making simple feature concatenation methods often difficult to achieve ideal performance.

In specific applications such as scene classification, these limitations are particularly obvious. Due to the significant difference in the contribution of audio and visual features to scene classification, it is difficult to make optimal use of complementary information of each modality by simple feature concatenation. This may lead to two serious consequences. On the one hand, the key features that are discriminative for the classification task may be ignored by the model. On the other hand, when there is low relevance between scene classification and the characteristics of the model, these features actually act as noise interference, and will eventually reduce the scene classification and the accuracy of the overall performance of the model.

To address the above challenges of multi-modal fusion, this paper proposes an audio and video scene classification method based on a cross-modal attention mechanism. The contribution of each modal feature is dynamically adjusted through the attention mechanism, and the key information of different modal data is adaptively mined to solve the problem of information collaboration between modalities.

3.2.1. Two-Stream Cross-Modal Attention Mechanism

To illustrate the proposed cross-modal attention mechanism, we take the ‘subway station’ scene as a concrete example. In this scene, the audio modality may capture sounds such as train announcements, footsteps, and platform door operations, while the visual modality captures platform layouts, crowd distributions, and signage. When the audio modality detects a train arrival announcement, the cross-modal attention mechanism $O_{a \rightarrow v}$ computes high attention weights for visual regions corresponding to platform edges and crowd movement areas. This enables the model to leverage audio cues to guide visual feature extraction, effectively achieving cross-modal information guidance. Conversely, when the visual modality detects crowded platform conditions, $O_{v \rightarrow a}$ guides the audio modality to focus more on announcement sounds and warning signals. This bidirectional information exchange allows the model to exploit complementary information from both modalities for more accurate scene classification.

The audio and video scene classification network based on cross-modal attention firstly extracts audio features A and video features V through the audio subnetwork and the video subnetwork. Due to the different dimensions of audio and video data in the original state, direct fusion will lead to information loss or not effective collaboration. Therefore, audio features A and video features V need to be projected to a unified dimension d_k through linear transformation, and the transformation formula is shown in Equations (2) and (3):

$$H_a = W_a A + b_a \quad (2)$$

$$H_v = W_v V + b_v \quad (3)$$

where W_a and W_v are the weight matrix, b_a and b_v are the bias vector.

The bias vector introduces a learnable constant term for the feature transformation to improve the data fitting ability of the model. Through this transformation, audio and video features are mapped to the same dimensional space. Cross-modal attention is achieved by multi-head attention, which is calculated in Equations (4) and (5):

$$O_{a \rightarrow v} = \text{MultiHeadAttention}(Q = H_a, K = H_v, V = H_v) \quad (4)$$

$$O_{v \rightarrow a} = \text{MultiHeadAttention}(Q = H_v, K = H_a, V = H_a) \quad (5)$$

Intuitively, $O_{a \rightarrow v}$ computes the attention that audio features pay to video features. A high attention weight indicates that certain audio cues are strongly correlated with specific visual regions—for example, in a ‘subway station,’ the sound of a train announcement (A) should direct the model’s visual attention to platform areas (V) where passengers are boarding. This allows audio modality to guide visual feature extraction.

Similarly, $O_{v \rightarrow a}$ computes the attention that video features pay to audio features. For instance, when the visual modality detects crowded conditions in a shopping mall (V), it guides the audio modality to focus more on crowd noise and ambient sounds (A). This bidirectional mechanism ensures that both modalities mutually enhance each other’s representations.

When calculating, $O_{a \rightarrow v}$ takes H_a as the query vector Q , and H_v is used as both the key vector K and the value vector V . The model measures the degree of attention that audio features pay to video features by calculating the attention score between audio features and video features, and then obtains the cross-modal attention output from audio to video. This process is conducive to the model to deeply explore the guiding effect of audio modality on video modality information attention. $O_{v \rightarrow a}$ takes H_v as the query vector Q , and H_a is used as both the key vector K and the value vector V , and the cross-modal attention output

from video to audio is calculated to achieve bidirectional information interaction between modalities.

3.2.2. Three-Stream Cross-Modal Attention Mechanism

Let the feature vectors of the fused audio and video be A and V respectively, and the fused feature F after the simple concatenation of audio and video in the previous text. To achieve multi-modal feature alignment, each modality feature is projected to a unified dimension d_k through linear transformation. Among the four options, 128, 256, 512, and 1024, under the same experimental conditions, the best results were obtained when d_k was set to 512 in this study.

$$\begin{cases} H_a = W_a A + b_a, \\ H_v = W_v V + b_v, \\ H_f = W_f F + b_f, \end{cases} \quad (6)$$

where W_a , W_v , and W_f are the learnable weight matrix, and b_a , b_v , and b_f are the bias term. Through this step, the audio, video and fusion features are mapped to a unified dimensional space, which lays a foundation for subsequent multi-modal feature interaction and ensures that information from different sources can be fused at the same scale. Cross-modal attention is realized by multi-head attention, and the attention formula of the video is shown in Equation (7):

$$\begin{cases} O_{v \rightarrow a} = \text{MultiHeadAttention}(Q = H_v, K = H_a, V = H_a), \\ O_{v \rightarrow f} = \text{MultiHeadAttention}(Q = H_v, K = H_f, V = H_f), \\ O_{v \rightarrow af} = \frac{O_{v \rightarrow a} + O_{v \rightarrow f}}{2}, \end{cases} \quad (7)$$

where $O_{v \rightarrow af}$ is the cross-modal output of the video modality.

