2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Self-supervised Neural Audio-Visual Sound Source Localization via Probabilistic Spatial Modeling

Yoshiki Masuyama[1,2], Yoshiaki Bando[1], Kohei Yatabe[2], Yoko Sasaki[1], Masaki Onishi[1], Yasuhiro Oikawa[2]

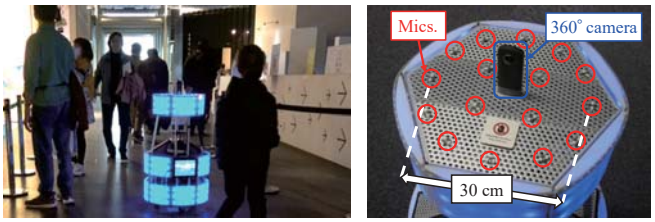
***Abstract*— Detecting sound source objects within visual observation is important for autonomous robots to comprehend surrounding environments. Since sounding objects have a large variety with different appearances in our living environments, labeling all sounding objects is impossible in practice. This calls for self-supervised learning which does not require manual labeling. Most of conventional self-supervised learning uses monaural audio signals and images and cannot distinguish sound source objects having similar appearances due to poor spatial information in audio signals. To solve this problem, this paper presents a self-supervised training method using 360° images and multichannel audio signals. By incorporating with the spatial information in multichannel audio signals, our method trains deep neural networks (DNNs) to distinguish multiple sound source objects. Our system for localizing sound source objects in the image is composed of audio and visual DNNs. The visual DNN is trained to localize sound source candidates within an input image. The audio DNN verifies whether each candidate actually produces sound or not. These DNNs are jointly trained in a self-supervised manner based on a probabilistic spatial audio model. Experimental results with simulated data showed that the DNNs trained by our method localized multiple speakers. We also demonstrate that the visual DNN detected objects including talking visitors and specific exhibits from real data recorded in a science museum.**

I. Introduction

Autonomous robots have to comprehend surrounding environments to decide their actions in our living circumstances. Major tasks of those robots are human-robot interaction [1] and security surveillance [2], which require localization and recognition of sound sources around the robots. While detection of the direction of arrivals (DoAs) of the sound has been performed based mainly on audio information obtained by a set of microphones [3]–[5], taking its correspondence to objects in the environment is also important for manipulating the robots. This requires to find sounding objects within visual observation, which is referred to audio-visual sound source localization (AV-SSL) in this paper.

The difficulty of AV-SSL arises from the excess diversity of the sounding objects. For instance, an autonomous robot interacting with an audience in a museum (as in Fig. 1) is exposed to sounds emitted from not only the target speaker but also the surrounding crowd and exhibits having very different appearances. Since recording and labeling every sound source object (not restricted to human) with all possible sound and appearance are impossible in practice, a strategy without manual supervision is necessary for learning the complicated relation between audio and visual observations.

[1]National Institute of Advanced Industrial Science and Technology, Japan
{masuyama.yoshiki, y.bando}@aist.go.jp
[2]Department of Intermedia Art and Science, Waseda University, Japan



(a) Robot in museum (b) Head of robot

Fig. 1: Autonomous robot called Peacock.

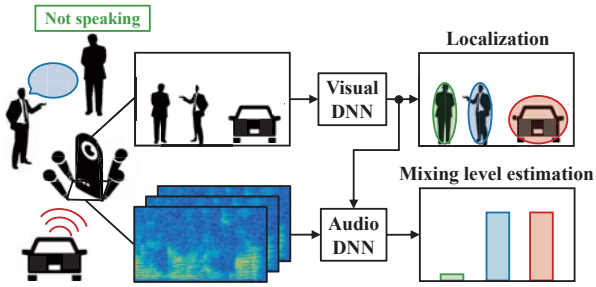


Fig. 2: Illustration of our multichannel AV-SSL system: sound source localization with mixing level estimation.

Self-supervised learning is one promising approach to AV-SSL as it does not require manual labeling. In the existing literature [6]–[8], two deep neural networks (DNNs), an audio network and a visual network, are utilized to relate the same event observed in both of the audio and visual recordings. They are trained so that their output features coincide with each other, where the data required for training is only monaural audio signals with the corresponding images [6], [7] (or videos [8]). These methods can learn various kinds of sounding objects contained in the training data and perform well for audio events having different appearances (e.g., a musical performance with different instruments). However, applying them to audio events consisting of multiple sound sources having similar appearances (e.g., people talking simultaneously, which is typical in our daily life) is not easy [9]. This should be because monaural audio signals do not contain spatial information which is a key to localize objects in visual observations.

In this paper, we present a self-supervised training method of audio and visual DNNs for AV-SSL using a 360° camera and a microphone array as illustrated in Fig. 2. The proposed method is composed of two DNNs: visual DNN for object detection and audio DNN for source number estimation. The visual DNN detects and localizes sound source candidates within the input image, where the candidates are allowed to include silent objects which should be ignored in AV-

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

SSL. Then, the audio DNN estimates the mixing levels (mixing proportion) of sound for the candidates to omit the silent objects as in Fig. 2. The key idea of our method is training both audio and visual DNNs based on a probabilistic spatial audio model, which allows us to train the two DNNs without manual labeling. This training is formulated as a probabilistic inference of a spatial audio model that has DoAs and mixing levels of sound sources as latent variables. In summary, the main contributions of this paper are as follows: (1) proposing a self-supervised training method for AV-SSL using multichannel audio signals to distinguish sound sources with similar appearances, (2) formulating a probabilistic framework including mixing levels as the latent variables to simultaneously estimate the number and locations of sound sources, (3) deriving an objective function for integrating audio and visual information based on a spatial audio model, and (4) experimentally confirming the applicability of the proposed method to real-world data recorded at a science museum using a robotic system.

