Self-Supervised Learning for Audio-Visual Relationships of Videos With Stereo Sounds

TOMOYA SATO, YUSUKE SUGANO, AND YOICHI SATO, (Senior Member, IEEE)
Institute of Industrial Science, The University of Tokyo, Meguro 153-8505, Japan
Corresponding author: Tomoya Sato (tomasato@iis.u-tokyo.ac.jp)
This work was supported by JST SPRING under Grant JPMJSP2108.

ABSTRACT Learning cross-modal features is an essential task for many multimedia applications such as sound localization, audio-visual alignment, and image/audio retrieval. Most existing methods mainly focus on the semantic correspondence between videos and monaural sounds, and spatial information of sound sources has not been considered. However, sound locations are critical for understanding the sound environment. To this end, it is necessary to acquire cross-modal features that reflect the semantic and spatial relationship between videos and sounds. A video with stereo sound, which has become commonly used, provides the direction of arrival of each sound source in addition to the category information. This indicates its potential to acquire a desired cross-modal feature space. In this paper, we propose a novel self-supervised approach to learn a cross-modal feature representation that captures both the category and location of each sound source using stereo sound as input. For a set of unlabeled videos, the proposed method generates three kinds of audio-visual pairs: 1) perfectly matched pairs from the same video, 2) pairs from the same video but with the flipped stereo sound, and 3) pairs from a different video. The cross-modal feature encoder of the proposed method is trained on triplet loss to reflect the relationship between these three pairs (1 > 2 > 3). We apply this method to cross-modal image/audio retrieval. Compared with previous audio-visual pretext tasks, the proposed method shows significant improvement in both real and synthetic datasets.

INDEX TERMS Computer vision, feature extraction, machine learning, self-supervised learning, audio-visual learning, cross-modal retrieval.

I. INTRODUCTION

With the surge in the availability of videos that contain both images and synchronized sounds, the acquisition of a useful representation of audiovisual characteristics has attracted strong attention. Audio-visual cross-modal features are necessary for practical applications such as cross-modal retrieval, alignment, and localization [1], [2], [3].

Previous work has successfully obtained cross-modal features by self-supervised learning based on the synchronization of images and sounds provided by videos. Several methods have been proposed that are based on the semantic relationship between images and sounds [4], as well as those based on the temporal alignment of videos and sounds [5]. However, these methods mainly focus on the semantic relationship between images and monaural sounds. We cannot utilize the spatial relationship typically represented by sound locations in these methods. Namely, in the previous methods, it is implicitly assumed that the main sound source appears in a large portion of the image. Therefore, these methods have difficulties in applying to videos where the sound source exists in a specific location or where multiple sound sources exist. For example, the video on the left in Figure 1 depicts that the main sound source is in a specific location but does not occupy a large area of the image. To perform appropriate feature extraction, we need to identify the location of the trumpet as being on the left side. Also, the right video in Figure 1 cannot be properly interpreted if the positions of the drum and piano are unknown. To acquire a feature extractor that can be applied to these videos, the arrival directions of sound sources need to be incorporated into the training.

Nowadays, video clips with stereo sounds can be easily acquired. They contain both semantic information and
horizontal spatial information of sound sources. Using stereo sound in self-supervised audio-visual learning leads to enriched feature representations that contain both the semantics and locations of sound sources. Indeed, several studies have utilized stereo sound in self-supervised audio-visual learning [6], [7]. However, these studies focus only on the spatial correspondence between images and stereo sounds but do not adequately capture the semantic information of sound sources. A self-supervised method for learning cross-modal features that capture both the semantics and location of sound sources for unlabeled videos has not yet been fully investigated.

To obtain a desired feature space, our main observation is that it is effective to apply the order constraint reflecting both audio-visual semantic and spatial information. Specifically, audio-visual pairs matching in both semantics and spatial layout must be closer than those matching only in semantics, while the semantically matched pairs must still be closer than those that do not even show a semantical match. Consider the example illustrated in Figure 2. For the target image, we adopt three types of stereo sound that are easily obtained from unlabeled audio-visual datasets: (a) sound taken from the same video as the image, (b) left-right flipped sound from the same video, and (c) sound from a different video. (a)–(c) represent the degree of semantic and spatial match. Similarly to previous work [4], it is reasonable to learn a feature space where the matched sound (a) becomes closer to the image feature than the unmatched sound (c). In addition, since sound (b) has at least the same semantics, the distance to the image should be longer than (a) but shorter than (c). By introducing this ordering constraint on the feature distances, the feature encoder can be trained to preserve both semantic and spatial information about the sound sources.

In this paper, we propose a self-supervised method to learn cross-modal audio-visual features on the basis of a novel loss function, Stereo Sound Ranking (SSR) loss. To force the ordered relationship between features, we use a triplet loss function [9] on different audio-visual pairs. For each image, the proposed method selects three stereo sounds: 1) corresponding, 2) flipped, and 3) mismatched. The SSR loss is designed so that the feature similarity to the image is in the same order, i.e., \(1 > 2 > 3\). When learning the image and sound encoder using the SSR loss, the network can extract cross-modal representations that contain the semantic and horizontal spatial information of sound sources. By these cross-modal feature representations, the proposed method can process the difficult cases mentioned above and illustrated in Figure 1. We evaluate the proposed method on a cross-modal image/sound retrieval scenario using both real and synthetic datasets. Experimental results demonstrate that the proposed method enables cross-modal retrieval with the semantic and spatial correspondences of the sound source information.