When calculating, the video feature H_v is taken as the query vector Q , while the audio features H_a and H_f serve as both the key vector K and the value vector V . The model computes the attention scores between the video feature and the audio features, as well as the fused features, to obtain the cross-modal attention output $O_{v \rightarrow af}$. Similarly, the attention calculation formula for audio can be derived as shown in Equation (8):

$$\begin{cases} O_{a \rightarrow v} = \text{MultiHeadAttention}(Q = H_a, K = H_v, V = H_v), \\ O_{a \rightarrow f} = \text{MultiHeadAttention}(Q = H_a, K = H_f, V = H_f), \\ O_{a \rightarrow vf} = \frac{O_{a \rightarrow v} + O_{a \rightarrow f}}{2}, \end{cases} \quad (8)$$

The attention formula of the fused features is shown in Equation (9):

$$\begin{cases} O_{f \rightarrow a} = \text{MultiHeadAttention}(Q = H_f, K = H_a, V = H_a), \\ O_{f \rightarrow v} = \text{MultiHeadAttention}(Q = H_f, K = H_v, V = H_v), \\ O_{f \rightarrow av} = \frac{O_{f \rightarrow a} + O_{f \rightarrow v}}{2}, \end{cases} \quad (9)$$

We note that the ‘third stream’ in this architecture is not an independent modality (e.g., language or text), but rather a fused feature representation $F = \text{Concat}(A, V)$ that serves as an additional information source to study whether introducing a pre-fused representation provides complementary benefits beyond direct cross-modal attention between the two original modalities.

3.3. Multi-Modal Gating Mechanism

The theoretical basis of gating mechanisms in the field of deep learning stems from the cross-fusion of neuroscience and control theory. From the perspective of neuroscience, the

gating mechanism simulates the principle of synaptic plasticity in the human brain, and realizes the dynamic control of information flow through a similar way to the regulation of neurotransmitters. In the framework of control theory, the gating mechanism acts as an adaptive feedback regulation system, which can automatically adjust the information transmission path according to the importance of input features. At the mathematical level, the gating unit usually uses a differentiable function such as Sigmoid or Softmax to construct a probabilistic weight assignment model, which can not only ensure gradient backpropagation, but also maintain computational efficiency. The mechanism, by thinning the weights in the feature space to realize feature selection, has the flexibility to adapt to changes in the data distribution.

The current theoretical development of the gating mechanism shows a multi-dimensional evolution trend. At the structural design level, the early single-gating architecture has developed into a hierarchical gating architecture, which has realized the leap from global control to local fine-grained regulation. In terms of computational paradigm, the fusion use of hard and soft gating breaks the boundary between discrete and continuous decision-making. In terms of theoretical depth, the latest research has begun to introduce the information bottleneck theory into gating design to optimize the feature selection process by maximizing the effective information compression. In terms of gating mechanism and model of concentration, sparse coding theory provides the profound inner link; for example, attention mechanism gating can be regarded as a special kind of space, and the gating weight itself has the natural features of sparse representation.

The theoretical advantages of the gating mechanism are mainly reflected in three aspects. Firstly, differential continuity ensures the gradient stability in end-to-end training. Secondly, the parameter efficiency makes it achieve a good balance between model complexity and performance. Most importantly, its dynamic adaptability enables it to automatically respond to the changing characteristics of input data. These features make it so that multi-modal fusion can perform time-series modeling of complex tasks, such as loose unique value. However, this mechanism still faces theoretical challenges such as gradient saturation and initialization sensitivity. With the development of new technologies such as differentiable architecture search, the gating mechanism is evolving towards a more intelligent and adaptive direction, and its theoretical system is continuously enriched and improved.

3.3.1. Two-Stream Attention Gating Mechanism

$$G_{av} = \sigma(\text{Concat}(O_{a \rightarrow v}, A)W_g^{av} + b_g^{av}) \quad (10)$$

$$G_{va} = \sigma(\text{Concat}(O_{v \rightarrow a}, V)W_g^{va} + b_g^{va}) \quad (11)$$

In the formula, $\text{Concat}(O_{a \rightarrow v}, A)$ is obtained by first concatenating the audio feature $O_{a \rightarrow v}$ obtained through cross-modal attention with the original audio feature A , then multiplying by the weight matrix W_g^{av} and adding the bias matrix b_g^{av} , and finally compressing the result to the range of 0 to 1 through the activation function σ to obtain the gating value G_{av} from audio to video. This value measures the importance of cross-modal information interaction between audio and video. The principle of G_{va} is similar to this. Based on the importance of features determined by the gating mechanism, the interactive information between different modalities is weighted to improve the quality of fusion features. The weighted output is given in Equations (12) and (13):

$$O'_{a \rightarrow v} = O_{a \rightarrow v} \odot G_{av} + A \odot (1 - G_{av}) \quad (12)$$

$$O'_{v \rightarrow a} = O_{v \rightarrow a} \odot G_{va} + V \odot (1 - G_{va}) \quad (13)$$

where $\odot$ denotes element-wise multiplication.