II. Related Work

SSL has typically been formulated as an audio problem using a set of microphones, e.g., physical-model-based [3]–[5] and DNN-based methods [10], [11]. Integration of audio and visual information has also been studied to go beyond the audio-only settings, which is briefly reviewed in this section. We also describe recent progress of self-supervised learning.

A. Audio-Visual Sound Source Localization (AV-SSL)

AV-SSL has been applied to various applications including speaker tracking [12], [13], traffic monitoring [14], and search-and-rescue tasks [15]. Since audio SSL and visual object detection have a similar purpose (i.e., finding the position of some objects), the standard strategy is to integrate methods for each task, which often requires a pre-trained DNN for visual object detection [16]. Because of this requirement of the pre-trained DNN, existing literature mainly focuses on human [12], [13] or the objects whose well-established dataset of labeled images are relatively easy to obtain. The focus of this paper is in a more general setting: the sounding objects to be detected include not only humans but also objects whose labeled dataset is difficult to obtain due to their excess variations. Therefore, supervised training as in [12], [13] cannot be applied to our situation.

B. Self-supervised Learning with Audio and Visual Data

Self-supervised learning allows us to avoid the requirement of labels on the training dataset, which is quite desirable property for AV-SSL aiming at various types of objects. Some early investigations have utilized monaural audio signals [6]–[8] which do not contain any spatial information. These methods are based on the contents in the observation, which requires differences in contents such as appearance and type of sound. A recent study distinguishes sound sources from the temporal motion of corresponding objects (e.g., movements of violins) [9]. Our proposal aims at another direction, which distinguishes sound sources with similar appearance based on the spatial information (contained in a 360° image and multichannel audio signal).

Some recent studies on AV-SSL utilize multichannel audio signals with self-supervised learning [17], [18]. These methods utilize teacher-student learning techniques, where a pre-trained visual DNN produces training targets of the audio DNN for localization. That is, while an audio DNN does not need manual supervision, these self-supervised methods require labeled data for visual DNN. In contrast, our method trains both audio and visual DNNs in a fully-self-supervised manner, i.e., no label is necessary for the training thanks to our probabilistic framework of the spatial audio model.

Meanwhile, spatial models have also been utilized for fully-self-supervised learning in each modality. For monocular depth estimation, self-supervised learning is an active research topic because oracle depth maps are not easy to obtain [19]. Self-supervised learning of neural sound source separation has also been investigated by using a spatial model of multichannel audio signals [20], [21]. These methods do not require clean source signals, which are often unavailable in the real recordings. They utilize a complex Gaussian mixture model (cGMM) [22]–[24] and train DNNs to maximize the marginal likelihood or evidence lower bound (ELBO) of the probabilistic model. Inspired from these existing works, our self-supervised method for AV-SSL aims to maximize the ELBO of the variant of a cGMM.

III. Self-supervised Training of Audio-Visual Sound Source Localization

In order to achieve localization of objects with similar appearance, we present a self-supervised training method for multichannel neural AV-SSL that uses pairs of 360° images and the corresponding multichannel audio mixtures. The self-supervised training is formulated in a probabilistic manner based on a spatial audio model called a cGMM as in [21].

A. Problem Specification

The problem setting of our self-supervised training for multichannel neural AV-SSL is defined as follows:

Training data:
N pairs of (1) 360° images $\mathbf{Y}^{(n)} \in \mathbb{R}_{+}^{I \times J \times 3}$ captured by a 360° camera, and (2) multichannel audio mixtures $\mathbf{x}_{tf}^{(n)} \in \mathbb{C}^{M}$ recorded by an M-channel microphone array.
Training targets:
(1) visual DNN $\mathscr{G}$ that takes a 360° image as input and estimates directions of sound source candidates, and
(2) audio DNN $\mathscr{H}$ that estimates mixing levels of sound source candidates from the estimated directions and multichannel audio mixture.
Assumptions:
(1) the configuration of the microphone array is given,
(2) movements of sources are negligibly small in T frames.

---

where $i = 1, \ldots, I$ and $j = 1, \ldots, J$ are respectively the vertical and horizontal indices for an image, and $t = 1, \ldots, T$ and $f = 1, \ldots, F$ are respectively the time and frequency indices for a multichannel audio spectrogram. The suffix (n)

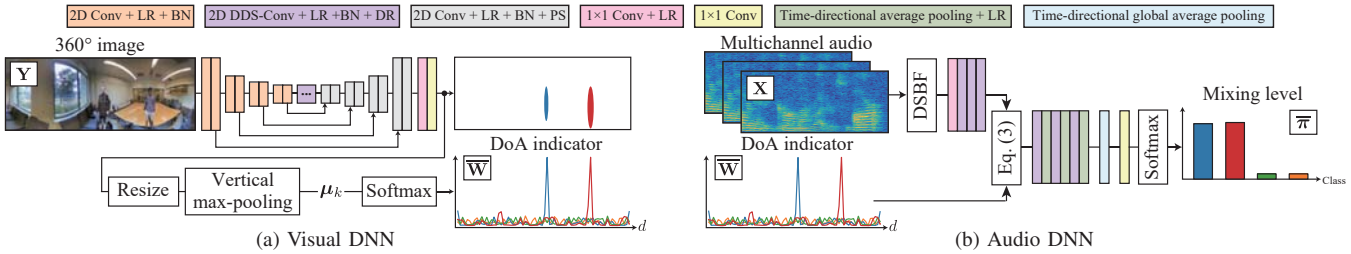


Fig. 3: Architectures of visual and audio DNNs. Here, LR, BN, DR, PS, and DDS-Conv indicate the leaky-ReLU, batch normalization, dropout, pixel shuffler, and dilated depthwise-separable convolution, respectively.

