MARGINNCE: ROBUST SOUND LOCALIZATION WITH A NEGATIVE MARGIN

Sooyoung Park[1,2]*, *Arda Senocak*[1]*, *Joon Son Chung*[1]

[1]Korea Advanced Institute of Science and Technology, South Korea
[2]Electronics and Telecommunications Research Institute, South Korea

ABSTRACT

The goal of this work is to localize sound sources in visual scenes with a self-supervised approach. Contrastive learning in the context of sound source localization leverages the natural correspondence between audio and visual signals where the audio-visual pairs from the same source are assumed as positive, while randomly selected pairs are negatives. However, this approach brings in noisy correspondences; for example, positive audio and visual pair signals that may be unrelated to each other, or negative pairs that may contain semantically similar samples to the positive one. Our key contribution in this work is to show that using a less strict decision boundary in contrastive learning can alleviate the effect of noisy correspondences in sound source localization. We propose a simple yet effective approach by slightly modifying the contrastive loss with a negative margin. Extensive experimental results show that our approach gives on-par or better performance than the state-of-the-art methods. Furthermore, we demonstrate that the introduction of a negative margin to existing methods results in a consistent improvement in performance.

Index Terms— audio-visual learning, audio-visual sound source localization, audio-visual correspondence, self-supervised learning

1. INTRODUCTION

We understand the world around us through multiple sensory signals. Among them, sight and sound signals are continuously used for our perception. To make this perception ability seamless, the human brain has developed to organize audio and visual modalities by associating or separating them. Thus, mimicking this ability is of great interest in order to have better learning algorithms. Audio-visual learning is explored with a variety of tasks such as audio-visual fusion and video understanding [1, 2, 3, 4, 5, 6], sound source localization [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20], audio spatialization [21, 22], and audio-visual sound separation [12, 23, 24, 25, 26].

In this work, we explore the sound source localization task. Human perception has the capability to easily find the objects or events that make the sound in the scene. We leverage the natural correspondence between how objects look and what sounds they make. Humans learn this correspondence without any specific training phase during their daily lives. Thus, accurately solving the sound localization problem in a self-supervised way is the main goal of this line of research. There have been vast efforts on self-supervised sound localization tasks recently. [7, 8, 9, 27] use the correspondence as a self-supervision proxy task in their training for sound localization. Further, [7, 9] use the attention mechanism to refine visual features with the obtained sound source localization predictions to learn the better correspondence between audio and visual signals. Hu *et al*. [10] incorporate a clustering approach in audio-visual samples to learn cross-modal correlation. More recently, starting with [14], the noise contrastive learning is adopted in sound localization methods [15, 18, 19, 20]. While [14] uses a hard negative mining approach to consider the background area in the positive image as a negative sample for contrastive learning, Senocak *et al*. [15] mine multiple semantically similar samples – hard positives – to use in contrastive learning. [19] extends the model of LVS [14] with aggressive data augmentations along with the geometrical consistency loss to give invariance and equivariance properties. [18] presents a multiple instance learning approach by focusing only on the most correspondent area on the image for audio-visual contrastive learning. Additionally, the initial self-supervised sound localization predictions are refined by mixing the object-guided activation maps from pre-trained visual encoders in the inference phase in [18]. [20] addresses the problem of simultaneous negative detection and localization by introducing an off-screen sound detection objective, allowing it to detect off-screen sound sources. Lastly, different from the aforementioned approaches that incorporate positive and negative samples together, [17] explores a negative-free learning approach in sound localization.

The correspondence between audio and visual signals undoubtedly plays a key role in general audio-visual learning and specifically in sound source localization as well. Contrastive learning approaches aim to align the features of the same instances while distinguishing the ones from different instances. Similarly, contrastive learning in the context of audio-visual learning leverages the natural correspondence and assigns the audio-visual pairs from the same source as positive and randomly selected mismatched pairs as negative because they are not related. While this seems plausible ideally, it leads to noisy correspondences in reality. These noisy correspondences can be in two forms - (1) Audio-visual signals from the same source may not be semantically related, uninformative to each other, (2) Negative pairs may contain semantically related information to the positive one due to the random selection in a batch. Morgado *et al*. [28] show that learning process is falsely guided because of these noisy correspondences when contrastive learning is used without careful consideration.