The residual connection can make the model easier to learn and avoid the gradient problem during deep network training. The residual connection and output are shown in Equations (14) and (15):

$$Z_a = \text{LayerNorm}(A + O'_{a \rightarrow v}) \quad (14)$$

$$Z_v = \text{LayerNorm}(V + O'_{v \rightarrow a}) \quad (15)$$

where *LayerNorm* denotes the layer normalization operation.

Fusing the gating, the output is shown in Equation (16):

$$\text{Output} = \text{Concat}(Z_a, Z_v) \quad (16)$$

After a series of processing steps, the fused audio and video features are transformed into specific classification results, as shown in Equation (17):

$$P(y) = \text{softmax}(W_c \cdot \text{Output}) \quad (17)$$

3.3.2. Three-Stream Attention Gating Mechanism

The contribution of each modal attention output is dynamically adjusted through the gating weight. The mathematical formula of the gating mechanism is shown in Equation (18):

$$\begin{cases} G_{vaf} = \sigma(\text{Concat}(O_{v \rightarrow af}, V)W_g^{vaf} + b_g^{vaf}), \\ G_{avf} = \sigma(\text{Concat}(O_{a \rightarrow vf}, A)W_g^{avf} + b_g^{avf}), \\ G_{fav} = \sigma(\text{Concat}(O_{f \rightarrow av}, F)W_g^{fav} + b_g^{fav}), \end{cases} \quad (18)$$

where σ is the Sigmoid function.

Generate gating values between 0 and 1. See Equation (19) for the gating output:

$$\begin{cases} O'_{v \rightarrow af} = O_{v \rightarrow af} \odot G_{vaf} + V \odot (1 - G_{vaf}), \\ O'_{a \rightarrow vf} = O_{a \rightarrow vf} \odot G_{avf} + A \odot (1 - G_{avf}), \\ O'_{f \rightarrow av} = O_{f \rightarrow av} \odot G_{fav} + F \odot (1 - G_{fav}), \end{cases} \quad (19)$$

where $\odot$ denotes element-wise multiplication.

The residual connection can make it easier for the model to learn and avoid the gradient problem during deep network training. The residual connection and output are shown in Equation (20):

$$\begin{cases} Z_v = \text{LayerNorm}(V + O'_{v \rightarrow af}), \\ Z_a = \text{LayerNorm}(A + O'_{a \rightarrow vf}), \\ Z_f = \text{LayerNorm}(F + O'_{f \rightarrow av}), \end{cases} \quad (20)$$

The same can be said for three-stream fusion: Z_v, Z_a, Z_f , as shown in Equation (21):

$$\text{Output} = \text{Concat}(Z_v, Z_a, Z_f) \quad (21)$$

The fused audio and video features are transformed into specific classification results, as shown in Equation (22):

$$P(y) = \text{softmax}(W_c \cdot \text{Output}) \quad (22)$$

where W_c is the weight matrix of the classification layer of the model.

4. Experiment with Related Settings

4.1. Experimental Environment

The experimental environment is based on Ubuntu 20.04 operating system, using and PyCharm as an integrated development environment, and the programming language is Python. The computing hardware configuration includes a vGPU with 32 GB of video memory, a 24-core AMD EPYC 7T83 processor, and 90 GB of RAM. The deep learning framework uses TensorFlow and CUDA 11.6 for GPU-accelerated computation. The detailed experimental environment configuration parameters are shown in Table 2.

Table 2. Experimental environment configuration.

Experimental Environment	Environment Configuration
GPU	vGPU-32 GB
CPU	24 vCPU AMD EPYC 7T83 64-Core Processor
Memory	90 GB
Operating environment	Ubuntu 20.04
Model framework	TensorFlow
CUDA version	11.6
Programming language	Python
Python	3.7
PyCharm	2021.2.2

4.2. Dataset Configuration

This study uses the TAU 2021 Urban Audio-Visual Dataset [24], released for Task 1B of the DCASE 2021 challenge, as the experimental dataset. The dataset consists of 10-s MP4 video and WAV audio files, with a total duration of approximately 34 h. Collected from eight European cities—including Amsterdam and Barcelona—between 2018 and 2019, the data are annotated according to ten typical urban scenes, such as public squares, street traffic, and airports. These scenes can be further categorized into three major types: indoor, outdoor, and transportation areas. Detailed information about the dataset is shown in Table 3.

Table 3. Details of datasets.

Scenario	Training Set	Test Set
Airport	697	281
Shopping malls	841	387
Subway stations	893	386
Pedestrian street	968	421
Public square	982	387
Street traffic	985	402
Streetcar	763	308
Buses	806	327
Metro	893	386
Parks	1006	386
Total	8645	3645

In the data preprocessing stage, Gammatone filter banks are employed to extract time-frequency features from the audio data, effectively preserving both the temporal and spectral characteristics of the sound signal. For video data processing, key frames are extracted using frame sampling techniques, which significantly reduces computational complexity while ensuring the representativeness of the extracted features.