is hereinafter omitted because the objective function and the networks are independently defined for each of the N data. Under the first assumption, we can obtain the steering vectors $\mathbf{b}_{fd} \in \mathbb{C}^M$ for potential directions $d = 1, \ldots, D$ and associate the directions of sounds and the locations of the corresponding objects. We set D potential directions at the same intervals on a horizontal plane for simplicity. Under the second assumption, this paper treats not a video stream but an image. The extension to handle a video stream is included in our future work. Here, we allow that there exist objects that do not produce sounds (e.g., persons not speaking), which is not considered in conventional audio-visual self-supervised learning. In addition, we also allow the number of sound sources is unknown.

B. System Overview of Multichannel Neural AV-SSL

Our multichannel neural AV-SSL system is composed of two DNNs: visual DNN $\mathscr{G}$ and audio DNN $\mathscr{H}$ (Fig. 3).

1) Visual DNN: The visual DNN $\mathscr{G}$ estimates the DoA indicator $\overline{\mathbf{W}} = [\overline{\mathbf{w}}_1, \ldots, \overline{\mathbf{w}}_K]^\mathsf{T}$, where K is the number of potential sound sources, and $\overline{\mathbf{w}}_k = [\overline{w}_{k1}, \ldots, \overline{w}_{kD}]^\mathsf{T} \in \mathbb{R}_+^D$ indicates how the direction d is dominated by source k:

$$\overline{\mathbf{w}}_k = \mathrm{softmax}(\boldsymbol{\mu}_k) = \mathrm{softmax}(\mathscr{G}(\mathbf{Y})_k), \tag{1}$$

where $\boldsymbol{\mu}_k = [\mu_{k1}, \ldots, \mu_{kD}]^\mathsf{T}$ is the output of $\mathscr{G}$. The main part of the visual DNN is similar to the U-Net architecture [25] that is followed by additional 1×1 convolutional layers, resize, and vertical max pooling as shown in Fig. 3-(a). The output of the U-Net module represents activation maps of source candidates within an input image. To associate the 2D activations to the 1D potential directions on a horizontal plane, the vertical max pooling is applied to the map.

2) Audio DNN: The audio DNN predicts the mixing levels $\overline{\boldsymbol{\pi}} \in \mathbb{R}_+^K$ from the observed M-channel audio mixture $\mathbf{X} \in \mathbb{C}^{T \times F \times M}$ and the predicted DoA indicators $\overline{\mathbf{W}}$:

$$\overline{\boldsymbol{\pi}} = \mathscr{H}(\mathbf{X}, \overline{\mathbf{W}}). \tag{2}$$

In the audio DNN $\mathscr{H}$ (Fig. 3-(b)), we first apply delay-and-sum beamforming (DSBF) for every potential direction d and extract their amplitude as $\mathbb{U}\big[\log(|\mathbf{b}_{fd}^\mathsf{H} \mathbf{x}_{tf}|)\big]$, where $\mathbb{U}$ is the frequency-wise mean and variance normalization. The beamforming results are converted to direction-wise features $\mathbf{u}_{cd} \in \mathbb{R}^T$ by convolutional block where the frequencies are treated as channels, and c denotes the feature index.

To associate the audio and visual information, the audio features related to directions $\mathbf{u}_{cd}$ are transformed to those related to sources $\mathbf{v}_{ck} \in \mathbb{R}^T$ by using DoA indicators $\overline{\mathbf{W}}$:

$$\mathbf{v}_{ck} = \sum_{d=1}^{D} \overline{w}_{kd} \mathbf{u}_{cd}. \tag{3}$$

The mixing levels $\boldsymbol{\pi}$ are then obtained by passing the source-wise feature $\mathbf{v}_{ck}$ to another convolutional block, time-directional global average pooling, and affine block.

3) Inference of Multichannel Neural AV-SSL System: At the inference, our AV-SSL system predicts DoAs in the following two steps. We first obtain DoA indicators $\overline{\mathbf{W}}$ and mixing levels $\overline{\boldsymbol{\pi}}$ by passing the observation through the visual and audio DNNs. Then, we determine whether each source candidate k actually produces sound by thresholding the DoA indicator and mixing level. Finally, the DoA of each source $d_k^\star \in \{1, \ldots, D\}$ is calculated as follows:

$$d_k^\star = \arg\max_d \ \overline{w}_{kd}. \tag{4}$$

C. Generative Model of Multichannel Audio Signal

To associate the spatial information in audio and visual observations, we formulate a cGMM-based spatial model of a multichannel mixture signal. The cGMM has been known to robustly work in real-world environments [22], [26]. As in the original cGMM, an observed signal $\mathbf{x}_{tf} \in \mathbb{C}^M$ is represented with K source signals $s_{tfk} \in \mathbb{C}$ by

$$\mathbf{x}_{tf} = \sum_{k=1}^{K} z_{tfk} \cdot \mathbf{a}_{fk} s_{tfk}, \tag{5}$$

where $z_{tfk} \in \{0, 1\}$ is a time-frequency (T-F) mask that satisfies $\sum_{k=1}^K z_{tfk} = 1$, and $\mathbf{a}_{fk} \in \mathbb{C}^M$ is the steering vector of source k at frequency f. The T-F mask z_{tfk} is introduced by assuming the source spectra sufficiently sparse in the T-F domain, and is assumed to follow a categorical distribution (denoted as Cat):

