# Video Segmentation and Tokenization for Model-Based Video Scene Classification

Qing Wang, Yajian Wang, Hang Chen, Shuxian Wang, Jun Du and Chin-Hui Lee

***Abstract*—In this paper, we propose a novel approach for segmenting and tokenizing a video scene recording into a sequence of cascade units, known as visual segment units and modeled with visual segment models (VSMs) for video scene classification (VSC). Specifically, the proposed VSM framework takes deep visual features extracted from pre-trained encoders as inputs and models the temporal interactions between segment units by hidden Markov models. Next, we use unit co-occurrence statistics to introduce relationships between VSM units within a video scene recording. Furthermore, the VSM approach is extended to an acoustic-visual variant, subsequently integrating itself into a deep learning-based multi-modal scene classification system. This combination serves to further exploit the complementary nature of audio and video data. By incorporating a set of visual segment units into modeling a video scene class, it captures both inter-class similarity and intra-class diversity, facilitating improved scene classification, especially within categories prone to confusion. Extensive experimental results on a benchmark published by the DCASE (Detection and Classification of Acoustic Scenes and Events) 2021 Challenge show that the proposed framework can effectively handle the confusion issue among similar video scenes. In addition, our multi-modal integration system achieves state-of-the-art performance in the audio-visual scene classification task in the DCASE 2021 Challenge, thereby demonstrating the effectiveness of our proposed approach.**

***Index Terms*—Video scene classification, visual segment model, pre-trained model, latent semantic analysis, deep learning**

## I. Introduction

AUTOMATIC scene classification (SC) stands as a crucial task in scene analysis, with its primary objective being the categorization of a given signal into a predefined class, thereby facilitating machine comprehension of its surroundings. This task has a variety of applications spanning various domains, including autonomous driving, smart surveillance, personal archiving [1], and robotic navigation [2]. One promising application of scene classification is found in smartphones, where it enables continuous environmental awareness, automatically adjusting call volume upon entry into a noisy scene.

Previous studies on scene classification are mainly based on separate audio and video domains. Early approaches for acoustic scene classification (ASC) task focused on designing proper features for the classifiers [3]–[5]. In [6], mel-frequency cepstral coefficients (MFCCs) were adopted as the spectral representation, with Gaussian mixture models (GMMs) [7] and support vector machines (SVMs) [8] serving as the designated classifiers. Bisot *et. al* [9] investigated different matrix factorization methods to perform feature learning for ASC, achieving significant performance improvement compared to previous handcrafted features. Recent developments have witnessed the integration of convolutional neural networks (CNNs) to solve the ASC task, demonstrating superior performance over conventional approaches [10]–[12].

Visual scene classification [13] has been a long-standing challenge in computer vision, which aims to classify an image or a video. Various approaches have been developed for image scene classification, including global attribute descriptors [14], patch feature encoding [15], spatial layout pattern learning [16], discriminative region detection [17], and the emergence of deep models [18]. Different from static images, video scene classification (VSC) is a more complicated task, as it involves analyzing the temporal evolution of visual content within videos. In [19], hidden Markov models (HMMs) were used to model the time-varying patterns exhibited by different scene categories in video data. Some works [20], [21] applied image scene classification methods to produce results for videos, where key frames were selected for classification. Recently, more powerful models have been proposed for video understanding, such as SlowFast Networks [22], Multiscale Vision Transformers (ViT) [23], and UniformerV2 [24].

Hearing and vision are the two primary senses through which humans perceive and comprehend their environment. With the rapid progress in artificial intelligence technologies, the field of audio-visual learning (AVL) [25] has witnessed significant advancements in both academic research and industrial applications. AVL combines information from two sensory modalities, thereby overcoming the limitations associated with single-modal tasks. Transfer learning-based methods have shown promising results in the audio-visual scene classification (AVSC) task [26], [27]. However, it is worth noting that many video encoders utilize pre-trained image models, which do not fully leverage the temporal information present in video clips, as they only take a single image frame as input. Self-supervised learning (SSL) techniques offer an alternative for audio-visual representation [28]–[30], achieving promising results in various multi-modal downstream tasks. However, since SSL models are typically trained on large-scale unlabeled datasets, they often face the risk of overfitting, particularly when these models have a high number of parameters and are applied to downstream tasks with low data resources.

One of the challenges in VSC is the similarity between different scene categories. In cases where these categories are easily confused, the static visual image of a single frame

Qing Wang, Yajian Wang, Hang Chen, Shuxian Wang and Jun Du are with University of Science and Technology of China, Hefei, Anhui, China. Chin-Hui Lee is with Georgia Institute of Technology, Atlanta, GA, USA. e-mail: qingwang2@ustc.edu.cn, yajian@mail.ustc.edu.cn, ch199703@mail.ustc.edu.cn, sxwang21@mail.ustc.edu.cn, jundu@ustc.edu.cn, chl@ece.gatech.edu. (Corresponding author: Jun Du.)

within the video can appear similar, making the contextual dependencies among visual frames captured by temporal information crucial for recognizing scenes correctly. In this study, we propose a novel visual segment model (VSM) based approach for video scene classification with the assumption that the overall visual characteristics of all scenes can be represented by a universal set of fundamental units. The VSM methodology is inspired by the acoustic segment model (ASM), which was first introduced in [31] to characterize fundamental speech sound units for automatic speech recognition (ASR). In an ASM-based framework, variable-length segments are modeled using GMM-HMMs to incorporate temporal information. Each training sample is then encoded into a sequence of segment units.

In this study, we focus on low-resource AVSC and propose a novel VSM-based approach to generate effective representations for modeling visual scenes. The method begins by extracting deep visual features from pre-trained encoders for each video clip, which are then divided into non-overlapping segments. These segments are designed to encapsulate coherent visual properties, with segment boundaries indicating visual discontinuities. Next, a video recording is represented as a temporal sequence of fundamental units via visual segment modeling, which involves initial segmentation and iterative training. Given the temporal interactions and relationships between frames in a video, HMMs serve as a straightforward way to model temporal dependencies. Then, feature vectors are obtained using latent semantic analysis (LSA) [32], followed by vector-based classification. Furthermore, we extend the VSM approach to an audio-visual version by leveraging complementary information from both modalities. We incorporate the acoustic-visual segment model (AVSM) framework into our previously proposed AVSC system [33] to fully exploit the synergy between the audio and video modalities. This method enhances our understanding of changes in visual features, ultimately improving the accuracy of classifying scenes that are easily confused.

The three major contributions are summarized as follows and respectively detailed in Sections V-B, V-C, and V-D:

- We propose a novel VSM-based hybrid classification approach, which consists of three essential modules: visual segment modeling, latent semantic analysis and vector-based classification. Visual segment modeling is at the core of our approach, which aims to translate a video recording into a sequence of discrete visual units.
- We further extend the proposed VSM-based framework to an AVSM-based approach and subsequently integrate it into a deep learning-based multi-modal scene classification system to fully exploit the synergy between the audio and video modalities.
- We conduct a comprehensive set of experiments to demonstrate the effectiveness of the VSM-based approach in obtaining high-quality video representations, which outperforms those obtained by deep models. Furthermore, our multi-modal fusion system achieves a state-of-the-art classification accuracy when evaluated on the development set of the DCASE (Detection and Classification of Acoustic Scenes and Events) 2021 Challenge.

## II. Related Works

### A. Bag-of-Words Representation

The bag-of-words (BoW) methodology, originally proposed for text categorization [34], [35], has been successfully adapted to the field of computer vision. Sivic and Zisserman [36] were among the pioneers in applying the BoW representation to object and scene retrieval by building a visual vocabulary. Csurka *et. al* [37] introduced a bag-of-keypoints approach to visual categorization, utilizing scale invariant feature transform (SIFT) descriptors. In [38], an enhanced bag-of-visual-word (BoVW) method was presented to represent visual content, involving the construction of informative visual words from representative keypoints and a novel approach to restructure the vector space model. Another notable contribution was made in [39], which proposed a BoVW model that combined both local and global features for high spatial resolution image scene classification. This model demonstrated superiority over state-of-the-art methods when evaluated on datasets like UC Merced and Google data sets of SIRI-WHU. Furthermore, Gidaris *et. al* [40] proposed a self-supervised approach for predicting the BoW representation. In [41], acoustic features were incorporated into the standard BoVW approach for movie shot representation, followed by a latent topic driving model designed for affective scene classification.

Most of these works mainly focused on the image domain and used BoW models to represent individual images. Our method differs from these previous works by introducing a segment model for building a visual inventory for video data.

### B. Visual Tokenization

Visual tokenization aims to map pixels into compact discrete tokens suitable for language models. Image tokenization methods, including VQ-VAE [42] and ViT-VQGAN [43], usually consist of a CNN encoder and a vector-quantization (VQ) bottleneck to obtain discrete tokens. Video tokenization is more challenging due to the temporal coherence of video data. As depicted in [44], C-ViViT employs a spatial transformer and a causal transformer to build a joint image-video tokenizer. The state-of-the-art video tokenizer is MAGVIT-v2 [45], which is designed with a lookup-free quantization approach to generate concise and expressive tokens for both videos and images.

Discrete tokens are generated by applying quantization techniques on individual feature maps from the CNN encoders. Our method is quite different from these methods. First, VQ is performed on segment-level features. Second, HMMs are used to improve tokenization by utilizing temporal information.