4.3. Experimental Parameter Settings

The following parameter configurations were adopted in this experiment: Acoustic and visual features were extracted using their respective audio and visual uni-modal feature extraction networks. For the multi-class classification task, the cross-entropy loss function was employed, and the Adam optimizer was used for algorithm optimization, with an initial learning rate of 0.01 and a weight decay coefficient of 0.1. The model was trained for 200 epochs, and the batch size was fixed at 32. The projection dimension d_k was set to 512, and the number of attention heads was 8. The dropout rate was set to 0.1. Detailed hyperparameter configurations are provided in Table 4.

Table 4. Basic hyperparameter table.

Parameter Names	Parameter Value
Audio dimensions	$64 \times 50 \times 32$
Image dimensions	$224 \times 224 \times 3$
Batch size	32
Learning rate	0.0001
Optimizer	Adam
Epoch	200
Projection dimension d_k	512
Number of attention heads	8
Dropout rate	0.1

4.4. Experimental Results and Analysis

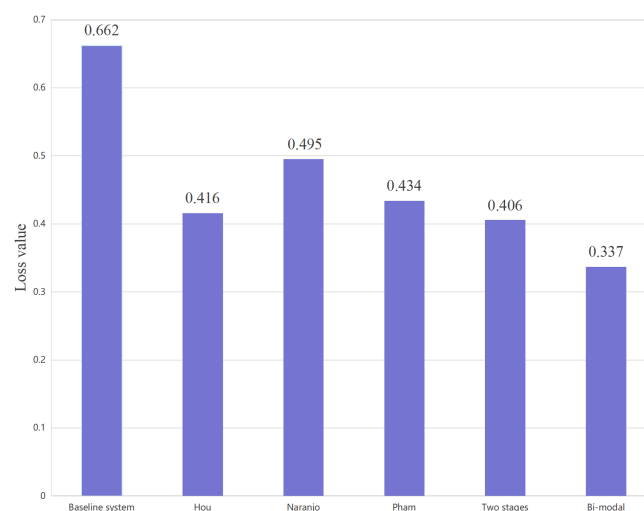
4.4.1. Comparison of Experimental Results and Analysis

Based on the results obtained on the test set, the two-stream fusion model and the three-stream fusion model proposed in this paper are compared and analyzed against (DCASE 2021 official), with the results being presented in Table 5. As a CNN-based dual-stream network, the model in Hou et al. [29] outperforms all other CNN-only approaches in the DCASE 2021 challenge. Naranjo et al. [30] presented a CRNN with Squeeze-excitation blocks, representing an attention-based extension of conventional CNN approaches. The data show that the bi-modal fusion model achieves an average accuracy of 89.3% in the video scene classification task, surpassing the baseline model's average accuracy of 77.0% and outperforming the best-performing comparison model ([25]) by 0.9 percentage points, thereby confirming the effectiveness of combining cross-modal attention with a gating mechanism. In dynamic scene classification, the advantages of the two-stream model are particularly significant. In the subway station scene, the proposed model achieves an accuracy of 96.2%, which is only 0.5 percentage points lower than that of the Naranjo et al. [30]. model. For the metro and bus scenes, the accuracies reach 94.1% and 94.4%, respectively, ranking among the top results across all compared models. In the pedestrian street scene, the model attains an accuracy of 92.9%, which is 20.2 percentage points higher than the baseline and 10 percentage points higher than the best-performing Pham model [31] among comparable methods. In terms of scene coverage, the bi-modal model achieves superior classification accuracy compared to mainstream methods in 8 out of the 10 scene types. In scenes with complex environmental characteristics, such as street traffic and public squares, the proposed model reaches accuracies of 91.5% and 86.1%, respectively—1.9 and 15 percentage points higher than the baseline—demonstrating stronger feature adaptability and robustness.

Table 5. Experimental environment configuration (%).

Scenario	Baseline System	Hou [29]	Naranjo [30]	Pham [31]	Tri-Modal	Bi-Modal
Airport	66.8	82.8	78.6	84.8	82.9	83.4
Shopping malls	72.6	86.7	91.8	89.6	87.4	86.1
Subway stations	80.8	95.1	97.7	96.7	95.3	96.2
Pedestrian street	72.7	81.8	80.2	82.9	92.5	92.9
Public square	71.1	70.4	78.2	77.4	81.5	86.1
Street traffic	89.6	94.1	95.7	93.7	90.0	91.5
Trams	73.1	73.3	74.3	81.0	73.8	76.2
Buses	85.9	83.5	86.9	92.3	94.2	94.4
Metro	80.4	88.1	82.6	88.7	96.8	94.1
Parks	77.2	100	99.4	97.3	93.2	92.4
Mean	77.0	85.6	86.5	88.4	88.8	89.3

According to the bar chart of loss value in Figure 5, the bi-modal-plus-gating-mechanism model showed the lowest loss value in the comparison experiment, which verified the effectiveness of the model design. The model performs dynamic and adaptive fusion of multi-modal features through the gating mechanism, which can automatically filter effective information during the training process and reduce the redundant interference caused by traditional feature fusion. This mechanism makes the model converge more stably in iterative optimization and effectively reduces the training fluctuation. From the perspective of performance correlation, the 89.3% classification accuracy achieved by the model and the lowest loss value form a validation closed loop: a low loss value indicates a high degree of model parameter optimization, while a high accuracy reflects its strong generalization ability. The two jointly prove the effectiveness of the model in the task of video scene understanding. In complex dynamic scenes, the gating mechanism realizes the accurate weighted fusion of information by quantifying the contribution of each modal feature, enhances the adaptability of the model to environmental changes, and thus improves the overall performance of scene classification.

