
Zengjie Song\textsuperscript{1} Yuxi Wang\textsuperscript{1,3} Junsong Fan\textsuperscript{1,2} Tieniu Tan\textsuperscript{1,2} Zhaoxiang Zhang\textsuperscript{1,2,3,}\textdagger
\textsuperscript{1}Center for Research on Intelligent Perception and Computing, NLPR, CASIA
\textsuperscript{2}University of Chinese Academy of Sciences (UCAS)
\textsuperscript{3}Centre for Artificial Intelligence and Robotics, HKISI_CAS
\{zengjie.song, wangyuxi2016, fanjunsong2016, zhaoxiang.zhang\}@ia.ac.cn, tnt@nlpr.ia.ac.cn

Abstract

Sound source localization in visual scenes aims to localize objects emitting the sound in a given image. Recent works showing impressive localization performance typically rely on the contrastive learning framework. However, the random sampling of negatives, as commonly adopted in these methods, can result in misalignment between audio and visual features and thus inducing ambiguity in localization. In this paper, instead of following previous literature, we propose Self-Supervised Predictive Learning (SSPL), a negative-free method for sound localization via explicit positive mining. Specifically, we first devise a three-stream network to elegantly associate sound source with two augmented views of one corresponding video frame, leading to semantically coherent similarities between audio and visual features. Second, we introduce a novel predictive coding module for audio-visual feature alignment. Such a module assists SSPL to focus on target objects in a progressive manner and effectively lowers the positive-pair learning difficulty. Experiments show surprising results that SSPL outperforms the state-of-the-art approach on two standard sound localization benchmarks. In particular, SSPL achieves significant improvements of 8.6\% cIoU and 3.4\% AUC on SoundNet-Flickr compared to the previous best. Code is available at: https://github.com/zjsong/SSPL.

1. Introduction

When strolling in a park brimming with life, you notice that the bird sitting on a twig is chirping; the puppy on the road ahead gives a little bark; and after a while an acquaintance may walk by and say hello to you friendly. Despite a short notice, humans own the excellent ability to associate the sounds they hear with the corresponding visual perception, and thus can localize and distinguish different sounding objects from one another.

To mimic humans’ such ability, in this work, we pay attention to the task of sound source localization in visual scenes, where the goal is to localize regions of the visual landscape that correlate highly with the audio cues. While handling this task is a long-standing challenge [17, 23], remarkable breakthroughs have been made until recent progresses on self-supervised audio-visual learning [1, 25, 26, 36, 38, 41]. These methods leverage the free supervision rooted in videos, \textit{e.g.}, the natural correspondence and/or temporal synchronization between audio and visual sources, to guide multi-modal feature extraction and alignment; then the similarity map between audio and visual features is usually employed to localize sounding objects. Among them, contrastive learning has particularly achieved impressive performance on this task [1, 5, 38, 41, 45].

Existing contrastive learning methods in this line of work...
can be cast into two categories: the first one is global-level contrastive learning (GLCL, Figure 1a) [1, 33, 38, 41], which commonly attracts audio and visual features extracted from the same video and repulses features from different videos; the other one is local-level contrastive learning (LLCL, Figure 1b) [5, 31, 45], which further compares audio feature with different visual feature components, even though they have correspondence at the video level. Generally, to perform contrastive learning, these methods randomly sample sounds to form negative pairs with the given video frame. However, this randomness can produce false negatives by sampling sounds that actually belong to the same category as the positive one, and thus hampering the model to align audio and visual features in semantic level. The misalignment as a result induces the learning process to build inaccurate audio-visual correspondence for localization.

We conduct a pilot experiment to illustrate the effect of such false negatives in Figure 2. Given image and sound from the same video (e.g., saxophone playing) as positive pair, other videos’ sounds, holding the same category as the positive one, are allowed to construct negative pairs in Figure 2a, while not allowed in Figure 2b. We keep the remaining training settings same for these two cases. During testing, consequently, the former case generates ambiguous localization on sounding objects (i.e., saxophone here) and the latter one not. Based on this observation, we take a step back and ask the questions: Do we really need negatives to develop self-supervised sound localization methods? Can the image-audio positive pair alone be used to achieve the same goal?

