Audio-visual scene classification via contrastive event-object alignment and semantic-based fusion

Yuanbo Hou
WAVES Research Group
Ghent University, Gent, Belgium
Yuanbo.Hou@UGent.be

Bo Kang
IDLAB
Ghent University, Gent, Belgium
Bo.Kang@UGent.be

Dick Botteldooren
WAVES Research Group
Ghent University, Gent, Belgium
Dick.Botteldooren@UGent.be

***Abstract*—Previous works on scene classification are mainly based on audio or visual signals, while humans perceive the environmental scenes through multiple senses. Recent studies on audio-visual scene classification separately fine-tune the large-scale audio and image pre-trained models on the target dataset, then either fuse the intermediate representations of the audio model and the visual model, or fuse the coarse-grained decision of both models at the clip level. Such methods ignore the detailed audio events and visual objects in audio-visual scenes (AVS), while humans often identify a scene through both audio events and visual objects within, and the congruence between them. To exploit the fine-grained information of audio events and visual objects in AVS, and coordinate the implicit relationship between audio events and visual objects, this paper proposes a multi-branch model equipped with contrastive event-object alignment (CEOA) and semantic-based fusion (SF) for AVSC. CEOA aims to align the learned embeddings of audio events and visual objects by comparing the difference between audio-visual event-object pairs. Then, visual objects associated with certain audio events and vice versa are accentuated by cross-attention and undergo SF for semantic-level fusion. Experiments show that: 1) the proposed AVSC model equipped with CEOA and SF outperforms the results of audio-only and visual-only models, i.e., the audio-visual results are better than the results from a single modality. 2) CEOA aligns the embeddings of audio events and related visual objects on a fine-grained level, and the SF effectively integrates both; 3) Compared with other large-scale integrated systems, the proposed model shows competitive performance, even without using additional datasets and data augmentation tricks.**

***Index Terms*—audio-visual scene classification, audio event, visual object, contrastive learning, semantic-based fusion, attention**

I. Introduction

Audio-visual scene classification (AVSC) aims to use both audio and visual modalities to classify a video recording into one of the predefined scene categories (such as metro station, airport, or street pedestrian). Compared with scene classification relying solely on audio or visual modality, AVSC is able to not only exploit the richer information from the data but also leverage the relationship between the two modals to achieve better accuracy. Recently, AVSC has attracted many interests due to its wide applications [1] [2] [3] [4].

Scene classification provides semantic information to effectively guide higher-level audio or visual content understanding. Prior works [5] [6] [7] [8] on scene classification are mainly based on either audio or image information, and the related tasks are called acoustic scene classification (ASC) or image scene classification (ISC), respectively. The ASC models in these works [5] [6] make decisions based on the clip-level information about scenes. In real life, an acoustic scene and the audio events took place within are naturally correlated. For example, in a park scene, birds singing and dogs barking are more likely to occur than keyboard sounds, where the later are often found in the office scenes. To exploit the inherent relationships between the coarse-grained scenes and corresponding fine-grained events, relation-guided ASC [9] coordinates scene-event relationships for the mutual benefit of scene and event recognition. For ISC models [10] [11], the input is usually an image or image sequence [12], and then the scene is recognized based on the rich object information, spatial layout information, as well as the relationship between the objects and layouts. ASC and ISC aim to understand scene semantic information from the perspective of human cognition based on either audio or visual information, while humans often use audio-visual information to distinguish various scenes.

Human naturally recognizes diverse scenes based on various objects they see and complex audio events they hear. Inspired by this simple observation, an increasing number of studies expect to jointly model audio-visual information within scenes. Recent works [13] [14] show that the joint learning of acoustic and visual features can bring additional benefits to AVSC. To exploit the audio-visual information simultaneously, a multi-modal system based on convolutional recurrent neural networks (CRNN) is presented in [15]. For better integration of audio-visual information, a multi-modal ensemble approach [16] enhanced by CLIP [17] with late fusion is proposed for AVSC. The above AVSC systems fine-tune pretrained audio and image models on target datasets, and then fuses the intermediate representations from audio and image models, or fuses the classification decisions of audio and image models. The decision fusion (DF) is also called late fusion, and intermediate fusion (IF) is also called early fusion [13]. However, no matter using IF or DF, the above AVSC systems did not fully leverage the rich information of audio events and visual objects in scenes, as well as the correlation between them. To exploit the fine-grained information of audio events and visual objects within coarse-grained scenes, this paper proposes a contrastive learning-based alignment for audio events and visual objects to model the detailed relationship between audio-visual information.