$$\mathbf{z}_{tf} = [z_{tf1}, \ldots, z_{tfK}]^\mathsf{T} \sim \mathrm{Cat}\left(\pi_1, \ldots, \pi_K\right), \tag{6}$$

where $\pi_k \in \mathbb{R}_+$ is the mixing level of source k that satisfies $\sum_k \pi_k = 1$. Each source signal s_{tfk} is assumed to follow a complex Gaussian distribution characterized with a power spectral density $\lambda_{tfk} \in \mathbb{R}_+$ as follows:

$$s_{tfk} \sim \mathcal{N}_\mathbb{C}\left(0, \lambda_{tfk}\right). \tag{7}$$

Based on these assumptions, the observation $\mathbf{x}_{tf}$ follows a multivariate complex Gaussian mixture distribution given by

$$\mathbf{x}_{tf} \sim \sum_{k=1}^{K} \pi_k \mathcal{N}_\mathbb{C}\left(\mathbf{0}, \lambda_{tfk} \mathbf{H}_{fk}\right), \tag{8}$$

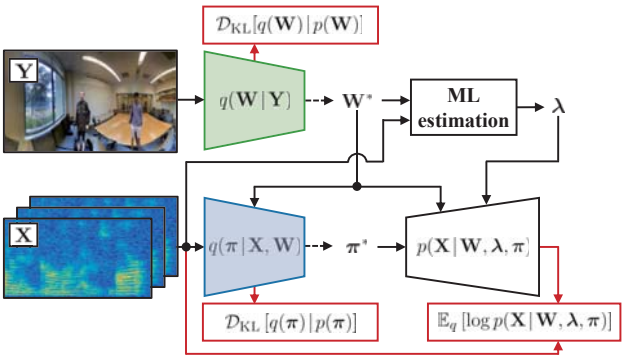


Fig. 4: Overview of our self-supervised training based on probabilistic spatial audio model. The green and blue blocks are the visual and audio DNNs, respectively.

where $\mathbf{H}_{fk} = \mathbb{E}[\mathbf{a}_{fk}\mathbf{a}_{fk}^{\mathsf{H}}] \in \mathbb{C}^{M\times M}$ is the spatial covariance matrix (SCM) of source k. To deal with the unknown number of sources, we encourage the shrinkage of redundant source classes by putting a Dirichlet distribution (denoted as Dir) [23] on the mixing levels:

$$[\pi_1,\ldots,\pi_K]^{\mathsf{T}} \sim \mathrm{Dir}\left(\alpha_0,\ldots,\alpha_0\right), \tag{9}$$

where $\alpha_0 \in \mathbb{R}_+$ is a hyperparameter.

To associate the multichannel audio observation $\mathbf{X}$ and DoA candidates $\mathbf{W}$ estimated by the visual DNN, we represent the SCM of source k by the weighted sum of template SCMs $\mathbf{G}_{fd} = \mathbf{b}_{fd}\mathbf{b}_{fd}^{\mathsf{H}}$ for potential directions d as follows:

$$\mathbf{H}_{fk} \approx \sum_{d=1}^{D} w_{kd}\mathbf{G}_{fd} + \epsilon\mathbf{I}, \tag{10}$$

where ϵ is a small number to ensure $\mathbf{H}_{fk}$ positive definite. To prevent the estimates of w_{kd} from taking exceedingly large values, we put the following log-normal prior on w_{kd}:

$$w_{kd} \sim \mathcal{LN}\left(0,\sigma_0^2\right), \tag{11}$$

where $\sigma_0 \in \mathbb{R}_+$ is a scale parameter.

D. Self-supervised Training Based on Variational Inference

The audio and visual DNNs are trained in a self-supervised manner so that they estimate the posterior distributions of $\boldsymbol{\pi}$ and $\mathbf{W}$, respectively (Fig. 4). We employ a Bayesian approach to treat the unknown number of sources in a unified framework [23]. Specifically, we represent the posterior distributions with the outputs of the DNNs, $\boldsymbol{\mu}_k = \mathscr{G}(\mathbf{Y})_k$ and $\overline{\boldsymbol{\pi}} = \mathscr{H}(\mathbf{X},\mathbf{W})$, as follows:

$$q(\mathbf{W}\mid\mathbf{Y}) = \prod_{k=1}^{K}\prod_{d=1}^{D}\mathcal{LN}\left(\mu_{kd},\sigma_k^2\right), \tag{12}$$

$$q(\boldsymbol{\pi}\mid\mathbf{X},\mathbf{W}) = \mathrm{Dir}(\beta\overline{\pi}_1,\ldots,\beta\overline{\pi}_K), \tag{13}$$

where $\sigma_k \in \mathbb{R}_+$ and $\beta \in \mathbb{R}_+$ are scale parameters representing the uncertainty and jointly optimized in this training. We train the two DNNs so that the variational posterior $q(\boldsymbol{\pi}|\mathbf{X},\mathbf{W})q(\mathbf{W}|\mathbf{Y})$ approximates the true posterior $p(\mathbf{W},\boldsymbol{\pi}|\mathbf{X},\boldsymbol{\lambda})$ by minimizing the Kullback–Leibler (KL) divergence between them. This minimization, which is indeed intractable, is conducted by maximizing the ELBO [23] defined as follows:

$$\mathcal{L} = \mathbb{E}_q\left[\log p(\mathbf{X}\,|\,\mathbf{W},\boldsymbol{\lambda},\boldsymbol{\pi})\right] - \mathcal{D}_{\mathrm{KL}}[q(\mathbf{W},\boldsymbol{\pi})\,|\,p(\mathbf{W},\boldsymbol{\pi})]\,. \tag{14}$$

