Appendix

Brief Introduction. The appendix is structured into four main sections: Algorithm, Experimental Settings, Supplementary Experiments, and Further Analysis. The main contents are as follows:

- A. **Algorithm:** Pseudo-codes and algorithm details.
- B. **Experimental Settings:** More detailed description of the datasets, backbones, metrics formula, implementation details and baselines.
- C. **Supplementary Experiments:** Results of cross-dataset evaluation, comparison of different predicted map, exploration of Warm-Up speed and its influence on the final results, false-positive rejection capability of Noise Filtering, investigation of hyperparameters for Filter, IPL, and EMA, effect of data augmentation and quality analysis (visualization).
- D. **Further Analysis:** Theoretical elaboration on the challenges faced by existing contrastive learning methods, and explanation of why contrastive learning alone cannot achieve precise localization.

A Algorithm

To make it more clear, Dual Mean-Teacher is specifically depicted in Algorithm 1.

Algorithm 1 Dual Mean-Teacher algorithm.

```
Input: $\mathcal{D}_u = \{(a_i, v_i)\}, \mathcal{D}_l = \{(v_i, a_i), \mathcal{G}_i\}$ {labeled data and unlabeled data.}
while not reach the maximum iteration do
    for $(a_i, v_i)$ in $\mathcal{D}_u$ do
        while not reach the convergency of Warm-Up do
            $\mathcal{L}_{\text{Warm-Up}} = \mathbb{E}_{(a_i,v_i)\sim\mathcal{D}_l} H(\mathcal{G}_i, \mathcal{P}_i^t)$ {Supervised learning on labeled data.}
        end while
        Get the pseudo-labels $\mathcal{M}_i^{t,A}, \mathcal{M}_i^{t,B}$ from dual teachers
        if $\mathrm{IoU}(\mathcal{M}_i^{t,A}, \mathcal{M}_i^{t,B}) \geq \tau$ then
            $\mathcal{IPL}(a_i, v_i) = \mathcal{M}_i^{t,A} \cdot \mathcal{M}_i^{t,B}$ {Compute Intersection of Pseudo-Labels (IPL).}
            $\hat{\mathcal{G}}_i = \mathcal{IPL}(a_i, v_i)$ {Update the pseudo-label $\hat{\mathcal{G}}_i$ of unlabeled data.}
            Add $(a_i, v_i)$ to new dataset $\mathcal{D}'_u$
        end if
    end for
    $\mathcal{D}_{mix} = \mathcal{D}_l \cup \mathcal{D}'_u$ {Mix the filtered unlabeled data and labeled data.}
    $\mathcal{L}_{\text{full}} = \left(\mathcal{L}_{\text{sup}}^A + \mathcal{L}_{\text{sup}}^B\right) + \lambda_u \left(\mathcal{L}_{\text{unsup}}^A + \mathcal{L}_{\text{unsup}}^B\right)$. {Students learning.}
    $\theta_m^t \leftarrow \beta\theta_{m-1}^t + (1-\beta)\theta_m^s$ {Students update teachers via EMA.}
end while
Return: Dual teachers and students model parameters.
```

NOTING TIPS:

Train. Warm-Up Stage is essentially a supervised learning. The performance gains of subsequent Unbiased-Learning Stage over Warm-Up Stage is actually the performance gains of our semi-supervised framework over vanilla supervised training on the same labelled dataset $\mathcal{D}_l$, which proves the validity of the proposed Dual Mean-Teacher, as shown in the main results in Table 1 and Table 2.

Inference. For the localization result of i_{th} audio-visual pair, we merge the outputs of the dual teachers to create a predicted map as below. Comparison of different predicted maps are described in C.2.

$$\mathcal{P}_i = \frac{1}{2}(\mathcal{P}_i^{t,A} + \mathcal{P}_i^{t,B}). \tag{1}$$

B Experimental Settings

B.1 Datasets

We have conducted our training and evaluation of the Progressing Teacher on two large-scale audio-visual datasets: Flickr-SoundNet and VGG-Sound, which consist of millions of unconstrained videos and $5,000$ and $5,158$ annotated samples, respectively. Each audio-visual pair is comprised of a single image frame from each video clip and an audio segment centered around it. The annotations are provided in the form of bounding boxes. The relevant information is presented in the Table 1.