Unlike previous studies on AVSC using IF or DF, this

paper aims to exploit the fine-grained information of events and visual objects contained in audio-visual scer coordinate and fuse audio-visual modalities informatio better joint modeling of the audio-visual scenes. There this paper proposes an AVSC model equipped with contr event-object alignment (CEOA) and semantic-based f (SF). The CEOA based on contrastive learning [18] to explore the event-object similarity within the intra scene and analyze the differences in the composition of events and visual objects between different scenes to enl the discriminability of the model for various scenes. Th aims to coordinate the information of audio events and objects after CEOA to generate the semantic-level audio- information required for the final scene classification.

The contributions of this paper are: 1) we propose contrastive learning-based event-object alignment to coordinate the relationship of fine-grained information between audio and visual modalities in scenes to assist AVSC; 2) To better fuse the audio-visual information after the alignment from CEOA, we propose SF based on cross-attention to derive the visual objects caused by audio events and the audio events caused by visual objects, and then fuse them. 3) Quantitative evaluation shows the proposed model achieves competitive performance when compared with other large-scale integrated systems, even without using additional datasets and data augmentation tricks. Visual analysis of the intermediate representations of the proposed model provides further justification for the model. This paper is organized as follows, Section II proposes the AVSC model. Section III describes the dataset, baseline, experimental setup, and analyzes results. Section IV gives conclusions.

II. Multi-branch AVSC model with CEOA and SF

The proposed AVSC model in Fig. 1 consists of an audio branch, a visual branch, contrastive event-object alignment (CEOA), and semantic-based fusion (SF). The audio branch and the visual branch generate global-level embeddings of audio events and visual objects in audio-visual clips, respectively. CEOA aligns embeddings of audio events and visual objects in scenes to enhance the model's ability to capture the relationship between audio-visual information. Then, semantic-level SF leverages the aligned audio-visual representations by applying cross-sensory attention and fusing the bi-modal event-object information to classify the scene.

A. The audio branch

The structure of the audio branch evolved from the Transformer [19], more specifically an Audio Spectrogram Transformer (AST) [20]. Convolution-free AST can be applied to audio spectrograms and is able to capture long-range global context information [20]. The input of the audio branch is the log-mel spectrograms [21] of a whole audio clip, and the output are probabilities of audio events that may be contained in this audio clip. The spectrograms containing acoustic features are split into a sequence of patches. Each patch is flattened and projected onto a lower dimensional embedding space via a linear projection layer. Referring to AST [20], the total number of Transformer encoder layers

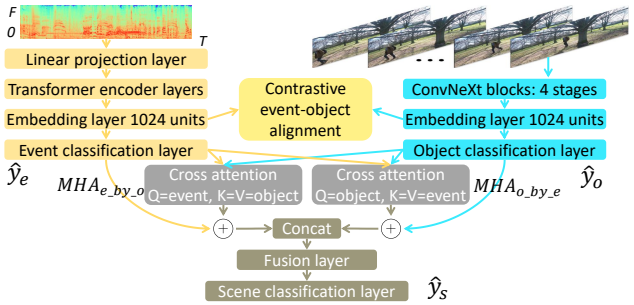

Fig. 1. The proposed multi-branch AVSC model with CEOA and SF.

in Fig. 1 is 12, and each layer has 12 heads for multi-head attention (MHA) [19]. The dimension of embedding in MHA is 768 and the dimension of each head is 64 ($768/12 = 64$), which are the same as those in [22]. The encoder layers are followed by a linear embedding layer with ReLU activation that maps the high-level representations of audio events to labels for classification. As the audio branch performs multi-label classification, binary cross-entropy (BCE) loss is used [9]. Denote the output of audio branch as $\hat{y}_e \in \mathbb{R}^{C_e}$, and the corresponding label as $y_e \in \mathbb{R}^{C_e}$, the loss can be defined as:

$$\mathcal{L}_e = -\sum_{i=1}^{C_e} y_{e_i} \log(\hat{y}_{e_i}) + (1 - y_{e_i}) \log(1 - \hat{y}_{e_i}) \quad (1)$$

where C_e is the total number of event classes in the dataset, $\hat{y}_{e_i} \in [0, 1]$ is the occurrence probability of the i-th event in the audio clip, and $y_{e_i} \in \{0, 1\}$ is the corresponding label.