To answer these questions, we propose Self-Supervised Predictive Learning (SSPL), a negative-free approach for sound source localization through explicit audio-visual positive mining. The predictive learning is embodied from two perspectives (Figure 1c): predicting across different visual views and predicting across audio and visual modalities. Given an image-audio pair from one video, the former perspective hypothesizes that if two different visual views of one video frame contain the same sounding objects, they should share a consistent correspondence with the given sound source. We achieve this consistency by the mutual prediction of two audio-visual representations. For the later perspective, we devise a predictive coding module (PCM) that uses visual features to iteratively predict audio ones, providing a coarse-to-fine way to automatically align features. To the best of our knowledge, this is the first attempt to apply the self-supervised negative-free method to the audio-visual task of sound localization.

Our main contributions can be summarized as follows:

- We propose a novel negative-free method to extend a self-supervised learning framework to the audio-visual data domain for sound localization, and show how it can effectively address the false negative sampling problem.
- We propose the predictive coding module for feature alignment, which enables the model to progressively attend to relevant visual features while ignoring information irrelevant to audio cues, boosting sound localization significantly.
- Comprehensive experiments demonstrate the effectiveness of the proposed approach, which achieves localization performance superior to the state-of-the-art on SoundNet-Flickr and VGG-Sound Source.

### 2. Related Work

**Self-Supervised Visual Representation Learning.** Self-supervised learning (SSL) has achieved remarkable breakthroughs on large computer vision benchmarks. Most of the current SSL methods [7, 9–11, 16, 21, 22, 39, 46] resort to the design of contrastive learning strategy [35]. These methods, at their core, transform one image into multiple views, and repulse different images (negatives) meanwhile attracting the same image’s different views (positives). Recently, several efforts have been made to further relieve the requirement of negatives and simplify the SSL framework beyond conventional contrastive learning, including BarlowTwins [49], W-MSE [15], BYOL [20], and SimSiam [12]. In SimSiam [12], researchers investigate the importance of simple Siamese architecture for unsupervised representation learning, and empirically show that the stop-gradient operation is critical for the network to prevent collapse, even without using momentum encoder [21] and large batches [9]. These advances in image representation learning provide insights for our work to develop effective audio-visual SSL method.

**Audio-Visual Representation Learning.** The vision and sound are usually two co-occurring modalities, which can naturally be used to derive supervisions for audio-visual...
Learning [4, 8, 30, 37, 44]. In [4], for instance, the visual features extracted from pretrained teacher networks act to guide the student network to learn more discriminative sound representation, and vice versa in [37]. Kornbar et al. [30] and Owens and Efros [36] leverage the synchronization between audio and visual streams to build negative samples and contrastive losses, obtaining versatile multisensory features, respectively. Several works also explore the audio-visual correspondence by feature clustering [2, 25, 27]. In general, these methods focus on learning task-agnostic representations, which work well on classification-related downstream tasks, such as action/scene recognition [2, 13, 30, 33, 36], audio event classification [2, 3, 25, 30, 34], video retrieval [13], etc. However, they are not customized for sound source localization, and as a result only achieve limited performance on this task [25, 27, 36].

**Sound Localization in Visual Scenes.** Early works to solve this task mainly rely on statistical modeling of the cross-modal relationship by using, for example, mutual information [17, 23] and canonical correlation analysis [28, 29]. However, these methods as shallow models only show advantages in simple audio-visual scenarios. By digging into the correspondence between deep audio and visual features, recent deep learning methods give promising solutions to this problem [1, 5, 18, 25, 38, 41, 45, 47]. For instance, Senocak et al. [41] employ a two-stream framework and an attention mechanism to compute sound localization map. Qian et al. [38] achieve the same goal by using the class activation map (CAM) derived from a weakly-supervised approach. In [25, 27], audio and visual features are clustered, respectively, and the assignment weights based on the distance between features and cluster centers are adopted to localize sounding objects. In addition to viewing image and sound extracted from different videos as negative pair, Chen et al. [5] and Lin et al. [31] propose to mine hard negatives within an image-audio pair, i.e., background regions that correlate lowly with the given sound are treated as extra hard negatives. Different from these negative-based works, we handle the same task by explicit positive mining, providing an effective alternative for sound localization.

### 3. Method

Figure 3 depicts the overall framework of our SSPL, which is a three-stream network, making a big difference with widely-used two-stream ones. The top and bottom streams first serve to extract deep visual features from different views of the same image. Then, they employ PCM and AM to integrate visual and audio features, where the discriminative audio feature is derived from the middle stream subnetwork. Subsequently, two audio-visual representations are enforced to be similar by self-supervised predicting with each other. The sound localization map is a natural consequence of representation learning and is generated in the AM. Note that the vanilla SSPL without PCM focuses on exploring audio-visual correspondence across different image views (Sec. 3.2), while the PCM component excels at aligning features across modalities (Sec. 3.3), and thus boosting localization performance further. We elaborate on and formulate each part in the following.