Since the first term of Eq. (14) is difficult to calculate analytically, we utilize Monte Carlo approximation whose gradient can be obtained by using reparameterization trick [27]:

$$\mathbb{E}_q\left[\log p(\mathbf{X}\,|\,\mathbf{W},\boldsymbol{\lambda},\boldsymbol{\pi})\right] \approx \sum_{t=1}^{T}\sum_{f=1}^{F}\log\sum_{k=1}^{K}\frac{\pi_k^*}{|\lambda_{tfk}\hat{\mathbf{H}}_{fk}|}\exp\left(-\frac{\mathbf{x}_{tf}^{\mathsf{H}}\hat{\mathbf{H}}_{fk}^{-1}\mathbf{x}_{tf}}{\lambda_{tfk}}\right) \tag{15}$$

where $\hat{\mathbf{H}}_{fk} = \sum_{d=1}^{D} w_{kd}^*\mathbf{G}_{fd}$ is a sampled SCM, and w_{kd}^* and π_k^* are samples of $w_{kd}^* \sim q(w_{kd}\mid\mathbf{Y})$ and $\pi_k^* \sim q(\pi_k\mid\mathbf{X},\mathbf{W}^*)$, respectively. The power spectral density λ_{tfk} is substituted with a maximum likelihood estimate:

$$\lambda_{tfk} = \frac{1}{M}\mathbf{x}_{tf}^{\mathsf{H}}\hat{\mathbf{H}}_{fk}^{-1}\mathbf{x}_{tf}. \tag{16}$$

The second term of Eq. (14) is calculated separately as

$$\mathcal{D}_{\mathrm{KL}}\left[q(\mathbf{W},\boldsymbol{\pi})\,|\,p(\mathbf{W},\boldsymbol{\pi})\right] = \mathcal{D}_{\mathrm{KL}}\left[q(\mathbf{W})\mid p(\mathbf{W})\right] + \mathbb{E}_{q(\mathbf{W})}\left[\mathcal{D}_{\mathrm{KL}}\left[q(\boldsymbol{\pi}|\mathbf{W})\mid p(\boldsymbol{\pi})\right]\right]. \tag{17}$$

The first term in Eq. (17) is calculated analytically as follows:

$$\mathcal{D}_{\mathrm{KL}}\left[q(\mathbf{W})\,|\,p(\mathbf{W})\right] = \sum_{k=1}^{K}\sum_{d=1}^{D}\frac{\mu_{kd}^2+\sigma_k^2-\sigma_0^2}{2\sigma_0^2} + \log\frac{\sigma_0}{\sigma_k}, \tag{18}$$

where μ_{kd} is the output of the visual DNN $\mathscr{G}$. The second term is approximately calculated by using the Monte Carlo sampling as follows:

$$\mathbb{E}_{q(\mathbf{W})}\left[\mathcal{D}_{\mathrm{KL}}\left[q(\boldsymbol{\pi}\,|\,\mathbf{W})\,|\,p(\boldsymbol{\pi})\right]\right] \approx \log\frac{\Gamma(\beta\overline{\pi}_\cdot)}{\Gamma(K\alpha_0)} - \sum_{k=1}^{K}\log\frac{\Gamma(\beta\overline{\pi}_k)}{\Gamma(\alpha_0)} + \sum_{k=1}^{K}(\beta\overline{\pi}_k-\alpha_0)\{\psi(\beta\overline{\pi}_k)-\psi(\beta\overline{\pi}_\cdot)\}, \tag{19}$$

where $\overline{\boldsymbol{\pi}}$ is calculated with the sampled DoA indicators $\mathbf{W}^*$, $\overline{\pi}_\cdot$ represents $\sum_{k=1}^{K}\overline{\pi}_k$, $\Gamma(\cdot)$ is the gamma function, and $\psi(\cdot)$ is the digamma function.

Based on Eqs. (15)–(19), we can approximately calculate the ELBO $\mathcal{L}$ and train the audio and visual DNNs so that the ELBO is maximized. This maximization can be conducted with stochastic gradient descent because all the above equations are differentiable. Note that this training does not use any labels or oracle signals and is conducted in a fully-self-supervised manner as a Bayesian inference.

IV. Experimental Evaluation with Simulated Data

We validated our self-supervised training method for multichannel neural AV-SSL with simulated indoor environments where multiple persons speak simultaneously.

A. Experimental Configuration

Inspired by a synthesis method [28], we generated pairs of 360° images and corresponding multichannel audio mixtures. We synthesized the images by utilizing indoor images from 2D-3D-S dataset [29] and person images from the Clothing Co-Parsing dataset [30]. More specifically, persons were located and rendered 0.5 m to 2.0 m from the camera at random. The multichannel audio signals, on the other

TABLE I: Localization performance in F-measure under the first condition in which all persons spoke.

# of mics.	$M=2$		$M=4$		$M=6$	
# of srcs.	$L=2$	$L=3$	$L=2$	$L=3$	$L=2$	$L=3$
MUSIC	–	–	0.80	0.70	0.84	0.78
SRP-PHAT	0.41	0.37	0.65	0.56	0.74	0.63
w/o Aud	0.68	**0.68**	0.85	0.82	**0.85**	0.82
Proposed	**0.74**	0.66	**0.86**	**0.83**	**0.85**	**0.84**

TABLE II: Localization performance in F-measure under the second condition in which not all persons spoke.