B. The visual branch

Motivated by the superior performance of convolutional neural networks (CNN) in image processing, the structure of the visual branch is referred to as ConvNeXt [23], which is a recent version of CNN that utilizes the key components that made Transformers work well. ConvNeXt uses depth-separable convolution, inverse bottleneck layer, Gaussian error linear unit GELU [23], and a larger convolution kernel (7x7) to increase the receptive field to extract richer features. To capture the contextual and dynamic features between images of an audio-visual scene stream, the image sequence is used as input to the visual branch. The output of the visual branch are the probabilities of the target objects that might be contained in the image sequence. Referring to ConvNeXt-Base [23], the visual branch contains 4 stages, the number of ConvNeXt blocks of 4 stages are (3, 3, 27, 3), respectively, the number of convolution channels of 4 stages is (128, 256, 512, 1024), where the number of channels doubles at each new stage. After ConvNeXt blocks, a linear embedding layer with ReLU activation maps the high-level representations of visual objects to labels. As the visual branch also performs multi-label classification, BCE loss is used. Denote the output of the visual branch as $\hat{y}_o \in \mathbb{R}^{C_o}$, and the corresponding visual object label as $y_o \in \mathbb{R}^{C_o}$, the loss of the visual branch can be defined as:

$$\mathcal{L}_o = -\sum_{i=1}^{C_o} y_{o_i} \log(\hat{y}_{o_i}) + (1 - y_{o_i}) \log(1 - \hat{y}_{o_i}) \quad (2)$$

where C_o is the total number of object classes in the dataset, $\hat{y}_{o_i} \in [0, 1]$ is the occurrence probability of the i-th object in the image sequence, $y_{o_i} \in \{0, 1\}$ is the corresponding label.

C. Contrastive event-object alignment (CEOA)

CEOA aims to model the relationship between audio and visual modal information, and to align the fine-grained audio-visual event-object information. Due to the complexity of real-life scenes, explicit event-object correlation in diverse scenes is often unknown from the input data. This paper expects to learn the relative distances of event-object pairs in different scenes indirectly through contrastive learning [24]. CEOA tries to embed and coordinate representations of audio events and corresponding visual objects into the same area of the latent space so that they can be aligned in the semantic space for cross-modal fusion in the fusion part of the model.

CEOA adopts the pairwise contrastive loss (PCL). The goal of PCL is to make the representations corresponding to positively correlated samples closer together, and the representations of samples that are less or negatively correlated farther apart. The PCL is guided by the gap between positive-positive (PP) and positive-negative (PN) pairs [25]. During the optimization of PCL, the model will automatically focus on expanding the distance between the PP pairs and PN pairs, so as to cluster embeddings in the positive pairs and align the corresponding embeddings of audio events and visual objects to achieve fine-grained alignment of audio-visual information. Given $P_e \in \mathbb{R}^{K\times 1024}$ and $N_e \in \mathbb{R}^{K\times 1024}$ are the weights of K audio events with the highest and lowest probabilities from the last classification layer in audio branch. $P_o \in \mathbb{R}^{K\times 1024}$ and $N_o \in \mathbb{R}^{K\times 1024}$ are the weights of K visual objects with the highest and lowest probabilities of the same input sample from the classification layer in visual branch. These weight matrices in final classification layers can be viewed as the core knowledge about targets learned by the model. For the audio branch, P_e is used as positive samples, P_o and N_o are the corresponding positive and negative samples, respectively. to jointly construct the event-to-object contrastive loss by PP pair $P_e P_o^T$ and PN pair $P_e N_o^T$.

$$\begin{aligned}\mathcal{L}_{e2o} &= -\ln(\text{mean}(\frac{e^{P_e P_o^T}}{e^{P_e P_o^T} + e^{P_e N_o^T}})) \\ &= -\ln(\text{mean}(\sigma(P_e P_o^T - P_e N_o^T))\end{aligned} \tag{3}$$

where σ is the logistic sigmoid: $\sigma(x) = 1/(1+e^{-x})$. For the visual branch, P_o is used as positive samples, P_e and N_e are the corresponding positive and negative samples, to build the object-to-evet contrastive loss:

$$\mathcal{L}_{o2e} = -\ln(\text{mean}(\sigma(P_o P_e^T - P_o N_e^T)) \tag{4}$$

The method of composing negative samples based on K components with the lowest probabilities is called the lowest K mode (LKM) in this paper. In addition to LKM, random K mode (RKM) can also be used. That is, randomly select K classes of events or objects from the weights that do not contain positive classes of events or objects. RKM will increase the difficulty of model learning and bring more challenges. For example, the K audio events with the lowest probabilities in LKM are the K audio events that the audio branch is most confident with the least likely to occur in the input clip, while the confidence for the randomly selected K audio events in RKM is ambiguous. So, N_e in RKM will have larger information uncertainty (larger entropy) than N_e in LKM. Greater entropy will bring more burdens and possibilities for model learning, the impact of K and 2 modes on model performance will be explored in the experiments.