#### 3.1. Unimodal Features of Audio and Vision

Let $I \in \mathbb{R}^{3 \times H_v \times W_v}$ and $a \in \mathbb{R}^{H_a \times W_a}$ denote a video frame and a corresponding audio signal from the same
video clip, respectively. Here the raw 1D audio waveform has been converted into the 2D spectrogram by Short-Time Fourier Transform (STFT), and therefore we use 2D CNNs to extract deep semantic features of audio modality like vision. In practice, we employ the off-the-shelf VGG16 [42] for frame processing (Enc_v) and the VGGish network [24] for spectrogram analysis (Enc_a), similar to [27]. The output feature map of the final convolution layer of VGG16 is treated as the original visual feature \( f_a \). We use layers before the final post-processing stage of VGGish to produce a high-level embedding as the original audio feature \( f_v \). These feature extraction processes are formulated as:

\[
\begin{align*}
&f_v = \text{Enc}_v(I), \quad f_v \in \mathbb{R}^{c_v \times h \times w}, \\
&f_a = \text{Enc}_a(a), \quad f_a \in \mathbb{R}^{c_a}.
\end{align*}
\]

**3.2. Predictive Learning across Visual Views**

This section details the vanilla SSPL, which contains the attention module to compute audio-visual representation from previously extracted features, and the self-supervised learning to guide model training via cross-view representation prediction.

Let \( I^1 \) and \( I^2 \) denote two randomly augmented views of the given image \( I \). The two views are respectively fed into the visual CNN, \( \text{Enc}_v \), to obtain spatial feature maps \( f_v^1 \) and \( f_v^2 \) as in Eq. (1). Besides, an audio feature vector \( f_a \) is derived from the audio signal \( a \) using Eq. (2). For simplicity, we use \( n \in \{1, 2\} \) to index different views.

**Attention Module (AM).** We adopt the normalized inner product (or cosine similarity) to measure the similarity between audio and visual features, as suggested by [3, 41]. Considering that \( f_a \) and \( f_v^n \) are from two heterogeneous modalities, we first transform \( f_a \) to be comparable with the visual feature via a nonlinear transformation, \( f_a = g(f_a) \in \mathbb{R}^{c_a} \), and then perform the similarity measurement. Formally, for the spatial location \( (i, j) \) in visual feature map \( f_v^n \), a similarity value is computed as follows:

\[
S^n(i, j) = \frac{\langle f_a, f_v^n(\cdot, i, j) \rangle}{\|f_a\|_2\|f_v^n(\cdot, i, j)\|_2^2}, \quad (i, j) \in [h] \times [w],
\]

where \( f_v^n(\cdot, i, j) \in \mathbb{R}^{c_v} \).

The similarity map \( S \in \mathbb{R}^{h \times w} \) plays two important roles in our method. On the one hand, it indicates the degree of correlation between each image location (after resized to image scale) and the given audio cues, and thus can serve as the sound localization map. On the other hand, it acts as an attention mechanism to weigh the original visual feature, resulting in the following audio-visual representation:

\[
f_{av}(k) = \sum_{i,j} \tilde{S}^n(i, j) f_v^n(k, i, j), \quad k \in \{1, \ldots, c_v\},
\]

\[
\tilde{S}^n = \frac{S^n - \min(S^n)}{\max(S^n) - \min(S^n)}.
\]

Here we scale the similarity values to \([0, 1]\) by min-max normalization [31]. This operation makes different feature elements more distinguishable, and performs better compared with the sigmoid and softmax scaling functions [38, 41] (see Table 5 for empirical comparisons). Since the \( f_{av} \in \mathbb{R}^{c_v} \) selects and integrates visual features that are more related to audio cues, we treat it as a multi-modal representation to advance subsequent learning.

**Self-Supervised Learning.** The learning procedure aims to make the two audio-visual representations similar. Our hypothesis is that two visual scenes containing the same sounding objects should consistently correspond to the same audio cues in semantic level. We follow SimSiam [12] to achieve this goal in the audio-visual setting.