Table 1: Datasets overview.

	All Labeled Data					Test Set					Labeled Split			
	small	medium	large	huge	total	small	medium	large	huge	total	train	val	test	total
Flickr-SoundNet	3	254	687	4056	5000	0	9	83	158	250	4250	500	250	5000
VGG-SoundSource	134	1796	1726	1502	5158	8	86	83	73	250	4250	500	250	5000

Furthermore, for the purpose of assessing the generalizability of our model, we have extended DMT to music domain (distribution), including: MUSIC-solo, MUSIC-duet, and MUSIC-Synthetic. The MUSIC dataset [1] comprises 685 untrimmed videos, encompassing 536 solo performances and 149 duet renditions, spanning across 11 distinct categories of musical instruments. The MUSIC-Synthetic [2, 3] is a multifaceted assemblage wherein four disparate solo audio-visual pairs of divergent classifications are randomly mixed, retaining solely two out of the four audio segments. This deliberate curation aligns aptly with the evaluation of discerningly sounding object localization.

B.2 Backbones: VGGish and SoundNet

For audio backbones, we employ pre-trained VGGish and SoundNet. VGGish is pre-trained on AudioSet as audio feature extractors. The raw 3s audio signal is resampled at 16kHz and further transformed into 96 × 64 log-mel spectrograms as the audio input. The output is 128D vector. SoundNet takes the raw waveform of the 3s audio clip as input and produces a 1401D vector as output, which concatenates the 1000D object-level feature and the 401D scene-level feature, which are both obtained from different conv8 layer. Our main focus is to train the nonlinear audio feature transformation function, g(·), which is instantiated with two fully connected networks and a ReLU layer, to transform the network output feature into a 512D representation.

B.3 Metrics: CIoU, MSE, F1 Score, Precision

We consider a set of audio-visual pairs as $\mathcal{D} = \{(v_i, a_i), \mathcal{G}_i\}$, where $\mathcal{G}_i$ is the ground-truth. We set $\mathcal{P}_i(\delta) = \{(x,y)|\mathcal{P}_i(x,y) > \delta\}$ is the foreground region of predicted map, and $\mathcal{G}_i(x,y) = \{(x,y)|\mathcal{G}_i(x,y) > 0\}$ is the foreground region of ground truth.

CIoU. The IoU of predicted map and ground truth can be calculated by:

$$IoU_i(\delta) = \frac{\sum_{x,y\in\mathcal{P}_i(\delta)} \mathcal{G}_i(x,y)}{\sum_{x,y\in\mathcal{P}_i(\delta)} \mathcal{G}_i(x,y) + \sum_{x,y\in\{\mathcal{P}_i(\delta)-\mathcal{G}_i\}} 1}. \tag{2}$$

In previous works, CIoU quantifies the proportion of samples with IoU value exceeding a predetermined threshold, typically set at 0.5.

MSE. MSE measures the difference between two maps on a pixel-wise basis, making it more suitable for evaluating dense prediction tasks than IoU. Other two metrics for small objects localization.

$$MSE_i = \frac{1}{HW}\sum_{x=1}^{W}\sum_{y=1}^{H}\left(\mathcal{P}_i(x,y) - \mathcal{G}_i(x,y)\right)^2. \tag{3}$$

Max-F1 and AP. To compute true positives, false positives and false negatives, we closely follow SLAVC [4]. Then we can compute the precision and recall:

$$\text{Precision} = \frac{|\mathcal{TP}|}{|\mathcal{TP}| + |\mathcal{FP}|}, \qquad \text{Recall} = \frac{|\mathcal{TP}|}{|\mathcal{TP}| + |\mathcal{FN}|}. \tag{4}$$

Then we compute F1 for all values of δ and report the Max-F1 score:

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}, \qquad \text{max-F1} = \max(\text{F1}). \tag{5}$$

Average Precision (AP) is the area under the precision-recall curve above. For a detailed calculation of max-F1 and AP, please refer to the SLAVC [4].