# of mics.	$M=2$		$M=4$		$M=6$	
# of srcs.	$L=2$	$L=3$	$L=2$	$L=3$	$L=2$	$L=3$
MUSIC	–	–	0.73	0.64	**0.85**	**0.77**
SRP-PHAT	0.43	0.36	0.66	0.54	0.73	0.63
w/o Aud	0.55	0.54	0.62	0.63	0.69	0.67
Proposed	**0.70**	**0.70**	**0.77**	**0.73**	0.78	0.74

hand, were generated by convoluting room impulse responses (RIRs) to monaural speech signals selected from the WSJ0 corpus [31]. Speech signals were cut to a 1.0-second clip randomly, and mixed at random powers uniformly chosen between -2.5 dB and 2.5 dB from a reference value. The RIR was generated by using the image method [32] where the reverberation time (RT_{60}) was chosen at random between 0.2 s and 0.4 s, and the room dimension was fixed to 5.0 m $\times$ 5.0 m $\times$ 3.0 m for simplicity. A circular microphone array with a diameter of 20 cm was located at the center of the room. Gaussian noise was added to the mixtures with the signal-to-noise ratio of 20 dB to imitate diffuse noise.

To confirm the effectiveness of the audio DNN, we generated two types of datasets. (1) The first one contains two or three persons in each image, and all persons speak. (2) The other one contains two to four persons, but only two or three speak for simulating the situation where several persons were actually not speaking. For each of dataset types, we generated three datasets with the different number of microphones $M \in \{2, 4, 6\}$. For all conditions, we generated 20000 pairs of simulated data for training and 1000 pairs for testing, respectively. The images were downsampled to 88×288 and the audio signals were sampled at 16-kHz.

The visual and audio DNNs were jointly trained by using the AdamW optimizer for 200 epochs with the learning rate of 1.0×10^{-3}. The hyperparameters of the cGMM K, α_0, and σ_0 were set to 4, 0.01 and 1.0, respectively. The multichannel spectrograms were calculated by the short-time Fourier transform with a Hann window whose length of 512 samples and time-shift of 160 samples. The number of potential directions was set to $D = 72$, and the steering vector for each direction $\mathbf{b}_{fd}$ was theoretically calculated under the plane-wave assumption. The thresholding parameter for $\overline{\pi}_k$ was set to 0.02. The source classes corresponding to the diffuse noise were omitted by thresholding $\overline{w}_{kd}$. The threshold was determined such that the F-measure [33] of localization results was maximized. The architectures of the audio and visual DNNs are summarized in Fig. 3. These parameters were determined experimentally.

TABLE III: Sound source number estimation performance in correct rate under the second condition.

# of mics.	$M=2$		$M=4$		$M=6$	
# of srcs.	$L=2$	$L=3$	$L=2$	$L=3$	$L=2$	$L=3$
MUSIC	–	–	0.59	0.27	**0.83**	**0.59**
SRP-PHAT	0.48	0.01	0.50	0.17	0.51	0.26
w/o Aud	0.37	0.42	0.54	0.42	0.56	0.51
Proposed	**0.52**	**0.58**	**0.72**	**0.45**	0.71	0.47

The proposed SSL was compared with two existing audio-only methods. One was the subspace-based method called multiple signal classification (MUSIC) [3], and the other was steered-response power phase transform (SRP-PHAT) [4]. Although these methods were proposed more than 10 years ago, they are still being used in the state-of-the-art systems [34]. For MUSIC, the parameter corresponding to the number of sound sources was set to 2 for $M = 4$ and 3 for $M = 6$. To validate the effectiveness of the audio DNN, we also evaluated a simplified version of our system in which the mixing level π_k was fixed to $1/K$, which is abbreviated here as w/o Aud.

B. Experimental Results

Table I shows the SSL performance in F-measure at the first condition in which all persons speak. Here, L is the number of actual sound sources. Our multichannel neural AV-SSL system was comparable to the conventional methods when $M = 6$. When the number of microphones decreases, the conventional methods were significantly degraded. On the other hand, our AV-SSL system retained its performance thanks to the visual information. In this condition, the proposed method and that without the audio DNN achieved similar performance except $M = 2$.

The SSL performance under the second condition is shown in Table II. Under this condition, our AV-SSL system should determine whether or not each person actually spoke. The performance of our system without the audio DNN was significantly decreased because it cannot distinguish persons who did not speak. On the other hand, the proposed system with the audio DNN still performed well. We also evaluated the correct rate of the source number estimation under this condition. The results are shown in Table III. The proposed system with the audio DNN outperformed that without the audio DNN except $L = 3$ with $M = 6$. These results show the importance of the audio DNN that determines whether or not each sound source candidate produces sound.

To confirm that the visual DNN detects the sound source objects properly, we visualized the activation map of the visual DNN (Fig. 3-(a)) under the second condition with $M = 6$. The results are shown in Fig. 5. Each colored area corresponds to each estimated sound source candidate, and the estimated mixing levels $\overline{\boldsymbol{\pi}}$ were shown in boxes. We can see that the DNN distinguished multiple persons properly with $L = \{2, 3\}$. In addition, we confirmed that the persons whose estimated mixing levels were 0.00 did not produce sounds in the left and center on the first row of Fig. 5.

Fig. 5: Visualization of activation maps obtained by visual DNN from our simulated indoor dataset. Each color corresponds to each sound source candidate. Estimated mixing levels are shown in the boxes.

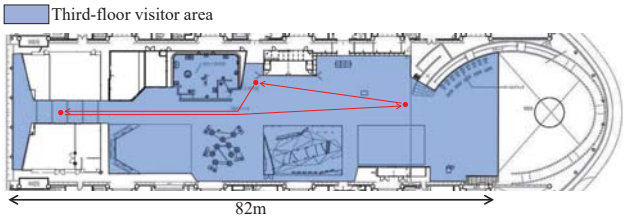


Fig. 6: Third floor of Miraikan. Peacock moved along red arrows while avoiding visitors.