Formally, we feed \( f_{av} \) into a MLP head to obtain the projection of corresponding view, \( z^n = \text{MLP}(f_{av}^n) \). Then a predictor head, denoted as \( \text{Pred} \), takes as input \( z^n \) to predict \( z^2 \) by minimizing the negative cosine similarity (NCS):

\[
\mathcal{L}_{NCS}(z^1, z^2) = -\langle \text{Pred}(z^1), z^2 \rangle / \|\text{Pred}(z^1)\|_2 \|z^2\|_2.
\]

To symmetrize the above loss, we also feed \( z^2 \) into \( \text{Pred} \) to estimate \( z^1 \), leading to another loss term \( \mathcal{L}_{NCS}(z^2, z^1) \). The total loss is therefore defined as:

\[
\mathcal{L}_{SSPL} = \frac{1}{2} \mathcal{L}_{NCS}(z^1, z^2) + \frac{1}{2} \mathcal{L}_{NCS}(z^2, z^1).
\]

However, as discussed in [12], directly minimizing the loss in Eq. (7) could easily induce representation collapse. To overcome this problem the stop-gradient (SG) operation is employed. That is, Eq. (6) is modified as \( \mathcal{L}_{NCS}(z^1, \text{SG}(z^2)) \), where \( z^2 \) is viewed as a constant such that branch on \( I^2 \) receives no gradient from \( z^2 \) through this loss term. Similarly we have \( \mathcal{L}_{NCS}(z^2, \text{SG}(z^1)) \), and the form in Eq. (7) is implemented as:

\[
\mathcal{L}_{SSPL} = \frac{1}{2} \mathcal{L}_{NCS}(z^1, \text{SG}(z^2)) + \frac{1}{2} \mathcal{L}_{NCS}(z^2, \text{SG}(z^1)).
\]

Note that we follow the SimSiam framework for its simplicity in the use of only positive pairs without representation collapse. However, our predictive learning strategy can be combined with other self-supervised learning methods, such as BYOL [20], W-MSE [15], and BarlowTwins [49]. We leave the potential extension for future works.

**3.3. Predictive Learning across Modalities**

In this section, we propose the PCM for audio and visual feature alignment, and continuously improving the localization performance of SSPL. The key idea inherits the spirit of predictive coding (PC) in neuroscience [40, 43, 48], which simulates the mechanism of information processing in visual cortex. In brief, PC uses feedback connections from a
higher-level area to a lower-level one to convey predictions of lower-level neural activities; it employs feedforward connections to carry the errors between the actual activities and the predictions; and the brain dynamically updates representations so as to progressively reduce the prediction errors. In our PCM (Figure 3), we treat the visual feature as a type of prior knowledge to predict the audio feature in an iterative manner. In the following, at the heart of PCM, we give the representation update rules of feedback and forward processes, respectively. The detailed derivations can be found in supplement.

Denote by $r_l(t), l \in \{1, \ldots, L\}, t \in \{0, \ldots, T\}$ the representation of the $l$-th layer of PCM network at time step $t$, and by $W_{l, l-1}$ the feedback connection weights from layer $l$ to layer $l-1$ (and vice versa for $W_{l-1,l}$).

The feedback process updates representations through a mechanism of layer-wise prediction generation. Concretely, at $l$-th layer the prediction, $p_l$, of representation, $r_l$, is first derived using the above layer’s representation, $r_{l+1}$. Then $r_l$ is updated with its previous state and the prediction, i.e., at time step $t$ we have:

$$p_l(t) = (W_{l+1,l})^T r_{l+1}(t),$$

$$r_l(t) \leftarrow \phi((1-b_l)r_l(t-1) + b_l p_l(t)), \quad (9)$$

where $\phi$ is a nonlinear activation function and $b_l$ serves as a positive scalar to balance two terms. The above update rules are executed from top layer $L$ to bottom layer 1 in sequence, and by setting $p_L(t) \equiv f_v^n$, we in fact achieve further feature extraction from visual source.

In feedforward process, representations are again modulated based on prediction errors emerged at each layer. Specifically, the representation $r_{l-1}$ and its prediction $p_{l-1}$ are often unequal, resulting in a prediction error $e_{l-1}$. The error signal contains unpredictable components of $r_{l-1}$, and is forwarded to higher level to correct the representation $r_l$. This leads to complementary update rules:

$$e_{l-1}(t) = r_{l-1}(t) - p_{l-1}(t), \quad (11)$$

$$r_l(t) \leftarrow \phi((1-b_l)r_l(t-1) + b_l p_l(t)), \quad (12)$$

where $r_0(t) \equiv f_a$ is the original audio feature, $p_0(t) = \phi((W_{1,0})^T r_1(t))$ refers to the prediction of $f_a$, and $a_l$ denotes a trade-off scalar like $b_l$.