Our contributions are twofold.

1) We propose a novel triplet loss function for self-supervised cross-modal feature learning of videos with stereo sounds. To our knowledge, this is the first work to enable learning of cross-modal representation containing both semantic and spatial information for video with stereo sound.

2) We demonstrate an application of the proposed method to cross-modal retrieval and show that we can successfully retrieve images and sounds with the same sound category and location.

The rest of this paper is organized as follows. In Section 2, we review related work and show the position of this study in audio-visual and self-supervised learning. In Section 3, we propose a method to acquire the feature space that captures both semantic and spatial relationships between images and stereo sounds. In Section 4, we evaluate the proposed method by audio-visual cross-modal retrieval. Finally, in Section 5, we conclude this paper.

II. RELATED WORK
A. SELF-SUPERVISED AUDIO-VISUAL LEARNING
Due to the difficulties of providing annotations for both visual and synchronized auditory information, self-supervised learning is often adopted for audio-visual representation learning. A wide variety of pretext tasks have been proposed to capture the audio-visual relationship. The most representative approach is the one classifying matched and unmatched audio-visual pairs utilizing the fact that visual and audio
inputs from the same video have the same semantic information [3], [4], [10], [11], [12], [13]. Other approaches include those that determine the temporal alignment of videos and sounds [5], [14], [15], and hybrid approaches that combine both tasks [16]. In addition to simple single-domain applications such as sound/image classification [10], [11], [16] and action recognition [5], [15], these works demonstrate the benefits of learned features in complex cross-domain applications such as sound localization [3], [12], [13], [14], [15], cross-modal retrieval [4], and sound separation [5]. However, the target of these prior works is limited to learning semantic cross-modal relationships. The learned features are hard to apply to difficult videos represented in Figure 1.

B. AUDIO-VISUAL SPATIAL FEATURE LEARNING
Several methods have been proposed to learn the spatial relationship between visual and auditory information. One approach is learning cross-modal features by reconstructing multi-channel sounds from pairs of video and monaural sound [8], [17], [18]. In addition to the usefulness of the reconstruction task itself, the learned cross-modal features can be applied to other applications including sound separation and localization.

Multi-channel sound has also been utilized in several methods. Gan et al. proposed to learn a vehicle localization network with stereo sound input using a pre-trained vision network as a teacher [6]. Yang et al. proposed a pretext task that identifies whether the left and right audio channels are flipped for video [7]. Through this pretext task, the spatial relationship between videos and stereo sounds is learned. Thus, the learned features are adopted for downstream tasks such as sound localization and separation. Also, Morgado et al. proposed a method for learning cross-modal features by aligning Anbisonics and 360° video [19]. However, these methods mainly consider the spatial audio-visual relationship and do not incorporate the semantic relationship. Our goal is to acquire both semantic and spatial audio-visual relationships from stereo-recorded videos in a unified manner.

C. TRIPLET LOSS FOR CROSS-MODAL FEATURE LEARNING
Triplet loss is a function defined in triplets of samples (anchor, positive, and negative) and guides the network to reduce the distance between the anchor and the positive samples while increasing the distance between the anchor and the negative samples. It has been widely used as a loss function for many tasks such as image retrieval [9], face recognition [20], and person re-identification [21]. It has been also adopted in various cross-modal learning methods, such as sound localization [3], [22], image-text retrieval [23], [24], [25], and image-sound retrieval [1], [26]. In contrast to these prior works to obtain audio-visual representation for videos with monaural sounds, our method adopts the triplet loss approach to videos with stereo sounds.

### Table 1. Audio-visual datasets. We employ FAIR-Play which is constructed with stereo-recorded videos of various sound categories.

| Datasets                  | Sound categories                           | Sound categories                           |
|--------------------------|--------------------------------------------|--------------------------------------------|
| videos with monaural sound | AudioSet [27], MUSIC [28], HIMV-200K [1]   | –                                          |
| videos with stereo sound  | FAIR-Play [8]                              | YouTube-ASMR [7], Auditory Vehicle Tracking dataset [6] |

D. AUDIO-VISUAL DATASET
There are a variety of datasets for audio-visual learning. Typical datasets are listed in Table 1. AudioSet [27] is a particularly popular dataset of images and monaural sounds. This is a large-scale dataset that contains various sound categories collected from YouTube and used to train audio-visual models [4], [5], [16]. Other datasets of images and monaural sounds include MUSIC [28] and HIMV-200K [1]. They focus on playing musical instruments and music videos, respectively, and include various categories of sound sources to train their networks. However, these datasets use monaural sounds that inherently lack spatial information of sounds. In this paper, because the target of analysis is audio-visual semantic and spatial relationships, these datasets are out of scope.