### C. Acoustic Segment Model

The ASM model was first proposed for ASR by Lee *et. al* [31]. In this work, an HMM was trained for each segment model representing acoustically similar sounds, aiming to capture the intra-segment variability within each sound class. Subsequently, the ASM approach has been applied to various audio classification tasks, including spoken language recognition [46], music genre classification [47], speaker recognition [48], and acoustic scene classification [49], [50]. Bai *et. al* [49]

proposed an ASM approach to represent an audio recording as a temporal sequence of basic sound units. Each unit was modeled using a GMM-HMM, and the method outperformed CNN in acoustic scene classification. In [51], an ASM model was utilized to generate guidance information and was combined with a two-stage attention network to identify salient frames for scene classification. Inspired by the idea of stop words in information retrieval, Hu *et. al* [50] proposed an ASM-based segment unit selection framework to remove frames carrying little information content for the ASC task.

### D. Audio-Visual Scene Classification

The goal of AVSC is to automatically categorize diverse types of audio-visual scenes, such as indoor versus outdoor scenes and urban versus rural scenes. This task typically involves analyzing a variety of features extracted from both the audio and visual components of a scene. In recent years, the accuracy of image classification has been significantly improved, largely due to the availability of large-scale image datasets like ImageNet [52] and Places [53]. Leveraging this progress, Hu *et. al* introduced an annotated dataset [26] consisting of geotagged aerial image-sound pairs. In their work, the authors proposed three transfer learning-based approaches that incorporated sound event knowledge into the visual scene recognition task. Additionally, Pham *et. al* [54], a wide range of deep learning (DL)-based models were investigated using different spectrogram features for crowded audio-visual scene classification. Their approach applied late fusion to integrate information from both audio and visual modalities, resulting in enhanced classification accuracy.

AVSC was first introduced in the DCASE 2021 Challenge [27], marking a significant development in the field and attracting much attention from the Audio and Acoustic Signal Processing (AASP) research community. The top-ranked system in the challenge [55] adopted a two-stage fine-tuning method to obtain robust visual representations from powerful pre-trained image models such as EfficientNet [56]. The authors in [57] proposed a multi-modal fusion method to address AVSC. It leveraged both audio and video models based on CNN variants, as well as CLIP-based networks [58] with multiple image encoders for feature extraction. A multi-branch model was introduced in [59], which was enhanced by contrastive event-object alignment and semantic-based fusion, resulting in competitive performance compared to state-of-the-art models in the AVSC domain. Unlike previous works that mainly employed CNN models, Zhou *et. al* [60] proposed an attentional graph convolutional network (AGCN) for AVSC. This AGCN was designed for structure-aware representation learning, providing an alternative approach to the task.

Indeed, many video encoders in recent works [27], [55], [57], [59], [61] have utilized pre-trained image models that do not fully exploit the temporal information present in video data, as they operate on individual image frames. In contrast, this paper introduces an innovative approach that leverages the inherent characteristics of scene transitions within video data. We introduce a visual segment model that is designed explicitly for semantic representation. This approach aims to address the limitations associated with relying solely on pre-trained image models for video analysis.

## III. VSM-based Hybrid Classification

In this section, we present our proposed VSM-based hybrid classification approach. The overall framework is shown in Fig. 1. It consists of three stages: visual segment modeling, latent semantic analysis and vector-based classification. In Section III-A, we elaborate on the VSM-based method for visual inventory generation. In Section III-B, we introduce the LSA technique to characterize each video clip as a feature vector. Finally, in Section III-C, we present the vector-based classification using a DNN classifier.

### A. Visual Segment Modeling

Similar to how speech utterances consist of phonemes, different video scene recordings are composed of fundamental units that exhibit internal correlations. Based on this understanding, we assume that the overall visual characteristics of video scene recordings can be well represented by a universal collection of fundamental units. This idea is similar to the notion that a video scene recording inherently contains a sequence of visual segments. Within scenes belonging to the same category, there exist dissimilar segments with differences in lighting, angles, objects, etc., which we call intra-class diversity. There also exist similar segments across categories, such as sky, buildings, etc., which denote inter-class similarity.

To solve these challenges, we propose a VSM-based method to build a visual inventory and then translate each video scene recording into a sequence of fundamental visual units defined in the inventory. The VSM training process consists of three key steps as shown in Fig. 1. (a). Deep models are adopted as visual feature extractors due to their powerful modeling ability in classification tasks. Unsupervised techniques, such as K-means clustering algorithm [62], are used in the initial segmentation stage to create an initial set of VSMs and transcripts. This stage utilizes changes in visual features to suggest potential boundaries between segments. With these initial VSMs, we can achieve a finer segmentation of each video scene recording through iterative training algorithms such as HMM training.

*1) Visual Feature Extraction:* The visual feature extractor $f_{\mathrm{v}}$ is derived from the pre-trained VGG-19 image encoder, which was trained on the on ImageNet [63]. This architecture consists of five convolution blocks, each followed by a max-pooling operation, and is further supplemented with two fully-connected (FC) layers, as illustrated in Fig. 2. Each convolution block contains several convolutional layers. Each convolutional layer (Conv) utilizes 2D kernels of size $3 \times 3$, along with batch normalization and rectified linear unit (ReLU) activation function. The number of channels increases sequentially from 64 to 128, 256, and 512 across the five convolution blocks. Max-pooling is performed using a $2 \times 2$ pixel window with a stride of 2. The two FC layers consist of 128 and 64 units, respectively, with ReLU non-linearities. The final FC layer performs $C$-way scene classification using softmax activation.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3595019



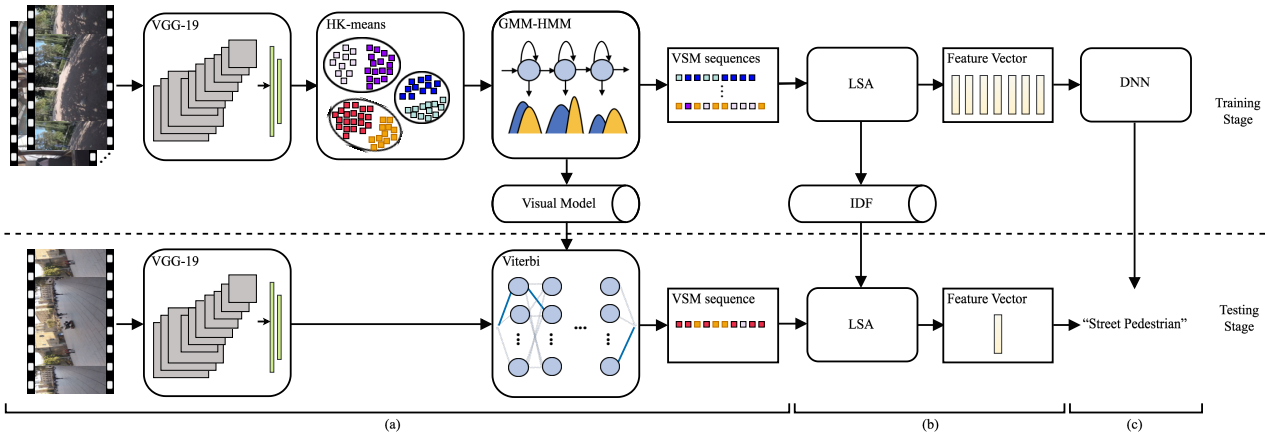



Fig. 1. The overall framework of the proposed VSM-based hybrid classification approach. (a) Visual segment modeling. (b) Latent semantic analysis. (c) Vector-based classification.

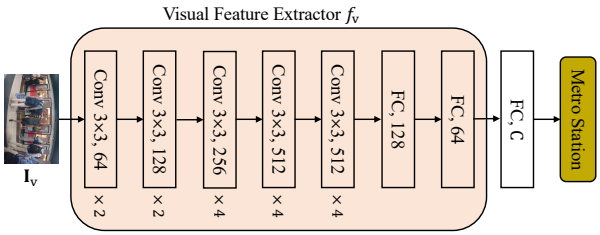



Fig. 2. An illustration of the visual feature extractor. The convolutional layer parameters are denoted as "Conv ⟨kenel size⟩, ⟨number of channels⟩".

Let the image corresponding to a video frame be $\mathbf{I}_\mathrm{v}$, where $\mathbf{I}_\mathrm{v} \in \mathbb{R}^{W_\mathrm{v} \times H_\mathrm{v} \times 3}$. The visual feature $\mathbf{E}_\mathrm{v} \in \mathbb{R}^{D_\mathrm{v}}$ is obtained as follows:

$$\mathbf{E}_\mathrm{v} = f_\mathrm{v}(\mathbf{I}_\mathrm{v}) \tag{1}$$

The convolutional layers in the visual feature extractor inherit their weights from the pre-trained VGG-19 network, while the FC layers are initialized with random weights. The visual feature extractor is then fine-tuned using a classification objective with the cross-entropy (CE) loss function:

$$\mathcal{L} = -\frac{1}{L}\sum_{l=1}^{L}\sum_{c=1}^{C} y_{l,c}\log\hat{y}_{l,c} \tag{2}$$

where $\hat{y}_{l,c}$ and $y_{l,c}$ are the posterior probability of the model output and the ground truth, respectively, for the $l$-th sample and the $c$-th category. $L$ is the number of samples in a mini-batch, and $C$ denotes the number of scene categories. The mini-batch size is set to 32, and the fine-tuning of the pre-trained weights is conducted with a relatively small learning rate of 1e-5. Then, 64-dimensional visual features are extracted from the fine-tuned extractor, as illustrated in Fig. 2.

*2) Initial Segmentation:* The initial segmentation process uses changes in visual features to obtain rough segment boundaries, which consists of two main steps: segmentation and quantization. The generation of initial VSM sequences is performed at the segment-level feature. In the segmentation step, an equal-segmentation method is used, where all training video data is divided into sequences of fixed-length segments with non-overlapping frames. Since the later VSM training involves an iterative procedure to redefine the optimal position of segment boundaries, we consider it is reasonable to sacrifice finer segmentation for faster processing at this stage. We calculate the arithmetic mean of visual features within a segment and use the resulting feature vector to represent the entire visual segment. The mean feature vector from each visual segment in every training recording is used to build a universal set of fundamental units. In this study, we conduct ablation experiments to determine the proper segment length.

