T-VSL: Text-Guided Visual Sound Source Localization in Mixtures

Supplementary Material

In the supplementary material, we present additional experimental analysis, which are summarized as follows:

- We highlight major differences between proposed T-VSL and concurrent CLIP-SSL [34] (Sec. A).
- We study one-stage training method in T-VSL in comparison with the proposed two-stage method (Sec. B).
- We present additional qualitative comparisons with SOTA methods in multi-source localization (Sec. C).

A. Significant Differences Between Concurrent CLIP-SSL and T-VSL

In comparison with the concurrent CLIP-based sound source localization framework CLIP-SSL [34], our proposed T-VSL has several significant differences for multi-source localization from mixtures. We highlight the major differences as follows:

1) **Use of Self-supervised Pre-trained Encoders.** CLIP-SSL used an off-the-shelf pre-trained mask-generator that relies on large-scale densely supervised pre-training on image segmentation datasets. However, sound source localization often demands complex spatio-temporal reasoning across audio and vision modalities, which is absent in image segmentation task. In contrast, we directly use self-supervised pre-trained AudioCLIP [17] model as our backbone, and introduced an weakly-supervised sound source localization framework. Therefore, our method inherently learns audio-visual correspondence for sound source localization without being limited by large-scale image segmentation supervised pre-training constraints as CLIP-SSL.

2) **Text Guidance as Weak Supervision to Noisy Audio and Vision.** The primary focus of CLIP-SSL is to replace the text query encoder of the supervised baseline with an audio encoder. However, environmental audio contains significant noises from background sources in contrast to cleaner text modality. In addition, presence of silent objects in visual scenes make the audio-visual correspondence more challening. Instead of replacing one modality (text with audio), the proposed T-VSL introduces a joint learning across audio, vision, and text modality utilizing the grounded tri-modal embedding space of AudioCLIP. Our approach particularly leverages the text modality as weak supervision to learn audio-visual correspondence in noisy mixtures, that effectively exploits all three modalities.

3) **Disentanglement of Multi-Source Mixtures.** The proposed T-VSL attempts to solve multi-source localization problem in an weakly supervised manner *without having access to single-source audio-visual cues*. Since both audio and visual modality contain noises from background sources, it is particularly challenging to learn their correspondence in multi-source scenarios. By leveraging the text representation of single-source sounds, we introduce multi-source audio-visual feature disentanglement for enhanced localization performance. In contrast, CLIP-SSL focuses on directly learning audio-visual correspondence following existing single-source baselines without explicitly tackling the multi-source localization problem. Such an approach often struggles in learning audio-visual correspondence in challenging multi-source mixtures, when single source sounds and corresponding visual objects are not available.

B. Comparison between one-stage and two-stage architectures in T-VSL

We introduce a two-stage architecture in the proposed T-VSL, that consists of audio-visual class instance detection followed by iterative localization of each sounding source present in the multi-source mixture. In general, this two-stage approach simplifies the problem of challenging multi-source localization by initially detecting common audio-visual instances present in both audio and visual modality. However, we also study an one-stage alternative of the proposed T-VSL following AVGN [30], by removing the audio-visual class instance detection stage, and by incorporating all N-class text embedding of sounding sources in both audio and visual conditioning blocks, without explicitly separating $K(K \leq N)$ class instances. We present the quantitative comparisons on VGGSound-Single and VGGSound-Duet datasets in Table 7. We note that the two-stage method in T-VSL achieves $+2.2$ IoU@0.5% and $+1.4$ CIoU@0.3% improvements on VGGSound-Single and VGGSound-Duet datasets, respectively, compared to its single-stage counterpart. We hypothesize that the proposed two-stage method greatly reduces the effect of background noises in challenging audio-visual correspondence learning from natural mixtures. In contrast, single-stage method introduces additional noises from background sources in audio-visual conditioning blocks for not explicitly suppressing background conditions.