This motivates us to design a sound localization method that is more robust to noisy correspondences. To this end, we introduce a "negative margin" in the contrastive learning loss function, InfoNCE [29], to reduce the effect of the noisy samples. Considering the standard InfoNCE loss with a zero margin or a positive margin, these noisy correspondences will be pulled or pushed falsely in the wrong direction, even more with a positive margin. It degrades the learning ability of the model. However, as discussed in [30], a negative margin can alleviate the effect of noisy correspondences by providing a looser decision boundary. Our experiments support our design, employing InfoNCE loss with a negative margin, by showing that this simple approach improves the performance of the sound localization performance on standard benchmarks.

We propose a new training loss rather than a new sound source localization architecture. To the best of our knowledge, this is the first study on the effect of margin value in contrastive learning loss for sound source localization. Our main contributions are summarized as follows: 1) We present a self-supervised sound source localization model that

*These authors contributed equally to this work.

uses a margin contrastive loss; 2) We demonstrate that using less strict decision boundaries in contrastive learning, only a simple extension of the contrastive loss function with a negative margin, gives on-par or better sound localization performance with state-of-the-art methods that use additional task-oriented strategies; 3) We further investigate that applying a negative margin contrastive loss into existing works consistently improves the performances and shows its merit.

2. APPROACH

2.1. Preliminaries

This part describes our contrastive learning method for the audio-visual sound source localization. Let the image frame $\mathbf{v}_i \in \mathbb{R}^{3\times H_v \times W_v}$ and the audio spectrogram $\mathbf{a}_i \in \mathbb{R}^{1\times H_a \times W_a}$ from the i-th clip $\mathbf{X}_i = \{\mathbf{v}_i, \mathbf{a}_i\}$. To learn audio-visual correspondence, we use the training method maximizing the similarity between the image representation map $\mathbf{V}_i = f_v(\mathbf{v}_i; \boldsymbol{\theta}_v) \in \mathbb{R}^{c\times h\times w}$ using an image encoder f_v and global audio representation $\mathbf{A}_i = f_a(\mathbf{a}_i; \boldsymbol{\theta}_a) \in \mathbb{R}^c$ using an audio encoder f_a. Then, we localize the sound source on the given image with an audio-visual response map $\boldsymbol{\alpha}_{ij} \in \mathbb{R}^{h\times w}$ obtained by using the pixel-wise cosine similarity between the image representation $\mathbf{V}_i$ and the globally summarized audio representation $\mathbf{A}_j$. We build our model based on recent works [14, 18]. As a baseline, we use LVS-based loss function [14] on top of the EZ-VSL [18] architecture. The objective function of our baseline is as follows:

$$S_{i,j} = \frac{1}{\left\|\sigma\left(\frac{\boldsymbol{\alpha}_{i,j}-\epsilon}{\beta}\right)\right\|_1} \left\langle \sigma\left(\frac{\boldsymbol{\alpha}_{i,j}-\epsilon}{\beta}\right),\ \boldsymbol{\alpha}_{i,j} \right\rangle, \tag{1}$$

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{S_{i,i}/\tau}}{e^{S_{i,i}/\tau} + \sum_{i\neq j} e^{S_{i,j}/\tau}}, \tag{2}$$

where $\langle\cdot,\cdot\rangle$ refers Frobenius inner product, σ is the sigmoid function for thresholding $\boldsymbol{\alpha}_{i,j}$, ϵ denotes the thresholding parameter, β is the temperature for thresholding, $S_{i,j}$ refers the spatial-wise averaged value of the thresholded audio-visual response map, and τ is the temperature for contrastive loss. In the inference stage, we can deduce where the sound is visually localized in the paired image using audio-visual response map $\boldsymbol{\alpha}_{ii}$ from the clip $\mathbf{X}_i$.

