DOA-Aware Audio-Visual Self-Supervised Learning for Sound Event Localization and Detection

Yoto Fujita*†, Yoshiaki Bando†, Keisuke Imoto‡†, Masaki Onishi†, and Kazuyoshi Yoshii*
*Graduate School of Informatics, Kyoto University, Japan, {fujita.yoto.26m@st.kyoto-u.ac.jp, yoshii@i.kyoto-u.ac.jp}
†National Institute of Advanced Industrial Science and Technology, Japan, {y.bando,onishi-masaki}@aist.go.jp
‡Faculty of Science and Engineering, Doshisha University, Japan, keisuke.imoto@ieee.org

***Abstract*—This paper describes sound event localization and detection (SELD) for spatial audio recordings captured by first-order ambisonics (FOA) microphones. In this task, one may train a deep neural network (DNN) using FOA data annotated with the classes and directions of arrival (DOAs) of sound events. However, the performance of this approach is severely bounded by the amount of annotated data. To overcome this limitation, we propose a novel method of pretraining the feature extraction part of the DNN in a self-supervised manner. We use spatial audio-visual recordings abundantly available as virtual reality contents. Assuming that sound objects are concurrently observed by the FOA microphones and the omni-directional camera, we jointly train audio and visual encoders with contrastive learning such that the audio and visual embeddings of the same recording and DOA are made close. A key feature of our method is that the DOA-wise audio embeddings are *jointly* extracted from the raw audio data, while the DOA-wise visual embeddings are *separately* extracted from the local visual crops centered on the corresponding DOA. This encourages the latent features of the audio encoder to represent both the classes and DOAs of sound events. The experiment using the DCASE2022 Task 3 dataset of 20 hours shows non-annotated audio-visual recordings of 100 hours reduced the error score of SELD from 36.4 pts to 34.9 pts.**

***Index Terms*—Sound event localization and detection, audio-visual contrastive learning, self-supervised learning**

I. Introduction

Sound event localization and detection (SELD) is a task that aims to estimate the activations, classes, and directions of arrival (DOAs) of sound events from multichannel audio recordings [1]. It is one of the foundations of computational intelligence for understanding acoustic environments. The current standard approach to this task is to train a deep neural network (DNN) in a supervised manner using pairs of audio recordings with ground-truth annotations [2]–[4]. In general, it is difficult to collect a sufficient amount of annotated audio data covering a wide range of acoustic environments.

A popular solution to this problem is data augmentation [5], [6]. Specifically, one can synthesize multichannel audio data by convolving source signals of various classes with room impulse responses (RIRs) that simulate various acoustic environments and DOA conditions. This approach is known to effectively improve the performance of SELD for many common sound events (e.g., music and speech). However, SELD for some complex sound events (e.g., wildlife sound) remains an open problem because it is difficult to isolate and capture their individual sound source signals.

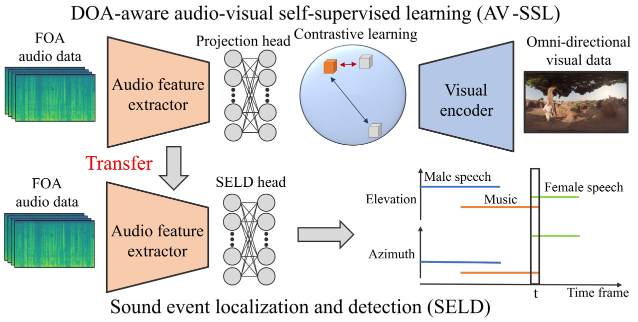

Fig. 1. AV-SSL of an audio feature extractor for SELD.

Another solution is to pretrain an audio feature extractor that constitutes the front end of a DNN used for SELD with audio-visual self-supervised learning (AV-SSL) (Fig. 1) [7]. This approach can make effective use of abundant virtual reality (VR) contents, each of which consists of multichannel audio data recorded by first-order ambisonics (FOA) microphones and 360° equirectangular visual data recorded by an omni-directional camera. Considering the cross-modal co-occurrence between the sounds and appearances of the same objects, one can jointly train audio and visual encoders in a contrastive fashion such that the audio and visual embeddings are made close to each other if they correspond to the same DOA and recording (positive sample) and far apart otherwise (negative sample). The front end of the trained audio encoder is then used for initializing the audio feature extractor for SELD.

Audio-visual spatial alignment (AVSA) [7] is one of the latest AV-SSL methods. It takes advantage of the FOA format, in which the single-channel audio signal with an arbitrary DOA can be computed from the observed FOA data. The audio embedding extracted from this enhanced signal are made close to the visual embedding extracted from the visual crop centered on the same DOA in the equirectangular visual data. However, the audio feature extractor trained in this way is insufficient for SELD because the DOA features of sound events cannot be extracted from the enhanced signal. On the other hand, the features useful for DOA estimation of sound events are not learnable with such non-DOA-aware contrastive learning equivalent in principle to AVC [8].