**Figure 5.** Bar chart of comparative experimental loss value.

4.4.2. Ablation Experimental Results and Analysis

Based on the experimental data in Table 6, the two-stream model with the gating mechanism achieved the best performance among all compared schemes, attaining an average accuracy of 89.3%. Compared with the two-stream model without gating, the proposed model yields a 1.2 percentage point improvement in accuracy, demonstrating a significant advantage in dynamic scene classification: the accuracies for subway station,

tram, and bus scenes reach 96.2%, 76.2%, and 94.4%, respectively. The gating mechanism contributes substantially to the performance improvement. In dynamic scenarios, the model achieves an average accuracy of 88.9%, further validating the effectiveness of the gating mechanism in enhancing the classification capability of the model.

Table 6. Comparison table of accuracy results of ablation experiments (%).

Scenario	Three-Stream Without Gating Mechanism	Two-Stream Without Gating Mechanism	Tri-Modal	Bi-Modal
Airport	80.1	82.3	82.9	83.4
Shopping mall	80.9	82.5	87.4	86.1
Subway stations	90.8	94.2	95.3	96.2
Pedestrian street	91.4	91.4	92.5	92.9
Public square	88.2	87.8	81.5	86.1
Street traffic	91.6	92.0	90.0	91.5
Trams	72.8	70.3	73.8	76.2
Buses	84.9	93.9	94.2	94.4
Metro	92.8	95.7	96.8	94.1
Parks	92.1	90.9	93.2	92.4
Mean	86.6	88.1	88.8	89.3

It can be seen from the comparison bar chart of model loss value in Figure 6 that the loss value of the three-stream model without gating unit is 0.460, the two-stream model without the gating unit is 0.405, the three-stream model with gating unit is 0.387, and the two-stream model with the gating unit is the lowest, 0.337. This shows that the gating unit can effectively reduce the model loss value. With the assistance of the gating unit, the training effects of the bi-modal and tri-modal models are improved, and the bi-modal-plus-gating-unit model has obvious advantages, which can more accurately regulate feature fusion, reduce training errors, and optimize model performance.

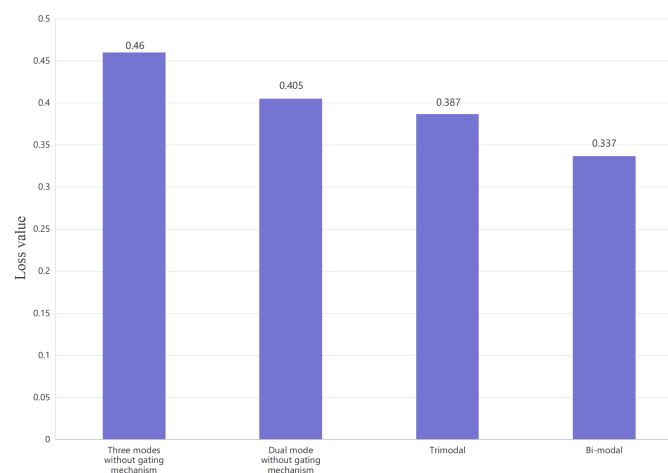


Figure 6. Bar chart of loss value in ablation experiment.

According to the precision and F1-score results for each scene of the ablation models presented in Table 7, the performance of both the two-stream and three-stream configurations is improved after the introduction of the gating mechanism, with increases in the average precision and F1-score values, which is consistent with the performance improvements discussed in the paper. This reflects the optimization effect of the gating mechanism on feature fusion. Meanwhile, the two-stream configuration outperforms the three-stream model in terms of both the mean precision and F1-score values as well as in most individual scenes, thereby validating the effectiveness of the reduced modality combination. These

results corroborate the accuracy experiments and fully demonstrate the critical role of the gating mechanism in improving the scene classification performance of the model.

Table 7. Comparison table of ablation experimental results (%).