B.4 Implementation details

In addition to the experimental settings mentioned in the main text, we used a batch size of 128. Warm-Up stage is trained for 6 epochs to achieve convergence, while the Unbiased-Learning stage is trained for 20 epochs. The learning rate for the image is set to 1e-4, and the weight for the contrastive loss λ_u is set to 1. An Exponential Moving Average (EMA) decay of 0.999 is applied. The Adam optimizer is used for training, and the training is conducted on two GPUs. Our supplementary experiments were conducted on the Flickr-10k or Flickr-144k dataset, which contains 4k annotations. The trained models were evaluated on the Flickr-SoundNet testset.

B.5 Baselines

- Attention 10k [5, 6] (CVPR$_{2018}$): introduce a dual-stream network and leverage an attention mechanism to capture the salient regions in semi-supervised or self-supervised environments.
- DMC [7] (CVPR$_{2019}$) : establish audio-visual clustering to associate sound centers with their corresponding visual sources.
- CoarsetoFine [8] (ECCV$_{2020}$) : leveraged a two-stage framework to capture cross-modal feature alignment between sound and vision.
- LVS [9] (CVPR$_{2021}$) : propose to mine hard negatives within an image-audio pair.

Table 2: Cross dataset performance. We train our model using the VGG-Sound 10k and 144k datasets and evaluate its performance on the Flickr-SoundNet dataset.

Trainset	Methods	Flickr testset	
		CIoU	AUC
VGG-Sound 10k	attention10k	52.20	50.20
	LVS	61.80	53.60
	EZVSL	65.46	54.57
	SLAVC	74.00	57.74
	SSPL	76.30	59.10
	SSL-TIE	77.04	60.36
	Ours($|\mathcal{D}_l| = 256$)	85.04 (80.08)	65.06 (60.14)
	Ours($|\mathcal{D}_l| = 2k$)	87.36 (81.60)	67.38 (61.26)
	Ours($|\mathcal{D}_l| = 4k$)	**88.20 (82.88)**	**67.56 (62.06)**
VGG-Sound 144k	attention10k	66.00	55.80
	LVS	71.90	58.20
	EZVSL	79.51	61.17
	SLAVC	80.00	61.68
	SSPL	76.70	60.50
	SSL-TIE	79.50	61.20
	Ours($|\mathcal{D}_l| = 256$)	87.04 (80.08)	64.72 (60.14)
	Ours($|\mathcal{D}_l| = 2k$)	88.32 (81.60)	67.78 (61.26)
	Ours($|\mathcal{D}_l| = 4k$)	**89.84 (82.88)**	**68.64 (62.06)**

- EZVSL [10] (ECCV$_{2022}$) : introduce a multi-instance contrastive learning framework that utilizes Global Max Pooling (GMP) to focus only on the most aligned regions when matching audio and visual inputs.
- SLAVC [4] (NeurIPS$_{2022}$) : adopts momentum encoders and dropout to address overfitting and silence issues in single-source sound localization.
- SSPL [11] (CVPR$_{2022}$) : propose a negative-free method to extend a self-supervised learning framework to the audio-visual data domain for sound localization
- SSL-TIE [12] (ACM-MM$_{2022}$): introduce a self-supervised framework with a Siamese network with contrastive learning and geometrical consistency.

C Comprehensive Experimental Results

C.1 Cross-dataset Evaluation

To further validate the generalization ability of DMT, we conducted cross-dataset validation experiments. The results in Table 2 show that DMT still stays ahead, confirming the high generalization ability of our model.

C.2 Different Predicted Map

In this section, we compare the accuracy of different predicted maps for sound localization. We evaluate individual predicted maps and a fused map as the final localization map, as defined by Eq. 1. Training is performed on the Flickr144k dataset using dual teacher results, as shown in Table 3. We find that fused predicted map from dual teachers with different backbones achieves better localization performance than from individual maps, which can be attributed to the fact that considering both localization results helps mitigate biases inherent in a single model.

Additionally, we assess the performance of teachers and students by comparing their fused predicted maps obtained during the same training session. The results, as shown in Table 3, indicate that **teachers outperform students**, which aligns with our expectations and further validates the effectiveness of our model.