2.2. Training

Based on the architecture and optimization mentioned above, we introduce marginNCE. In general, margin loss has been applied to make strict decision boundaries among the embedding space by adding a positive margin to the distance between two different embeddings to increase discriminability. However, audio-visual learning may suffer from the faulty positive problem because of the possibility that the image and paired audio are not semantically aligned, and the faulty negative problem due to random sampling in batch configurations [28]. Therefore, as in [30], we apply a looser decision boundary by using a negative margin m on (2) to alleviate the effect of noisy correspondences from the forementioned two problems. We simply modify the contrastive loss with a margin. The proposed objective function, marginNCE, is as follows:

$$\mathcal{L}_{marginNCE} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{(S_{i,i}-m)/\tau}}{e^{(S_{i,i}-m)/\tau} + \sum_{i\neq j} e^{S_{i,j}/\tau}}. \tag{3}$$

Method	VGG-SS cIoU ↑	VGG-SS AUC ↑	Flickr-SoundNet cIoU ↑	Flickr-SoundNet AUC ↑
Attention [7]$_{\text{CVPR18}}$	18.50	30.20	66.00	55.80
LCBM [16]$_{\text{WACV22}}$	32.20	36.60	-	-
LVS [14]†$_{\text{CVPR21}}$	30.30	36.40	72.40	57.80
LVS [14]$_{\text{CVPR21}}$	34.40	38.20	71.90	58.20
HardPos [15]$_{\text{ICASSP22}}$	34.60	38.00	76.80	59.20
SSPL(w/o PCM) [17]$_{\text{CVPR22}}$	27.00	34.80	73.90	60.20
SSPL(w/ PCM) [17]$_{\text{CVPR22}}$	33.90	38.00	76.70	60.50
EZ-VSL(w/o OGL) [18]$_{\text{ECCV22}}$	35.96	38.20	78.31	61.74
SSL-TIE [19]$_{\text{ACM MM22}}$	**38.63**	**39.65**	79.50	61.20
Ours	38.25	39.06	**83.94**	**63.20**
EZ-VSL(w/ OGL.) [18]$_{\text{ECCV22}}$	38.85	39.54	83.94	63.60
Ours (w/ OGL)	**39.78**	**40.01**	**85.14**	**64.55**

Table 1. **Quantitative results on the VGG-SS and SoundNet-Flickr test sets**. All models are trained with 144K samples from VGG-Sound and tested on VGG-SS and SoundNet-Flickr. † is the result of the model released on the official project page.

3. EXPERIMENTS

3.1. Datasets and Evaluation Metrics

Datasets. We train our method on VGGSound [31] and SoundNet-Flickr-Training set provided by [7, 9]. VGGSound is an audio-visual dataset containing around 200K videos. SoundNet-Flickr-training set is the subset of SoundNet-Flickr [32] and it has 144K samples. After training, the sound localization performance is tested with VGG-SS [14] and SoundNet-Flickr-Test [7] datasets. These evaluation sets have bounding box annotations of sound sources for ∼5K and 250 samples, respectively.
Evaluation metrics. We measure the sound localization performance with two commonly used metrics: 1) Consensus Intersection over Union (cIoU) [7] measures the localization accuracy between the ground-truth and the prediction with intersection over union approach. 2) Area Under Curve (AUC) measures the area under the cIoU curve plotted by various threshold values from 0 to 1.

3.2. Implementation Details

Following the common practice in earlier sound localization methods [14, 15, 18, 19], we use the center frame of the video with the corresponding 3 seconds audio segment around that frame as input data during training on the VGGSound dataset. In the SoundNet-Flickr dataset, frames are given with the paired audio. Input images for training are in the size of 224×224. We use 16kHz sampling rate for audios in both datasets. We transform audios to log spectrograms with the size of 257×200. Similar to [14, 15, 18, 19], ResNet18 is used as a backbone network for each modality. We set the hyperparameters as $\epsilon = 0.65$, $\beta = 0.03$, and $\tau = 0.07$. Unless it is mentioned explicitly, we adopt the value of -0.2 as a margin in our experiments. We use adam optimizer with the weight decay. We train our model for 20 epochs.