PCM conducts the two distinct processes alternatively while all layers’ representations are progressively refined so as to reduce the prediction error. Subsequently, we use a $1 \times 1$ convolution to transform the top layer representation at last time step, $r_L(T)$, to a new visual feature, $f_v^n$, with the same dimension of $f_v^0$. Consequently the vanilla SSPL method in Sec. 3.2 can be enhanced by feeding $f_v^n$, instead of $f_v^0$, into the AM to compute audio-visual representation (i.e., Eqs. (3) and (4)).

4. Experiments

4.1. Datasets and Evaluation Metrics

SoundNet-Flickr [4]. This dataset consists of more than 2 million videos from Flickr. We use a 3s audio clip around the middle frame of the whole audio, and the accompanied video frame to form an image-audio pair. Following [5,41], we train models with two random subsets of 10k and 144k image-audio pairs, respectively, and perform evaluation on the 250 annotated pairs provided by [41]. Note that the location of the sound source in each test frame is given by 3 separate bounding boxes, each of which is obtained by a different annotator.

VGG-Sound [6] and VGG-Sound Source [5]. VGG-Sound dataset contains over 200k video clips that are divided into 300 sound categories. Similar to [5], we conduct training with 10k and 144k image-audio pairs randomly sampled from this dataset, respectively. For fair comparisons with recent works [1,5,41], we evaluate models on the VGG-Sound Source (VGG-SS) benchmark with 5k annotated image-audio pairs collected by [5]. Compared with SoundNet-Flickr benchmark that spans about 50 sounding object classes, VGG-SS has 220 classes and thus providing a more challenging scenario for sound localization task.

We focus on and reimplement two related methods, Attention [41] and HardWay [5] (SOTA on this task), which could be representatives of GLCL- and LLCL-based approaches, respectively. We denote our method without using PCM by SSPL (w/o PCM), and the version equipped with PCM by SSPL (w/ PCM). Additionally, we employ consensus Intersection over Union (cIoU) and Area Under Curve (AUC) as evaluation metrics, and report cIoU scores with threshold 0.5 in experiments, same as [5,31,38,41].

4.2. Implementation Details

We use VGG16 [42] pretrained on ImageNet [14] and VGGish [24] pretrained on AudioSet [19] as visual and audio feature extractors, respectively. The visual input is an image of size $256 \times 256 \times 3$, on which we perform the data augmentation pipeline: random cropping with $224 \times 224$ resizing and random horizontal flip. The raw 3s audio signal is re-sampled at 16kHz and further transformed into $96 \times 64$ log-mel spectrograms as audio input, and the audio output feature $f_a$ is a 128D vector. The nonlinear audio feature transformation function $g(\cdot)$ of SSPL is instantiated with a single two-layer network as in [41]: FC(512)-ReLU-FC(512). We closely follow SimSiam [12] to set the projection and prediction MLPs. For PCM, we mainly adopt Conv-MaxPool-GELU layers in the feedback pathway, and Upsample-DeConv-GELU layers in the feedforward counterpart. The weights of two feature extractors are kept frozen during training, and we optimize the rest of the model with AdamW [32]. We utilize the early stop-
ping strategy to avoid overfitting in all cases. More setting
details (e.g., learning rate and batch size) are in supplement.

### 4.3. Comparisons with State-of-the-art Methods

We first compare SSPL with recent methods on the
SoundNet-Flickr test set in Table 1. We observe that when
trained by 10k Flickr samples, the vanilla SSPL (w/o PCM)
performs favorably against the two competing methods,
HardWay [5] and ICL [31], while the enhanced SSPL (w/ PCM)
outperforms the previous best [31] by a large margin (0.710 vs. 0.743, around 5% improvement). In the
Flickr144k training case, SSPL (w/ PCM) increases performance by 8.6% cIoU and 3.4% AUC compared to HardWay, establishing a new state-of-the-art on this benchmark. These results demonstrate that SSPL without relying on negatives is feasible and effective for sound localization.