It is expected that the observations of the same fundamental unit will be close in a certain metric, while observations of different units will be clearly separated, assuming that the features extracted from the video signal capture slowly changing visual characteristics. Therefore, methods like VQ [64] and GMM-HMM [49] can be employed to cluster visual segments into a small number of categories. This results in an initial collection of VSM units representing the whole visual space, as characterized by the training data. In the quantization step, three methods are adopted in this study: K-means clustering, hierarchical K-means (HK-means) clustering, and GMM-HMM based method.

K-means clustering [62] is one of the most widely used unsupervised algorithms for clustering analysis. Its main goal is to group similar data points into a user-defined number of clusters. In this study, the K-means algorithm is used to determine a set of $J$ centroids by minimizing the Euclidean distance using segment-level visual feature vectors. Video segments are then clustered into a small number of $J$ classes, where each class represents a VSM unit. As a result, each video recording is transformed into an initial VSM sequence by identifying the closest centroid for each visual segment.

The HK-means algorithm is a hybrid clustering method that combines both hierarchical clustering and K-means clustering. In our VSM-based framework, we incorporate the simple HK-means algorithm as described in [65] to speed up the clustering process and obtain stable centroids. In this work, we choose two hierarchy levels with $l = 2$. At each level of the clustering hierarchy, we adjust the number of clusters to $k_1$ and $k_2$. Using the standard K-means algorithm, we first

divide all segment-level visual features into $k_1$ clusters. Then, we iteratively cluster each set of points from the previous step into $k_2$ clusters. As a result, we generate a total of $J = k_1 \times k_2$ classes to represent the whole visual space, where the centroids are used to map video clips into initial VSM sequences.

The GMM-HMM based method, which is adopted for the initial segmentation and clustering of acoustic features in the ASM-based ASC framework [49], is also used as an alternative to cluster visual segments using hidden Markov chains. We use $C$ GMM-HMMs, each with a left-to-right HMM topology, to model $C$ types of video scenes. Each scene category is represented by a GMM-HMM with $S$ hidden states. The parameters of the GMM-HMMs are updated via the Baum-Welch algorithm [66]. All the hidden states of the GMM-HMMs collectively form a set of $J = C \times S$ VSM units jointly. Then, each video recording is decoded into a sequence of hidden states, forming what is known as a VSM sequence.

*3) Iterative Training:* With the initial VSM transcripts obtained in the previous stage, we can perform an iterative HMM training process to generate more refined segmentation results. Specifically, each VSM unit is characterized by an HMM using the training data and corresponding initial transcripts, where each HMM consists of two states. The probability density function for each state is represented by a GMM. After training the GMM-HMMs, a new VSM sequence is re-estimated for each video recording through Viterbi decoding. These re-estimated VSM sequences are then used as new transcripts for further GMM-HMM training. In general, a small number of re-estimation iterations is sufficient to achieve convergence. In this study, we use five iterations of HMM training. Finally, each video scene recording is transformed into a sequence of VSM units specified in the set of HMMs.

### B. Latent Semantic Analysis

We now represent each video clip as a sequence of VSM units, similar to how a transcription of a text document is represented as a sequence of words. This allows us to use text classification techniques such as LSA, an algebraic model for information retrieval proposed by Dumais *et. al* [67]. LSA uses statistical computation on large corpora of texts to extract the underlying semantic relationships between words. In this study, we adopt the LSA technique to transform a VSM sequence into a feature vector.

The goal of LSA is to represent the training set as a term-document matrix, in which each row corresponds to a unique VSM term and each column corresponds to a video recording. Given a collection of visual units denoted as $\mathrm{S} = \{\mathrm{s}_0, \mathrm{s}_1, ..., \mathrm{s}_{\mathrm{J}-1}\}$, where $J$ denotes the number of VSM units, the characteristics of the training set, which contains $M$ video recordings, can be represented as a term-document matrix $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_M)$. In this matrix, each component $\mathbf{h}_m$ describes the statistics related to both uni-gram terms and bi-gram terms. Due to the large number of non-existent bi-gram terms in the data set, the resulting dimension of $\mathbf{h}_m$, denoted as $N$, is much smaller than the theoretical value $J \times (J+1)$. The $n$-th VSM term in the $m$-th vector, denoted as $h_{n,m}$, contains the term frequency (TF) weighted by inverse document frequency (IDF), which is given as follows:

$$h_{n,m} = TF_{n,m} \times IDF_n \tag{3}$$

where TF represents the frequency with which a VSM term from its row appears in a scene recording denoted by its column. Meanwhile, IDF quantifies the informativeness of the VSM term. The term frequency of the $n$-th VSM term within the $m$-th scene recording, denoted as $TF_{n,m}$, and the inverse document frequency of the $n$-th VSM term, denoted as $IDF_n$, are expressed as follows:

$$TF_{n,m} = \frac{P_{n,m}}{\sum_{n=1}^{N} P_{n,m}} \tag{4}$$

$$IDF_n = \log \frac{M+1}{Q_n+1} \tag{5}$$

where $P_{n,m}$ denotes the number of occurrences of the $n$-th VSM term within the $m$-th scene recording. Concurrently, $M$ represents the total number of scene recordings in the training set, while $Q_n$ denotes the aggregate count of occurrences of the $n$-th VSM term throughout the entire training set.

Term weighting schemes, such as TF-IDF, provide a measure of indexing power. On one hand, terms that appear frequently in a few documents but rarely in others have high indexing power for those specific documents. On the other hand, terms that appear very frequently across all documents offer little indexing power. Examples of terms with maximum indexing power include proper nouns, such as names of individuals and countries, while function words like “a” and “the” have minimal indexing power.

In the testing phase, we firstly calculate the TF values of each transcribed video recording using Eq. (4), where the IDF values are obtained in the training phase. Then, the term-document matrix for test set, denoted as $\mathbf{H}_{\mathrm{test}}$, is calculated using Eq. (3) and used for classification.

### C. Vector-Based Classification

Using the LSA technique, each video recording is transformed into a feature vector, making VSC a vector-based classification problem. Several vector-based classifiers, such as SVM [68] and artificial neural network (ANN) [69], have been developed in the field. In this study, the feature vectors are fed into a DNN classifier. The DNN usually consists of five layers of nodes. The input layer accepts feature vectors, with the number of nodes equal to the total number of uni-gram and bi-gram terms present in the dataset. Three hidden layers consist of 1024, 512, and 128 nodes, respectively. For a classification task with $C$ classes, the output layer contains $C$ nodes. The ReLU and softmax activation functions are applied to the hidden and output nodes, respectively.

## IV. AVSM-based Multi-modal Classification

We expand upon the proposed VSM approach to create an AVSM approach by exploiting the complementarity of both audio and video modalities. In addition, we design a fusion model that takes both deep features and semantic features as input for scene category prediction.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3595019



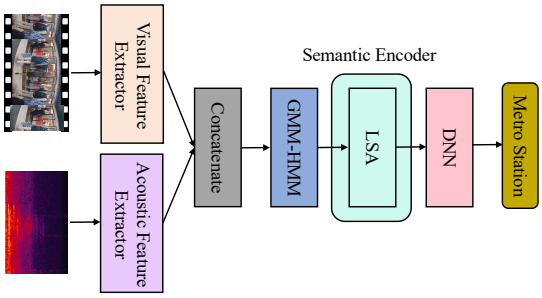



Fig. 3. The overall framework of the AVSM-based hybrid classification.

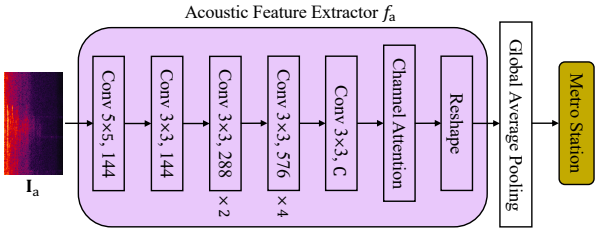



Fig. 4. An illustration of the acoustic feature extractor. The convolutional layer parameters are denoted as "Conv ⟨kernel size⟩, ⟨number of channels⟩".

### A. AVSM Model

As shown in Fig. 3, the proposed AVSM model consists of three stages, similar to the VSM model, but there are two main differences. First, whereas the VSM model utilizes only visual features as input, the AVSM model incorporates both visual and acoustic features to make use of the complementarity between video and audio data. Second, considering the enhanced representation of multi-modal features, we use GMM-HMM for the initial clustering of segments, without the need for iterative training in the AVSM model.

The visual feature extractor in the AVSM model keeps the same weights as in the VSM model, as shown in Fig. 2. For acoustic feature extraction, a fully convolutional neural network (FCNN) is adopted, which has demonstrated promising performance for the ASC task in our previous work [12]. The acoustic feature extractor $f_{\mathrm{a}}$, depicted in Fig. 4, consists of nine stacked convolutional layers and a channel attention module [70]. Each convolutional layer (Conv) is followed by batch normalization and a ReLU activation function. Dropout is used to alleviate overfitting. The number of channels in the last convolutional layer is equal to the number of scene classes.