3.3. Quantitative Results

3.3.1. Comparison with the State-of-the-art Methods

In this section, we compare our method with existing sound source localization approaches. Specifically, we provide the results in two settings by following previous works [14, 18, 19]: 1) Training on VGGSound-144K and testing on VGG-SS and SoundNet-Flickr test sets 2) Training on SoundNet-Flickr-Training-144K and testing on SoundNet-Flickr test set. All the models are trained and tested on the same amount of data. Results are shown in Table 1 and Table 2. Our model outperforms prior work on the SoundNet-Flickr test set regardless of datasets it is trained,

Method	cIoU ↑	AUC ↑
Attention[7]$_{\text{CVPR18}}$	66.00	55.80
DMC[10]$_{\text{CVPR19}}$	67.10	56.80
LVS [14]†$_{\text{CVPR21}}$	67.20	56.20
LVS [14]$_{\text{CVPR21}}$	69.90	57.30
HardPos [15]$_{\text{ICASSP22}}$	75.20	59.70
SSPL(w/o PCM) [17]$_{\text{CVPR22}}$	69.90	58.00
SSPL(w/ PCM) [17]$_{\text{CVPR22}}$	75.90	61.00
EZ-VSL(w/o OGL) [18]$_{\text{ECCV22}}$	71.89	58.81
SSL-TIE [19]$_{\text{ACM MM22}}$	81.50	61.10
Ours	**84.74**	**63.08**
EZ-VSL(w/ OGL) [18]$_{\text{ECCV22}}$	83.13	63.06
Ours(w/ OGL)	**85.54**	**64.27**

Table 2. **Quantitative results on the SoundNet-Flickr test set.** All models are trained and tested on the SoundNet-Flickr dataset. † is the result of the model from the official project page.

Test Set	Training Set	Method	cIoU ↑	AUC ↑
SoundNet-Flickr	VGGSound 144k	LVS [14]$_{\text{CVPR21}}$	71.90	58.20
		EZ-VSL(w/o OGL) [18]$_{\text{ECCV22}}$	78.31	61.74
		Ours	**83.94**	**63.20**
		EZ-VSL(w/ OGL) [18]$_{\text{ECCV22}}$	83.94	63.60
		Ours(w/ OGL)	**85.14**	**64.55**
VGG-SS	SoundNet-Flickr 144k	LVS [14]$_{\text{CVPR21}}$	26.95	34.30
		EZ-VSL(w/o OGL) [18]$_{\text{ECCV22}}$	29.39	35.53
		Ours	**34.45**	**37.35**
		EZ-VSL(w/ OGL) [18]$_{\text{ECCV22}}$	38.62	39.20
		Ours(w/ OGL)	**39.41**	**39.81**

Table 3. **Quantitative results for cross-dataset evaluation.**

4.44% cIoU and 1.46% AUC when trained on VGGSound and 3.24% cIoU and 1.98% AUC when trained on SoundNet-Flickr. However, it achieves slightly lower accuracy compared to [19] on the VGG-SS test when trained on VGGSound. We would like to highlight that existing approaches use additional task-specific strategies such as; SSL-TIE [19] incorporates aggressive augmentations and transformation together with additional geometrical consistency loss and background suppression, SSPL [17] attaches an explicit sub-module called PCM to reduce the effect of background noise, and EZ-VSL [18] refines their initial localization results by using object guidance (OGL) in the inference stage. In contrast, our model only uses a simple approach that extends the training objective with a negative margin and it still gives an on-par or better performance with the existing state-of-the-art methods. As aforementioned, EZ-VSL proposes a refinement of audio-visual sound localization with object-guided localization (OGL). To make a fair comparison with EZ-VSL, we also report the performance of our method with OGL, and the results are shown in the bottom part of the tables. Note that our method gives an on-par performance with EZ-VSL (OGL) even *without using OGL*. We do not use OGL in our architecture in the remainder of this paper unless it is directly compared with EZ-VSL (OGL).

3.3.2. Cross-Dataset Audio-Visual Localization

As expected, the best results are typically obtained when training and testing are done on the same dataset. Here, we present the cross-dataset generalization performance where the datasets used for training and testing are different. Table 3 shows the quantitative results where the model is trained on VGGSound-144K and SoundNet-Flickr-144K, and tested on SoundNet-Flickr and VGG-SS test sets respectively. As the results show, our model has better generalization ability and it outperforms all the other methods in this task.