D. Semantic-based fusion (SF)

SF aims to perform cross-modal fusion of audio and visual embeddings to generate semantic-level audio-visual representations after fine-grained alignment by CEOA. To consider the possible interactions and correlations of audio events and visual objects in diverse scenes, this paper proposes SF based on the multi-head attention (MHA) [19], which is calculated on a set of queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$).

$$\begin{aligned}MHA(\mathbf{Q},\mathbf{K},\mathbf{V}) &= Concat(head_1,...,head_h)\mathbf{w}^O \\ where \quad head_i &= A(\mathbf{Q}\mathbf{w}_i^Q, \mathbf{K}\mathbf{w}_i^K, \mathbf{V}\mathbf{w}_i^V),\end{aligned} \tag{5}$$

$$A(\mathbf{Q}\mathbf{w}_i^Q, \mathbf{K}\mathbf{w}_i^K, \mathbf{V}\mathbf{w}_i^V) = \Phi(\mathbf{Q}\mathbf{w}_i^Q {\mathbf{K}\mathbf{w}_i^K}^T/\sqrt{d})\mathbf{V}\mathbf{w}_i^V$$

Referring to settings of Transformer [19], Φ is softmax function, the total number of attention heads h is 8, learnable weights $\{\mathbf{w}_i^Q, \mathbf{w}_i^K, \mathbf{w}_i^V\} \in \mathbb{R}^{1\times d}$ and $d = 64$, $\mathbf{w}^O \in \mathbb{R}^{(h*d)\times 1}$. When $\mathbf{K}$ and $\mathbf{V}$ are core event embeddings from the final event classification layer ($\{\mathbf{K},\mathbf{V}\} \in \mathbb{R}^{C_e\times 1}$), and $\mathbf{Q}$ is core object embeddings from the final object classification layer ($\mathbf{Q} \in \mathbb{R}^{C_o\times 1}$), the output of MHA can be viewed as visual objects embeddings caused by audio events, which is denoted as $MHA_{o\_by\_e} \in \mathbb{R}^{C_o\times 1}$. In this, $\mathbf{Q}\mathbf{w}_i^Q {\mathbf{K}\mathbf{w}_i^K}^T \in \mathbb{R}^{C_o\times C_e}$ can be regarded as the transformation matrix from audio space to visual space. In contrast, when $\mathbf{K}$ and $\mathbf{V}$ are core object embeddings ($\{\mathbf{K},\mathbf{V}\} \in \mathbb{R}^{C_o\times 1}$), and $\mathbf{Q}$ is core event embeddings ($\mathbf{Q} \in \mathbb{R}^{C_e\times 1}$), the output of MHA can be viewed as audio events embeddings caused by visual objects, which is denoted as $MHA_{e\_by\_o} \in \mathbb{R}^{C_e\times 1}$. Next, as shown in Fig. 1, $MHA_{o\_by\_e}$ is added with object embeddings $\hat{y}_o$ to produce audio-visual enriched objects embeddings, $MHA_{e\_by\_o}$ is added with event embeddings $\hat{y}_e$ to produce audio-visual enriched events embeddings. Then, the event and object embeddings are concatenated together to form audio-visual semantic embeddings, and the fusion layer with ReLU activation maps the audio-visual embeddings into scene classes. As scene classification performs single-label multi-class classification, cross-entropy loss [9] is used between the output $\hat{y}_s \in \mathbb{R}^{C_s}$ and the scene label $y_s \in \mathbb{R}^{C_s}$,

$$\mathcal{L}_s = -\sum\nolimits_{i=1}^{C_s} y_{s_i} \log(\hat{y}_{s_i}) \tag{6}$$

where C_s is the total number of scene classes in the dataset, $\hat{y}_{s_i} \in [0,1]$ is the occurrence probability of the i-th scene in the input clip, and $y_{s_i} \in \{0,1\}$ is the corresponding label.

In the training phase, to help the model comprehensively consider the fine-grained information of audio events ($\mathcal{L}_e$), visual objects ($\mathcal{L}_o$), contrastive event-to-object ($\mathcal{L}_{e2o}$), contrastive object-to-event ($\mathcal{L}_{o2e}$), and coarse-grained global information of audio-visual scene ($\mathcal{L}_s$) within the input clips, the final loss of the proposed model can be defined as:

$$\mathcal{L} = \lambda_1\mathcal{L}_e + \lambda_2\mathcal{L}_o + \lambda_3\mathcal{L}_{e2o} + \lambda_4\mathcal{L}_{o2e} + \lambda_5\mathcal{L}_s \tag{7}$$

where λ_i is the scale factor of each loss and defaults to 1.

III. Experiments and results

A. Dataset, Baseline, Experiments Setup, and Metrics

TAU Audio-Visual Urban Scenes 2021 development dataset [26] used in this paper consists of 12291 10-seconds clips totaling 34.14 hours and contains 10 different classes of audio-visual scenes. This real-life dataset does not contain labels for audio events nor visual objects. AST[1] [20] and ConvNeXt[2] [23] are used to tag each clip with pseudo labels indicating the probability of audio events and visual objects for model training. Since AST and ConvNeXt are trained on Audioset [27] (527 classes) and ImageNet-1K [28] (1000 classes), respectively, the number of classes of audio events and visual objects in pseudo labels is 527 and 1000. We set the thresholds (0.0365 and 0.9216) for the occurrence of audio events and visual objects to binarize the probabilities output by pre-trained models into hard labels consisting of 1 and 0.

