# Self-Supervised Learning for Cross-Modal Retrieval based on Sound Category and Location

Tomoya Sato, Yusuke Sugano, Yoichi Sato
The University of Tokyo*

## 1. Introduction

The task of obtaining feature representations shared among data of different modalities has been studied for many multimedia application scenarios, such as cross-modal retrieval and grounding. In particular, because a strong connection between vision and hearing such as the cocktail party effect has been demonstrated in the cognitive psychology field, various studies have been conducted on the relationship between images and sound, including sound localization and sound separation. In audio-visual representation learning, self-supervised learning is often adopted because it is difficult to annotate videos by considering both visual and auditory information. For learning audio-visual relationships without annotations, pretext tasks based on semantic or temporal correspondences have been proposed [1, 6]. However, in these methods, the monaural sound is enough to capture these correspondences, making the analysis of spatial information such as sound location unaddressed.

From videos containing stereo sound, there is potential for extracting cross-modal features that capture the sound category and the location of the sound source. Learning such cross-modal features have practical applications including image-to-sound or sound-to-image retrieval that preserves both semantic and spatial information. This is because stereo sound contains both semantic and horizontal spatial information, i.e., what kind of object the sound source is and where the sound comes from. However, while some prior works have used videos with stereo sound for self-supervised cross-modal feature learning [2, 9], they have focused only on spatial information, which makes the representation of semantic information in the learned features insufficient.

This paper proposes a method for learning a feature space in which the distance between an image feature and a stereo audio feature represents both semantic and spatial similarity. In learning this feature space, we consider three types of stereo sound for an image taken from unlabelled videos: (a) sound corresponding to the image semantically and spatially, (b) sound not corresponding to the image spatially but having the same semantics, and (c) sound not corresponding to the image semantically. Here, we represent the spatial mismatch by flipping left and right channels of stereo sounds. Specifically, (a) is extracted from the original video, which has the same sound categories and locations,

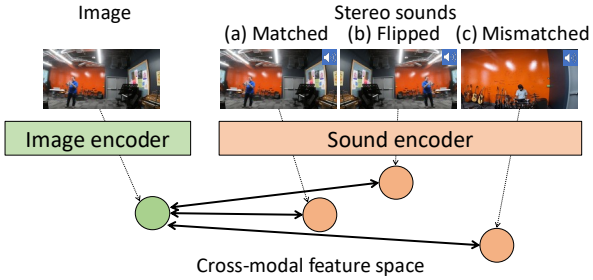


Figure 1. Assuming a dataset of unannotated videos, there could be three types of stereo sounds paired with an image.By learning the distance between these features as shown in this figure, we can obtain the semantic and spatial relationships between them.

(b) left-right flipped sound from the original video, which has the same categories but flipped locations, and (c) from different video, which has different sound categories. Our key observation is that the semantic and spatial similarity of the stereo sound for the image should be aligned in the order (a), (b), (c). Especially (b) is closer to the image than (c) because we assume that audio-visual semantic mismatch has a greater impact than the spatial mismatch. We illustrate an example in Fig. 1. For the image of the trumpet player on the left, we have (a) trumpet sound from left (Matched), (b) trumpet sound from the right (Flipped), and (c) drums sound (Mismatched). Then, (a) is placed in the closest distance to the image because it corresponds to the image semantically and spatially. Also, (b) is placed closer to the image than (c) because (b) has at least the same sound category, trumpet, while (c) represents drum sound which does not even semantically correspond to the image. By introducing the order constraint on the distances among image and sound features, the feature encoder can be trained to preserve both semantic and spatial information about the sound sources.

In this work, we propose a self-supervised method to learn cross-modal audio-visual features based on a novel loss, stereo sound ranking (SSR) loss. To enforce the ordered relationship between features, we employ a triplet loss function on image and sound inputs. Experimental results demonstrate that our method enables novel cross-modal retrieval with both semantic and spatial correspondences of the sound sources.

## 2. Proposed Method

Our method aims at learning a cross-modal feature representation that captures both semantic and spatial relation-

*{tomosato,sugano,ysato}@iis.u-tokyo.ac.jp

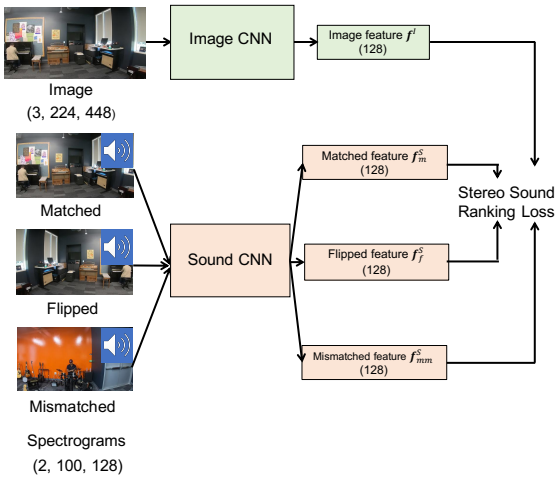


Figure 2. Proposed network. Features obtained from Image and Sound CNNs are input to the proposed loss function, SSR loss.

ships among images and stereo sounds. To this end, we propose to learn the feature space so that the order of three stereo sound features to the image feature is (a) Matched, (b) Flipped, and (c) Mismatched, as shown in Fig. 1.