Following [5], we also train on VGG-Sound using respec-
tive 10k and 144k data pairs, which enables SSPLs to
achieve the top two localization performance in both set-
tings. As discussed in [5], the sounding objects are often
visible in video clips from VGG-Sound, revealing that our
method can benefit from the improved data quality. What’s
more, the performance of SSPL is significantly boosted by
PCM, especially in the 10k’s setting (11% improvement for
Flickr10k and 9% for VGG-Sound10k). This illustrates the
advantage of PCM for facilitating sound localization.

We further evaluate SSPL on the newly released VGG-

<table>
<thead>
<tr>
<th>Method</th>
<th>Training set</th>
<th>cIoU ↑</th>
<th>AUC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention [41]</td>
<td>Flickr10k</td>
<td>0.442</td>
<td>0.461</td>
</tr>
<tr>
<td>DMC [25]</td>
<td>Flickr10k</td>
<td>0.414</td>
<td>0.450</td>
</tr>
<tr>
<td>CALV [27]</td>
<td>Flickr10k</td>
<td>0.500</td>
<td>0.492</td>
</tr>
<tr>
<td>MSSL [38]</td>
<td>Flickr10k</td>
<td>0.522</td>
<td>0.496</td>
</tr>
<tr>
<td>AVObject [19]</td>
<td>Flickr10k</td>
<td>0.546</td>
<td>0.504</td>
</tr>
<tr>
<td>DSOL [26]</td>
<td>Flickr10k</td>
<td>0.566</td>
<td>0.515</td>
</tr>
<tr>
<td>HardWay [5]</td>
<td>Flickr10k</td>
<td>0.615</td>
<td>0.535</td>
</tr>
<tr>
<td>ICL [31]</td>
<td>Flickr10k</td>
<td>0.710</td>
<td>0.580</td>
</tr>
<tr>
<td>SSPL (w/o PCM)</td>
<td>Flickr10k</td>
<td>0.671</td>
<td>0.556</td>
</tr>
<tr>
<td>SSPL (w/ PCM)</td>
<td>Flickr10k</td>
<td>0.743</td>
<td>0.587</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Training set</th>
<th>cIoU ↑</th>
<th>AUC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention [41]</td>
<td>VGG-Sound10k</td>
<td>0.522</td>
<td>0.502</td>
</tr>
<tr>
<td>HardWay [5]</td>
<td>VGG-Sound10k</td>
<td>0.647</td>
<td>0.560</td>
</tr>
<tr>
<td>SSPL (w/o PCM)</td>
<td>VGG-Sound10k</td>
<td>0.699</td>
<td>0.572</td>
</tr>
<tr>
<td>SSPL (w/ PCM)</td>
<td>VGG-Sound10k</td>
<td>0.763</td>
<td>0.591</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Training set</th>
<th>cIoU ↑</th>
<th>AUC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention [41]</td>
<td>VGG-Sound144k</td>
<td>0.723</td>
<td>0.605</td>
</tr>
<tr>
<td>HardWay [5]</td>
<td>VGG-Sound144k</td>
<td>0.735</td>
<td>0.590</td>
</tr>
<tr>
<td>SSPL (w/o PCM)</td>
<td>VGG-Sound144k</td>
<td>0.739</td>
<td>0.602</td>
</tr>
<tr>
<td>SSPL (w/ PCM)</td>
<td>VGG-Sound144k</td>
<td>0.767</td>
<td>0.605</td>
</tr>
</tbody>
</table>

Table 1. Quantitative localization results on SoundNet-Flickr test set. “∗” denotes our reproduction, and “†” indicates improved reproduction vs. original papers (see supplement).

<table>
<thead>
<tr>
<th>Method</th>
<th>Training set</th>
<th>cIoU ↑</th>
<th>AUC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention [41]</td>
<td>VGG-Sound144k</td>
<td>0.319</td>
<td>0.370</td>
</tr>
<tr>
<td>HardWay [5]</td>
<td>VGG-Sound144k</td>
<td>0.297</td>
<td>0.357</td>
</tr>
<tr>
<td>SSPL (w/o PCM)</td>
<td>VGG-Sound144k</td>
<td>0.270</td>
<td>0.348</td>
</tr>
<tr>
<td>SSPL (w/ PCM)</td>
<td>VGG-Sound144k</td>
<td>0.339</td>
<td>0.380</td>
</tr>
</tbody>
</table>

Table 2. Quantitative localization results on VGG-SS test set.