Datasets of images and stereo sounds have also been constructed. FAIR-Play is a dataset of playing various musical instruments recorded with a camera and a binaural microphone [8]. It is used for several audio-visual spatial feature learning [8], [18], [29]. YouTube-ASMR is a dataset that efficiently collects stereo-recorded videos from YouTube by targeting only ASMR videos for training the network in [7]. Auditory Vehicle Tracking dataset contains street videos with stereo sound for vehicle localization [6]. Among these datasets, we employ FAIR-Play, which contains various sound categories, as our training dataset. This is because variations in sound categories are necessary to evaluate the proposed method to acquire general audio-visual semantic and spatial correspondences. However, FAIR-Play has a problem of low visual diversity because it is recorded by limited performers in one music room. To solve this problem, we construct PseudoMUSIC, an additional synthetic dataset of images and stereo sounds described in Section IV-A1.

III. PROPOSED METHOD
A. OVERVIEW
Our goal is to learn a representation of the cross-modal features between images and stereo sounds to acquire semantic and spatial information about sound sources. To this end, we introduce a novel loss function, called Stereo Sound Ranking (SSR) loss, to learn a feature space where a spatially and semantically matched image/stereo sound pair becomes closer than a spatially incongruent pair (flipped stereo sound) using one triplet loss. Also, a spatially incongruent pair becomes closer than a semantically incongruent pair.
(different stereo sound) by the other triplet loss. By jointly applying these triplet losses, the SSR loss enables to represent a feature space for stereo-recorded videos.

Our method consists of an image and sound CNN. They extract the image and stereo sound features separately from the inputs (Figure 3). To learn the semantic and spatial audio-visual relationships, the two CNNs are trained to preserve the distance order of the stereo sound features for the image feature. We consider three types of stereo sound for each image: 1) stereo sound taken from the same video as the image, 2) the same stereo sound with the left and right audio channels flipped (spatially incongruent), and 3) stereo sound from a different video (semantically incongruent). We obtain the image feature $f^I$ from the input image using the image CNN, and the sound features $f^S_m$ (matched), $f^S_f$ (flipped), and $f^S_m$ (mismatched) from these three types of stereo sound using the sound CNN. The SSR loss function takes all four (one image and three sounds) features and aligns the order of the feature distances between the three audio-visual pairs. As a result, the set of stereo-recorded videos is projected onto a cross-modal feature space in which both image and stereo sound features are interrelated. The proposed network is an extension of triplet networks for audio-visual learning to represent audio-visual semantic and spatial relationships that can solve difficult cases illustrated in Figure 1. In the following sections, we describe the details of the SSR loss function and the network architecture.

**B. STEREO SOUND RANKING LOSS**

Given the image feature $f^I$ and the stereo sound features $f^S_m$, $f^S_f$, and $f^S_m$, the SSR loss function first calculates the Euclidean distances $d_m$, $d_f$, and $d_mm$ between $f^I$ and each of $f^S_m$, $f^S_f$, and $f^S_m$ (Figure 4). Distances $d_m$, $d_f$, and $d_mm$ denote the similarities of stereo sound features to the image feature. To represent the relationship among $f^I$, $f^S_m$, $f^S_f$, and $f^S_m$, the SSR loss $L_{SSR}$ is then given as

$$L_{SSR} = L_{m&f} + L_{f&mm},$$

where

$$L_{m&f} = \max(0, d_m + \alpha_{m&f} - d_f),$$

$$L_{f&mm} = \max(0, d_f + \alpha_{f&mm} - d_mm).$$

The parameter $\alpha_{m&f}$ is the separation margin of $f^S_m$ and $f^S_f$. Similarly, triplet loss $L_{f&mm}$ requires that $d_f$ be smaller than $d_mm$, and the parameter $\alpha_{f&mm}$ is the separation margin between $f^S_f$ and $f^S_m$. With $L_{m&f}$ and $L_{f&mm}$, we have the three distances in the desired order ($d_m < d_f < d_mm$).

**C. NETWORK ARCHITECTURE**

The image and sound CNNs are based on the ResNet-18 architecture [30] commonly used to extract features from images and spectrograms in audio-visual learning [7], [8], [13]. The image CNN takes an RGB image of $448 \times 224$ resolution as input. The feature after the fourth residual block is flattened and passed through a fully-connected (FC) layer. Finally, the 128-dimensional image feature vector is obtained after L2 normalization.
The input to the sound CNN is a two-channel 257 × 100 log-spectrogram calculated from a one-second stereo sound. The dimension of the first convolutional layer is reduced from three to two. The same L2 normalization is applied to the sound feature, which results in a 128-dimensional vector.

D. IMPLEMENTATION DETAILS

To construct a training input, we randomly select two videos from the training dataset. We extract an image and a one-second stereo sound from one video. We make the flipped sound by flipping the left and right channels of the given sound. In addition, we extract a one-second stereo sound from the other video. These three stereo sounds are combined with the image as the training input.

The input RGB image is first resized to 480 × 240. Then, as data augmentation, we randomly crop 448 × 224 images. The color and intensity of the image are randomly changed in the range of 0.7 to 1.3.

As sound input, we calculate the log-spectrogram of a one-second sound resampled at 16kHz. Short-time Fourier transform (STFT) is applied by a 25-ms Hann window with a 10-ms hop and an FFT size of 512. We stack the log-spectrograms of the left and right audio channels. The input size becomes 2 × 100 × 257.