Table 3: Results of different inference strategies.

	CIoU	AUC
Student A	86.20	66.16
Student B	86.80	66.84
Fused Students	88.60	68.56
Teacher A	87.20	67.57
Teacher B	87.60	67.98
Fused Teachers	**90.40**	**69.36**

C.3 Effect of Warm-Up Stage

This section focuses on the analysis of convergence speed and the influence of the Warm-Up performance on the final results.

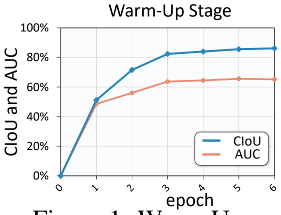


Figure 1: Warm-Up.

Convergence Speed Initially, we investigate the convergence speed of the Warm-Up stage with varying amounts of labeled data, as depicted in Figure 1. Notably, all supervised models exhibit rapid convergence within a specific number of epochs. Furthermore, as the quantity of data increased, the convergence speed decreases while simultaneously achieving higher levels of performance.

Effect of Warm-Up Performance. Subsequently, we investigate how the Warm-Up performance affects final results by experimenting with models that achieved different levels of convergence using the same amount of data. Training is performed on the Flickr144k dataset using dual teacher results, as presented in Table 4. The results indicate that better performance of Warm-Up stage leads to better final model performance, which can be attributed to higher-quality pseudo-labels and improved noise filtering, reducing confirmation bias. Conversely, the model exhibits the poorest performance in the absence of Warm-Up stage.

Table 4: Effect of Warm-Up Performance.

Warm-Up		Final	
CIoU	AUC	CIoU	AUC
0	0	84.32	64.52
51.20	48.62	87.28	67.18
71.60	56.08	89.04	68.26
86.20	**65.56**	**90.40**	**69.36**

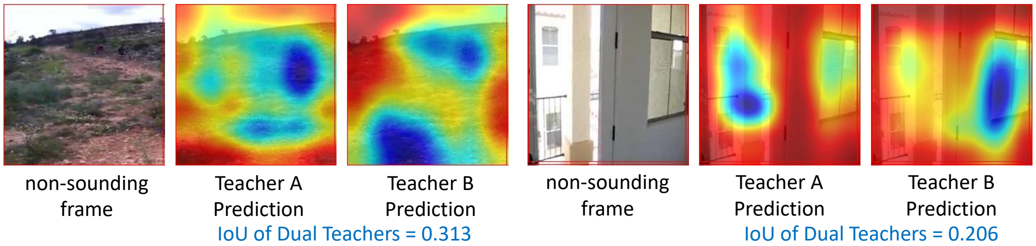


Figure 2: False-Positive Rejection Capability of Noise Filtering.

Overall, supervised audio-visual source localization demonstrates ease of convergence without requiring excessive training resources. Moreover, our proposed semi-supervised model consistently outperforms the supervised model by approximately 3% in terms of absolute performance, validating its effectiveness.

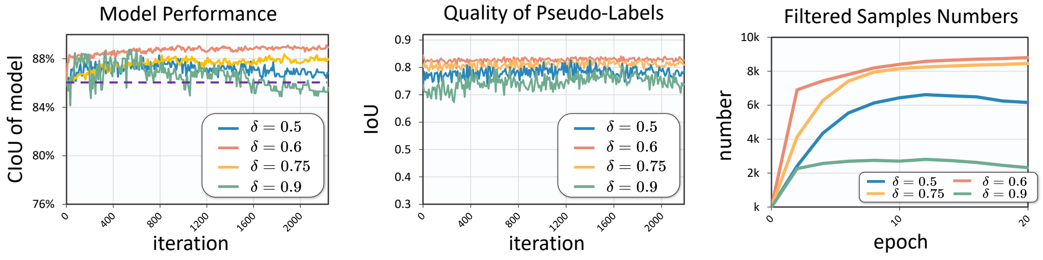
C.4 False-Positive Rejection Capability of Noise Filtering

After analyzing the filtered-out samples, we observed that the two independent teachers exhibit disagreement in localizing non-sounding objects. In such cases, the IoU falls significantly below the threshold, enabling the Dual Teachers to identify and reject non-sounding samples, which can be considered as false positives, as illustrated in Figure 2. Additionally, different filter thresholds represents different levels of filtering strictness, as detailed in Section C.6.