In this paper, we propose a DOA-aware extension of AVSA. Our method differs from the conventional AVSA [7] in that it jointly extracts audio embeddings over a DOA grid from raw

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

FOA audio data without DOA-wise signal enhancement. This encourages the latent features of the audio encoder to represent both the classes and DOAs of sound events. In addition, this paper also tackles one of the remaining challenges in AVSA that the cross-modal co-occurrence between sound and appearance does not hold for visible silent objects and occluded sound objects. In general, meaningful sound events exist only in a small number of DOAs, which limits this *local* contrastive learning based on the DOA-wise similarity between audio and visual embeddings.

To mitigate this problem, we also propose *global* contrastive learning based on recording-wise audio-visual similarity obtained by averaging the DOA-wise similarities over all the DOAs. Our goal is to maximize the similarity when the audio and visual embeddings correspond to the same recording (positive) and minimize it otherwise (negative). To encourage the audio encoder to extract DOA information as the latent features, we introduce a data augmentation technique that randomly spatially rotates only the equirectangular visual data to generate negative samples from the same recording.

II. Related Work

This section reviews existing SELD and AV-SSL methods.

A. Sound event localization and detection (SELD)

The modern approach to SELD is to train a DNN in a supervised manner. For instance, the activations of sound events are estimated with a DNN, while the DOAs are estimated geometrically [2]. Both the activations and DOAs can be estimated using a DNN [3]. Recently, an end-to-end approach to SELD has been proposed for directly estimating a DOA vector whose length corresponds to the duration of sound events in the Cartesian coordinate system [4].

To improve the robustness of SELD with a limited amount of annotated audio data, data augmentation tequniques such as SpecAugment [9] and Mixup [10] can be employed. Particularly for audio data provided in the FOA format, rotation-based augmentation is known to be effective [11]. Multichannel audio data with ground-truth annotations can be synthesized by convolving audio samples of various classes with arbitrary transfer functions [5], [12]. Alternatively, a general-purpose audio model [13] pretrained with a large dataset of audio signals called AudioSet [14] in an unsupervised manner can be used for sound event detection [15].

B. Audio-visual self-supervised learning (AV-SSL)

At the heart of AV-SSL is contrastive learning, which aims to train audio and visual encoders based on the cross-modal co-occurrence between the sounds and appearances of objects. The information encoded in the audio and visual embeddings may vary depending on the design of positive and negative samples. AVC [8], for example, uses standard video recordings and considers a pair of audio and visual data from the same recording as a positive sample and a pair of those from different recordings as a negative pair. The audio and visual embeddings are thus encouraged to represent the features of sound event classes. AVSA [7] uses spatial video recordings, which are originally made for VR application that requires DOA-based audio-visual rendering. AVSA, however, is essentially identical to AVC because it can be regarded as AVC for *non-spatial* video standard recordings obtained by decomposing *spatial* video recordings in a DOA-wise manner.

Audio-visual temporal synchronization (AVTS) [16] is another AVC-like self-supervised method. A key feature of AVTS is that it incorporates audio and visual data in the same recording, but at different moments in time, to negative pairs. The embeddings are thus encouraged to represent both the classes and activations of sound events.

III. Proposed Method

This section describes two variants of DOA-aware AV-SSL to improve SELD. The first variant employs DOA-wise contrastive learning, and the other employs recording-wise contrastive learning (Fig. 2). In both variants, an audio feature extractor $\mathcal{A}$ is used to transform raw FOA data into the latent audio features that represent sound event classes and DOAs. In this study, K discrete points on the Fibonacci lattice [17] are considered as potential DOAs. Each DOA $k \in [1, K]$ is defined by an azimuth angle θ_k and an elevation angle ϕ_k. A projection head $\mathcal{H}$ is then used to *jointly* convert the latent features to the audio embeddings over the DOA grid. A visual encoder $\mathcal{V}$, in contrast, is used to *separately* convert visual crops of a 360° equirectangular visual data over the DOA grid to the visual embeddings. Let I be the total number of recordings.

In DOA-wise contrastive learning, we maximize the local similarity between DOA-wise audio and visual embeddings when they correspond to the same DOA (positive sample), and minimize it otherwise (negative sample). In recording-wise contrastive learning, in contrast, we maximize global similarity obtained by averaging the local similarities over the DOA grid when the audio and visual embeddings correspond to the same recording (positive), and minimize it otherwise (negative).

Once $\mathcal{A}$, $\mathcal{H}$, and $\mathcal{V}$ are trained jointly in a self-supervised manner, $\mathcal{A}$ is connected to another head $\mathcal{H}'$ for SELD and the entire network is fine-tuned in a supervised manner.