Scenario	Comparison of Precision (%) Results for Each Type of Scenario				Comparison of F1-Score (%) Results for Each Type of Scenario			
	Three-Stream Without Gating Unit	Two-Stream Without Gating Unit	Three-Stream Plus Gating Unit	Two-Stream Plus Gating Unit	Three-Stream Without Gating Unit	Two-Stream Without Gating Unit	Three-Stream Plus Gating	Two-Stream Gating Unit
Airport	94.1	85.1	88.2	89.5	86.5	83.7	85.5	86.4
Shopping malls	87.1	87.5	89.3	88.3	88.9	90.7	92.2	92.1
Subway stations	81.0	92.6	87.7	90.1	85.9	92.0	90.0	91.5
Pedestrian street	89.7	87.3	89.3	89.7	88.9	87.6	85.2	87.9
Street traffic	90.2	92.5	95.3	94.9	90.9	92.3	92.6	93.2
Trams	71.9	73.6	80.3	78.0	72.4	71.9	76.9	77.1
Buses	80.3	84.5	82.3	83.0	82.5	89.0	87.8	88.3
Metro	88.4	91.0	93.9	95.9	90.6	93.3	95.3	95.0
Parks	97.0	96.8	96.3	97.0	94.5	93.7	94.8	94.6
Mean	86.9	88.0	88.9	89.5	86.6	88.0	88.7	89.3

4.4.3. Visualization Results and Analysis

The confusion matrix of the audio and video scene classification method based on the cross-modal attention mechanism on the test set is shown in Figure 7.

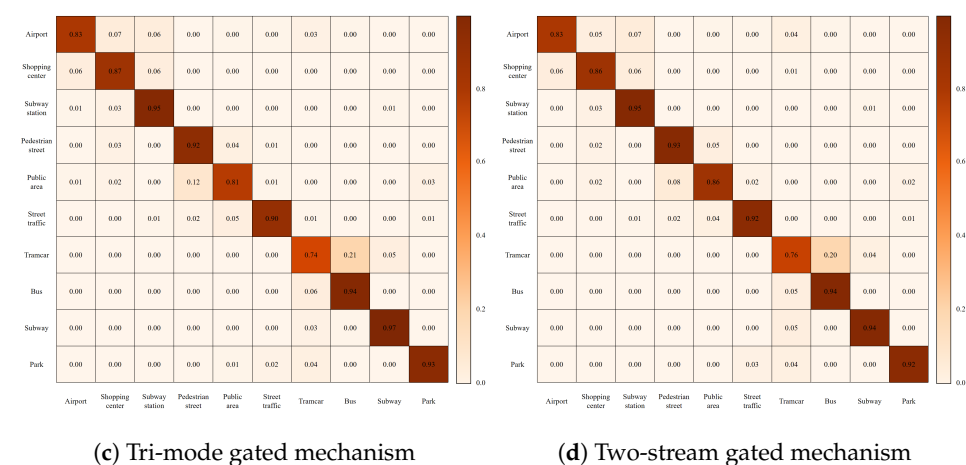
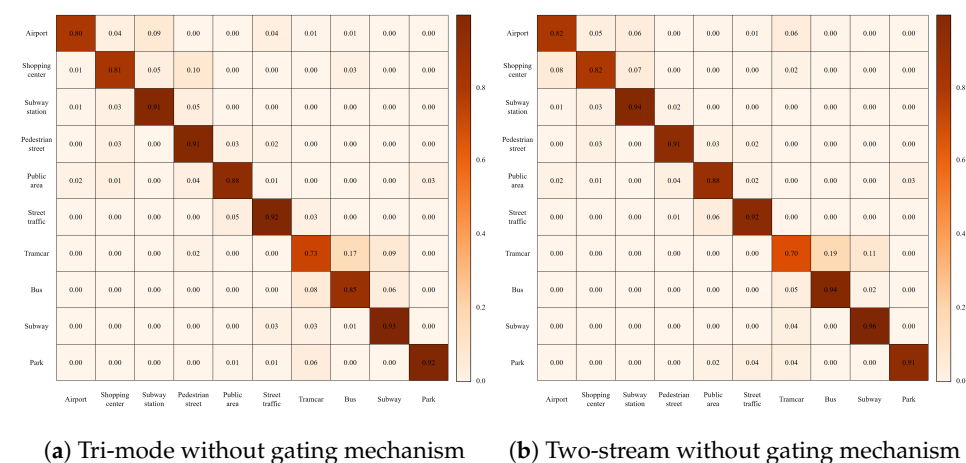


Figure 7. Confusion matrix for audio and video scene classification.

By comparing the confusion matrix, it was found that after adding the gating mechanism, the proportion of correct classification of the three-stream model and two-stream model in multiple scenes was significantly improved. In the two-stream model, the accuracy of shopping mall scene increased, and the probability of confusion with other scenes decreased, indicating that the gating mechanism is better than the model without the gating mechanism in the utilization of features. In the tram scene, although the confusion degree with the bus is unchanged, the overall confusion degree is reduced. In the three-stream network, the overall classification accuracy is improved, and the classification accuracy of each scene is increasing.