The network consists of around 24.0M parameters and is trained using the Adam optimizer with a batch size of 32 on one NVIDIA Quadro GV100 GPU. The computation time for the two datasets mentioned in Section IV-A1 is as follows. The network is trained for 1000 epochs in FAIR-Play which takes around 45 hours, and for 200 epochs in PseudoMUSIC which takes around 19 hours. The learning rate is set to 0.0001. For the loss margins, we set $\alpha_{m&f} = \alpha_{f&mm} = 0.2$.

IV. EXPERIMENTS

In this section, we report the experimental results on cross-modal retrieval to evaluate whether the proposed approach can acquire better feature representation containing both semantic and spatial audio-visual relationships. We present both quantitative and qualitative results for image-to-sound/sound-to-image retrieval. We also visualize the network attention and learned feature spaces for further analysis.

A. EVALUATION DETAILS

1) DATASETS

We train the proposed network on two datasets consisting of images and stereo sounds. The first one is FAIR-Play, a real-world dataset. Using this data set, we demonstrate the effectiveness of the proposed network for real videos. However, the evaluation in FAIR-Play alone does not guarantee the generality of the acquired features by the proposed method because FAIR-Play is recorded with limited performers in a single music room and has low visual diversity. To alleviate this problem, we construct a new dataset called PseudoMUSIC as the second dataset. This dataset is synthetic and designed to have a high diversity that includes various performers and backgrounds. The details of the two datasets are given as follows.

a: FAIR-PLAY [8]

FAIR-Play consists of 1,871 ten-second videos of one or more persons playing musical instruments. These are videos where the sound source comes from specific directions or videos where multiple sound sources exist, to which existing methods are difficult to apply. There are nine main musical instruments: banjo, cello, drum, guitar, harp, piano, trumpet, ukulele, and upright bass. In several cases, there are additional percussion instruments, such as tambourine, and several videos also contain singing voices. The dataset is randomly divided into training, validation, and test subsets with a ratio of 80%/10%/10%. Because there is no ground-truth annotation of sound category and location on FAIR-Play, we manually annotate the test subsets with category labels and their bounding boxes. The instrument types previously mentioned are used as category labels. For this purpose, from the test split, we further select items where all sound sources emit sound. Finally, we obtain 352 annotated image/audio pairs, containing those that are horizontally flipped in both images and sounds for additional data.

b: Pseudomusic

We built PseudoMUSIC following previous work on generating images and their corresponding spatial sounds [29]. In this method, image patches representing sound sources are placed in specific directions of a background image, and synchronized monaural sounds are converted to spatial sounds from those directions using head-related impulse response (HRIR) [31] to generate audio-visual pairs.

The image patches and synchronized monaural sounds are collected from the MUSIC dataset [28]. First, we extract the image and its corresponding one-second monaural sound for 14 instruments in the MUSIC dataset: accordion, acoustic guitar, bassoon, congas, drum, electric bass, flute, guzheng, piano, pipa, trumpet, ukulele, violin, xylophone. Then, following the previous work [29], we generate 500 audio-visual pairs for each instrument. They are divided into 400/50/50 for the train/validation/test subsets, respectively. In addition, we prepare a dataset for multiple sound sources. We first enumerate a set of two instruments: accordion-electric bass, accordion-piano, acoustic guitar-electric bass, acoustic guitar-xylophone, bassoon-flute, bassoon-pipa, congas-guzheng, congas-violin, drum-piano, drum-trumpet, flute-guzheng, pipa-trumpet, ukulele-violin, and ukulele-xylophone. We generate 500 audio-visual pairs for each element of this set. Finally, we obtain 11,200/1,400/1,400 data for the train/validation/test subsets, respectively.

Cropping patches from the images is performed by the masks both containing the performers and instruments taken as the union of the human masks and the bounding boxes of the instruments. For the human masks, we leverage the semantic segmentation model of a fully convolutional
network [32] with the ResNet-50 backbone. For instrument detection, we use the object detector of the Faster RCNN trained as in [33]. The area of each patch is resized to 160 × 160 and then randomly resized in the range of 0.95 to 1.05. The background images for the patches are randomly selected from the ADE20K dataset [34] with each image resized to 240 × 480. Furthermore, following previous work [29], the azimuth angles of sound sources are selected from −60 to 60 degrees. We show several examples of PseudoMusic in Figure 5.

2) BASELINE METHODS
Throughout the experiments, Proposed indicates the proposed method and is compared with the following baseline methods.

a: MISMATCH (MONO) [4]
Mismatch (Mono) learns the features by performing binary classification to identify whether the image and the monaural sound inputs are from the same video using the Euclidean distance between the image and the sound features. This is a representative method for learning the relationship between images and monaural sounds [4]. We verify the necessity of adopting stereo sound to acquire the features representing semantic and spatial relationships by comparing this method with Proposed.

b: MISMATCH (STEREO)
Mismatch (Stereo) replaces the sound input in Mismatch (Mono) with stereo sound. The comparison between this method and Proposed shows the necessity of spatially incongruent audio-visual pairs to obtain the desired feature space.

c: FLIP
As a replacement for the pretext classification task in Mismatch (Stereo), Flip uses binary classification to determine whether the left and right channels of the stereo sound input are flipped, which is inspired by previous audio-visual spatial learning [7]. Contrary to Mismatch (Stereo), this method presents the necessity of semantically incongruent audio-visual pairs.

d: FLIP/MISMATCH
In Flip/Mismatch, we calculate the loss function using \( L_{\text{m&mm}} = \max(0, d_m + a_{\text{m&mm}} - d_{\text{mm}}) \) instead of \( L_{\text{m&mm}} (a_{\text{m&mm}} = 0.2) \). With this loss function, the network is trained only with the policy of bringing semantically and spatially matched pairs closer than spatially mismatched pairs and semantically mismatched pairs. Namely, the comparison between Proposed and this method shows whether the proposed order constraint (Match>Flip>Mismatch) plays an important role in learning the feature space.