Furthermore, we analyzed the visual results of some noisy samples, as depicted in Figure 5. One can observe that frames without distinguishable sound objects or sounds that cannot be accurately represented by a bounding box (*e.g.*, wind sounds) can be easily identified through the inconsistency between the predictions of the two teachers.

C.5 Hyper-parameters for Filter, IPL, and EMA

Effect of Pseudo-Labeling Threshold. The threshold δ is used to convert the predicted map into a binary map, as described in Eq.(6). In this section, we analyze the impact of different thresholds on pseudo-labels and the model. Training is conducted on the Flickr10k dataset. Figure 3 shows the results. A small delta value (*e.g.* $\delta = 0.5$) creates a large foreground area, introducing excessive noise and causing performance degradation as training progresses. On the other hand, A large value of δ (*e.g.* $\delta = 0.9$) indicates a small foreground area, causing the intersection between Dual Teachers to be minimal and resulting in samples being falsely rejected as noise, thus disturbing the model. Therefore, we choose $\delta = 0.6$ as the optimal threshold for our final selection.



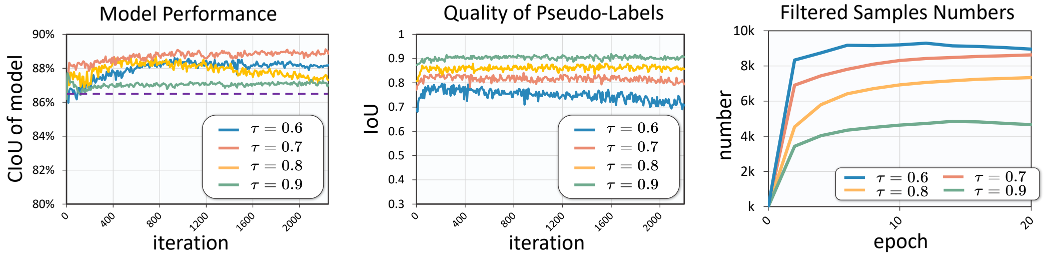
(a) Model performance.

(b) Quality of pseudo-labels.

(c) Number of filtered samples.

Figure 3: Results of various δ.

Effect of Filtering Threshold. In Section 4.2, we employ a confidence threshold, denoted as τ, to filter out noisy samples, which are more likely to be false-positive instances. We evaluate the effect of different threshold values τ. As shown in Figure 4, As the threshold value τ increases from 0 to 0.9, the number of accepted samples decreases. However, setting a very high threshold (e.g., $\tau = 0.9$) leads to unsatisfactory results due to the limited number of accepted samples, reducing the available information from unlabeled data. Conversely, using a low threshold (e.g., $\tau = 0.6$) introduces a confirmation bias from noisy samples, hindering favorable outcomes. Upon analysis, we discover that the performance shows little variation between threshold values of $\tau = 0.7$ and $\tau = 0.8$, indicating a balance between unlabeled information and bias within the 0.7-0.8 range. As a result, we opt for $\tau = 0.7$ as the preferred threshold for our final selection.



(a) Model performance. (b) Quality of pseudo-labels. (c) Number of filtered samples.

Figure 4: Results of various τ.

Effect of EMA Rates We also examine the model performance with various exponential moving average (EMA) decay values, denoted as β, ranging from 0.9 to 0.999, and present the results of the teachers in Table 5. We observe that a smaller EMA decay leads to a faster update rate, lower CIoU, and higher variance. Conversely, a larger EMA decay value results in slower learning for the teachers. Therefore, we select an appropriate EMA decay value of $\beta = 0.999$ to strike a balance between the update rate and the stability of the learning process.

Table 5: Results on various EMA β.