A. Audio encoding

Let $\mathbf{X}_i^{(a)} \in \mathbb{R}^{(4+3)\times F\times T_a}$ be the multichannel audio spectrogram of spatial recording $i \in [1, I]$ obtained by stacking the mel spectrograms of the four-channel audio signals of the FOA format and the three intensity spectrograms on the orthogonal axes, where F is the number of mel frequency bins, and T_a is the number of frames. A series of frame-wise audio latent features denoted by $\mathbf{Y}_i^{(a)} \triangleq \{\mathbf{y}_{it}^{(a)}\}_{t=1}^{T_a}$ is obtained with the audio feature extractor $\mathcal{A}$ as follows:

$$\mathbf{Y}_i^{(a)} \leftarrow \mathcal{A}\left(\mathbf{X}_i^{(a)}\right). \tag{1}$$

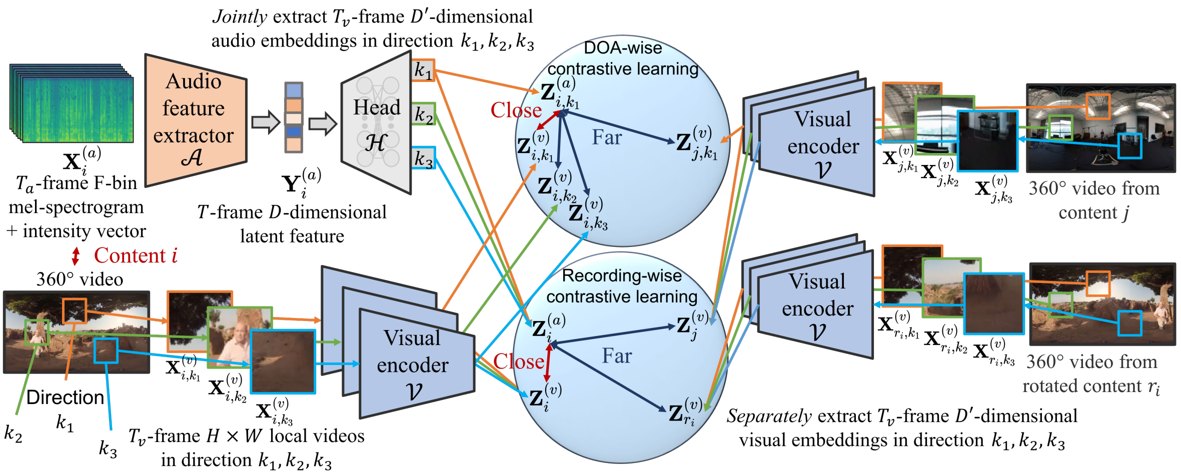
Jointly extract T_v-frame D'-dimensional
audio embeddings in direction k_1, k_2, k_3
Audio feature extractor $\mathcal{A}$
Head $\mathcal{H}$
k_1
k_2
k_3
$\mathbf{X}_i^{(a)}$
T_a-frame F-bin mel-spectrogram + intensity vector
$\mathbf{Y}_i^{(a)}$
T-frame D-dimensional latent feature
Content i
360° video
Direction
k_1
k_2
k_3
$\mathbf{X}_{i,k_1}^{(v)}$
$\mathbf{X}_{i,k_2}^{(v)}$
$\mathbf{X}_{i,k_3}^{(v)}$
T_v-frame $H \times W$ local videos in direction k_1, k_2, k_3
Visual encoder $\mathcal{V}$
DOA-wise contrastive learning
$\mathbf{Z}_{i,k_1}^{(a)}$
Close
Far
$\mathbf{Z}_{i,k_1}^{(v)}$
$\mathbf{Z}_{i,k_2}^{(v)}$
$\mathbf{Z}_{i,k_3}^{(v)}$
$\mathbf{Z}_{j,k_1}^{(v)}$
Recording-wise contrastive learning
$\mathbf{Z}_i^{(a)}$
$\mathbf{Z}_j^{(v)}$
$\mathbf{Z}_i^{(v)}$
$\mathbf{Z}_{r_i}^{(v)}$
Visual encoder $\mathcal{V}$
$\mathbf{X}_{j,k_1}^{(v)}$
$\mathbf{X}_{j,k_2}^{(v)}$
$\mathbf{X}_{j,k_3}^{(v)}$
360° video from content j
$\mathbf{X}_{r_i,k_1}^{(v)}$
$\mathbf{X}_{r_i,k_2}^{(v)}$
$\mathbf{X}_{r_i,k_3}^{(v)}$
360° video from rotated content r_i
Separately extract T_v-frame D'-dimensional
visual embeddings in direction k_1, k_2, k_3

Fig. 2. The proposed DOA-wise and recording-wise DOA-aware contrastive learnings for pretraining the audio feature extractor.

A series of DOA- and frame-wise audio embeddings $\mathbf{Z}_i^{(a)} \triangleq \{\mathbf{z}_{ikt}^{(a)}\}_{k=1,t=1}^{K,T_v}$ is then obtained with the projection head $\mathcal{H}$:

$$\{\mathbf{z}_{ikt}^{(a)}\}_{k=1}^{K} \leftarrow \mathcal{H}\left(\mathbf{y}_{it}^{(a)}\right), \tag{2}$$

where the features $\mathbf{y}_{it}^{(a)}$ at each frame t are independently transformed for temporally localizing information, followed by an adaptive average pooling to match the visual data of T_v frames.