Figure 2 shows an overview of the network. This consists of the Image and Sound CNNs which are designed by following a previous work [9]. The image and stereo sound features are 128-dimensional vectors obtained from Image and Sound CNNs with L2 normalization. $\boldsymbol{f}^I$ indicates the image feature, and $\boldsymbol{f}^S_{\mathrm{m}}; \boldsymbol{f}^S_{\mathrm{f}}; \boldsymbol{f}^S_{\mathrm{mm}}$ indicate three types of stereo sound features.

### 2.1. Stereo Sound Ranking (SSR) loss

We propose Stereo Sound Ranking (SSR) loss based on triplet loss [7] to learn the feature space that preserves semantic and spatial relationships among images and stereo sounds. Figure 3 shows the design of SSR loss.

Given the image feature $\boldsymbol{f}^I$ and the stereo sound features $\boldsymbol{f}^S_{\mathrm{m}}; \boldsymbol{f}^S_{\mathrm{f}}; \boldsymbol{f}^S_{\mathrm{mm}}$, SSR loss first computes the Euclidean distances $d_{\mathrm{m}}; d_{\mathrm{f}}; d_{\mathrm{mm}}$ of three feature pairs between $\boldsymbol{f}^I$ and each of $\boldsymbol{f}^S_{\mathrm{m}}; \boldsymbol{f}^S_{\mathrm{f}}; \boldsymbol{f}^S_{\mathrm{mm}}$. SSR loss $L_{\mathrm{SSR}}$ is then given as

$$L_{\mathrm{SSR}} = L_{\mathrm{m\&f}} + L_{\mathrm{f\&mm}} + L_{\mathrm{sc}} \tag{1}$$

where

$$L_{\mathrm{m\&f}} = \max(0; d_{\mathrm{m}} + \quad_{\mathrm{m\&f}} \quad d_{\mathrm{f}}); \tag{2}$$

$$L_{\mathrm{f\&mm}} = \max(0; d_{\mathrm{f}} + \quad_{\mathrm{f\&mm}} \quad d_{\mathrm{mm}}); \tag{3}$$

$$L_{\mathrm{sc}} = \max(0; d_{\mathrm{f}} \quad \quad_{\mathrm{sc}}): \tag{4}$$

(2)–(4) represent the following constraints, respectively. (2) gives the constraint that $d_{\mathrm{m}} < d_{\mathrm{f}}$. The parameter $_{\mathrm{m\&f}}$ is a margin for room of separation between $\boldsymbol{f}^S_{\mathrm{m}}$ and $\boldsymbol{f}^S_{\mathrm{f}}$. Similarly, (3) represents $d_{\mathrm{f}} < d_{\mathrm{mm}}$, and the parameter $_{\mathrm{f\&mm}}$ is a margin for room of separation between $\boldsymbol{f}^S_{\mathrm{f}}$ and $\boldsymbol{f}^S_{\mathrm{mm}}$. With

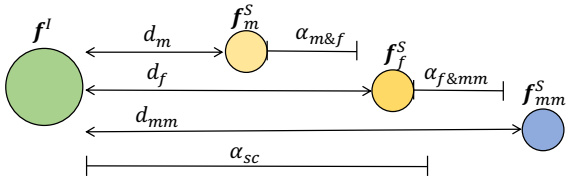


Figure 3. Design of Stereo Sound Ranking loss. We give an constraint to loss: $d_{\mathrm{m}} < d_{\mathrm{f}} < d_{\mathrm{mm}}$.

(2) and (3), we have the three distances in the desired order ($d_{\mathrm{m}} < d_{\mathrm{f}} < d_{\mathrm{mm}}$). However, it is not guaranteed that the stereo sound features of the same semantics reside within a certain range from the image feature. To solve this semantic constraint, we define (4). By this equation, $d_{\mathrm{f}}$ is limited within a certain range $_{\mathrm{sc}}$.

### 2.2. Implementation Details

To construct the training input, we first randomly take two videos with stereo sounds from the dataset. We extract an image and a stereo sound from one video. We also make the flipped sound by flipping the left and right channels of the sound. Also, we extract sound from the other video. These three stereo sounds are combined with the image into one input.

To extract an image from a video, we randomly sample one-second video segments from each 10-second video of the training dataset. The middle frame of the one-second video segment is used as an input RGB image after being resized to 480 $\times$ 240 resolution. As data augmentation, we randomly crop 448 $\times$ 224 images. The color and intensity are randomly changed in the range of 0.7 to 1.3.

We calculate the log-scaled mel-spectrogram of the one-second sound at 16kHz. Short-time Fourier transform (STFT) is applied by a 25 ms Hann window with 10 ms hop, and FFT size of 512. We stack the mel-spectrograms of left and right audio channels, and the input size becomes 2 $\times$ 100 $\times$ 128.

## 3. Experiments

In this section, we report experimental results on cross-modal retrieval to show that the proposed approach can acquire better feature representations containing both semantic and spatial audio-visual information.