Computational Cost on Baselines: Each baseline method consists of around 24.0M parameters because the structures of the image CNN and sound CNN occupying a major part of the network are shared with Proposed. The settings of the optimizer, batch sizes, and GPUs are the same for all the baseline methods as in Proposed.

As for the computation time, Flip/Mismatch adopts the same parameters as Proposed except for its loss function. Namely, the network of Flip/Mismatch is trained for 1000 epochs in FAIR-Play and 200 epochs in PseudoMUSIC, which indicates that computation time for each dataset is equivalent to Proposed. On the contrary, Mismatch (mono), Mismatch (stereo), and Flip are binary classification tasks to identify whether the inputs of the image and the sound correspond. These baselines require twice the number of epochs to provide positive and negative sounds for each image input as Proposed which provides both positive and negative sounds for the image inputs. Therefore, Mismatch (mono), Mismatch (stereo), and Flip are configured to be trained for 2000 epochs in FAIR-Play which takes around 42 hours and 400 epochs in PseudoMUSIC for around 15 hours.

3) EVALUATION METRICS
In the following evaluations, nDCG@K [35] is used as the main evaluation metric. Given a retrieval result, DCG@K is defined as \( \text{DCG@K} = \sum_{i=1}^{K} s(Q, R_i) / \log_2(i + 1) \), where \( s(Q, R_i) \) represents the score of the \( i \)-th ranked retrieval item \( R_i \) for query \( Q \). nDCG@K is calculated by normalizing DCG@K with the ideal DCG@K; the maximum value when the retrieval results are ordered by similarity scores. As a result, nDCG@K has a range of values \([0, 1]\), and a higher value indicates a better retrieval result. Here, \( K = 5 \), and we use following three definitions (a)–(c) for the score \( s(Q, R) \).

a: CatLoc SCORE
We introduce the CatLoc score to take into account both the categories and locations of sound sources. An example of the calculation is shown in Figure 6. We assume an image–sound retrieval task, where we obtain a query image \( Q \) and an image corresponding to a retrieved sound \( R \). As shown
in Figure 6, we divide each image into three regions: left, center, and right. We make a mapping of each region to an instrument contained in the region. If there is no sound source, No source is assigned as shown in Figure 6. An instrument is judged to be in a region if the center of the x-coordinate of its corresponding bounding box is within the region. Then, as an extension to the cosine similarity between $Q$ and $R$, $s(Q, R)$ for CatLoc is calculated as follows.

$$s(Q, R) = \frac{\sum_{p \in \{l,c,r\}} \text{sim}(Q_p, R_p)}{\sqrt{N_Q \cdot N_R}},$$

where $l/c/r$ represents left/center/right, $Q_p$ and $R_p$ represent the sound category of $Q$ and $R$ in each region, respectively, and $N_Q$ and $N_R$ are the numbers of sound sources in $Q$ and $R$, respectively. Note that the same process is performed for sound-to-image retrieval.

**b: CAT SCORE**

We also introduce the Cat score, which is based only on the category information of sound sources similar to an existing cross-modal retrieval metric. To reflect only the information of the category, the query image ($Q$) and the image corresponding to the retrieved sound ($R$) are converted to a set of sound sources contained in each image. Then, for each sound source in $Q$, the similarity to the most similar sound source in $R$ is calculated. We calculate $s(Q, R)$ for Cat as

$$s(Q, R) = \frac{\sum_{q \in S_Q} \max_{r \in S_R} \text{sim}(q, r)}{\sqrt{N_Q \cdot N_R}},$$

where $S_Q$ and $S_R$ represent the sound categories in $Q$ and $R$, respectively, and $N_Q$ and $N_R$ are the numbers of sound sources in $Q$ and $R$, respectively.

**c: LOC SCORE**

For further analysis, we also examine how the proposed method and the baselines perform when we only consider the location of the sound sources. For this reason, we introduce the Loc score that only considers the locations of sounds. $Q$ and $R$ are divided into three regions (left, center, right) as in CatLoc, and converted into a vector representing the number of sources in each region. Loc is calculated as the cosine similarity between these vectors.

**Similarity Calculation Between Sound Categories:** In each score, we define the similarity of sound categories following the previous audio-visual cross-modal retrieval [4]. A tree representing the similarities among the instruments is constructed using the ontology presented in the AudioSet dataset [27]. The trees in each dataset are shown in Figures 7 and 8. The similarity score of sound categories $\text{sim}(i_1, i_2)$ is calculated as $\text{sim}(i_1, i_2) = C - d(i_1, i_2)$, where $C$ represents the maximum distance in the tree and $d(i_1, i_2)$ represents the distance between categories of $i_1$ and $i_2$ in the tree.