C. Qualitative Comparisons

We present additional qualitative comparisons of the proposed T-VSL with SOTA methods on challenging multi-source localization in Figure 4. We note that the proposed T-VSL consistently achieves superior performance by generating more precise localization maps compared to other SOTA baselines, which follows our prior observation. In addition, T-VSL can selectively isolate the non-sounding

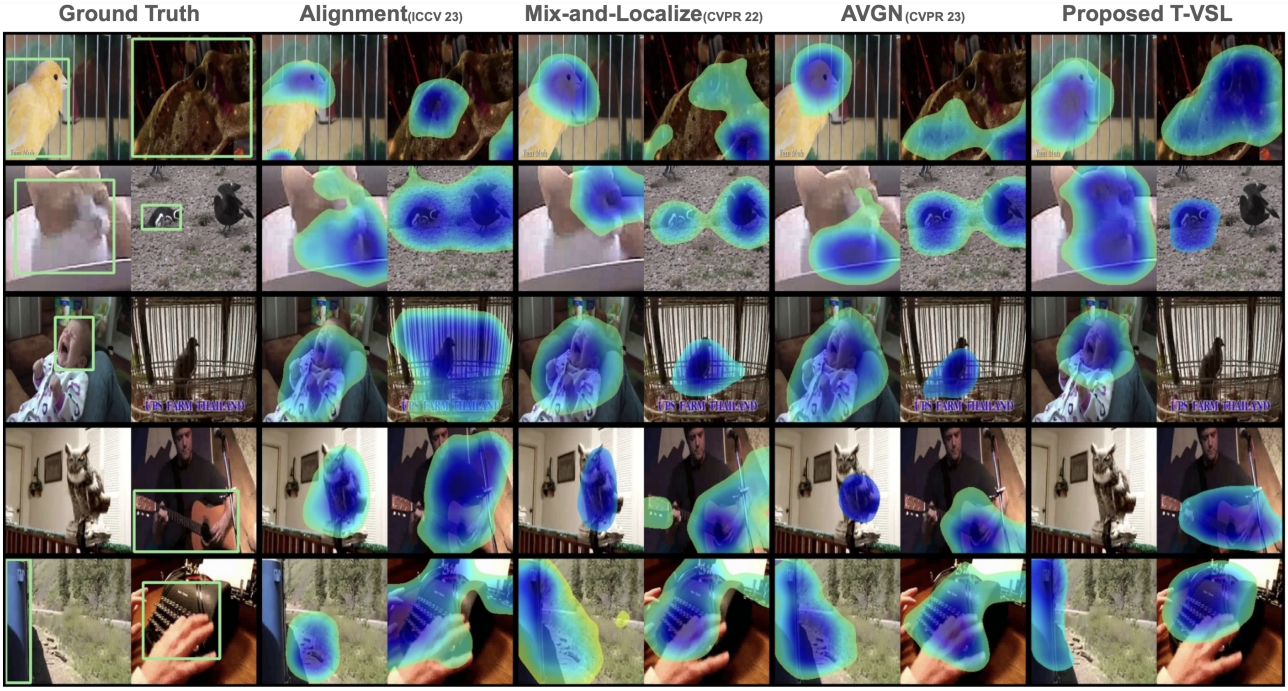

Figure 4. We present additional qualitative comparisons on challenging multi-source localization with SOTA single and multi-source baseline methods. Here, blue color represents high-attention values to the sounding object, and red color represents low-attention values. Similar to our prior observation, the proposed T-VSL generates more precise localization maps for sounding sources by selectively isolating the sounding regions from the background.

Stage Number	Method	VGGSound-Single		VGGSound-Duet	
		AP(%)	IoU@0.5(%)	CAP(%)	CIoU@0.3(%)
Single-Stage	AVGN CVPR23	44.1	49.6	31.9	37.8
Single-Stage	T-VSL	46.8	51.5	33.7	38.7
Two-Stage	T-VSL	**48.1**	**53.7**	**35.7**	**40.1**

Table 7. Quantitative comparison on single and multi-stage architectures on VGGSound-Single and VGGSound-Duet datasets. Same AudioCLIP encoders are used. Proposed two-stage method generates superior performance compared to its single-stage counterpart.

silent objects, where other baselines struggle in isolating the silent objects present in the surroundings. These qualitative results demonstrate the effectiveness of T-VSL in disentangling multi-source mixtures as well as in learning audio-visual correspondence for sound source localization.