B. Visual encoding

Let $\mathbf{X}_i^{(v)} \triangleq \{\mathbf{x}_{ikt}^{(v)}\}_{k=1,t=1}^{K,T_v}$ be the series of DOA-wise local images cropped from the 360° equirectangular visual data of recording $i \in [1, I]$, where $\mathbf{x}_{ikt}^{(v)} \in \mathbb{R}^{H \times W}$ is a local image that corresponds to DOA $k \in [1, K]$ at time $t \in [1, T]$, H and W are the height and width of each image (cropping size), and T_v is the number of frames. Note that $\mathbf{x}_{ikt}^{(v)}$ is centered on DOA k on the Gnomonic projection [18] of the original spatial visual data. A series of DOA- and frame-wise visual embeddings $\mathbf{Z}_i^{(v)} \triangleq \{\mathbf{z}_{ikt}^{(v)}\}_{k=1,t=1}^{K,T_v}$ is obtained with the visual encoder $\mathcal{V}$ as follows:

$$\{\mathbf{z}_{ikt}^{(v)}\}_{t=1}^{T_v} \leftarrow \mathcal{V}\left(\{\mathbf{x}_{ikt}^{(v)}\}_{t=1}^{T_v}\right), \tag{3}$$

where the same visual encoding is independently applied to each DOA k to obtain the embeddings that represent the classes of visible sound objects.

C. Self-supervised learning (pretraining)

We describe the two variants of contrastive learning.

1) Similarity measures: Since the audio and visual data of the same spatial recording usually have the DOA-wise correspondence, we define the similarity between recordings i and j for each DOA k as the cosine similarity as follows:

$$\mathcal{S}_{\text{DOA}}(\mathbf{Z}_{ik}^{(a)}, \mathbf{Z}_{jk}^{(v)}) = \frac{1}{T_v} \sum_{t=1}^{T_v} \frac{\mathbf{z}_{ikt}^{(a)\mathsf{T}} \mathbf{z}_{jkt}^{(v)}}{\|\mathbf{z}_{ikt}^{(a)}\| \|\mathbf{z}_{jkt}^{(v)}\|}, \tag{4}$$

where $\mathbf{Z}_{ik}^{(a)} \triangleq \{\mathbf{z}_{ikt}^{(a)}\}_{t=1}^{T_v}$ and $\mathbf{Z}_{ik}^{(v)} \triangleq \{\mathbf{z}_{ikt}^{(v)}\}_{t=1}^{T_v}$.

We here focus on the InfoNCE loss [19] defined as follows:

$$\mathcal{I}\left(\mathbf{Z}, \mathbf{Z}_p, \mathbf{U}_n\right) = -\frac{\exp\left(\mathcal{S}(\mathbf{Z}, \mathbf{Z}_p)/\tau\right)}{\sum_{\mathbf{Z}_n \in \mathbf{U}_n} \exp\left(\mathcal{S}(\mathbf{Z}, \mathbf{Z}_n)/\tau\right)}, \tag{5}$$

where $\mathbf{Z}$ is an anchor, $\mathbf{Z}_p$ is a positive sample, $\mathbf{U}_n$ is a set of negative samples, and τ is a temperature hyperparameter. Minimizing $\mathcal{I}$ encourages the similarity of $\mathbf{Z}$ to $\mathbf{Z}_p$ to be larger than to $\mathbf{U}_n$ in a contrastive manner.

2) DOA-wise contrastive learning: The loss $\mathcal{L}_{\text{DOA}}$ used in this variant is calculated as follows:

$$\mathcal{L}_{\text{DOA}} = \sum_{i,k=1}^{I,K} \mathcal{I}\left(\mathbf{Z}_{ik}^{(a)}, \mathbf{Z}_{ik}^{(v)}, \{\mathbf{Z}_{jk}^{(v)}\}_{j=1}^{I} \cup \{\mathbf{Z}_{ik'}^{(v)}\}_{k'=1}^{K}\right), \tag{6}$$

where the similarity between the audio and visual embeddings is maximized when they correspond to the same DOA.