To compare the performance of the model with other models on the same benchmark, this paper uses the baseline in DCASE 2021 Task 1 Subtask B (T1B) [13] as baseline, and further compares the proposed model with other methods from different perspectives on the same AVSC task. To facilitate the comparison with other methods, the training/testing split of the dataset follows the default split of T1B. Similar to all other comparison methods using weights from pre-trained models, the proposed model also uses part of the weights from AST [20] and ConvNeXt [23] during training.

For training, log mel-bank energy with 128 banks [21] is used as acoustic features, which is extracted by STFT with Hamming window length of 25 *ms* and a hop size of 10 *ms* between the window. An image sequence consisting of one image per second is input to the visual branch. For a 10-second scene clip, the input to the visual branch is an image sequence consisting of 10 images. A batch size of 16 and AdamW optimizer [29] with learning rate of 5e-6 are used to minimize the losses in the proposed model. To prevent over-fitting, dropout [30] and normalization are used. Systems are trained on a single card Tesla V100-SXM2-32GB for 100 epochs. The Logloss and average accuracy (Acc.) [31] are used as metrics. Larger Acc. and lower Logloss indicate better performance. More details, please see the homepage[3].

B. Results and Analysis

Difference between LKM and RKM. There are 2 modes in CEOA when selecting negative samples, LKM based on the lowest occurrence probabilities and RKM based on random screening. Compared with LKM, RKM increases the learning difficulty of the model for fine-grained event-object pairs due to the greater uncertainty of negative components. Table I shows the effects of the two modes on model performance.

As the value of K increases, the classification accuracy of the model in Table I in both modes improves. That is, the gradual increase in the influence of contrastive learning in model training is beneficial to the model's recognition of fine-grained event-object pairs in audio-visual scenes, which in turn helps to improve the model's ability to identify differences between different scenes. When K increases to a certain value, increasing the value of K will not bring more benefits to the scene analysis ability of the model. The models in LKM and RKM achieve the best performance when K is 15 and 10, respectively, where the performance of models in RKM are slightly inferior to that in LKM. The reason may be that the randomness of negative components in RKM brings more challenges and difficulties to the learning of the model, making it difficult for the model to achieve a balance between extracting and coordinating audio-visual event-object representations and efficiently capturing information for scene classification. However, it is worth noting that in LKM, varying K values obviously affects the model performance, while different K values in RKM do not have a much different impact on the model. This may be due to the fact that the negative components in event-object pairs of RKM are randomly selected, which makes the model less sensitive to the size of the contrastive pairs in aligning event-object representations. Based on the results in Table 2, LKM and $K = 15$ will be used in subsequent experiments.

TABLE I
ACC. (%) OF THE PROPOSED AVSC MODEL IN DIFFERENT MODES.

K value	*1*	*5*	*10*	*15*	*20*	*25*	*30*
LKM	*89.32*	*90.12*	*90.83*	***91.58***	*91.38*	*91.00*	*90.91*
RKM	*91.00*	*91.11*	***91.30***	*91.27*	*91.22*	*91.08*	*91.05*

Gain of adding CEOA and SF to the model. The proposed AVSC model aims to use the contrastive learning-based CEOA to align the fine-grained event-object information within audio-visual scenes to coordinate the relationship between the audio-visual information, and then utilize the attention-based SF to collaboratively fuse audio-visual representations across modalities. Table II summarizes the gain of the proposed modules for the model learning capability. Compared to the basic backbone, both the contrast learning-based CEOA and the semantic-level attention-based fusion proposed in this paper improves the classification ability of the model for audio-visual scenes. Among them, CEOA has a slightly larger improvement on the model performance. This also illustrates the benefit of mining the fine-grained event-object information contained in diverse scenes for recognition. Finally, the backbone equipped with CEOA and SF achieves better results.

TABLE II
THE ABLATION STUDY OF THE PROPOSED MODULES.

#	*Backbone*	*SF*	*CEOA*	*Acc. (%)*	*Logloss*
1	✔	✗	✗	88.42	0.439
2	✔	✔	✗	90.34	0.390
3	✔	✗	✔	90.75	0.357
4	✔	✔	✔	**91.58**	**0.259**

Weighting scale factors in the loss. Audio-visual scene naturally contains both audio and visual information, and the proposed CEOA also contains two kinds of contrastive information from the perspectives of audio events and visual objects. In training, different coefficients of loss components

[1]AST with 0.459 mAP: https://github.com/YuanGongND/ast
[2]https://download.pytorch.org/models/convnext_base-6075fbad.pth
[3]Homepage: https://github.com/Yuanbo2020/Contrastive-AVSC

represent the importance of their corresponding targets in the overall model performance. Different combinations of coefficients often imply different concerns of the model in the learning process. Table III summarizes the model performance with several combinations of loss weights. It shows the effect of changing the ratio between different components of the loss on the utilization of diverse information in the training.