We use log-Mel filter bank (LMFB) features of the audio data, with delta and delta-delta operations, forming the input $\mathbf{I}_{\mathrm{a}} \in \mathbb{R}^{T_{\mathrm{i}} \times F_{\mathrm{i}} \times C_{\mathrm{i}}}$. The acoustic feature $\mathbf{E}_{\mathrm{a}}$ is derived as follows:

$$\mathbf{E}_{\mathrm{a}} = f_{\mathrm{a}}(\mathbf{I}_{\mathrm{a}}) \tag{6}$$

where $\mathbf{E}_{\mathrm{a}} \in \mathbb{R}^{T_{\mathrm{a}} \times D_{\mathrm{a}}}$. To ensure that the frames of the acoustic and visual features are matched, we set $T_{\mathrm{i}} = 600$ for audio data with 10-second duration, and the time pooling size of the first convolutional layer is set to 2. We use $F_{\mathrm{i}} = 128$ and $C_{\mathrm{i}} = 6$ by default. The frequency pooling sizes for the first, second, fourth, and eighth convolutional layers are set to 2.

The channel attention module is an improved version of the Squeeze-and-Excitation block, incorporating two aggregation strategies to generate global spatial information. Suppose the input to the channel attention module is represented by $\mathbf{X}_{\mathrm{a}} \in \mathbb{R}^{T_{\mathrm{a}} \times F_{\mathrm{a}} \times C}$, the global spatial information, denoted as $\mathbf{c}_1$ and $\mathbf{c}_2$, can be obtained as follows:

$$\mathbf{c}_1 = \mathrm{GAP}(\mathbf{X}_{\mathrm{a}}) \tag{7}$$

$$\mathbf{c}_2 = \mathrm{GMP}(\mathbf{X}_{\mathrm{a}}) \tag{8}$$

where $\mathrm{GAP}(\cdot)$ and $\mathrm{GMP}(\cdot)$ denote global average pooling and global max pooling, respectively. Then, a simple gating mechanism with a sigmoid activation function is employed:

$$\boldsymbol{\alpha} = \sigma(\mathbf{W}_{21}\mathrm{g}(\mathbf{W}_{11}\mathbf{c}_1) + \mathbf{W}_{22}\mathrm{g}(\mathbf{W}_{12}\mathbf{c}_2)) \tag{9}$$

where $\mathrm{g}(\cdot)$ is the ReLU activation function, $\mathbf{W}_{11}, \mathbf{W}_{12} \in \mathbb{R}^{\frac{C}{r} \times C}$, and $\mathbf{W}_{21}, \mathbf{W}_{22} \in \mathbb{R}^{C \times \frac{C}{r}}$. Parameter $r$ denotes the reduction ratio. The output of the channel attention module is generated as follows:

$$\hat{\mathbf{X}}_{\mathrm{a}} = \mathbf{X}_{\mathrm{a}} \cdot \boldsymbol{\alpha} \tag{10}$$

Different from the visual feature extractor, the acoustic feature extractor is trained from scratch with a mini-batch size of 256. The initial learning rate is set to 1e-3 and will be reduced by $50\%$ if the accuracy does not improve for 20 consecutive epochs. We perform a reshaping operation on the output of the channel attention, resulting in the final acoustic features with a dimension equal to $D_{\mathrm{a}} = F_{\mathrm{a}} \times C = 80$.

To exploit the complementarity between video and audio data, we concatenate the features from both modalities. With this more robust acoustic-visual feature representation, we generate initial AVSM sequences using GMM-HMMs. These sequences are then fed into an algebraic model called LSA to extract latent semantic structures between AVSM units. Finally, a DNN with five FC layers is used for scene classification. The DNN classifier in the AVSM model shares the same parameters as those in the VSM model, with the only exception being the number of input layers.

### B. Acoustic-Visual-Semantic Fusion Model

Although AVSM sequences are derived from both acoustic and visual features, they are actually encoded in different formats. To better integrate these diverse modalities, we introduce an acoustic-visual-semantic fusion model (AVSFM) to generate unified representations of scene recordings. As shown in Fig. 5, the AVSFM model consists of three separate preprocessing modules and a shared DNN module.

*1) Preprocessing Module:* Audio and video data exhibit different characteristics. For example, audio is a one-dimensional (1D) continuous signal, while video is a three-dimensional (3D) continuous signal. Unlike audio and video data, an AVSM sequence is a 1D discrete signal with segment boundaries. Therefore, we introduce three separate preprocessing modules, namely acoustic, visual and semantic encoders.

Specifically, we extract LMFB features with delta and delta-delta operations for audio. For the acoustic encoder, we use a variant of the acoustic feature extractor shown in Fig. 4, where the time pooling size and frequency pooling size are both set to 2 for the first, second, fourth, and eighth convolutional

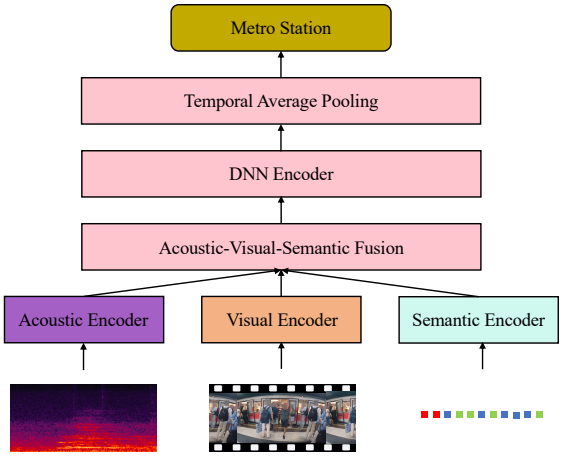



Fig. 5. An illustration of the AVSFM model, consisting of an acoustic preprocssing module, a visual preprocessing module, a semantic preprocssing module, and a shared DNN module.

layers. For the visual encoder, we use VGG-19 and flatten the visual embedding corresponding to each video frame into a 1D vector using 2D max pooling. In addition, we incorporate AVSM sequences into the classification model by introducing a semantic encoder. As shown in Fig. 3, we use the algebraic model known as LSA as our semantic encoder.

The acoustic, visual, and semantic encoders are trained separately. We extract $39 \times 128 \times 6$ LMFB features for a 1-second audio clip, which are then used to train the acoustic FCNN encoder shown in Fig. 4. Therefore, the acoustic embedding is computed as $\mathbf{z}_{\mathrm{a}} \in \mathbb{R}^{10\times 160}$ for audio with length of 10 seconds. The visual encoder is trained using all images in the video, with each image having a size of $224 \times 224 \times 3$. We select the first and fifteenth video frames within each second and combine their visual embeddings to obtain a representation for a 1-second video clip. The visual embedding is denoted as $\mathbf{z}_{\mathrm{v}} \in \mathbb{R}^{10\times 512}$ for video with length of 10 seconds. As shown in Fig. 3, we train the semantic encoder with AVSM sequences generated by GMM-HMMs. For audio-visual data with length of 10 seconds, we extract the semantic embedding $\mathbf{z}_{\mathrm{s}} \in \mathbb{R}^{1\times 200}$, with the number of AVSM units equal to 40.

*2) Backbone Network:* To effectively integrate different modalities for scene classification, we adopt an acoustic-visual-semantic fusion module that concatenates the vector representations of the acoustic, visual, and semantic data. The multi-modal representation $\mathbf{z}_{\mathrm{m}}$ is obtained by combining the acoustic embedding $\mathbf{z}_{\mathrm{a}}$, visual embedding $\mathbf{z}_{\mathrm{v}}$ and semantic embedding $\mathbf{z}_{\mathrm{s}}$ as follows:

$$\mathbf{z}_{\mathrm{m}} = \mathrm{Concat}(\mathbf{z}_{\mathrm{a}}, \mathbf{z}_{\mathrm{v}}, \mathbf{z}_{\mathrm{s}}) \tag{11}$$

where $\mathrm{Concat}(\cdot)$ denotes the concatenation operation. To synchronize the semantic embedding with the acoustic and visual embeddings in the time dimension, we repeat the semantic embedding appropriately. The resulting multi-modal representation $\mathbf{z}_{\mathrm{m}} \in \mathbb{R}^{10\times 872}$ is fed into a DNN encoder to learn discriminative features. This encoder consists of four FC layers with 512, 128, 64, and $C$ nodes, where $C$ denotes the number of scene categories. Finally, we make predictions through a temporal average pooling layer.

## V. Experiments

We evaluate the VSM-based hybrid classification method on two neural network models with significant differences in architectural design: VGG-19 [63] and UniformerV2 [24]. VGG-19 relies on convolution operations for feature extraction, which provides good local feature capturing capability but lacks the ability to model temporal information in videos. UniFormerV2, on the other hand, is a powerful video model that combines image-pretrained ViTs with an efficient UniFormer design. Besides, we conduct experiments on a spatiotemporal feature learning model named TimeSformer [71] and an audio-visual self-supervised learning model named MAViL [30].

### A. Experimental Setup

**Dataset.** The experiments were conducted on the TAU Audio-Visual Scenes 2021 (TAU AVS) dataset [27], designed for low-resource audio-visual scene classification. This dataset serves as the development set for the DCASE 2021 Challenge, specifically designed for AVSC. This dataset comprises synchronized audio and video recordings, each spanning 10 seconds, from 12 European cities. It contains a total of 7908 training clips, 740 validation clips, and 3645 test clips, all recorded in binaural format. The audio clips have a sampling rate of 48kHz and a resolution of 24-bit. The video clips have a frame rate of 30 frames per second (fps). The dataset comprises 10 scene classes: airport, shopping mall, metro station, pedestrian street, public square, street traffic, tram, bus, metro, and urban park. The evaluation metric used in this study is macro-average accuracy.

**VSM training on VGG-19.** In the training process of VSM on VGG-19, visual features with a size of 300 × 64 are used for data with 10-second duration. For vector-based classification, a DNN classifier is trained with a learning rate of 0.05 and a mini-batch size of 128. The parameters of the DNN are updated using the stochastic gradient descent (SGD) algorithm. The number of training epochs is set to 100. A dropout rate of 0.5, 0.3, and 0.3 is applied at the three hidden layers.