	β	CIoU	AUC
Flickr 10k	0.9	86.48	65.16
	0.99	88.64	66.94
	0.999	**88.80**	**67.81**
Flickr 144k	0.9	87.84	85.82
	0.99	89.92	68.86
	0.999	**90.40**	**69.36**

C.6 Effect of Data Augmentation

We evaluate the effect of RandAug [13] on a supervised model on 4k labeled data, as shown in Table 6. Without data augmentation, the model exhibits significant over-fitting. With RandAug, this issue is mitigated, which indicates that RandAug serves not only as a means of consistency regularization but also as a method to enhance the model's generalization performance.

Table 6: Results of data augmentation (*i.e.*, RandAug.).

	Trainset		Testset	
	CIoU	AUC	CIoU	AUC
w/o RandAugment	88.20	67.82	84.80	60.44
w/ RandAugment	87.68	67.54	86.20	65.56

C.7 IPL on Different Object Size

We assess the adaptability of IPL to various object sizes, and compare with existing methods, two teachers with DMT. Table 7 results highlight prior methods' diminishing performance with smaller objects, while DMT consistently excels across all size subsets. This enhancement is attributed to Filtering and IPL synergy. Under the filtering mechanism, only highly similar pseudo-labels can contribute to model training. This keeps the intersection of pseudo-labels consistently aligned with object sizes. If pseudo-labels decrease significantly, IoU declines, excluding noisy samples from training. Moreover, in the second-stage training, we use labeled data to prevent size bias and ensure unbiased treatment of objects of all sizes.

Table 7: Performance across various sizes of sounding objects.

Size	SLAVC		teacher1		teacher2		DMT	
	MSE ↓	IoU ↑	MSE ↓	IoU ↑	MSE ↓	IoU ↑	MSE ↓	IoU ↑
small	0.705	2.10	0.213	2.58	0.183	2.26	**0.205**	**2.65**
medium	0.235	22.00	0.156	12.47	0.176	12.28	**0.164**	**33.50**
large	0.427	48.11	0.202	55.32	0.221	54.68	**0.212**	**55.50**
huge	0.358	61.64	0.212	66.84	0.217	66.26	**0.215**	**67.70**

C.8 How to avoid model collapse?

There are diversity and individuality between two teachers, as in Q2, which helps to prevent two teachers convergence to one model. The noisy filter module of DMT selects 'stable samples' via consensus and assigns high-quality pseudo-labels with IPL, such spirit has been validated by prior work that 'stable samples' could help avoid model collapse. Two teachers are first trained in Warm-Up stage for better initialization. Moreover, in stage-2, we also include supervised training on labeled data and contrastive learning on unlabeled data, the two objectives would ensure the model possesses robust localization capabilities over the course of stage-2. The results in Table 8 validate each component to avoid model collapse.

Table 8: Model collapse results. $\mathcal{A}$, $\mathcal{B}$ denotes augmentation and backbone.

method	DMT	same $\mathcal{A}$	same $\mathcal{B}$	w/o annotation in stage-2	same $\mathcal{A}$ & $\mathcal{B}$ w/o annotation
CIoU	**90.4**	87.2	85.4	81.6	7.2

C.9 Quality Analysis

We present the visual localization results of DMT in Figure 5. It effectively locates objects of different sizes, distinguishes them from the background by clear boundaries, and demonstrates some multi-object localization capability. Notably, DMT learns semantic information and can precisely localize specific sound-producing regions instead of the entire object. For example, in the third row of the Figure 5 on the right, it accurately locates the mouth of a person rather than the entire person.