5. Conclusions

In real-world urban scene classification tasks, environmental information usually contains rich multi-modal signals, in which audio and visual data each carry complementary scene feature representation capabilities. Previous studies are mostly limited to a single modality, which cannot fully capture the complete information of complex urban scenes, and independent modal representation cannot effectively capture the multi-dimensional characteristics of the environment, resulting in limited classification performance. Aiming at the above problems, this paper studies the fusion mechanism of multi-modal features, and proposes a multi-modal fusion method based on a cross-modal attention mechanism for scene classification to solve the problem of insufficient modal interaction of audio and video data. Firstly, a bidirectional attention mapping is established between audio and video modalities, and the correlation weights between modalities are calculated by a QKV mechanism. Then, a learnable gating unit is used to dynamically adjust the contribution of each modality. Finally, the residual structure and gating mechanism are used to ensure the effective complementarity of the fused features and retain the integrity of the original modality features to avoid information loss. Experiments show that on the TAU Urban Audio-Visual Scenes 2021 dataset, the accuracy of audio classification is improved by 3.8% to 68.9%, the accuracy of video classification is improved by 18.5% to 84.2%, and the accuracy of two-stage audio–video fusion is improved by 9.8% to 86.8%. The accuracy of cross-modal attention method is improved by 12.3% to 89.3%.

Author Contributions: M.X.: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing—original draft, Data curation. G.Y.: Conceptualization, Methodology, Validation, Supervision, Writing—review and editing, Project administration. Y.D.: Supervision, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The TAU Urban Audio-Visual Scenes 2021, a multi-modal public dataset released by Tampere University, was specifically designed for DCASE 2021 Challenge Task 7—sound source localization in audio-visual urban scenes. <https://zenodo.org/records/4767103> (accessed on 7 July 2026).

Acknowledgments: The authors have never utilized ChatGPT (the GPT-5.5 version of OpenAI, San Francisco, CA, USA) for language editing and grammar correction. During the writing of any part of this paper, no artificial intelligence or AI-assisted tools were employed.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Baltrusaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 423–443. <https://doi.org/10.1109/TPAMI.2018.2798607>.
- Mulimani, M.; Venkatesh, S.; Koolagudi, S.G. Acoustic Event and Scene Classification: A Review. *SN Comput. Sci.* **2025**, *6*, 54. <https://doi.org/10.1007/s42979-024-03592-9>.
- Lei, H.; Chen, N. Audio-Visual Scene Classification Based on Multi-modal Graph Fusion. In *Interspeech 2022*; International Speech Communication Association (ISCA): Grenoble, France, 2022; pp. 4157–4161. <https://doi.org/10.21437/interspeech.2022-741>.
- Ata-Ur-Rehman.; Ullah, H.S.; Farooq, H.; Khan, M.S.; Owais, H. Multi-Modal Anomaly Detection by Using Audio and Visual Cues. *IEEE Access* **2021**, *9*, 30587–30603. <https://doi.org/10.1109/ACCESS.2021.3059519>.
- Lin, J.; Chen, J.; Peng, K.; He, X.; Li, Z.; Stiefelhofen, R.; Yang, K. EchoTrack: Auditory Referring Multi-Object Tracking for Autonomous Driving. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 18964–18977. <https://doi.org/10.1109/TITS.2024.3437645>.
- Apostolidis, K.; Abeber, J.; Cuccovillo, L.; Mezaris, V. Visual and audio scene classification for detecting discrepancies. In *Proceedings of the 3rd ACM International Workshop on Multimedia AI Against Disinformation*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 30–36. <https://doi.org/10.1145/3643491.3660287>.
- Aryal, N.; Lee, S.W. Frequency-based CNN and attention module for acoustic scene classification. *Appl. Acoust.* **2023**, *210*, 109411. <https://doi.org/10.1016/j.apacoust.2023.109411>.
- Stowell, D.; Giannoulis, D.; Benetos, E.; Lagrange, M.; Plumbley, M.D. Detection and Classification of Acoustic Scenes and Events. *IEEE Trans. Multimed.* **2015**, *17*, 1733–1746. <https://doi.org/10.1109/TMM.2015.2428998>.
- Kamarajugadda, R.; Battula, R.; Borra, C.R.; Durga, H.; Bypilla, V.; Reddy, S.S.; Khan, F.F.; Bhavanam, S. Optimizing avian species recognition with MFCC features and deep learning models. *Int. J. Inf. Technol.* **2024**, *16*, 4621–4626. <https://doi.org/10.1007/s41870-024-02108-1>.
- Rambabu, D.; Swetha, G.; Datla, R.; Chalavadi, V.; Mohan, C.K. RSZero-CSAT: Zero-Shot Scene Classification in Remote Sensing Imagery using a Cross Semantic Attribute-guided Transformer. In *2024 International Joint Conference on Neural Networks (IJCNN)*; IEEE: New York, NY, USA, 2024; pp. 1–8.
- Sun, X.; Gao, J.; Yuan, Y. Alignment and Fusion Using Distinct Sensor Data for Multimodal Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5626811. <https://doi.org/10.1109/TGRS.2024.3406697>.
- Owens, A.; Efros, A.A. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. *arXiv* **2018**, arXiv:1804.03641. <https://doi.org/10.48550/arXiv.1804.03641>.
- Yin, W.; Kann, K.; Yu, M.; Schütze, H. Comparative Study of CNN and RNN for Natural Language Processing. *arXiv* **2017**, arXiv:1702.01923. <https://doi.org/10.48550/arXiv.1702.01923>.
- Yu, Y.; Si, X.; Hu, C.; Zhang, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput.* **2019**, *31*, 1235–1270. https://doi.org/10.1162/neco_a_01199.
- Dey, R.; Salem, F.M. Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks. In *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*; IEEE: New York, NY, USA, 2017.
- Yin, M.; Liang, T.; Zeng, X.; Shi, D. Multimodal Multi-Scale Temporal Enhancement Network for Audio-Visual Scene Classification. *IEEE Trans. Audio Speech Lang. Process.* **2026**, *34*, 1944–1951. <https://doi.org/10.1109/TASLPRO.2026.3678801>.
- Cantarini, M.; Gabrielli, L.; Mancini, A.; Squartini, S.; Longo, R. A3CarScene: An audio-visual dataset for driving scene understanding. *Data Brief* **2023**, *48*, 109146.
- Brousmiche, M.; Rouat, J.; Dupont, S. Multimodal Attentive Fusion Network for audio-visual event recognition. *Inf. Fusion* **2022**, *85*, 52–59. <https://doi.org/10.1016/j.inffus.2022.03.001>.
- Wang, Y.; Liu, Y.; Huang, W.; Ye, X.; Jiang, M. Two-Stage Fusion-Based Audiovisual Remote Sensing Scene Classification. *Appl.-Sci.* **2023**, *13*, 21. <https://doi.org/10.3390/app13211890>.
- Duan, H.; Xia, Y.; Zhou, M.; Tang, L.; Zhu, J.; Zhao, Z. Cross-modal Prompts: Adapting Large Pre-trained Models for Audio-Visual Downstream Tasks. In *Proceedings of the Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, 10–16 December 2023*.
- Sahu, S.; Goyal, P. Leveraging Local Temporal Information for Multimodal Scene Classification. In *ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2022; pp. 1830–1834. <https://doi.org/10.1109/icassp43922.2022.9747429>.
- Zhou, L.; Zhou, Y.; Qi, X.; Hu, J.; Lam, T.L.; Xu, Y. Feature pyramid attention network for audio-visual scene classification. *CAAI Trans. Intell. Technol.* **2025**, *10*, 359–374. <https://doi.org/10.1049/cit2.12375>.
- Wang, Q.; Du, J.; Zheng, S.; Li, Y.; Wang, Y.; Wu, Y.; Hu, H.; Yang, C.; Siniscalchi, S.; Wang, Y. A Study on Joint Modeling and Data Augmentation of Multi-Modalities for Audio-Visual Scene Classification. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*; IEEE: New York, NY, USA, 2022; pp. 453–457. <https://doi.org/10.1109/ISCSLP57327.2022.10038206>.
- Wang, S.; Mesaros, A.; Heittola, T.; Virtanen, T. A curated dataset of urban scenes for audio-visual scene analysis. *arXiv* **2021**, arXiv:2011.00030. <https://doi.org/10.48550/arXiv.2011.00030>.