Test Class	Method	Margin	cIoU ↑	AUC ↑
Heard 110	LVS [14]$_{\text{CVPR21}}$	-	28.90	36.20
	EZ-VSL(w/o OGL) [18]$_{\text{ECCV22}}$	-	31.86	36.19
	Ours	-0.2	**36.35**	**37.92**
	Ours	0.0	34.40	37.38
	Ours	+0.2	34.83	37.50
	EZ-VSL(w/ OGL) [18]$_{\text{ECCV22}}$	-	37.25	38.97
	Ours(w/ OGL)	-0.2	**38.07**	**39.39**
Unheard 110	LVS [14]$_{\text{CVPR21}}$	-	26.30	34.70
	EZ-VSL(w/o OGL) [18]$_{\text{ECCV22}}$	-	32.66	36.72
	Ours	-0.2	**37.90**	**39.17**
	Ours	0.0	36.74	38.39
	Ours	+0.2	36.15	38.44
	EZ-VSL(w/ OGL) [18]$_{\text{ECCV22}}$	-	39.57	39.60
	Ours(w/ OGL)	-0.2	**40.58**	**40.30**

Table 4. **Comparison results on open-set audio-visual localization experiments trained and tested on the splits of [14, 18].**

Method	Margin	cIoU ↑	AUC ↑
LVS [14]$_{\text{CVPR21}}$	0.0	33.99	37.76
LVS-marginNCE	-0.2	**34.80**	**38.17**
LVS-marginNCE	-0.3	**35.73**	**38.52**
EZ-VSL(w/o OGL) [18]$_{\text{ECCV22}}$	0.0	37.57	38.70
EZ-VSL(w/o OGL)-marginNCE	-0.2	**38.70**	**39.26**
Ours	0.0	36.93	38.58
Ours-marginNCE	-0.2	**38.25**	**39.06**

Table 5. **Generalization of the marginNCE on different baselines.** All models are trained with 144K samples from VGG-Sound and tested on VGG-SS.

3.3.3. Open-Set Audio-Visual Localization

Another assessment we can conduct on the generalization ability of our model is to evaluate the model in open-set settings where testing samples come from categories that are not used during self-supervised training. Here, following the train/test splits of previous works [14, 18], the model is trained with randomly selected 110 categories from VGGSound. Then, the evaluation is done on two test sets: 1) **Heard** shares the same categories with the training set, and 2) **Unheard** contains 110 disjoint categories from the training set. These categories are never seen and heard by the model during training.

Results are shown in Table 4. Three phenomena can be observed from this table. First, our method, regardless of the value used for margin, outperforms compared methods in both heard and unheard setups. Second, similar to EZ-VSL, ours also performs better on unheard categories than heard ones. This shows the generalization ability of our method. Third, we see that the negative margin works best among the other margins used in the unheard scenario, which is a real open-set setup. This observation is similar to the findings of [33] that a negative margin is more proper than a positive or zero margin for open-set scenarios. We can also see that while positive margin performance is higher than a zero margin in the heard scenario, it is the opposite in unheard setup which positive margin hurts the discriminability of unseen categories as discussed in [33].

3.3.4. Generalization of the Negative Margin on Different Baselines

To demonstrate that negative marginNCE is generally applicable to the other sound localization methods, we conduct experiments with the most recent methods by extending their loss with a negative margin. All of these baselines use InfoNCE loss in their architectures. We train each baseline with their publicly released codes on VGGSound-144K and test on the VGG-SS dataset for a fair comparison. Results in Table 5 show

Dataset	margin 0.2		margin 0.0		margin -0.1		margin -0.2		margin -0.3		margin -0.4	
	cIoU ↑	AUC ↑	cIoU ↑	AUC ↑	cIoU ↑	AUC ↑	cIoU ↑	AUC ↑	cIoU ↑	AUC ↑	cIoU ↑	AUC ↑
VGG-SS	37.05	38.56	36.93	38.58	37.51	38.89	38.25	39.06	37.65	38.84	36.93	38.47
SoundNet-Flickr	82.73	62.44	82.73	62.64	81.93	62.46	83.94	63.20	83.53	63.70	84.74	63.00

Table 6. **Accuracy w.r.t. Different Margins.** The results show that performance is improved by setting an appropriate negative margin.