V. Real-data Analysis with Autonomous Robot in Science Museum

We demonstrate our self-supervised method with real data recorded by an autonomous robot called Peacock (Fig. 1).

A. Experimental Setup

This experiment was conducted as a special event in the National Museum of Emerging Science and Innovation (Miraikan)[1] in Japan. Peacock [35], [36] has demonstrated its autonomous navigation on the third floor of the museum (Fig. 6), where more than a thousand people visit every day. We put a 16-ch microphone array and 360° camera on the top of the robot (Fig. 1-(b)) and recorded audio-visual data of the visitors and exhibits around the robot. To encourage various reactions of the visitors, we demonstrated SSL by putting an LED tape that indicates DoAs estimated by MUSIC while the robot moved around the floor. The audio signals were recorded at 16-kHz and 24-bit sampling with an A/D converter called RASP-ZX (Systems In Frontier Corp.), and the videos were recorded at 15 fps and a resolution of 720×1280 with a 360° camera called RICOH THETA S. The recording was conducted six days in the period from December 2019 to February 2020, and the robot has demonstrated seven hours for each day.

Our AV-SSL system was trained from the first four days of the recorded data. The recordings were split into 1.0-second clips, and we used the top 5000 clips having large sound levels as training data because most of the clips had low sound level. The hyperparameters of our method were changed from those in Sec. IV for adapting the real-world environment. The hyperparameters K, α_0, and σ_0 were set to 8, 1.0, and 1.0, respectively. The ϵ was set to 0.1 because the sounds were not stable.

[1]https://www.miraikan.jst.go.jp/en/

B. Experimental Results

Fig. 7 shows inference examples for the recordings that were not used in the training. The examples on the top row show that multiple visitors are localized in individual source classes. These results demonstrate that the proposed method was successfully trained to localize source candidates in a real environment. In the all examples on the middle row, the right regions were localized. There was an exhibit producing large mechanical sounds in this region. Since we took a self-supervised approach, our method was successfully trained to localize such a non-obvious sound source without any supervision. As shown in the bottom row, our method did not always succeed in localization. Although the two visitors were successfully localized in the left example of this row, the right visitor was not localized in the other examples, and the left visitor was localized as two source candidates in the middle example. This problem would be solved by utilizing the continuity of source movements in a video sequence. Please see the video in the supplementary material that includes not only observed images but also corresponding audio recordings.

VI. Conclusion

In this paper, we presented a self-supervised training of audio and visual DNNs for AV-SSL using 360° images and multichannel audio signals. Our method trains both DNNs to maximize the ELBO of the probabilistic spatial audio model, which does not require labeled images or source direction as supervision. Our simulation results confirmed that the trained DNN detected each person separately and determined whether or not each person speaks. We also demonstrated the applicability of the proposed method by using data recorded in a science museum. These results indicate the effectiveness of using multichannel audio signals for audio-visual self-supervised learning. Our future work includes utilizing a video stream to track moving sound source objects.

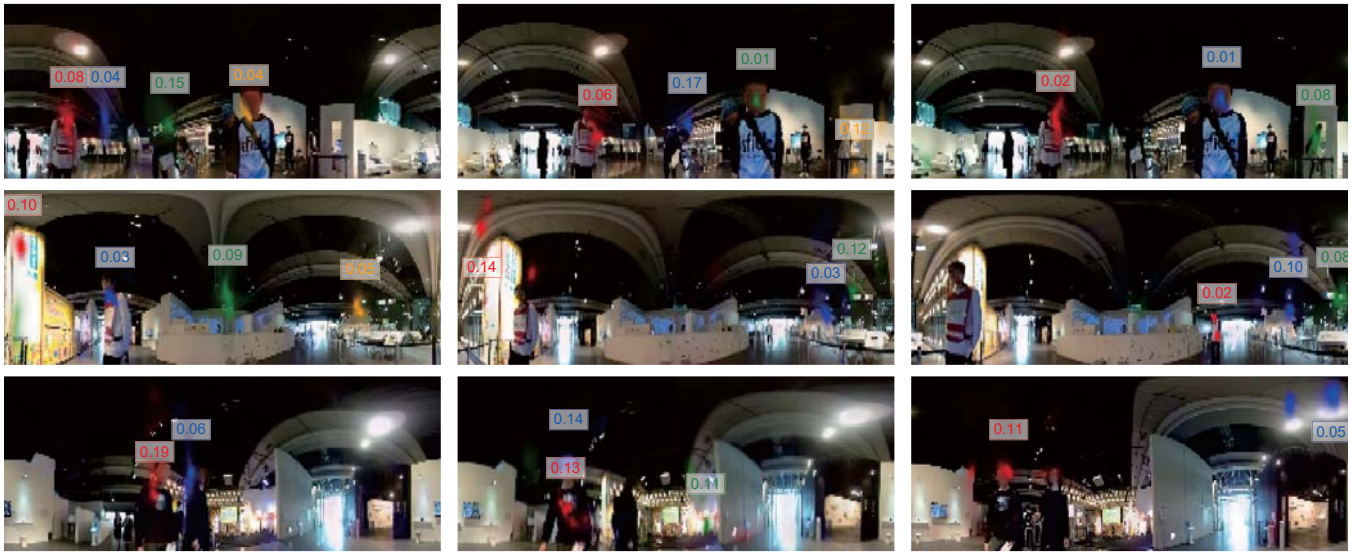


Fig. 7: Visualization of localization results for recordings in Miraikan. Each row displays excerpts from a single video stream at 2-second intervals. Faces of visitors are blurred for their privacy.