3) Recording-wise contrastive learning: The loss $\mathcal{L}_{\text{REC}}$ used in this variant is calculated as follows:

$$\mathcal{L}_{\text{REC}} = \sum_{i=1}^{I} \mathcal{I}\left(\mathbf{Z}_i^{(a)}, \mathbf{Z}_i^{(v)}, \{\mathbf{Z}_j^{(v)}\}_{j=1}^{I} \cup \{\mathbf{Z}_{r_i}^{(v)}\}\right), \tag{7}$$

where $\mathbf{Z}_{r_i}^{(v)}$ is a series of visual embeddings that corresponds to the visual data from the rotated recording r_i, which is obtained by rotating the spatial information of the recording i. The recording-wise similarity used in $\mathcal{L}_{\text{REC}}$ is obtained by averaging the DOA-wise similarities over all the directions:

$$\mathcal{S}_{\text{REC}}(\mathbf{Z}_i^{(a)}, \mathbf{Z}_j^{(v)}) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{S}_{\text{DOA}}(\mathbf{Z}_{ik}^{(a)}, \mathbf{Z}_{jk}^{(v)}). \tag{8}$$

D. Supervised learning (fine-tuning)

Using annotated data, the audio feature extractor $\mathcal{A}$ is fine-tuned for SELD based on activity-coupled Cartesian DOA representation (ACCDOA) [20]. Specifically, multiple sound events of the same class can be separately assigned to different tracks. Let $\hat{\mathbf{P}} \in \mathbb{R}^{T' \times C \times N \times 3}$ be a multi-ACCDOA vector of T' frames, C classes, N tracks given by

$$\hat{\mathbf{P}} \leftarrow \mathcal{H}' \left(\mathcal{A} \left(\mathbf{X} \right) \right), \tag{9}$$

where $\mathbf{X} \in \mathbb{R}^{(4+3) \times F \times T'_a}$ is the multichannel spectrogram computed in the same way as $\mathbf{X}_i^{(a)}$, where T_a is the number of frames. The projection head $\mathcal{H}'$ consists of several fully-connected layers, followed by an adaptive average pooling to suit the target time resolution T'. Since the SELD label for frame t, class c, and track n is given as a pair of the activity $a^*_{tcn} \in \{0, 1\}$ and the Cartesian DOA vector $\mathbf{R}_{tcn}$, the target ACCDOA vector is given by

$$\mathbf{P}^*_{tcn} = a^*_{tcn} \mathbf{R}_{tcn}. \tag{10}$$

The overall network is trained to minimize:

$$\mathcal{L}^{\mathrm{PIT}} = \frac{1}{TC} \sum_{t,c=1}^{T',C} \min_{\alpha \in \mathrm{Perm}(ct)} \mathcal{L}^{\mathrm{ACCDOA}}_{\alpha,tc}, \tag{11}$$

$$\mathcal{L}^{\mathrm{ACCDOA}}_{\alpha,tc} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{MSE} \left(\mathbf{P}^*_{\alpha,tcn}, \hat{\mathbf{P}}_{tcn} \right), \tag{12}$$

where $\alpha \in \mathrm{Perm}(t)$ is a possible frame-level permutation of M tracks at the frame t and $\mathrm{Perm}(t)$ is a set of all possible permutations. $\mathrm{MSE}(\cdot, \cdot)$ is the mean square error function.

IV. Evaluation

This section reports a comparative experiment conducted for evaluating the proposed AV-SSL methods with the two variants of contrastive learning.

A. Network configuration

The STFT spectrograms with a shifting interval of 480 samples and a window size of 1024 were converted to the mel spectrograms with $F = 64$ mel bins. The number of DOAs K was 220.

We used a ResNet-Conformer [12] for the audio feature extractor $\mathcal{A}$ as shown in Figs. 3 (a)–(c). The architecture of the projection head $\mathcal{H}$ consisted of two linear+SiLU layers followed a linear layer which output dimension was 220 × 128 as shown in Figs. 3 (d). The FOA audio data were resampled at a sampling rate of 24 kHz and split into 2-second clips in the training.

The visual encoder $\mathcal{V}$ consisted of 9-layer $R(2+1)D$ convolution layers [21] as shown in Fig. 4. The input equirectangular visual data were extracted at the time resolution T_v of 16 (8 Hz). The field of view ψ of visual crops for each direction was 40°, and the resolution $H \times W$ was 16×16. The horizontal and vertical flipping was applied to the visual crops as data augmentation.

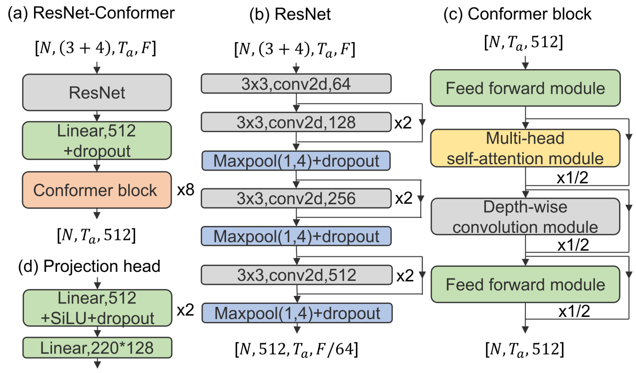

Fig. 3. The architecture of the audio encoder.