**VSM training on UniformerV2.** We first fine-tune the UniFormerV2-B/16 using the TAU Audio-Visual Scenes 2021 dataset to get the DL-based video model, which is pre-trained on Kinetics-400 [72] and operates on 8 × 224 × 224 video clips. The DL-based UniformerV2 is fine-tuned for 10 epochs with a learning rate of 1e-5 and then generates $L$=4×14×14 spatiotemporal tokens for a 10-second long video clip. Visual features with a size of 784 × 768 are used to train VSM models. For vector-based classification, a DNN classifier is trained with a learning rate of 0.001 and a mini-batch size of 128. The parameters of the DNN are updated using the Adam algorithm with a total epoch number of 200.

**Initial segmentation and iterative training.** We adopt three methods for initial segmentation. In K-means clustering, the parameter $J$ is set to the number of VSM units. When using the HK-means algorithm, we choose two hierarchy levels. The number of clusters in the first and second level is

TABLE I
AN ACCURACY (%) COMPARISON FOR VIDEO SCENE CLASSIFICATION AMONG VSM SYSTEMS FROM V1 TO V9 ON THE VALIDATION SET ($\mathrm{ACC_d}$) OF THE TAU AVS DATASET. ATTRIBUTES FOR COMPARISON INCLUDE: 1) INITIAL SEGMENTATION; 2) NUMBER OF VSM UNITS; 3) SEGMENT LENGTH.

| System | Initial Segmentation | | | Number of VSM Units | | | | | Segment Length | | | $\mathrm{ACC_d}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | K-means | GMM-HMM | HK-means | 20 | 40 | 60 | 80 | 100 | 2 | 3 | 6 | |
| V1 | ✓ | - | - | - | - | - | ✓ | - | - | ✓ | - | 79.0 |
| V2 | - | ✓ | - | - | - | - | ✓ | - | - | ✓ | - | 80.7 |
| V3 | - | - | ✓ | - | - | - | ✓ | - | - | ✓ | - | 80.7 |
| V4 | - | - | ✓ | ✓ | - | - | - | - | - | ✓ | - | 80.2 |
| V5 | - | - | ✓ | - | ✓ | - | - | - | - | ✓ | - | 80.9 |
| V6 | - | - | ✓ | - | - | ✓ | - | - | - | ✓ | - | 81.6 |
| V7 | - | - | ✓ | - | - | - | - | ✓ | - | ✓ | - | 79.7 |
| V8 | - | - | ✓ | - | - | ✓ | - | - | ✓ | - | - | 80.4 |
| V9 | - | - | ✓ | - | - | ✓ | - | - | - | - | ✓ | 81.2 |

set to $k_1 = 20$ and $k_2 = J/k_1$, respectively. In GMM-HMM, the number of hidden states in each HMM is determined by $S = J/C$, where $J$ is the number of VSM units and $C$ is the number of scene categories. The emission probability of each state is modeled by a GMM with 80 mixtures. During the iterative training step, each HMM consists of two states, and each state is associated with 40 Gaussian mixtures. With a relatively small number of clusters, hidden states, and Gaussian mixtures, the computational complexity of the proposed VSM approach on the TAU AVS dataset would not be excessively high.

**Uni-modal DL models.** For comparison, we train uni-modal audio and video models. The DL-based audio model (Audio-DL) employs a FCNN to capture local temporal dependencies and is trained from scratch with 1-second long data. The DL-based video model (Video-DL) is fine-tuned with three networks: VGG-19, TimeSformer, and UniformerV2. Initialized with weights learned from Kinetics-400 [72], both TimeSformer and UniformerV2 are fine-tuned on 8 × 224 × 224 video clips. An additional DNN module that share the same parameters as shown in Fig. 5 is employed for VGG-19, while for TimeSformer and UniformerV2, only a linear layer is employed. Both TimeSformer and UniformerV2 are trained for 10 epochs with a learning rate of 1e-5 and a mini-batch size of 10 in the fine-tuning stage. When evaluating the FCNN and VGG-19, we split the test data into ten non-overlapping segments and obtain the final prediction by averaging the results of each segment according to [27]. VGG-19 cannot utilize temporal information with image frames as input.

**AVSM and AVSFM training.** To ensure the temporal synchronization between the acoustic and visual features during AVSM training, acoustic features are extracted with a size of 300 × 80 (as shown in Fig. 4). The network architecture and training hyperparameters of the DNN classifier are the same as those set in VSM. The parameters of Audio-DL FCNN and Video-DL VGG-19 are used to initialize the acoustic and visual encoders in AVSFM, respectively. It is worth noting that the acoustic feature extractor in Fig. 3 is trained with 10-second long data, while the acoustic encoder in Fig. 5 is trained with 1-second long data. When training the AVSFM model, the weights of the acoustic and visual encoders are fine-tuned. The models are trained using the PyTorch toolkit, and the Adam optimizer is used during training.

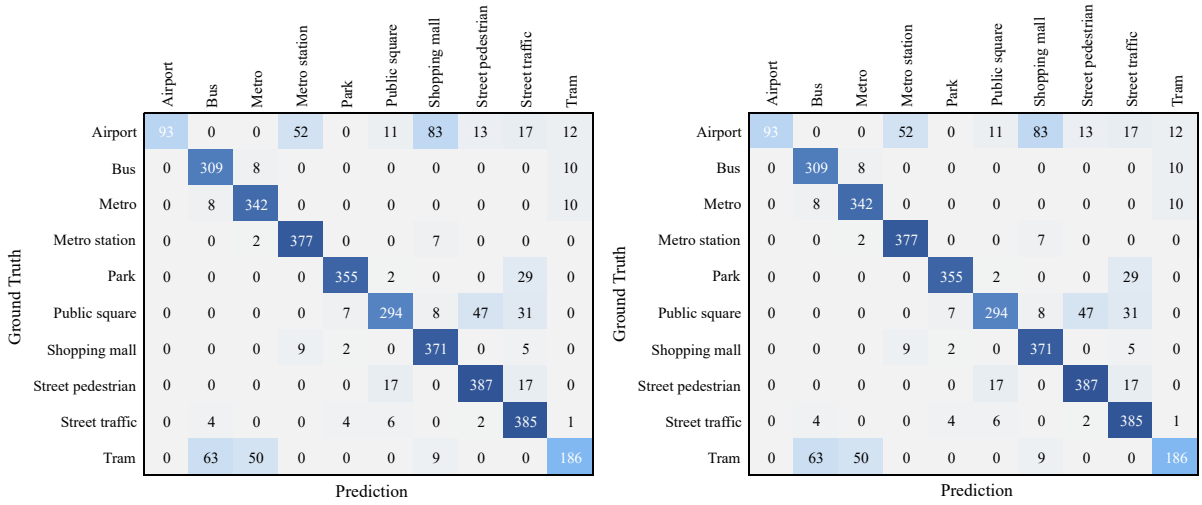
**MAViL training.** MAViL [30], [73] is a self-supervised learning model that achieves state-of-the-art performance on AudioSet [74] and VGGSound [75]. The MAViL pre-trained on AudioSet is adopted in this study. Following [30], Mel-frequency spectrogram with dimension of 1024 × 128 is extracted for the audio, and RGB frames with a size of 8 × 224 × 224 are extracted for the video. During fine-tuning, MAViL is trained for 60 epochs, with the number of warm-up epochs equal to 4. A Half-cycle cosine decay learning rate schedule is used according to [30], and the base learning rate is set to 0.001. The mini-batch size is set to 10.

### *B. Effectiveness of VSM-Based Approach*

To evaluate the effectiveness of each module, we conduct ablation experiments with different configurations related to VSM training. The performance comparison for video scene classification on the TAU Audio-Visual Scenes 2021 dataset among VSM models is presented in Table I.

**Ablations on initial segmentation.** The purpose of initial segmentation is to divide video segments with similar characteristics into the same cluster, thereby obtaining a rough collection of visual units known as VSM units. In order to evaluate the effect of different initialization methods, we design three VSM systems (denoted as V1, V2 and V3) as shown in Table I. When GMM-HMM or HK-means are used, the models (V2 and V3) demonstrate better and similar performance. This shows that an appropriate initialization can help build a more accurate visual inventory, which leads to higher classification accuracy. Given that HK-means clustering is faster than GMM-HMM, we use HK-means for initial segmentation in the VSM training process.

**Ablations on the number of VSM units.** A key parameter in modeling visual segments is the number of VSM units, which indicates the number of clusters for all video segments ($J$ as mentioned in Section III-A). The number of VSM units determines the coverage of the visual space. Too few VSM units would not sufficiently capture the variation of visual events in feature vectors, while too many would lead to redundant information and increased computational complexity. To investigate the impact of the number of VSM units, we design systems V3 to V7, as shown in Table I. Notably, we observe




(a) Video-DL (b) VSM

Fig. 6. Confusion matrices for the Video-DL and VSM models using VGG-19 on the test set of the TAU AVS dataset.

TABLE II
CLASS-WISE ACCURACY (%) COMPARISON FOR VIDEO-DL AND VSM MODELS ON THE TEST SET ($\mathrm{ACC_t}$) OF THE TAU AVS DATASET.