**B. PERFORMANCE OF RETRIEVAL IN VIEW OF SOUND CATEGORIES AND LOCATIONS**

We perform image-to-sound and sound-to-image retrieval tasks and evaluate the performance on the basis of both categories and locations of sound sources using CatLoc. Table 2 shows the performance of the proposed method and all baseline methods. We show the scores for both the cases where the query has a single sound source (single) and where it has multiple sound sources (multiple).

In FAIR-Play, we see that Proposed outperforms the baselines. Specifically, we claim the following superiorities of Proposed over baselines.

**Over Mismatched (Mono) and Mismatched (Stereo)**

Proposed outperforms Mismatched (Mono) and Mismatched (Stereo) in both cases of single and multiple sound sources in the query. This indicates that Proposed captures the locations of sounds in addition to the sound categories.

**Over Flip**

The score of Proposed is higher than that of Flip. This also indicates that Proposed captures both the categories and locations of the sound sources.

**Over Flip/Mismatch**

Proposed is superior to Flip/Mismatch. This indicates that we cannot obtain the desired feature space simply by increasing the similarity between...
TABLE 2. Results of cross-modal retrieval in FAIR-Play and PseudoMUSIC using CatLoc. We adopt nDCG@5 as the metric for retrieval evaluation. “image-sound” represents the result of image-to-sound retrieval and “sound-image” represents the result of sound-to-image retrieval. “single” and “multiple” indicate the result when the query contains only a single sound source and multiple sound sources, respectively.

|                  | FAIR-Play | PseudoMUSIC |
|------------------|-----------|-------------|
|                  | image-sound | sound-image | image-sound | sound-image |
|                  | single | multiple | single | multiple | single | multiple | single | multiple |
| Mismatch (Mono)   | 0.401 | 0.481 | 0.421 | 0.498 | 0.108 | 0.147 | 0.084 | 0.138 |
| Mismatch (Stereo) | 0.439 | 0.514 | 0.433 | 0.516 | 0.241 | 0.234 | 0.213 | 0.215 |
| Flip              | 0.197 | 0.319 | 0.209 | 0.321 | 0.163 | 0.165 | 0.176 | 0.155 |
| Flip/Mismatch     | 0.668 | 0.696 | 0.679 | 0.717 | 0.371 | 0.374 | 0.357 | 0.364 |
| Proposed          | 0.727 | 0.768 | 0.707 | 0.782 | 0.511 | 0.525 | 0.446 | 0.463 |

![FIGURE 9. Examples of cross-modal retrieval based on the categories and locations of sound sources. The first row is queries and the other rows are the top5 retrieved results. (a) and (b) represent the results of FAIR-Play and (c) and (d) represent those in PseudoMUSIC. (a) and (c) are examples of a single sound source, and (b) and (d) are those of multiple sound sources.](image)

matched audio-visual pairs compared with mismatched pairs as in the Flip/Mismatch pretext task. The order constraint of Proposed is beneficial in capturing the categories and locations of sound sources.

We also provide the results of PseudoMUSIC. In PseudoMUSIC, Proposed has a higher CatLoc than other methods, although the score values are lower than those of FAIR-Play because of the difficulty of the retrieval itself due to the diversity of the dataset. This indicates that the proposed method generally enables audio-visual feature extraction on the basis of the categories and locations of sound sources.

Figure 9 shows several retrieval results using Proposed. Columns (a) and (b) show the retrieval results in FAIR-Play and columns (c) and (d) are those in PseudoMUSIC. We can see that Proposed extracts the image and sound features that represent the cello on the left in (a). Furthermore, from (b), Proposed can extract features representing the categories and locations of multiple sound sources (the drum on the left and the piano on the right). The cases shown in (c) and (d) show that Proposed is effective for extracting features from videos with high visual diversity, including various performers and backgrounds.

C. PERFORMANCE OF RETRIEVAL IN VIEW OF SOUND CATEGORIES

The baseline methods Mismatch (Mono) and Mismatch (Stereo) are expected to perform best in the evaluation using Cat. Therefore, to fairly evaluate whether Proposed captures sound categories, we compare the Cat scores of Proposed and the baselines, including Mismatch (Mono) and Mismatch (Stereo). Table 3 shows the performance of Proposed and all baseline methods using Cat.

In FAIR-Play, Proposed scores highest in most cases, suggesting that Proposed is successful in extracting features on the basis of the semantic information of sound sources. However, Mismatch (Mono) and Mismatch (Stereo) obtain scores comparable to Proposed. This is because the diversity of FAIR-Play is low, and even learning the pattern of the recording room results in a certain performance in FAIR-Play.