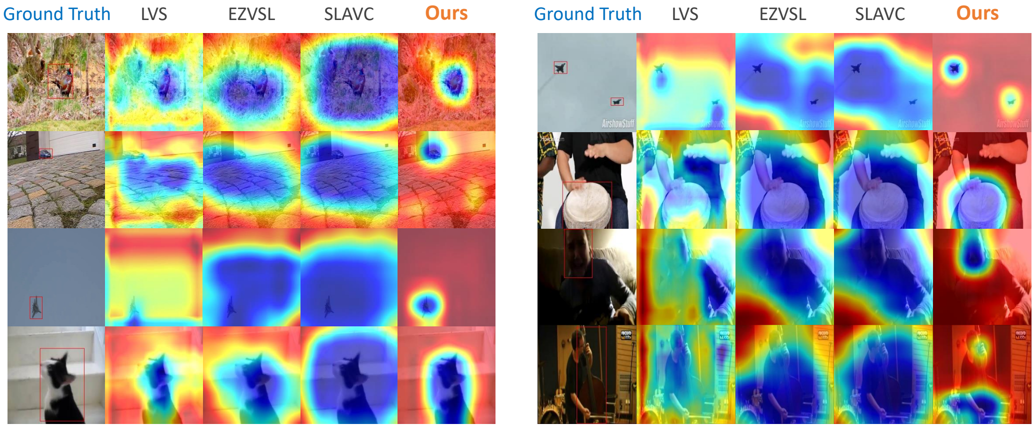


Figure 5: Visualizations of various methods.

D Further Analysis: Limitations in Existing AVSL and DMT

Based on the formula of contrastive loss, we can observe that the core idea of existing contrastive learning methods is to match the visual frames and corresponding audio clips within the same video as a whole. The audio-visual pairs from the same video are considered positive pairs, while the frames and audio clips from different videos are considered negative pairs. The contrastive loss aims to maximize the similarity between positive samples and minimize the similarity between negative samples. The differences among existing self-supervised methods lie in the selection of the similarity function $s(\cdot)$ and the positive-negative sample pairs.

$$\mathcal{L}_{\text{unsup}} = -\mathbb{E}_{(a_i,v_i)\sim\mathcal{D}_u}\left[\log\frac{\exp(s(g(a_i),f(v_i))/\tau_t)}{\sum_{j=1}^{n}\exp\left(s\left(g(a_i),f(v_j)\right)/\tau_t\right)}+\log\frac{\exp(s(f(v_i),g(a_i))/\tau_t)}{\sum_{j=1}^{n}\exp\left(s\left(f(v_i),g(a_j)\right)/\tau_t\right)}\right].$$

D.1 Global and Local Information

In the given formula, different methods employ different match functions $s(\cdot)$ to compute the distance or similarity between positive samples. For instance, Attention10k [5, 6] uses the Euclidean distance, LVS [9] utilizes the Frobenius inner product, and EZVSL [10] applies Global Max Pooling:

$$
\begin{aligned}
\text{Attention10k:}\quad s(\cdot) &= \left\| f_{att}(v_i) - g(a_i) \right\|_2 , \\
\text{LVS:}\quad s(\cdot) &= \frac{1}{|\hat{m}_{ip}|} \left\langle \hat{m}_{ip}, \mathrm{sim}\left(f(v_i), g(a_i)\right) \right\rangle , \\
\text{EZVSL:}\quad s(\cdot) &= \max \mathrm{sim}\left(f(v_i), g(a_i)\right) , \\
\text{SLAVC:}\quad s(\cdot) &= \sum_{x,y} \rho\left(\frac{1}{\tau}\mathrm{sim}\left(g^{\mathrm{loc}}\left(a_i\right), f^{\mathrm{loc}}\left(v_i\right)\right)\right) \cdot \rho\left(\frac{1}{\tau}\mathrm{sim}\left(g^{\mathrm{avc}}\left(a_i\right), f^{\mathrm{avc}}\left(v_i\right)\right)\right) .
\end{aligned}
$$

All of these functions capture the overall matching degree between audio and global visual representations. However, after the computation of $s(\cdot)$, the model loses the positional information of the two-dimensional visual representation. This positional information is crucial for fine-grained localization tasks.

D.2 Position-Aware Contrastive Loss

We refer to the methods that incorporate position information as 'position-aware'. In the above formulas, we can observe that the distances or similarities between samples are calculated in a position-aware manner. For example, in the Attention10k [5, 6] method, the attention mechanism f_{att} takes into account the positional information. Similarly, in LVS [9], the foreground mask $\hat{m}_{ip}$ distinguishes the background as hard negatives, incorporating the positional context. EZVSL [10] uses the maximum value to capture the positional information, while SLAVC [4] incorporates localization information. Taking LVS [9] as an example, it specifically treats the background of the image as hard negatives, effectively leveraging the positional cues for discrimination and learning.