SS benchmark and report results in Table 2. Because in
this challenging benchmark the sounding object categories
are more diverse and the number of test samples is greater
than those of SoundNet-Flickr [5], the performance of all
methods drops severely compared with the results in Ta-
ble 1. While SSPL (w/o PCM) still outperforms Attention
by a large margin, it does not overtake HardWay. We
attribute this to the limitation of vanilla SSPL on dealing
with background noise (see Sec. 4.4 for an empirical com-
parison). However, by combining with feature alignment
module, SSPL (w/ PCM) yields performance better than the
state-of-the-art HardWay, especially by a substantial gap in
the 10k’s scenario (0.277 vs. 0.314, over 13% gain). This
verifies the superiority of the enhanced SSPL.

To address diverse demands for sound localization fine-
ness, we compute cIoU scores with various thresholds as
shown in Figure 4. The proposed method, SSPL (w/ PCM),
again consistently surpasses the state-of-the-art (HardWay)
under all thresholds.

### 4.4. Qualitative Analysis

We provide visualized localization results in Figure 5.
We observe that Attention [41] is prone to overlook tar-
get objects (e.g., the first and second rows in Figure 5a)
and cover unrelated background details (e.g., ground and
sky). Since localization map also visualizes similarities be-
tween audio and visual features, the inaccurate localization
indicates that Attention (random sampling of negatives) has
the potential to misalign features. Although HardWay [5]
Figure 5. **Qualitative comparisons.** In each panel, the first column shows images accompanied with annotations, and remaining columns represent the predicted localization of sounding objects. Here the attention map or similarity map produced by different methods is visualized as the localization map. Note that for SoundNet-Flickr the bounding boxes are derived from multiple annotators.

![Visualisation on SoundNet-Flickr test set](image1)

![Visualisation on VGG-SS test set](image2)

|     | Pre-train | Stop-grad | $T$ | cIoU $\uparrow$ | AUC $\uparrow$
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td></td>
<td></td>
<td></td>
<td>0.141</td>
<td>0.147</td>
</tr>
<tr>
<td>(b)</td>
<td>✓</td>
<td></td>
<td></td>
<td>0.382</td>
<td>0.432</td>
</tr>
<tr>
<td>(c)</td>
<td></td>
<td>✓</td>
<td></td>
<td>0.570</td>
<td>0.511</td>
</tr>
<tr>
<td>(d)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.671</td>
<td>0.556</td>
</tr>
<tr>
<td>(e)</td>
<td>✓</td>
<td>✓</td>
<td>1</td>
<td>0.655</td>
<td>0.562</td>
</tr>
<tr>
<td>(f)</td>
<td>✓</td>
<td>✓</td>
<td>3</td>
<td>0.719</td>
<td>0.584</td>
</tr>
<tr>
<td>(g)</td>
<td>✓</td>
<td>✓</td>
<td>5</td>
<td>0.743</td>
<td>0.587</td>
</tr>
</tbody>
</table>

Table 3. **Ablation on training strategies.** “Pre-train” represents whether using the ImageNet-pretrained backbone to extract visual features, and $T$ denotes the recursive cycles for iterative computing in PCM during training.

4.5. **Ablation Study**

In this section, we delve deeper into SSPL by conducting extensive ablation studies. Unless otherwise specified, all experiments are performed on SoundNet-Flickr dataset.

**Training Strategy.** As discussed in prior art [12], a simple Siamese network without using negative samples can easily suffer from the problem of representation collapse. In this regard we evaluate key factors of SSPL that facilitate audio-visual learning. In Table 3a we train the model from scratch while removing the stop-gradient operation, which indeed causes collapse in our practice. The variant with only pre-training strategy (Table 3b) improves performance because of the better parameter initialization, but it does not avoid collapsed solution yet. Adding stop-gradient alone during training (Table 3c) can obtain obvious gains, and the combination with pre-training (Table 3d) further boosts cIoU to 0.671, which is the default configuration of vanilla SSPL.

Based on above configuration, we perform additional ablation on the recursive cycles for representation updates in PCM. The performance slightly drops by 2% as conducting feedback and feedforward representation updates (Eqs. (9) to (12)) only once (Table 3e). This is because one computing step is not enough for PCM to reduce prediction errors between audio and visual features, and such non-negligible errors could degrade the subsequent learning. However, by increasing recursive cycles, SSPL can harvest significant performance improvements (nearly 10% in Table 3f and over 13% in Table 3g, respectively).

In summary, the results demonstrate that stop-gradient also works in our audio-visual setting to prevent collapse; and that both pre-training and PCM induce the model to learn effectively so as to promote localization accuracy.