In addition, significant improvement can be observed with Proposed compared to Mismatch (Mono) and Mismatch (Stereo) in PseudoMUSIC. Because PseudoMUSIC contains...
TABLE 3. Results of cross-modal retrieval in FAIR-Play and PseudoMUSIC using Cat. We adopt the same setting as Table 2.

|                  | FAIR-Play | PseudoMUSIC |
|------------------|-----------|-------------|
|                  | image-sound | sound-image | image-sound | sound-image |
|                  | single    | multiple   | single    | multiple   |
| Mismatch (Mono)  | 0.923     | 0.838      | 0.938     | 0.844      |
| Mismatch (Stereo)| 0.958     | 0.852      | 0.964     | 0.848      |
| Flip             | 0.415     | 0.430      | 0.413     | 0.444      |
| Flip/Mismatch    | 0.887     | 0.770      | 0.866     | 0.773      |
| Proposed         | **0.968** | **0.869**  | **0.956** | **0.860**  |

**TABLE 4.** Results of cross-modal retrieval in FAIR-Play and PseudoMUSIC using Loc. We adopt the same setting as Table 2.

|                  | FAIR-Play | PseudoMUSIC |
|------------------|-----------|-------------|
|                  | image-sound | sound-image | image-sound | sound-image |
|                  | single    | multiple   | single    | multiple   |
| Mismatch (Mono)  | 0.425     | 0.805      | 0.443     | 0.827      |
| Mismatch (Stereo)| 0.454     | 0.837      | 0.445     | 0.834      |
| Flip             | 0.444     | 0.644      | 0.440     | 0.645      |
| Flip/Mismatch    | 0.739     | 0.865      | 0.760     | 0.850      |
| Proposed         | **0.742** | **0.893**  | **0.729** | **0.894**  |

Diverse images with stereo sounds, Proposed effectively focuses on sound sources by learning the semantic and spatial relationships.

D. PERFORMANCE OF RETRIEVAL IN VIEW OF SOUND LOCATIONS

The baseline method Flip is designed to perform best in the evaluation using Loc. For a fair evaluation of whether Proposed captures sound locations, we compare Proposed and the baselines using Loc. The performance of Proposed and all baseline methods using Loc are shown in Table 4. In FAIR-Play, Proposed has higher scores than baselines in most cases, suggesting that Proposed extracts features by focusing on the locations of the sound sources. In PseudoMUSIC, Proposed also shows high scores.

It is interesting that Mismatch (Stereo) obtains high Loc in PseudoMUSIC. This indicates that Mismatch (Stereo) learns the locations, not the semantics of the sound sources. This is caused by the pretext task of Mismatch (Stereo). This task identifies whether the input image and sound are from the same video. When the sound input is monaural, the network learns if the image and sound have the same semantics. When the sound input is stereo sound, the network can alternatively use sound locations instead of semantic information to determine the match between the image and sound inputs. This is conspicuous in PseudoMUSIC where the locations of sound sources are easier to generalize than the semantics. In contrast, Proposed properly maps features by its order constraint. Therefore, we can conclude the effectiveness of Proposed in acquiring the relationship between images and stereo sounds.

E. VISUALIZATION OF NETWORK ATTENTION

To qualitatively prove that the proposed method successfully focuses on sound sources for feature extraction, we visualize the region where the network is paying attention using the previous sound localization method [8]. Specifically, for an input image of 224 × 448, we use a mask of size 32 × 32 to nullify the image regions. Then the distance between the feature of the nullified image and the corresponding stereo sound feature is calculated. If this distance increases over the original one, the nullified area should represent the sound source because the sound source is critical in calculating the distance. By shifting the mask in the input image, we obtain a 7 × 14 (= 224 × 448/32 × 32) map. This map indicates whether the network focuses on the sound sources.

Several examples of attention maps are shown in Figure 10. In FAIR-Play and PseudoMUSIC, we see that the generated attention maps successfully capture the sound sources, indicating that the proposed method focuses on sound sources.

F. VISUALIZATION OF FEATURE SPACES

We additionally visualize the feature spaces to qualitatively evaluate whether the learned features capture both semantic and spatial information. In Figures 11 and 12, we show results of applying t-SNE [36] to image and sound features from FAIR-Play and PseudoMUSIC, respectively. We visualize image and sound feature spaces by the proposed method in two ways: visualization of categories and locations. First, we show the visualization of sound categories. In Figures 11
FIGURE 11. Visualization of image and sound features of FAIR-Play. There are three labels, Category (Single): sound categories for the single sound source, Category (Multiple): sound categories for multiple sound sources, and Location: sound locations. In the labels for Location, we use vectors representing the number of sound sources in each region. For example, label 101 denotes that the left, center, and right regions have one, zero, and one sound source, respectively.

In FAIR-Play, we see clear cluster structures in the visualization of categories for single sound sources (left column). For multiple sounds (center column), while several categories do not have enough samples to form clusters, categories with enough samples, such as drum-piano and cello-piano, form clusters. These indicate that the proposed method captures the semantic information of sound sources in FAIR-Play. For the visualization of the sound source locations (right column), we see that multiple local clusters are formed for each pattern of the sound location, although the clusters for several patterns are not very clear. Therefore, this visualization indicates that the proposed method also performs audio-visual learning based on the locations of sound sources in FAIR-Play.

In PseudoMUSIC, sound features form cluster structures in the visualization of sound categories. Also, although the image features do not form distinguished clusters compared with the sound features, multiple local clusters are formed for each label. These visualizations show that the network effectively learns the categories of sound sources by the proposed method in PseudoMUSIC. Furthermore, in both image and sound feature spaces, the proposed method forms clusters in the visualization of sound locations. These indicate that the
proposed method captures the locations of sound sources in PseudoMUSIC.