$$
\begin{aligned}
P_i &= \frac{1}{|\hat{m}_{ip}|} \left\langle \hat{m}_{ip}, \mathrm{sim}(g(a_i), f(v_i)) \right\rangle , \\
N_i &= \frac{1}{|\mathbf{1} - \hat{m}_{in}|} \left\langle \mathbf{1} - \hat{m}_{in}, \mathrm{sim}(g(a_i), f(v_i)) \right\rangle + \frac{1}{hw} \sum_{j \neq i} \left\langle \mathbf{1}, \mathrm{sim}(g(a_i), f(v_j)) \right\rangle , \\
\mathcal{L}_{unsup} &= -\frac{1}{k} \sum_{i=1}^{k} \left[\log \frac{\exp\left(P_i\right)}{\exp\left(P_i\right) + \exp\left(N_i\right)} \right] .
\end{aligned}
$$

where, $\hat{m}_{ip}$ is the mask of foreground, which strongly relies on the initialization of the model. According to the formula, both the positive (P_i) and negative (N_i) samples in the training process are influenced by the initial values of the foreground mask $\hat{m}_{ip}$. This implies that the model's localization results are heavily dependent on the initialization.

D.3 Initialization

The different matching mechanisms, represented by the function $s(\cdot)$, rely on the initialization of the entire visual model, specifically the pre-trained ResNet-18 [14, 15], where the average of the pixel-wise features is taken as the initial result at epoch 0. This initialization result serves as the basis for the computation of position-aware components, such as the attention mechanism or Global Max Pooling (GMP). Subsequently, during the model's training, these initial localization results are reinforced and refined. However, if the initial localization results are inaccurate (which is often the case), subsequent training may have difficulty detecting and correcting these inaccuracies. As a result, the errors may accumulate over time without being effectively addressed, leading to degraded performance.

D.4 False Positives, False Negetives and Multi-Source

From the contrastive learning formula, it is apparent that contrastive learning assumes the presence of sound-producing objects in the visual input and enforces alignment between highly confident visual regions and their corresponding audio features. However, pure contrastive learning, without the incorporation of additional modules, cannot directly reject non-sounding samples. Recently, some works have recognized this limitation and started to investigate the presence of sound-producing objects in images and tackle the task of multi-source sound localization. Examples of such works include DSOL [2], IER [16], and AVGN [17].

Furthermore, due to the absence of class labels during the selection of positive and negative samples, visual-audio pairs belonging to the same sound-producing object class but originating from different videos are still treated as negative samples, resulting in a false negatives issue. Several methods have emerged to address this problem, as highlighted in [18, 19].

In addition, the commonly used matching mechanism, Global Max Pooling, is suitable only for single-source localization since it focuses solely on the region with the highest confidence, neglecting other potential sound-producing objects.

These three aforementioned challenges cannot be effectively resolved solely through simple models or algorithms without positional annotations. Therefore, they have become prominent research areas that are currently receiving considerable attention.

D.5 Limitations of DMT

DMT does not involve class information, so it struggles to localize among fine-grained objects due to poor discriminative ability. By incorporating category signals, models could better implement fine localization. Besides, DMT could not handle multi-object localization well. We will devise specialized components to address this issue.

References

[1] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018.

[2] 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. *Advances in Neural Information Processing Systems*, 33:10077–10087, 2020.

[3] Di Hu, Yake Wei, Rui Qian, Weiyao Lin, Ruihua Song, and Ji-Rong Wen. Class-aware sounding objects localization via audiovisual correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9844–9859, 2021.

[4] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. In *Advances in Neural Information Processing Systems*, 2022.

[5] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018.

[6] 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. *TPAMI*, 43(5):1605–1619, 2019.

[7] Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 18:351–376, 2021.

[8] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 292–308. Springer, 2020.

[9] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021.

[10] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. *arXiv preprint arXiv:2203.09324*, 2022.

[11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3222–3231, 2022.

[12] Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3742–3753, 2022.

[13] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[16] Xian Liu, Rui Qian, Hang Zhou, Di Hu, Weiyao Lin, Ziwei Liu, Bolei Zhou, and Xiaowei Zhou. Visual sound localization in the wild by cross-modal interference erasing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1801–1809, 2022.

[17] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[18] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12945, 2021.

[19] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. 2020.