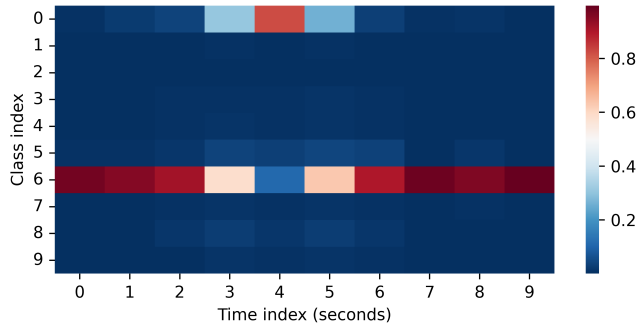
| Class | $\mathrm{ACC_t}$ | |
|---|---|---|
| | Video-DL | VSM |
| Airport | 33.1 | 60.1 |
| Bus | 94.5 | 94.5 |
| Metro | 95.0 | 96.7 |
| Metro station | 97.7 | 96.6 |
| Park | 92.0 | 89.1 |
| Public square | 76.0 | 81.7 |
| Shopping mall | 95.9 | 95.1 |
| Street pedestrian | 91.9 | 90.0 |
| Street traffic | 95.8 | 95.8 |
| Tram | 60.4 | 69.8 |
| Overall | 85.0 | 88.0 |

that the visual segment model trained with a moderate number of VSM units shows great effectiveness in scene classification. Specifically, using 60 VSM units yields the highest accuracy of 81.6% on the validation set. It is worth noting that in system V6, HK-means is used to generate initial VSM sequences, where the number of clusters at the two levels is set to $k_1 = 20$ and $k_2 = 3$.

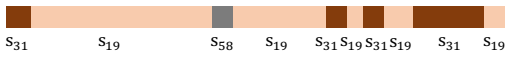
**Ablations on segment length.** The generation of VSM sequences is based on segment-level features. To determine the optimal segment length, we design VSM systems V6, V8, and V9, as shown in Table I. As the segment length increases from 2 to 3 image frames, performance improves from 80.4% to 81.6%, which indicates that longer segments may contain distinctive visual features beneficial for scene classification. However, when we further increase the segment length to 6 image frames, performance drops to 81.2%. This suggests that segments comprising too many image frames may include rapidly changing visual characteristics, making them unsuitable as fundamental units for modeling.

**Confusion matrix.** From the experimental results in Table I, we can see that the VSM system V6 achieves the best accuracy on the validation set. The class-wise accuracy comparison of V6 and the video model using VGG-19 on the test set is shown in Table II. We can find that the VSM model outperforms the Video-DL model in general, with an accuracy improvement of 27.0% for the airport category and 9.4% for the tram category. To further analyze the prediction results, we show the confusion matrix in Fig. 6. It can be observed that the VSM model is more suitable to handle easily confused categories, such as airport and tram. As shown in Fig. 6, the Video-DL model misclassifies samples of the airport category into metro station and shopping mall, while the VSM model correctly classifies them.

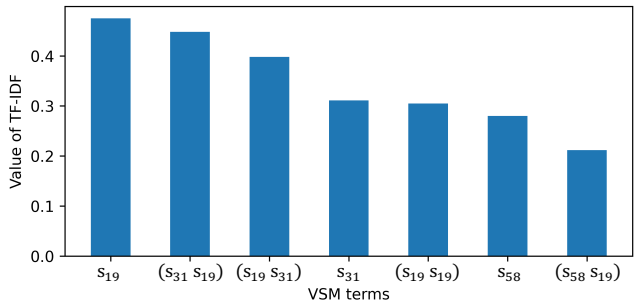
**Case study.** Fig. 7 shows the intermediate prediction results of the Video-DL and VSM models for a test sample from the airport scene (Class '0'). The sample is misclassified by the Video-DL model as the shopping mall scene (Class '6'), but correctly classified by the VSM model. As shown in Fig. 7(a), nine out of ten 1-second long segments are classified by the Video-DL model as the shopping mall category, which leads to the wrong prediction class for this sample when using temporal average pooling. The main reason for this misclassification may be the existence of similar image frames in both video scene categories. Besides, the decoded VSM sequence of this sample with the explicit segment length is shown in Fig. 7(b). The sample is transcribed by the visual inventory consisting of 60 VSM units, labeled from '$\mathrm{s}_0$' to '$\mathrm{s}_{59}$'. Several uni-gram and bi-gram VSM terms with high TF-IDF values of the airport sample are shown in Fig. 7(c). It can be seen that '$\mathrm{s}_{19}$', '$\mathrm{s}_{31}$', and their bi-gram terms achieve higher values, indicating their strong indexing power. We calculated the VSM terms with the top five highest TF-IDF values for the test samples labeled airport and shopping mall categories, respectively. For the airport scene, the five VSM terms are '$\mathrm{s}_{59}$', '($\mathrm{s}_{19}$ $\mathrm{s}_{19}$)', '$\mathrm{s}_{38}$', '($\mathrm{s}_{59}$ $\mathrm{s}_{59}$)', and '$\mathrm{s}_{19}$'. For the shopping mall scene, the five VSM terms are '$\mathrm{s}_{30}$', '$\mathrm{s}_5$', '$\mathrm{s}_{15}$', '($\mathrm{s}_5$ $\mathrm{s}_{31}$)', and '($\mathrm{s}_{31}$ $\mathrm{s}_{30}$)'. It shows that different VSM terms are used to characterize these




(a) The posterior probability predicted by the Video-DL model.




(b) The VSM sequence generated by the visual segment model.




(c) The TF-IDF value of several VSM terms.

Fig. 7. Intermediate prediction results of a test sample from the airport scene by the Video-DL and VSM models based on VGG-19.

two types of video scenes. The VSM terms with high indexing power for this sample better match the airport scene category, leading to the correct prediction. This proves the effectiveness of the VSM-based approach in video scene classification.

**Generalization of VSM.** To compare with other state-of-the-art models and demonstrate the generalization ability of the VSM-based method, we conducted several experiments using two ViT variants, namely TimeSformer [71] and UniformerV2 [24]. Experimental results are shown in Table III. Compared to VGG-19, the Video-DL models achieve significant improvements when using TimeSformer and UniFormerV2 as backbone networks, mainly because they can learn temporal dependencies. Based on the visual features extracted by the best UniFormerV2-DL, applying the VSM-based hybrid classification method (as shown in the fifth row) results in further accuracy improvement, demonstrating the generalizability of our VSM approach. Additionally, when training the model by combining both the deep features extracted by Video-DL and the fuzzy semantic features extracted by VSM, as shown in Fig. 5 without the acoustic encoder, we observe an accuracy of 92.7%. This indicates the strong complementarity between deep features and the term-document matrix.

**Visualization of VSM units.** To enhance comprehension of the proposed method, we provide visualization of representative video frames from several VSM units, as shown in Fig. 8. Fig. 8.(a) represents a video frame from a park scene clustered as VSM unit '$s_4$'. Fig. 8.(b) and Fig. 8.(c) show two video frames from the VSM units '$s_{19}$' and '$s_{30}$', respectively. These

TABLE III
AN ACCURACY (%) COMPARISON FOR VSC AMONG SEVERAL VISUAL SYSTEMS ON THE TEST SET OF THE TAU AVS DATASET.

| System | Network | Visual Encoder | Semantic Encoder | $ACC_t$ |
|---|---|---|---|---|
| Video-DL | VGG-19 | ✓ | - | 85.0 |
| Video-DL | TimeSformer | ✓ | - | 89.5 |
| Video-DL | UniformerV2 | ✓ | - | 91.7 |
| VSM | VGG-19 | - | ✓ | 88.0 |
| VSM | UniformerV2 | - | ✓ | 92.1 |
| Video-DL + VSM | UniformerV2 | ✓ | ✓ | 92.7 |

(a) $s_4$ (b) $s_{19}$ (c) $s_{30}$ (d) $s_{47}$

Fig. 8. Visualization of representative video frames from several VSM units.

two video frames are from airport and shopping mall scenes, respectively, which is consistent with the related analysis in Fig. 7. The fourth figure denotes a representative video frame from the VSM unit '$s_{47}$' labeled as the bus scene class.

### C. Effectiveness of AVSM-Based Approach

**Ablations on AVSM.** By exploiting the synergy of audio and video modalities, we extend the proposed VSM-based hybrid classification to an AVSM-based approach. Different from the VSM model that needs an iterative training procedure, the AVSM model only uses an initial clustering method GMM-HMM to build the audio-visual inventory. Table IV shows the experimental results of AVSM systems trained with different settings. To select the number of AVSM units, we design systems S1 to S4. We find that the audio-visual segment model performs better when using 40 AVSM units, which is fewer than the number of VSM units. To evaluate the effect of segment length, we design systems S1, S5, and S6. It is observed that a moderate-long segment consisting of 2 image frames and 4 acoustic frames is more appropriate for modeling as an audio-visual fundamental unit. By comparing Table II and IV, we find that the AVSM system improves performance by 3.2% (from 88.0% to 91.2%) on the test set with fewer fundamental units and a shorter segment length, which may benefit from the audio modality.