TABLE III
THE EFFECT OF DIFFERENT λ_i VALUES ON THE AVSC TASK.

#	λ_1	λ_2	λ_3	λ_4	λ_5	*Acc. (%)*	*Logloss*
1	1	1	1	1	1	91.58	0.259
2	0.5	1	1	1	1	91.82	0.258
3	0.5	0.5	1	1	1	92.20	0.237
4	0.25	0.5	1	1	1	93.06	0.226
5	0.25	0.5	0.25	1	1	93.71	0.222
6	0.25	0.5	0.25	0.5	1	**94.10**	**0.192**
7	0.25	0.5	0.01	0.01	1	93.99	0.193
8	0.05	0.25	0.05	0.25	1	93.44	0.254
9	0.01	0.5	0.01	0.01	1	93.52	0.209
10	0.01	0.1	0.01	0.1	1	93.33	0.246
11	0.01	0.01	0.05	0.1	1	94.02	0.193
12	0.005	0.1	0.025	0.1	1	93.41	0.219

The fusion of cross-modal information of different types in Table III aims to maintain as much as possible the recognition ability of the audio model and visual model for audio events and visual objects, respectively, while making full use of the implicit relationship between audio events and visual objects, so that they both contribute to the model's ability of classifying scenes. Several combinations of coefficients are explored to select the optimal ratio between the information of audio events, visual objects, event-object pairs and audio-visual scenes for a better AVSC model. Finally, giving maximum weight to $\mathcal{L}_s$ and secondary weight to the information related to visual objects ($\mathcal{L}_o$, $\mathcal{L}_{o2e}$), while absorbing the information related to audio events ($\mathcal{L}_e$, $\mathcal{L}_{e2o}$), makes the best result of # 6 in Table III. This combination of coefficients gives a stronger emphasis on the visual object information, which means that the visual object information is more important than audio event information in the proposed model for the AVSC task.

TABLE IV
PERFORMANCE OF THE MODEL ON DIFFERENT MODAL INFORMATION.

#	*Audio*	*Visual*	*Acc. (%)*	*Logloss*
1	✔	✗	73.55	0.871
2	✗	✔	88.86	0.518
3	✔	✔	**94.10**	**0.192**

Comparison of single-modal and multi-modal models. Compared with the single-modal audio or visual model, the audio-visual model can utilize both modal information, and understand differences in the description of the same target from the perspectives of different modalities. In Table IV, the audio-visual model that fuses fine-grained information from cross-modal achieves better results, the audio-only model performs the worst, while the result of the visual-only model is slightly better. This is consistent with the trend reflected in Table III, that is, visual object information is more valuable than audio event information in the proposed model for the AVSC task. The reason for this phenomenon may be that in this dataset, the audio clips between different scenes do not sound very different, and most of the audio clips are full of noise, making it difficult for even a human to effectively distinguish target scenes by relying only on audio clips. However, the differences between visual objects in different scenes are obvious, a park with trees and a lake is clearly different from an airport with monitors and escalators. Therefore, using visual information can effectively distinguish between different scenes in this AVSC task, which also leads to the result that the visual object information plays a more important role than audio events.

Comparison with prior methods. Table V shows the performance of the proposed model and other systems on the same AVSC task. The proposed model does not use any data augmentation (Aug.) methods, nor does it use any ensemble methods. So, for a fair comparison, we extract the best single model result of other systems directly from the T1B website.

Table V includes published results of the top 5 teams in T1B competition, all systems listed in Table V use the weights of pre-trained models involved. Since the result of a single model could not be found in the paper [16] of # 6, we implemented their system according to their settings in [16]. Except for Baseline and the proposed model do not use data augmentation, other systems use diverse audio or visual data augmentation methods. The system of # 7 uses an additional large scene dataset Places365 for training. To compare the model performance on the same dataset ImageNet, we select the result based on ImageNet in # 7. The model in # 9 is trained with both ImageNet and Places365, and also uses a two-stage fine-tuning strategy. In contrast, the proposed model, which does not involve data augmentations, achieves similar results to that of # 9, which uses additional datasets and multiple data augmentations. That is, even without using additional datasets and data augmentation tricks, the proposed model shows competitive performance.

TABLE V
COMPARISON OF AUDIO-VISUAL SCENE CLASSIFICATION RESULTS OF DIFFERENT SYSTEMS ON THE SAME DATASET.