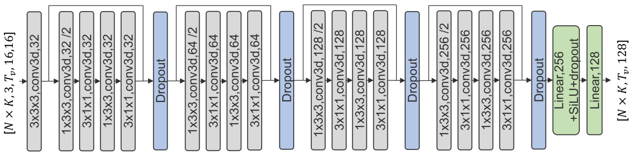

Fig. 4. The architecture of the visual encoder.

B. Pretraining

We used the YT-360 dataset [7] for pretraining. It contains VR contents collected from YouTube, each of which consists of a synchronized pair of FOA audio data and equirectangular visual data, including 246 hours of 10-second-long recordings on diverse genres including music and sports. After removing recordings with some missing channels, 104 hours of the training data and 20 hours of the validation data were used.

By the proposed AV-SSL methods, the ResNet-Conformer was trained for 100 epochs using AdamW optimizer [22] with a batch size of 4, a learning rate of 10^{-4}, a weight decay of 10^{-5}. Rotated recordings for the negative sample were obtained by randomly rotating the equirectangular visual data around the z-axis. The temperature hyper-parameter τ was 0.1. The SpecAugment [9] and the dropout with a rate of 0.1 were used. The model having the lowest validation loss was used for the downstream SELD task.

The conventional AV-SSL method called AVC [8] was also evaluated as a baseline for comparison. Specifically, the loss $\mathcal{L}_{\mathrm{AVC}}$ was calculated with the recording-wise similarity $\mathcal{H}(\{\mathbf{z}_{it}^{(a)}\}_{t=1}^{T_v}, \{\mathbf{z}_{it}^{(v)}\}_{t=1}^{T_v})$, where $\mathbf{z}_{it}^{(a)} \in \mathbb{R}^{128}$ was obtained by max-pooling the DOA-wise visual embeddings $\{\mathbf{z}_{ikt}^{(v)}\}_{k=1}^{K}$ and $\mathbf{z}_{it}^{(a)} \in \mathbb{R}^{128}$ was directly obtained from the output layer $\mathcal{H}$ with output dimension 128.

As in [7], curriculum learning was introduced for the proposed methods, where two audio feature extractors $\mathcal{A}$ were first trained with AVC for 50 epochs, then trained with the two different proposed AV-SSLs for 50 epochs.

C. Fine-tuning

The SELD performance was evaluated on the STARSS22 and Synth1 datasets [5]. The STARSS22 is a dataset of actual recorded FOA data annotated with the activations, classes, and

TABLE I
CLASS-WISE ACTIVITIY IN THE STARSS22 DATASET [5].

	Fem. speech	Male speech	Clap	Phone	Laugh	Dom. sounds	Footsteps	Door	Music	Music. instr.	Faucet	Bell	Knock
Frame coverage (% total frames)	20.4	37.6	0.7	1.4	2.7	17.9	1.3	0.6	29.4	4.0	1.7	1.5	0.1

TABLE II
EVALUATION RESULT FOR THE TWO DATASETS.

Fine-tuning dataset	Pretraining method	$ER_{\leq 20°}$ ↓	$F_{\leq 20°}$ ↑	LE ↓	LR ↑	$SELD$ ↓
STARSS22+Synth	None	0.53	48.9 %	18.2°	68.7 %	0.364
	AVC[8]	0.52	49.7 %	17.9°	69.0 %	0.359
	AV-SSL with DOA-wise contrastive learning	**0.51**	50.5 %	17.9°	**70.1 %**	0.351
	AV-SSL with recording-wise contrastive learning	**0.51**	**51.6 %**	**17.1°**	69.0 %	**0.349**
STARSS22	None	0.65	**37.9 %**	**22.1°**	**58.4 %**	**0.452**
	AVC[8]	**0.63**	37.7 %	22.4°	55.2 %	0.458
	AV-SSL with DOA-wise contrastive learning	0.68	35.8 %	22.9°	53.7 %	0.478
	AV-SSL with recording-wise contrastive learning	0.67	36.3 %	23.2°	56.3 %	0.467

TABLE III
CLASS-WISE INCREASE OR DECREASE OF THE THREE SELD METRICS BY EACH PRETRAINING.