**T-SNE visualization of hidden embedding.** In Fig. 9, we show the t-SNE visualization [76] of the hidden layer embedding for scene classification on the test set of VSM and AVSM models. Generally, the embedding of both VSM and AVSM models are clustered for different video scene

TABLE IV
AN ACCURACY (%) COMPARISON FOR AVSC AMONG AVSM SYSTEMS FROM S1 TO S6 ON THE VALIDATION ($ACC_d$) AND TEST ($ACC_t$) SETS OF THE TAU AVS DATASET, RESPECTIVELY. ATTRIBUTES FOR COMPARISON INCLUDE: 1) NUMBER OF AVSM UNITS; 2) SEGMENT LENGTH.

| System | Number of AVSM Units | | | | Segment Length | | | $ACC_d$ | $ACC_t$ |
|---|---|---|---|---|---|---|---|---|---|
| | 40 | 60 | 80 | 100 | 1 | 2 | 3 | | |
| S1 | ✓ | - | - | - | ✓ | - | - | 87.4 | 91.0 |
| S2 | - | ✓ | - | - | ✓ | - | - | 87.4 | 90.7 |
| S3 | - | - | ✓ | - | ✓ | - | - | 85.3 | 90.9 |
| S4 | - | - | - | ✓ | ✓ | - | - | 85.9 | 90.8 |
| S5 | ✓ | - | - | - | - | ✓ | - | 87.9 | 91.2 |
| S6 | ✓ | - | - | - | - | - | ✓ | 86.1 | 90.0 |

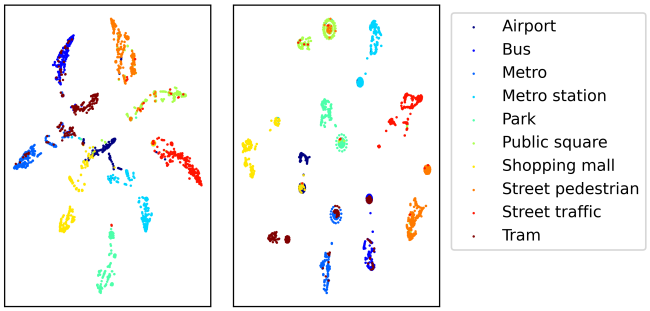



Fig. 9. Embedding visualization of the VSM (left) and AVSM (right) models for scene classification.

categories. However, the classification accuracy of VSM model is unsatisfactory for scene categories with similar video features. For example, 'airport' and 'metro station' categories are clustered together, as are 'airport' and 'shopping mall' categories. This clustering results in a low accuracy of 60.1% for the airport class in Table II. Considering that there may be differences in the sounds recorded in these video scene recordings, we propose using the audio modality to differentiate between various scene categories, thus forming an AVSM system. As illustrated in Fig. 9, the embedding of the 'airport' class is more distinctly separated from that of the 'metro station' and 'shopping mall' classes in AVSM system, resulting in a improved accuracy of 84.7% for the airport class.

**Comparison of uni-modal and multi-modal systems.** Table V lists the accuracy comparison for scene classification among several uni-modal and multi-modal systems on the TAU AVS dataset. 'Audio-DL' denotes the deep learning-based audio model using FCNN as the acoustic encoder. 'Video-DL' denotes the deep learning-based video model using VGG-19 as the visual encoder. 'AV-DL' combines Audio-DL with Video-DL, corresponding to the model in Fig. 5 with the semantic encoder removed. Several observations can be made. Firstly, the Video-DL outperforms the Audio-DL by a large margin, benefiting from the well pre-trained image encoder. We also experimented with pre-trained audio encoders on Audioset [74], such as PANN [77] and CNN [78]. However, neither of them performs better than the FCNN model trained from scratch. One possible reason is the pre-trained models do not perform well enough on AudioSet currently. The mean average precision (mAP) of PANN and CNN on AudioSet are 0.439 and 0.314 respectively, while the classification accuracy of VGG-19 network on ImageNet is 82.7%.

TABLE V
AN ACCURACY (%) COMPARISON FOR SCENE CLASSIFICATION AMONG SEVERAL UNI-MODAL AND MULTI-MODAL SYSTEMS ON THE TEST SET ($ACC_t$) OF THE TAU AVS DATASET.

| System | Acoustic Encoder | Visual Encoder | Semantic Encoder | $ACC_t$ |
|---|---|---|---|---|
| Audio-DL | FCNN | - | - | 74.3 |
| Video-DL | - | VGG-19 | - | 85.0 |
| AV-DL | FCNN | VGG-19 | - | 92.2 |
| MAViL | ViT-B | ViT-B | - | 91.3 |
| MAViL* | ViT-B | ViT-B | - | 86.4 |
| AVSM | - | - | ✓ | 91.2 |
| AVSFM | FCNN | VGG-19 | ✓ | 93.0 |

* Fixing pre-trained weights of MAViL.

TABLE VI
AN ACCURACY (%) COMPARISON WITH STATE-OF-THE-ART METHODS FOR THE AVSC TASK ON THE DEVELOPMENT SET ($ACC_t$) OF THE DCASE 2021 CHALLENGE.

| System Idx. | Acoustic Encoder | Visual Encoder | Semantic Encoder | $ACC_t$ |
|---|---|---|---|---|
| 0 [27] | OpenL3 | OpenL3 | - | 84.8 |
| 1 [60] | ResNet | ResNet | - | 91.6 |
| 2 [57] | EfficientNet | CLIP ViT | - | 93.3 |
| 3 [61] | CNN&Transformer | CvT | - | 93.9 |
| 4 [59] | Transformer | ConvNeXt | - | 94.1 |
| 5 [79] | DCNN | EfficientNetV2 | - | 94.6 |
| 6 Ours | FCNN | ResNeSt | LSA | 94.8 |

Secondly, MAViL performs slightly worse than the AV-DL model. The likely reason is that the large number of parameters in MAViL leads to overfitting when training on low-resource audio-visual scene classification task. When fixing the pre-trained weights of MAViL, accuracy drops from 91.3% to 86.4%. Thirdly, by incorporating both audio and video modalities, AVSM can further enhance performance, yielding an accuracy only marginally lower than that of the AV-DL model. This underscores the complementary nature of video and audio modalities. It is observed that MAViL and AVSM achieve very similar results, specifically 91.3% and 91.2%, respectively. Fourthly, by integrating the AVSM model with the AV-DL model, the resulting AVSFM model improves performance from 92.2% to 93.0%. Notably, the AVSFM model's accuracy for the airport class escalates from 84.7% to 90.7% compared to the AVSM model, demonstrating the synergistic effect of deep learning-based and fuzzy semantics-based features.

### *D. A Comparison with State-of-the-art Methods*

We compare our method with the state-of-the-art methods on the audio-visual scene classification task of the DCASE 2021 Challenge. The results shown in Table VI are from single systems without model ensemble. Note that we evaluate this subsection on 1-second long clip instead of 10-second long clip according to official setup. System Idx.0 is the baseline, which uses OpenL3 network as both the acoustic and visual encoders. Most multi-modal systems use both acoustic and visual encoders, which are shown from system Idx.1 to

Idx.5 in Table VI. The majority of the visual encoders use pre-trained image models without considering the temporal relationship in video. The state-of-the-art system [79] utilizes long-term scalogram as the acoustic feature. Our proposed system achieves the best performance of 94.8% by introducing an LSA-based semantic encoder, resulting in a 0.2% accuracy improvement over [79]. By exploiting the inherent characteristics of scene transitions within video data, the proposed VSM approach produces a sequence of segment units that are subsequently input into the semantic encoder.

## VI. Conclusion

This study primarily focuses on visual segment modeling for low-resource video scene classification. The proposed VSM-based approach is designed to translate each video recording into a sequence of fundamental, units under the assumption that the overall visual characteristics of all scene categories can be represented by a universal inventory of visual units. The inventory is generated through initial segmentation and iterative training, followed by an LSA technique to transform the VSM sequence into a fuzzy semantics-based feature vector. In addition to the VSM-based approach, we further extend it to the AVSM-based approach by leveraging the complementary nature of audio and video modalities. Furthermore, we integrate deep learning-based feature and fuzzy semantics-based feature to propose an audio-visual-semantic fusion model architecture. By evaluating our proposed approaches on the data set of the DCASE 2021 Challenge, we have greatly improved the performance of audio-visual scene classification, achieving state-of-the-art results using a single system.

In future work, we intend to explore the use of GPT-4 [80], a powerful large language model, to generate video attributes with clear semantics. This auxiliary information can be combined with the fuzzy semantics-based feature generated by the proposed VSM-based approach.

## VII. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62401533, and in part by the National Natural Science Foundation of China under Grant No. 62171427.

## References

[1] C. Landone, J. Harrop, and J. Reiss, "Enabling access to sound archives through integration, enrichment and retrieval: The EASAIER project," in *Proc. Int. Soc. Music Inf. Retrieval*, 2007, pp. 159–160.

[2] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2006, pp. 885–888.

[3] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1142–1158, 2009.

[4] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1684–1689, 2012.

[5] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 142–153, 2014.

[6] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. IEEE Eur. Signal Process. Conf.*, 2016, pp. 1128–1132.

[7] D. A. Reynolds *et al.*, "Gaussian mixture models," *Encyclopedia of biometrics*, vol. 741, no. 659-663, 2009.

[8] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, 1998.

[9] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature learning with matrix factorization applied to acoustic scene classification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 6, pp. 1216–1229, 2017.

[10] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 2, pp. 379–393, 2017.

[11] Z. Ren, Q. Kong, J. Han, M. D. Plumbley, and B. W. Schuller, "CAA-Net: Conditional atrous CNNs with attention for explainable device-robust acoustic scene classification," *IEEE Trans. Multimedia*, vol. 23, pp. 4131–4142, 2020.

[12] H. Hu, C.-H. H. Yang, X. Xia, X. Bai, X. Tang, Y. Wang, S. Niu, L. Chai, J. Li, H. Zhu *et al.*, "A two-stage approach to device-robust acoustic scene classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 845–849.

[13] L. Xie, F. Lee, L. Liu, K. Kotani, and Q. Chen, "Scene recognition: A comprehensive survey," *Pattern Recognit.*, vol. 102, p. 107205, 2020.

[14] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, pp. 145–175, 2001.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[16] Y. Jiang, J. Yuan, and G. Yu, "Randomized spatial partition for scene recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 730–743.

[17] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, and X. Jiang, "Learning discriminative and shareable features for scene classification," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 552–568.

[18] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, 2018.

[19] J. Huang, Z. Liu, and Y. Wang, "Joint scene classification and segmentation based on hidden Markov model," *IEEE Trans. Multimedia*, vol. 7, no. 3, pp. 538–550, 2005.

[20] M. Israël, E. L. van den Broek, P. van der Putten, and M. J. den Uyl, "Automating the construction of scene classifiers for content-based video retrieval," in *Proc. Int. Workshop Multimedia Data Mining*. Seattle, WA, USA, 2004, pp. 38–47.