### 3.1. Datasets

We train the proposed network on two stereo-recorded video datasets. Both datasets are randomly divided into training, validation, and test subsets with a ratio of 80%/10%/10%.

**FAIR-Play [3]** The dataset consists of 1,871 10-second videos of people playing musical instruments.There are 9 main musical instruments.

**YouTube-ASMR [9]** The dataset consists of 30,000 10-second ASMR videos. It is mainly composed of sounds emitted from the human face (voice, chewing sound) and sounds made by touching or hitting objects.

Because there is no ground-truth annotation of sound category and location on these datasets, we manually annotate the test subsets with category labels and their bounding boxes. We further select items from test subsets that have clear sound sources. Especially in YouTube-ASMR, we select only videos whose sound sources are male or female voices. We obtain 352 annotated image/sound pairs, including the ones that are horizontally flipped in both image and sound in FAIR-Play and 548 pairs in YouTube-ASMR.

### 3.2. Baseline Methods

Throughout the experiments, *Proposed*, the proposed method, is compared with the following baseline methods.

**Mismatch Classification [1]** This method performs binary classification identifying whether the visual and sound inputs are from the same video using the Euclidean distance between the image and sound features. The learned features capture audio-visual semantic relationships. Image and Sound CNNs are the same as *Proposed*.

**Flip Classification** As a replacement of the pretext classification task in *Mismatch Classification*, this method performs the binary classification whether the left and right channels of stereo sound are flipped or not [9] to learn audio-visual spatial relationships.

**Flip/Mismatch** This method uses both tasks of *Mismatch Classification* and *Flip Classification* in a multi-task learning manner. The Euclidean distance between image and sound features is calculated and fed into two separate branches for each task.

**Mismatch Distance** This corresponds to *Proposed* without $L_{\text{m\&f}}$, i.e., without considering the relationship with the flipped sound.

**Flip Distance** Similarly, this corresponds to the proposed method without $L_{\text{f\&mm}}$, i.e., without considering the relationship with the mismatched sound.

**CCA** This method uses weights obtained from existing large datasets. Image features are 4096-dimensional vectors extracted from the last hidden layer of VGG-16 trained on ImageNet [8]. Sound features are 128-dimensional vectors extracted from VGGish, which is CNN trained on AudioSet [4]. They are aligned to 128-dimensions by Canonical Correlation Analysis (CCA) and applied to retrieval.

### 3.3. Evaluation Metrics of Cross-Modal Retrieval

We use nDCG@K [5] as the evaluation metric of cross-modal retrieval. It evaluates top K retrieved items, with a value between 0 and 1, where higher values indicate better results. K is set to 5.

In calculating nDCG, we define the score that evaluates both sound category and location of retrieved items as follows. Note that in comparison with *CCA* the score based only on the category of the sound sources is used because the sound input in *CCA* is monaural sound and has only semantic information.

First, we define the similarity of sound categories in each dataset. In FAIR-Play, a tree representing the hierarchical meaning of instruments is constructed using the ontology in AudioSet [4]. The similarity is calculated using the distance between categories of the query and the retrieved item in this tree [1]. In YouTube-ASMR with a binary (male or female) category, the similarity is set to 1 for the same categories and 0 for the different categories. Then, we use the following two definitions for the score of retrieved items. *Category score* is used for comparison with *CCA*.

**Category score** To reflect only the category information, the similarity of sound categories is used as is.

**Category+location score** This score additionally takes into account the location of the sound source. We define the location of the sound source of the query $l_q$. This indicates whether the x-coordinate value of the center of the corresponding bounding box is on the left, center, or right when the image is divided into three parts. Let $c_q$ be the category of the sound source of the query. Similarly, we define $l_i; c_i$ for the retrieved items. Then, we compute the score as $_{l_q;l_i}$ $\text{sim}(c_q; c_i)$, where is Kronecker 's delta and $\text{sim}(c_q; c_i)$ is the similarity of sound categories.

### 3.4. Evaluation of Retrieval Performance

We perform both image-to-sound and sound-to-image cross-modal retrieval tasks to evaluate the performance based on both sound source categories and locations using *Category+location score*.

Table 1 shows the performances of *Proposed* and baseline methods except *CCA*. In FAIR-Play, we see that the score of *Proposed* is higher than that of *Mismatch Classification*, implying that our pretext task can learn the audio-visual spatial relationship better than the prior pretext task. We also see that *Proposed* achieved a higher score than *Flip Classification*, which indicates that *Proposed* captures the semantic information more effectively than simply performing left-right flip classification. Furthermore, the higher score than that of *Flip/Mismatch* indicates that simply combining these two classification tasks is not enough for the feature representation learning. The score of *Flip/Mismatch* is lower than that of *Mismatch*. This is partly because treating the *Flip* and *Mismatch* branches equally has a negative impact on learning the distance between features. In addition, *Proposed* outperforms *Mismatch Distance* and *Flip Distance*, which indicates that both $L_{\text{m\&f}}$ and $L_{\text{f\&mm}}$ play important roles in SSR loss.

Retrieval on YouTube-ASMR is, on the other hand, fun-