#	System	Audio Backbone	Visual Backbone	Aug. Method	*Acc (%)*
1	Baseline [13]	OpenL3	OpenL3	*None*	77.0
2	WaveTransformer [32]	OpenL3	OpenL3	3 types	79.5
3	CRNN [15]	SE-Net	VGG16	1 type	90.0
4	2-stage classifier [33]	fsFCNN	TimeSformer	8 types	91.5
5	2-stream model [34]	OpenL3	ResNet50	3 types	91.7
6	CLIP variants [16]	EfficientNet	CLIP ViT	6 types	93.3
7	AVSM [35]	VGGish	ResNet50	8 types	93.5
8	CNN_Transformer [36]	VGGish	Transformer	4 types	93.9
9	2-stage fine-tuning [37]	ResNet	EfficientNet	5 types	**94.1**
10	Proposed model	Transformer	ConvNeXt	*None*	**94.1**

Visual analysis. To gain concise insights into the event-object cross-modal alignment, Fig. 2 intuitively visualizes the core knowledge (weights from the classification layers) about audio events and visual objects learned by the model with CEOA using UMAP [38]. Different kinds of audio events and visual objects are interleaved and regularly distributed in Fig. 2. (Emergency vehicle (*sound*); Ambulance (*object*)), (Grey wolf; Canidae, wolves), (Mosquite; Wing), and (Clock; Analog clock) are effectively individually clustered. In addition, similar audio events and visual objects are located closely, like (Male singing; CD players, Speakers), (Mechanical fan, Squeak; Recreational vehicle, Tram). The efficient aggregation of various event-object pairs in Fig. 2 illustrates that con-

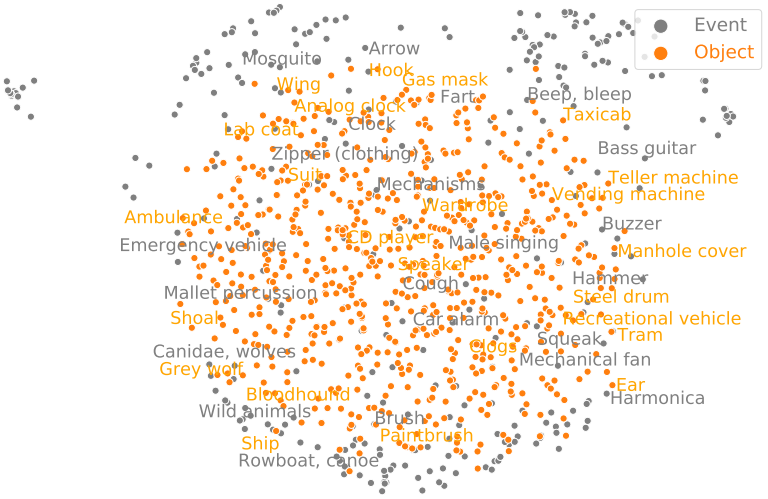

Fig. 2. Visualization of the learned weights of events and objects from the event classification layer and object classification layer, respectively. Please visit the homepage to interactively view the label for each point in detail.

trastive learning-based CEOA aligns fine-grained information of audio events and visual objects across modalities.

IV. CONCLUSION

To exploit the fine-grained information of audio-visual events-object in diverse real-life scenes and to coordinate the implicit relationship between audio events and visual objects, this paper proposes a multi-branch AVSC model equipped with CEOA for event-object alignment and SF for cross-modal fusion. Experiments show that 1) the proposed contrastive learning-based CEOA aligns fine-grained information of audio events and visual objects, and SF successfully fuses cross-modal information; 2) The influence of contrastive learning in model training is beneficial to the model's recognition of fine-grained event-object pairs in audio-visual scenes; 3) Even without using additional datasets and data augmentation tricks, the proposed model shows competitive performance.

REFERENCES

[1] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 257–267, 2007.

[2] C. C. Tan, Y. Jiang, and C. Ngo, "Towards textually describing complex video contents with audio-visual concept classifiers," in *Proceedings of 19th ACM International Conference on Multimedia*, 2011, p. 655–658.

[3] R. Kojima, O. Sugiyama, and K. Nakadai, "Audio-visual scene understanding utilizing text information for a cooking support robot," in *IEEE/RSJ IROS*, 2015, pp. 4210–4215.

[4] B. Kühn, B. Schauerte, R. Stiefelhagen, and K. Kroschel, "A modular audio-visual scene analysis and attention system for humanoid robots," in *Proc. 43rd Int. Symp. Robotics (ISR)*, 2012.

[5] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE SPM*, vol. 32, no. 3, pp. 16–34, 2015.

[6] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *24th EUSIPCO*, 2016, pp. 1128–1132.

[7] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[8] Q. Liu, R. Hang, H. Song, and Z. Li, "Learning multiscale deep features for high-resolution satellite image scene classification," *IEEE Transactions on GRS*, vol. 56, no. 1, pp. 117–126, 2017.