[21] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 712–727, 2008.

[22] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6202–6211.

[23] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 6824–6835.

[24] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, L. Wang, and Y. Qiao, "Uniformerv2: Unlocking the potential of image vits for video understanding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 1632–1643.

[25] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *Int. J. Automat Comput.*, vol. 18, pp. 351–376, 2021.

[26] D. Hu, X. Li, L. Mou, P. Jin, D. Chen, L. Jing, X. Zhu, and D. Dou, "Cross-task transfer for geotagged audiovisual aerial scene recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 68–84.

[27] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *Porc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 626–630.

[28] B. Shi, W.-N. Hsu, K. Lakhotia, and M. Abdelrahman, "Learning audio-visual speech representation by masked multimodal cluster prediction[," in *Proc. Int. Conf. Learn. Representations*, 2022.

[29] B. Mocanu, R. Tapu, and T. Zaharia, "Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning," *Image Vis. Comput.*, vol. 133, p. 104676, 2023.

[30] P.-Y. Huang, V. Sharma, H. Xu, C. Ryali, Y. Li, S.-W. Li, G. Ghosh, J. Malik, C. Feichtenhofer *et al.*, "Mavil: Masked audio-video learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.

[31] C.-H. Lee, F. K. Soong, and B.-H. Juang, "A segment model based approach to speech recognition," in *Porc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1988.

[32] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Processes*, vol. 25, no. 2-3, pp. 259–284, 1998.

[33] Q. Wang, J. Du, S. Zheng, Y. Li, Y. Wang, Y. Wu, H. Hu, C.-H. H. Yang, S. M. Siniscalchi, Y. Wang, and C.-H. Lee, "A study on joint modeling and data augmentation of multi-modalities for audio-visual scene classification," in *Proc. Int. Symp. Chinese Spoken Lang. Process.*, 2022, pp. 453–457.

[34] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. Eur. Conf. Mach. Learn.*, 1998, pp. 137–142.

[35] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *Proc. AAAI Workshop Learn. Text Categorization*, vol. 752, no. 1. Madison, WI, 1998, pp. 41–48.

[36] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 3, 2003, pp. 1470–1470.

[37] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Stat. Learn. Comput. Vis.*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.

[38] K. Kesorn and S. Poslad, "An enhanced bag-of-visual word vector space model to represent visual content in athletics images," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 211–222, 2011.

[39] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, 2016.

[40] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, and M. Cord, "Learning representations by predicting bags of visual words," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6928–6938.

[41] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 523–535, 2010.

[42] A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf

[43] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu, "Vector-quantized image modeling with improved VQGAN," in *Proc. Int. Conf. Learn. Representations*, 2022.

[44] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, and D. Erhan, "Phenaki: Variable length video generation from open domain textual descriptions," in *Proc. Int. Conf. Learn. Representations*, 2022.

[45] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu, A. G. Hauptmann *et al.*, "Language model beats diffusion–tokenizer is key to visual generation," *arXiv preprint arXiv:2310.05737*, 2023.

[46] B. Ma, H. Li, and C.-H. Lee, "An acoustic segment modeling approach to automatic language identification," in *Proc. Eur. Conf. Speech Commun. Technol.*, 2005.

[47] J. Reed and C.-H. Lee, "A study on music genre classification based on universal acoustic models." in *Proc. Int. Soc. Music Inf. Retrieval*, 2006, pp. 89–94.

[48] Y. Tsao, H. Sun, H. Li, and C.-H. Lee, "An acoustic segment model approach to incorporating temporal information into speaker modeling for text-independent speaker recognition," in *Porc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2010, pp. 4422–4425.

[49] X. Bai, J. Du, Z.-R. Wang, and C.-H. Lee, "A hybrid approach to acoustic scene classification based on universal acoustic models." in *Proc. INTERSPEECH*, 2019, pp. 3619–3623.

[50] H. Hu, S. M. Siniscalchi, Y. Wang, X. Bai, J. Du, and C.-H. Lee, "An acoustic segment model based segment unit selection approach to acoustic scene classification with partial utterances," in *Proc. INTERSPEECH*, 2020, pp. 1201–1205.

[51] X. Bai, J. Du, J. Pan, H.-s. Zhou, Y.-H. Tu, and C.-H. Lee, "High-resolution attention network with acoustic segment model for acoustic scene classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 656–660.

[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 248–255.

[53] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, 2017.

[54] L. Pham, D. Ngo, T. Nguyen, P. Nguyen, T. Hoang, and A. Schindler, "An audio-visual dataset and deep learning frameworks for crowded scene classification," in *Proc. Int. Conf. Content Multimedia Indexing*, 2022, pp. 23–28.

[55] S. Wang, T. Heittola, A. Mesaros, and T. Virtanen, "Audio-visual scene classification: analysis of DCASE 2021 challenge submissions," in *Proc. Detection Classication Acoust. Scenes Events*, 2021, pp. 45–49.

[56] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[57] S. Okazaki, Q. Kong, and T. Yoshinaga, "A multi-modal fusion approach for audio-visual scene classification enhanced by CLIP variants," in *Proc. Detection Classication Acoust. Scenes Events*, 2021, pp. 95–99.

[58] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[59] Y. Hou, B. Kang, and D. Botteldooren, "Audio-visual scene classification via contrastive event-object alignment and semantic-based fusion," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, 2022, pp. 1–6.

[60] L. Zhou, Y. Zhou, X. Qi, J. Hu, T. L. Lam, and Y. Xu, "Attentional graph convolutional network for structure-aware audiovisual scene classification," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–15, 2023.

[61] Y. Yang and Y. Luo, "Scene classification using acoustic and visual feature," *Tech. Rep., DCASE 2021 Challenge*, 2021.

[62] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *J. R. Stat. Soc. C-Appl.*, vol. 28, no. 1, pp. 100–108, 1979.

[63] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.

[64] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, 1980.

[65] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7464–7473.

[66] D. Elworthy, "Does baum-welch re-estimation help taggers?" in *Proc. Conf. Appl. Natural Lang. Process.*, 1994, pp. 53–58.

[67] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman, "Using latent semantic analysis to improve access to textual information," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 1988, pp. 281–285.

[68] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, 2002.

[69] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1998.

[70] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[71] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 813–824.

[72] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

[73] Y. Tseng, B. Berry, Y.-T. Chen, I.-H. Chiu, H.-H. Lin, M. Liu, P. Peng, Y.-J. Shih, H.-Y. Wang, H. Wu, P.-Y. Huang, C.-M. Lai, S.-W. Li, D. Harwath, Y. Tsao, A. Mohamed, C.-L. Feng, and H.-y. Lee, "Av-superb: A multi-task evaluation benchmark for audio-visual representation models," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2024, pp. 6890–6894.

[74] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 776–780.

[75] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGGSound: A large-scale audio-visual dataset," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 721–725.

[76] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.

[77] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.

[78] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 131–135.
[79] C. Chen, M. Wang, and P. Zhang, "Audio-visual scene classification using a transfer learning based joint optimization strategy," *arXiv preprint arXiv:2204.11420*, 2022.
[80] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

**Qing Wang** (Member, IEEE) received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2012 and 2018, respectively. From July 2018 to February 2020, she worked at Tencent company on single-channel speech enhancement. From March 2020 to February 2023, she was a Postdoctor at USTC. She is currently an Associate Researcher at USTC. Her research interests include speech enhancement, robust speech recognition, acoustic scene classification, sound event localization and detection.

**Yajian Wang** received the B.Eng. degree from the School of Information Science and Engineering, Shandong University, Qingdao, China, in 2020, and the M.Sc. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2023. Her research interests include acoustic scene classification, and audio-visual scene classification.

**Hang Chen** (Member, IEEE) received his B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2018 and 2024, respectively. He is currently a Postdoctoral Researcher with USTC. His research interests include audio-visual speech enhancement and recognition.

**Shuxian Wang** received the B.Eng. degree from the School of Information Science and Engineering, Shandong University, Qingdao, China, in 2021, and the M.Sc. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2024. Her research interests include acoustic scene classification, and unsupervised anomalous sound detection.

**Jun Du** (Senior Member, IEEE) received B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. From 2009 to 2010, he was with iFLYTEK Research as a team leader, working on speech recognition. From 2010 to 2013, he joined Microsoft Research Asia as an associate researcher, working on handwriting recognition and OCR. Since 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing, USTC. He has authored or coauthored more than 150 papers. His main research interests include speech signal processing and pattern recognition applications. He is an associate editor for the IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing and a Member of the IEEE Speech and Language Processing Technical Committee. He was the recipient of the 2018 IEEE Signal Processing Society Best Paper Award. His team won several champions of the CHiME-4/CHiME5/CHiME-6 Challenge, the SELD Task of the DCASE Challenge, and the DIHARD-III Challenge.

**Chin-Hui Lee** (Fellow, IEEE) is a professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology. Before joining academia in 2001, he had 20 years of industrial experience, ending at Bell Laboratories, Murray Hill, NJ, USA, as a Distinguished Member of Technical Staff and the Director of the Dialog Systems Research Department. He has authored or coauthored more than 600 papers and 30 patents and has been cited more than 80000 times for his original contributions, with an h-index of 80 on Google Scholar. He has received numerous awards, including the Bell Labs President's Gold Award in 1998. He also won SPS's 2006 Technical Achievement Award for Exceptional Contributions to the Field of Automatic Speech Recognition. In 2012, he was invited by the ICASSP to give a plenary talk on the future of speech recognition. In the same year, he was awarded the ISCA Medal in scientific achievement for pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition. He is also a Fellow of ISCA.