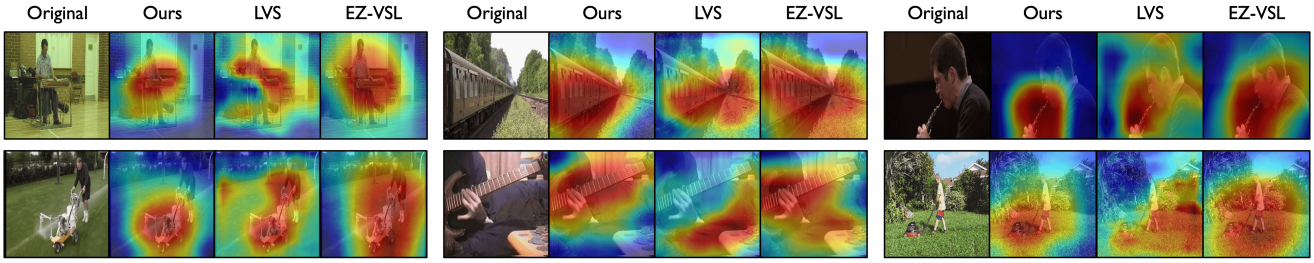

Fig. 1. **Sound localization results on VGG-SS and comparison with the state-of-the-art methods [14, 18].**

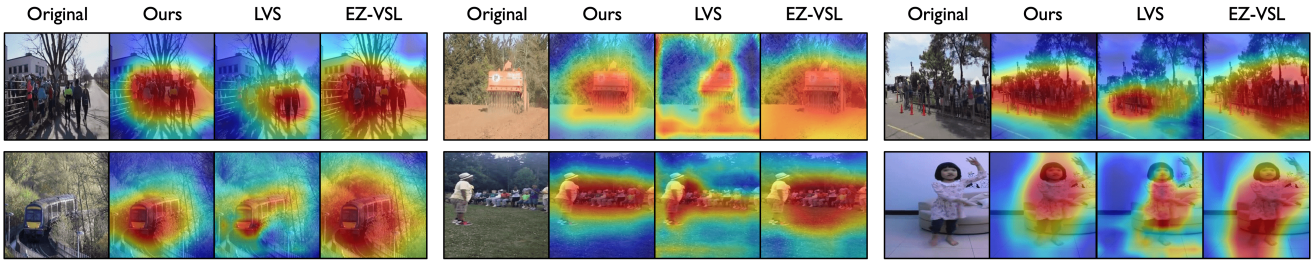

Fig. 2. **Sound localization results on SoundNet-Flickr and comparison with the state-of-the-art methods [14, 18].**

that using a negative margin consistently improves upon all baselines. It even helps EZ-VSL to get the state-of-the-art performance on this experimental setup (compare the results with Table 1).

3.3.5. Comparison of the Different Decision Margins

We conduct an ablation study to explore the accuracy of our method w.r.t. different margin values. We show the performance of our model in Table 6 when it is trained on VGGSound-144K and tested on VGG-SS and SoundNet-Flickr with different margins. As we expect, we get higher cIoU accuracy when the margin is set to negative than zero or positive margins on both test sets. Similarly, AUC performance is also higher with negative margins on both datasets.

3.4. Qualitative Results

In this section, we visualize our sound localization results on VGG-SS and SoundNet-Flickr and compare them with other existing methods [14, 18]. More results, including failure cases, are available at https://sites.google.com/view/marginssl.

VGG-SS: We provide the sound localization results of VGG-SS samples in Figure 1. Our results are more accurate and compact than the other methods. Our qualitative results show that our method handles the co-occurring class-related backgrounds or objects better than the other methods. Green grass background/land or humans often co-occur in the images of "lawn mowers". While our method localizes the lawn mower accurately, the response maps of the other methods contain human and green grass areas as well. As seen in the "man plays a flute" example (first row and third column), our method only focuses on the location of the flute, not on the man. However, EZ-VSL contains the area where the human head exists. A similar trend can be also seen in the example of "man plays a slide guitar" (first row and first column).

SoundNet-Flickr: Our results in Figure 2 depict more accurate localization responses in comparison to the recent methods in the SoundNet-Flickr test set as well. We notice that our method gives more accurate results for the scenes with the "crowd". While LVS results can not cover the entire crowd, EZ-VSL results generally contain a larger area than the crowd itself.