[9] Y. Hou, B. Kang, W. V. Hauwermeiren, and D. Botteldooren, "Relation-guided acoustic scene classification aided with event embeddings," in *International Joint Conference on Neural Networks*, 2022.

[10] X. Cheng, J. Lu, J. Feng, B. Yuan, and J. Zhou, "Scene recognition with objectness," *Pattern Recognition*, vol. 74, pp. 474–487, 2018.

[11] H. J. Kim and J. Frahm, "Hierarchy of alternating specialists for scene recognition," in *Proceedings of the ECCV*, 2018, pp. 451–467.

[12] K. G. Derpanis, M. Lecce, K. Daniilidis, and R. P. Wildes, "Dynamic scene understanding: The role of orientation features in space and time in scene classification," in *CVPR*, 2012, pp. 1306–1313.

[13] S. Wang, A. Mesaros, T. Heittola, et al., "Audio-visual scene classification: Analysis of DCASE 2021 Challenge submissions," in *Proceedings of DCASE Workshop*, 2021, vol. 38, pp. 788–792.

[14] L. Pham, A. Schindler, M. Schutz, J. Lampert, S. Schlarb, and R. King, "Deep learning frameworks applied for audio-visual scene classification," in *Data Science–Analytics and Applications*, pp. 39–44. 2022.

[15] J. Naranjo-Alcazar, S. Perez-Castanos, M. Cobos, et al., "Squeeze-excitation convolutional recurrent neural networks for audio-visual scene classification," in *Proceedings of DCASE Workshop*, 2021, pp. 16–20.

[16] S. Okazaki, Q. Kong, and T. Yoshinaga, "A multi-modal fusion approach for audio-visual scene classification enhanced by clip variants," in *Proceedings of DCASE Workshop*, 2021, pp. 95–99.

[17] A. Radford, J. W. Kim, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[18] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *ICASSP*, 2021, pp. 375–379.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, et al., "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[20] Y. Gong, Y. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. of INTERSPEECH*, 2021, pp. 571–575.

[21] A. Bala, A. Kumar, and N. Birla, "Voice command recognition system based on MFCC and DTW," *IJEST*, vol. 2, no. 12, pp. 735–742, 2010.

[22] H. Touvron, M. Cord, et al., "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021, pp. 10347–10357.

[23] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," *Proceedings of CVPR*, pp. 11976–11986, 2022.

[24] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein, "Cdpam: Contrastive learning for perceptual audio similarity," in *ICASSP*, 2021, pp. 196–200.

[25] P. Avgoustinakis et al., "Audio-based near-duplicate video retrieval with audio similarity learning," in *ICPR*, 2021, pp. 5828–5835.

[26] S. Wang, A. Mesaros, T. Heittola, et al., "A curated dataset of urban scenes for audio-visual scene analysis," in *ICASSP*, 2021, pp. 626–630.

[27] J. F. Gemmeke, D. PW. Ellis, et al., "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017, pp. 776–780.

[28] J. Deng, W. Dong, R. Socher, L. Li, k. Li, and Li F., "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[29] L. Ilya and H. Frank, "Decoupled weight decay regularization," in *ICLR*, 2019.

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, et al., "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[31] Q. Kong, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "DCASE 2018 challenge surrey cross-task convolutional neural network baseline," in *Proc. of DCASE 2018 Workshop*, 2018, pp. 217–221.

[32] A. Triantafyllopoulos, K. Drossos, et al., "A multimodal wavetransformer architecture conditioned on openl3 embeddings for audio-visual scene classification," Tech. Rep., DCASE2021 Challenge, 2021.

[33] Y. Wang and S. liang, "Bit submission for DCASE 2020 challenge task1," Tech. Rep., DCASE2021 Challenge, 2021.

[34] Y. Hou, Y. Tan, Y. Chang, T. Huang, S. Li, X. Shao, and D. Botteldooren, "CNN-based dual-stream network for audio-visual scene classification," Tech. Rep., DCASE2021 Challenge, 2021.

[35] Q. Wang, S. Zheng, et al., "A model ensemble approach for audio-visual scene classification," Tech. Rep., DCASE2021 Challenge, 2021.

[36] Y. Yang and Y. Luo, "Scene classification using acoustic and visual feature," Tech. Rep., DCASE2021 Challenge, 2021.

[37] M. Wang, C. Chen, Y. Xie, H. Chen, et al., "Audio-visual scene classification using transfer learning and hybrid fusion strategy," Tech. Rep., DCASE2021 Challenge, 2021.

[38] L. McInnes, J. Healy, N. Saul, et al., "Umap: Uniform manifold approximation and projection," *JOSS*, vol. 3, no. 29, pp. 861, 2018.