Fine-tuning dataset	Pretraining method	Δ Metrics	Dom. sounds	Door	Music	Music. instr.	Bell	Knock
STARSS22+Synth	AVC [8]	$\Delta F_{\leq 20°}$	+0.03	−0.02	−0.02	+0.02	+0.04	+0.01
		$\Delta(1 - \bar{LE}/180)$	+0.01	+0.01	−0.01	0.0	+0.00	−0.0
		ΔLR	+0.02	−0.01	+0.05	+0.01	+0.08	−0.0
	DOA-wise	$\Delta F_{\leq 20°}$	+0.05	+0.08	+0.02	−0.12	+0.11	−0.1
		$\Delta(1 - \bar{LE}/180)$	+0.0	+0.0	−0.01	−0.03	+0.02	−0.0
		ΔLR	+0.07	+0.07	+0.09	−0.02	+0.01	−0.05
	recording-wise	$\Delta F_{\leq 20°}$	+0.1	+0.06	−0.04	−0.01	+0.12	−0.03
		$\Delta(1 - \bar{LE}/180)$	+0.01	+0.01	−0.02	−0.0	+0.02	−0.0
		ΔLR	+0.04	−0.01	+0.02	−0.0	+0.06	−0.0
STARSS22	AVC [8]	$\Delta F_{\leq 20°}$	−0.02	+0.09	+0.0	−0.05	−0.26	−0.18
		$\Delta(1 - \bar{LE}/180)$	−0.01	+0.01	+0.0	−0.01	−0.09	+0.01
		ΔLR	+0.07	+0.07	−0.05	+0.06	−0.17	−0.38
	DOA-wise	$\Delta F_{\leq 20°}$	−0.0	+0.04	+0.02	−0.14	−0.2	−0.07
		$\Delta(1 - \bar{LE}/180)$	−0.0	+0.0	−0.01	−0.05	−0.06	+0.02
		ΔLR	+0.06	+0.07	−0.03	+0.06	−0.1	−0.31
	recording-wise	$\Delta F_{\leq 20°}$	+0.03	+0.0	−0.0	−0.1	−0.16	−0.2
		$\Delta(1 - \bar{LE}/180)$	−0.0	+0.02	+0.02	−0.02	−0.05	−0.0
		ΔLR	+0.05	+0.14	−0.14	−0.06	−0.09	−0.46

DOAs of 13 sound events every 0.1 seconds, where the class labels follow the Audioset ontology [14]. This dataset consists of 2.9 hours of training data and 2 hours of validation data. Note that some classes have very low frame coverages in STARSS22 as shown in Table I. The Synth1 is a dataset of synthesized FOAs annotated in the same way as STARSS22, obtained by remixing source signals from FSD50K [23] with room RIRs from TAU-SRIR database [24]. Two datasets of different sizes constructed from these datasets were used for fine-tuning. The first one was a dataset made up of all training data from STARSS22 and Synth1 (STARSS22+Synth1). The second one consisted only of all training data from STARSS22 (STARSS22). A dataset that consists of all validation data from STARSS22 and Synth1 was used as the validation dataset in both conditions. The original FOA data were split every 5 seconds, and they were sampled uniformly over all 13 classes to deal with the class imbalance in STARSS22.

Each pretrained model was fine-tuned with a head $\mathcal{H}'$ to output 3-track multi-ACCDOAs [20]. The architecture of $\mathcal{H}'$ was the same as the head $\mathcal{H}$ shown in the Sec. IV-A except that the output dimension was changed to $13 \times 3 \times 3$. AdamW with a learning rate of 10^{-4} and a weight decay of 10^{-5} was used for training. The models were trained for 1000 epochs. For data augmentation, the input FOA was randomly rotated. The dropout rate was 0.1 for ResNet-Conformer and 0.05 for $\mathcal{H}'$.

Models pretrained with our methods were compared with a non-pretrained model and a model pretrained with the conventional AVC method based on SELD score [1] on the validation dataset. The SELD score is obtained by averaging over four metrics as follows:

$$SELD = \frac{1}{4}\left(ER_{\leq 20^\circ} + (1 - F_{\leq 20^\circ}) + \frac{LE}{180} + (1 - LR)\right),$$

where the smaller score indicates better performance. $ER_{\leq 20^\circ}$ is the location-dependent error rate, where the prediction is counted as a true positive only if the estimated class activity is correct and the distance between the estimated and reference DOAs is smaller than 20°. $F_{\leq 20^\circ}$, LE and LR are the averages of class-wise metrics $F_{c,\leq 20^\circ}$, LE_c and LR_c, respectively. $F_{c,\leq 20^\circ}$ is the location-dependent F1-score for class c. LE_c is the average DOA estimation error over only the correctly detected events of class c. LR_c is the location-independent recall for class c. LE_c and LR_c can be considered to be the performance of detection and localization, respectively. The performance of each model was measured by averaging the metrics over the epochs with the first to tenth highest SELD scores.

D. Experimental results

Table II shows the overall SELD performances in terms of the five metrics, i.e., $ER_{\leq 20^\circ}$, $F_{\leq 20^\circ}$, LE, LR, and SELD score for each dataset and each pretraining method. In the STARSS22+Synth1 dataset, both the AV-SSL with DOA-wise contrastive learning and the AV-SSL with recording-wise contrastive learning improved the SELD scores by about 1.5 pts, while the conventional AVC improved only about 0.5 pts. This result indicates that our proposed approach is more suitable for the pretraining of SELD than AVC.

Table III shows the performance gaps in the class-wise metrics, $F_{c,\leq 20^\circ}$, LE_c and LR_c, between the non-pretrained model and each pretrained model, where all the metrics were transformed to the range of $[0, 1]$ for readability, and only some of the classes required for the following discussion were picked up due to the page limitation. The proposed AV-SSLs degraded the SELD scores more than AVC when only a few hours of imbalanced data were used for training (STARSS22). This would be partly due to the domain mismatch between YT-360 and STARSS22 and the class imbalance of STARSS22. As for the knock class, which was not included in the pretraining dataset (YT-360), for example, the detection performance LR_c got particularly worse. While in the classes related to home sounds (e.g., domestic sounds and door), which frequently appear in YT-360, the detection performance was improved.