4. CONCLUSION

In this paper, we concentrate on the problem of self-supervised sound source localization with contrastive learning. We identify noisy correspondences due to the assumption of a natural correspondence between audio and visual signals in contrastive learning. With the motivation that looser decision boundaries can alleviate the effect of these noisy correspondences on training, we suggest a simple extension of the contrastive loss function with a negative margin without bells and whistles. Our experiments support our design by showing on-par or state-of-the-art performance on standard benchmarks. We further demonstrate that the proposed negative margin is applicable to any existing approach with contrastive loss and their performances are consistently improved. Therefore, audio-visual sound source localization studies can benefit from our work.

5. ACKNOWLEDGMENTS

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government, [22ZH1200, The research of the basic media·contents technologies].

6. REFERENCES

[1] Bruno Korbar, Du Tran, and Lorenzo Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *NeurIPS*, 2018.

[2] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *Proc. ICCV*, 2019.

[3] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer, "Audiovisual slowfast networks for video recognition," *arXiv preprint arXiv:2001.08740*, 2020.

[4] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani, "Listen to look: Action recognition by previewing audio," in *Proc. CVPR*, 2020.

[5] Soo-Whan Chung, Hong-Goo Kang, and Joon Son Chung, "Seeing voices and hearing voices: Learning discriminative embeddings using cross-modal self-supervision," in *Proc. Interspeech*, 2020.

[6] Weiyao Wang, Du Tran, and Matt Feiszli, "What makes training multi-modal classification networks hard?," in *Proc. CVPR*, 2020.

[7] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon, "Learning to localize sound source in visual scenes," in *Proc. CVPR*, 2018.

[8] Relja Arandjelović and Andrew Zisserman, "Objects that sound," in *Proc. ECCV*, 2018.

[9] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon, "Learning to localize sound sources in visual scenes: Analysis and applications," in *IEEE TPAMI*, 2021.

[10] Di Hu, Feiping Nie, and Xuelong Li, "Deep multimodal clustering for unsupervised audiovisual learning," in *Proc. CVPR*, 2019.

[11] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin, "Multiple sound sources localization from coarse to fine," in *Proc. ECCV*, 2020.

[12] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman, "Self-supervised learning of audio-visual objects from video," in *Proc. ECCV*, 2020.

[13] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou, "Discriminative sounding objects localization via self-supervised audiovisual matching," in *NeurIPS*, 2020.

[14] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman, "Localizing visual sounds the hard way," in *Proc. CVPR*, 2021.

[15] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon, "Learning sound localization better from semantically similar samples," in *Proc. ICASSP*, 2022.

[16] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon, "Less can be more: Sound source localization with a classification model," in *Proc. WACV*, 2022.

[17] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang, "Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes," in *Proc. CVPR*, 2022.

[18] Shentong Mo and Pedro Morgado, "Localizing visual sounds the easy way," in *Proc. ECCV*, 2022.

[19] Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang, "Exploiting transformation invariance and equivariance for self-supervised sound localisation," in *in Proc. ACMM*, 2022.

[20] Shentong Mo and Pedro Morgado, "A closer look at weakly-supervised audio-visual source localization," in *NeurIPS*, 2022.

[21] Pedro Morgado, Nuno Vasconcelos, Timothy Langlois, and Oliver Wang, "Self-supervised generation of spatial audio for 360° video," in *NeurIPS*, 2018.

[22] Karren Yang, Bryan Russell, and Justin Salamon, "Telling left from right: Learning spatial correspondence of sight and sound," in *Proc. CVPR*, 2020.

[23] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," in *ACM Transactions on Graphics*, 2018.

[24] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba, "The sound of motions," in *Proc. ICCV*, 2019.

[25] Ruohan Gao and Kristen Grauman, "Co-separating sounds of visual objects," in *Proc. ICCV*, 2019.

[26] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel P. W. Ellis, and John R. Hershey, "Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds," in *Proc. ICLR*, 2021.

[27] Andrew Owens and Alexei A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. ECCV*, 2018.

[28] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos, "Robust audio-visual instance discrimination," in *Proc. CVPR*, 2021.

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[30] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy, "Delving into inter-image invariance for unsupervised visual representations," in *IJCV*, 2022.

[31] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "VGGSound: A large-scale audio-visual dataset," in *Proc. ICASSP*, 2020.

[32] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "Soundnet: Learning sound representations from unlabeled video," in *NeurIPS*, 2016.

[33] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu, "Negative margin matters: Understanding margin in few-shot classification," in *Proc. ECCV*, 2020.