25. Pham, L.; Ngo, D.; Nguyen, T.; Nguyen, P.; Hoang, T.; Schindler, A. An Audio-Visual Dataset and Deep Learning Frameworks for Crowded Scene Classification. In *CBMI '22: Proceedings of the 19th International Conference on Content-based Multimedia Indexing*; Association for Computing Machinery: New York, NY, USA, 2022; pp. 23–28. <https://doi.org/10.1145/3549555.3549568>.
26. Suzuki, K.; Itoyama, K.; Nishida, K.; Nakadai, K. Audio-Visual Class Association Based on Two-stage Self-supervised Contrastive Learning towards Robust Scene Analysis. In *2023 IEEE/SICE International Symposium on System Integration (SII)*; IEEE: New York, NY, USA, 2023; pp. 1–6. <https://doi.org/10.1109/SII55687.2023.10039379>.
27. Sung-Bin, K.; Senocak, A.; Ha, H.; Owens, A.; Oh, T. Sound to Visual Scene Generation by Audio-to-Visual Latent Alignment. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*; IEEE: New York, NY, USA, 2023; pp. 6430–6440. <https://doi.org/10.1109/CVPR52729.2023.00622>.
28. Gadzicki, K.; Khamsehashari, R.; Zetzsche, C. Early vs Late Fusion in Multimodal Convolutional Neural Networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*; IEEE: New York, NY, USA, 2020; pp. 1–6. <https://doi.org/10.23919/FUSION45008.2020.9190246>.
29. Hou, Y.; Tan, Y.; Chang, Y.; Huang, T.; Li, S.; Shao, X.; Botteldooren, D. *CNN-Based Dual-Stream Network for Audio-Visual Scene Classification*; Technical Report; DCASE2021 Challenge: Barcelona, Spain, 2021.
30. Naranjo-Alcazar, J.; Perez-Castanos, S.; Cobos, M.; Ferri, F.J.; Zuccarello, P. *Task 1B DCASE 2021: Audio-Visual Scene Classification with Squeeze-Excitation Convolutional Recurrent Neural Networks*; Technical Report; DCASE2021 Challenge: Barcelona, Spain, 2021.
31. Pham, L.; Schindler, A.; Schutz, M.; Lampert, J.; King, R. *DCASE 2021 Task 1B: Technique Report*; Technical Report; DCASE2021 Challenge: Barcelona, Spain, 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.