Another reason for the degradation would be that a large amount of background music data independent from the visual data are included in YT-360. Background music was considered to have a negative impact on the localization of sound events related to music because it did not spatially correspond to the paired visual data. In fact, the localization performance LE_c was significantly degraded by the pretraining for the musical instruments and bell classes.

V. Conclusion

We proposed two variants of pretraining an audio feature extractor useful for SELD using spatial audio-visual recordings. To obtain the latent audio features representing not only the classes but also the DOAs of sound events, the audio encoder takes the FOA data as input, and outputs the audio embeddings over the DOA grid. The transfer learning with a sufficient amount of data showed the effectiveness of the proposed AV-SSLs as pretraining for SELD. For future work, one should deal with the deterioration in the SELD performance when sufficient labeled data is unavailable. One of the promising approaches would be to prepare the dataset of spatial audio-visual recordings covering various domains with good correspondence between audio and visual data.

.

Acknowledgment

This work was supported in part by NEDO, JSPS KAKENHI Nos. 20H00602, 19H04137, and ISHIZUE 2023 of Kyoto University.

References

[1] A. Politis *et al.*, "Overview and evaluation of sound event localization and detection in DCASE 2019," *IEEE/ACM TASLP*, vol. 29, pp. 684–698, 2020.

[2] T. N. T. Nguyen *et al.*, "A sequence matching network for polyphonic sound event localization and detection," in *IEEE ICASSP*, 2020, pp. 71–75.

[3] S. Adavanne *et al.*, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE JSTSP*, vol. 13, no. 1, pp. 34–48, 2018.

[4] K. Shimada *et al.*, "ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *IEEE ICASSP*, 2021, pp. 915–919.

[5] A. Politis *et al.*, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *arXiv preprint arXiv:2206.01948*, pp. 1–6, 2022.

[6] Q. Wang, L. Chai, H. Wu, *et al.*, "The nerc-slip system for sound event localization and detection of dcase2022 challenge," DCASE2022 Challenge, Tech. Rep., Jun. 2022.

[7] P. Morgado *et al.*, "Learning representations from audio-visual spatial alignment," *NeurIPS*, vol. 33, pp. 4733–4744, 2020.

[8] R. Arandjelovic *et al.*, "Look, listen and learn," in *IEEE ICCV*, 2017, pp. 609–617.

[9] D. S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, pp. 1–6, 2019.

[10] N. Takahashi *et al.*, "Deep convolutional neural networks and data augmentation for acoustic event detection," in *INTERSPEECH*, 2016, pp. 2982–2986.

[11] L. Mazzon *et al.*, "First order ambisonics domain spatial augmentation for dnn-based direction of arrival estimation," in *DCASE Workshop*, 2019, p. 154.

[12] Q. Wang, J. Du, H. Wu, J. Pan, F. Ma, and C. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *CoRR*, vol. abs/2101.02919, 2021. arXiv: 2101.02919. [Online]. Available: https://arxiv.org/abs/2101.02919.

[13] Y. Gong *et al.*, "SSAST: Self-supervised audio spectrogram transformer," in *AAAI*, vol. 36, 2022, pp. 10 699–10 709.

[14] J. F. Gemmeke *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," in *IEEE ICASSP*, 2017, pp. 776–780.

[15] R. Scheibler *et al.*, "Sound event localization and detection with pre-trained audio spectrogram transformer and multichannel separation network," in *DCASE Workshop*, 2022, pp. 1–5.

[16] B. Korbar *et al.*, "Cooperative learning of audio and video models from self-supervised synchronization," *NeurIPS*, vol. 31, pp. 7774–7785, 2018.

[17] Á. González, "Measurement of areas on a sphere using fibonacci and latitude–longitude lattices," *Mathematical Geosciences*, vol. 42, pp. 49–64, 2010.

[18] E. W. Weisstein, "Gnomonic projection," *https://mathworld. wolfram. com/*, 2001.

[19] A. Oord *et al.*, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, pp. 1–13, 2018.

[20] K. Shimada *et al.*, "Multi-ACCDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *IEEE ICASSP*, 2022, pp. 316–320.

[21] L. Perotin *et al.*, "CRNN-based joint azimuth and elevation localization with the ambisonics intensity vector," in *IEEE IWAENC*, 2018, pp. 241–245.

[22] I. Loshchilov *et al.*, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[23] E. Fonseca *et al.*, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM TASLP*, vol. 30, pp. 829–852, 2021.

[24] A. Politis *et al.*, *TAU spatial room impulse response database (TAU-SRIR DB)*, 2022.