**Augmentation.** We investigate the influence of various image augmentations on localization. As shown in Table 4, with the random crop baseline, our method can already achieve reasonable performance, indicating that object scales really matter in SSPL. However, except for horizontal flip (over 30% and 8% improvements on two datasets, respectively), randomly combining other augmentations...
The runner-up. Parameters used to generate different augmentations are provided in supplement.

<table>
<thead>
<tr>
<th>Augmentation</th>
<th>SoundNet-Flickr</th>
<th>VGG-SS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cIoU ↑</td>
<td>AUC ↑</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crop (baseline)</td>
<td>0.514</td>
<td>0.499</td>
</tr>
<tr>
<td>+ Horizontal flip</td>
<td>0.671</td>
<td>0.556</td>
</tr>
<tr>
<td>+ Vertical flip</td>
<td>0.667</td>
<td>0.551</td>
</tr>
<tr>
<td>+ Translation</td>
<td>0.643</td>
<td>0.541</td>
</tr>
<tr>
<td>+ Rotation</td>
<td>0.639</td>
<td>0.543</td>
</tr>
<tr>
<td>+ Grayscale</td>
<td>0.610</td>
<td>0.535</td>
</tr>
<tr>
<td>+ Color jittering</td>
<td>0.679</td>
<td>0.560</td>
</tr>
<tr>
<td>+ Gaussian blur</td>
<td>0.619</td>
<td>0.533</td>
</tr>
</tbody>
</table>

Table 4. Ablation on image augmentations. Bold indicates the best and Underline the runner-up. Parameters used to generate different augmentations are provided in supplement.

<table>
<thead>
<tr>
<th>Scaling method</th>
<th>cIoU ↑</th>
<th>AUC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReLU [38]</td>
<td>0.353</td>
<td>0.424</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>0.647</td>
<td>0.547</td>
</tr>
<tr>
<td>Softmax [41]</td>
<td>0.667</td>
<td>0.554</td>
</tr>
<tr>
<td>ReLU + Softmax [41]</td>
<td>0.574</td>
<td>0.531</td>
</tr>
<tr>
<td>Min-Max Norm.</td>
<td>0.671</td>
<td>0.556</td>
</tr>
</tbody>
</table>

Table 5. Ablation on scaling methods.

Further Analysis of PCM. We empirically clarify the remarkable ability of PCM to boost sound localization. Since PCM features an iterative computing procedure, we inspect the performance of SSPL (w/ PCM) with different iterations in Figure 6. We observe that the localization accuracy tends to increase given more iterative computations, especially at the initial three time steps. To understand why this is the case, we look into attention maps from some test samples, as shown in Figure 7. PCM infers different visual representations with varying time steps (1 through 5), which are further used by AM to yield different attention maps. Attention is less definitive (light red on sounding objects) and/or inaccurate (crimson on backgrounds) at early time steps. At later time steps, however, the model corrects itself to pay more definitive and accurate attention to the objects of interest. Adjusting attention in such a coarse-to-fine manner is particularly helpful to address ambiguous cases, where the object’s appearance may be similar to backgrounds (e.g., the second row in Figure 7).

5. Conclusion and Future Works

In this work, we have developed a self-supervised audio-visual learning method, SSPL, that improves visual sound localization performance by explicit positive mining. A three-stream network, as well as its training strategy, was designed to explore correspondence between sound and video frame from the same video clip. We further proposed PCM to align audio and visual features via cross-modal feature prediction, which boosts localization accuracy significantly. Our approach shows promising performance on sound localization task, especially achieving the new state-of-the-art on SoundNet-Flickr benchmark.

While SSPL excels at single sound source localization, it is not applicable to localize multiple sound sources in unconstrained videos [38], which is still a challenge for the community. A potential solution is to develop weakly- or semi-supervised methods. We leave it for future works.

Acknowledgements. This work was supported in part by the Major Project for New Generation of AI (No. 2018AAA0100400), in part by the National Natural Science Foundation of China (Nos. 61836014, U21B2042, 62072457, 62006231, and 61976174), and in part by the Project funded by China Postdoctoral Science Foundation (No. 2021M703489).
References


[16] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In CVPR, pages 3299–3309, 2021. 2


[38] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In ECCV, pages 292–308, 2020. 1, 2, 3, 4, 5, 6, 8


[41] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In CVPR, pages 4358–4366, 2018. 1, 2, 3, 4, 5, 6, 8