Acknowledgements The authors would like to thank Mr. Yusuke Date and Dr. Yu Hoshina for their support in the experiment in Miraikan. This study was partially supported by JSPS KAKENHI No. 18H06490 for funding.

REFERENCES

[1] H. G. Okuno et al., "Human-robot interaction through real-time auditory and visual multiple-talker tracking," in *Proc. of IEEE/RSJ IROS*, 2001, vol. 3, pp. 1402–1409.
[2] P. Jung-Hyun et al., "A design of mobile robot based on network camera and sound source localization for intelligent surveillance system," in *Proc. of ICCAS*, 2008, pp. 674–678.
[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on AP*, vol. 34, no. 3, pp. 276–280, 1986.
[4] M. S. Brandstein et al., "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. of IEEE ICASSP*, 1997, vol. 1, pp. 1520–6149.
[5] F. Grondin and J. Glass, "Fast and robust 3-D sound source localization with DSVD-PHAT," in *Proc. of IEEE/RSJ IROS*, 2019, pp. 5352–5357.
[6] R. Arandjelovic et al., "Look, listen and learn," in *Proc. of IEEE ICCV*, 2017, pp. 609–617.
[7] R. Arandjelovic et al., "Objects that sound," in *Proc. of ECCV*, 2018, pp. 435–451.
[8] A. Owens et al., "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. of ECCV*, 2018, pp. 631–648.
[9] H. Zhao et al., "The sound of motions," in *Proc. of IEEE ICCV*, 2019, pp. 1735–1744.
[10] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. of IEEE ICASSP*, 2016, pp. 405–409.
[11] W. He et al., "Deep neural networks for multiple speaker detection and localization," in *Proc. of IEEE ICRA*, 2018, pp. 74–79.
[12] D. Gatica-Perez et al., "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Trans. on ASLP*, vol. 15, no. 2, pp. 601–616, 2007.
[13] Y. Ban et al., "Exploiting the complementarity of audio and visual data in multi-speaker tracking," in *Proc. of IEEE ICCV Workshop*, 2017, pp. 446–454.
[14] T. Wang et al., "Multimodal and multi-task audio-visual vehicle detection and classification," in *Proc. of IEEE AVSS*, 2012, pp. 440–446.
[15] L. Wang and othres, "Audio-visual sensing from a quadcopter: dataset and baselines for source localization and sound enhancement," in *Proc. of IEEE/RSJ IROS*, 2019, pp. 5320–5325.
[16] Y. Ban et al., "Variational Bayesian inference for audio-visual tracking of multiple speakers," *arXiv:1809.10961*, 2019.
[17] C. Gan et al., "Self-supervised moving vehicle tracking with stereo sound," in *Proc. of IEEE ICCV*, 2019, pp. 7053–7062.
[18] G. Irie et al., "Seeing through sounds: Predicting visual semantic segmentation results from multichannel audio signals," in *Proc. of IEEE ICASSP*, 2019, pp. 3961–3964.
[19] T. Zhou et al., "Unsupervised learning of depth and ego-motion from video," in *Proc. of IEEE CVPR*, 2017, pp. 1851–1860.
[20] L. Drude et al., "Unsupervised training of neural mask-based beamforming," in *Proc. of INTERSPEECH*, 2019, pp. 1253–1257.
[21] Y. Bando et al., "Deep Bayesian unsupervised source separation based on a complex Gaussian mixture model," in *Proc. of IEEE MLSP*, 2019, pp. 1–6.
[22] T. Higuchi et al., "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. of IEEE ICASSP*, 2016, pp. 5210–5214.
[23] T. Otsuka et al., "Unified auditory functions based on Bayesian topic model," in *Proc. of IEEE/RSJ IROS*, 2012, pp. 2370–2376.
[24] T. Otsuka et al., "Bayesian nonparametrics for microphone array processing," *IEEE/ACM Trans. on ASLP*, vol. 22, no. 2, pp. 493–504, 2014.
[25] O. Ronneberger et al., "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. of MICCAI*, 2015, pp. 234–241.
[26] J Barker et al., "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. of IEEE ASRU*, 2015, pp. 504–511.
[27] M. Figurnov et al., "Implicit reparameterization gradients," in *Proc. of NeurIPS*, 2018, pp. 441–452.
[28] D. Dwibedi et al., "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in *Proc. of IEEE ICCV*, 2017, pp. 1301–1310.
[29] I. Armeni et al., "Joint 2D-3D-Semantic data for indoor scene understanding," *arXiv: 1702.01105*, 2017.
[30] W. Yang et al., "Clothing Co-Parsing by joint image segmentation and labeling," in *Proc. of IEEE CVPR*, 2014, pp. 3182–3189.
[31] D. B. Paul et al., "The design for the Wall Street Journal-based CSR corpus," in *Proc. of ACL Workshop on SNL*, 1992, pp. 257–362.
[32] J. Allen et al., "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
[33] K. Furukawa et al., "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *Proc. of IEEE/RSJ IROS*, 2013, pp. 3943–3948.
[34] C. Evers et al., "The LOCATA challenge: Acoustic source localization and tracking," *arXiv:1909.01008*, 2019.
[35] Y. Sasaki and J. Nitta, "Long-term demonstration experiment of autonomous mobile robot in a science museum," in *Proc. of IEEE IRIS*, 2017, pp. 304–310.
[36] A. Kanezaki et al., "GOSELO: Goal-directed obstacle and self-location map for robot navigation using reactive neural networks," *IEEE RAL*, vol. 3, no. 2, pp. 696–703, 2018.