V. CONCLUSION
In this paper, we proposed a novel self-supervised approach to learn the semantic and spatial audio-visual relationships of stereo-recorded videos. The key idea is to utilize the ordered relationship among images and stereo sounds. To this end, we have proposed the SSR loss to learn image and sound feature encoders. We have applied the proposed method to cross-modal retrieval on the basis of both categories and locations of sound sources in stereo-recorded datasets. Furthermore, we have shown the effectiveness of the proposed method over baseline methods. The visualization of network attention qualitatively shows that the obtained features are appropriately focused on sound sources and that the proposed method is also effective for sound localization. Therefore, we can conclude that the proposed method is suitable for feature extraction from videos with stereo sounds.

REFERENCES
[1] S. Hong, W. Im, and H. S. Yang, “CBVRM: Content-based video-music retrieval using soft intra-modal structure constraint,” in Proc. ACM Int. Conf. Multimedia Retr., Jun. 2018, pp. 353–361.
[2] T. Halperin, A. Ephrat, and S. Peleg, “Dynamic temporal alignment of speech to lips,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 3980–3984.
[3] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon, “Learning to localize sound source in visual scenes,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4358–4366.
[4] R. Arandjelovic and A. Zisserman, “Objects that sound,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 435–451.
[5] A. Owens and A. A. Efros, “Audio-visual scene analysis with self-supervised multisensory features,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 631–648.
[6] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba, “Self-supervised moving vehicle tracking with stereo sound,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 7053–7062.
[7] K. Yang, B. Russell, and J. Salamon, “Telling left from right: Learning spatial correspondence of sight and sound,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9932–9941.
[8] R. Gao and K. Grauman, “2.5D visual sound,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 324–333.
[9] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, “Learning fine-grained image similarity with deep ranking,” in Proc. CVPR, Jun. 2014, pp. 1386–1393.

[10] Y. Aytar, C. Vondrick, and A. Torralba, “SoundNet: Learning sound representations from unlabelled video,” in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 902–909.

[11] R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in Proc. IEEE Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 609–617.

[12] T. Oya, S. Iwase, R. Natsume, T. Itazuri, S. Yamaguchi, and S. Morishima, “Do we need sound for sound source localization?” in Proc. Asian Conf. Comput. Vis. (ACCV), Nov. 2020, pp. 1–17.

[13] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman, “Localizing visual sounds the hard way,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 16867–16876.

[14] N. Khosravan, S. Ardestchi, and R. Puri, “On attention modules for audio-visual synchronization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops, Jun. 2019, pp. 25–28.

[15] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang, “Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning,” in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 3884–3892.

[16] B. Korbar, D. Tran, and L. Torresani, “Cooperative learning of audio and video models from self-supervised synchronization,” in Proc. Adv. Neural Inf. Process. Syst., vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 7763–7774.

[17] T. L. P. Morgado, N. Vasconcelos, and O. Wang, “Self-supervised generation of spatial audio for 360 video,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops, Jun. 2019, pp. 362–372.

[18] H. Zhou, X. Xu, D. Lin, X. Wang, and Z. Liu, “Sep-stereo: Visually guided stereophonic audio generation by associating source separation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 52–69.

[19] P. Morgado, Y. Li, and N. Vasconcelos, “Learning representations from audio-visual spatial alignment,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 4733–4744.

[20] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 815–823.

[21] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person re-identification by multi-channel parts-based CNN with improved triplet loss function,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1335–1344.

[22] L. Zha and E. Rahtu, “Visually guided sound source separation and localization using self-supervised motion representations,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2022, pp. 2171–2181.

[23] L. Wang, Y. Li, and S. Lazebnik, “Learning deep structure-preserving image-text embeddings,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 5005–5013.

[24] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, “Triplet-based deep hashing network for cross-modal retrieval,” IEEE Trans. Image Process., vol. 27, no. 8, pp. 3893–3903, Aug. 2018.

[25] L. Wang, Y. Li, J. Huang, and S. Lazebnik, “Learning two-branch neural networks for image-text matching tasks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 394–407, Feb. 2019.

[26] D. Zeng, Y. Yu, and K. Oyama, “Deep triplet neural networks with cluster-CCA for audio-visual cross-modal retrieval,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 16, no. 3, pp. 1–23, Aug. 2020.

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

[28] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, “The ACCV 2015 Face Database,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 570–586.

[29] X. Xu, H. Zhou, Z. Liu, B. Dai, X. Wang, and D. Lin, “Visually informed binaural audio generation without binaural audios,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 15485–15494.

[30] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[31] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, “The CIPIC HRIT database,” in Proc. IEEE Workshop Appl. Signal Process. Audio, Oct. 2001, pp. 99–102.

[32] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 3431–3440.

[33] R. Gao and K. Grauman, “Co-separating sounds of visual objects,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 3879–3888.

[34] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, “Scene parsing through ADE20K dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 633–641.

[35] J. Kärvelin and J. Kekäläinen, “Cumulated gain-based evaluation of IR techniques,” ACM Trans. Inf. Syst., vol. 20, no. 4, pp. 422–446, Oct. 2002.

[36] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, pp. 2579–2605, Nov. 2008.