Supplementary Material for Audio-Visual Grouping Network for Sound Localization from Mixtures

Shentong Mo
Carnegie Mellon University

Yapeng Tian*
University of Texas at Dallas

In this supplementary material, we provide the significant differences between our AVGN and the recent work, GroupViT [3], more experiments on the depth of transformer layers and grouping strategies. In addition, we validate the effectiveness of learnable audio-visual class tokens in learning disentangled audio-visual representations and report qualitative visualization results of localization maps.

1. Significant Difference from GroupViT and AVGN

When compared to GroupViT [3] on image segmentation, there are three significant distinct characteristics of our AVGN for addressing sound localization from mixtures, which are highlighted as follows:

1) **Constraint on Audio-Visual Category Tokens.** The major difference is that we have learned disentangled audio-visual class tokens for each sound source, *e.g.*, 37 audio-visual category tokens for 37 categories in the VGGSound-Instruments benchmark. During training, each audio-visual class token does not learn semantic overlapping information among each other, where we apply the cross-entropy loss $\sum_{i=1}^{C} \text{CE}(\mathbf{h}_i, \mathbf{e}_i)$ on each category probability $\mathbf{e}_i$ with the disentangled constraint $\mathbf{h}_i$. However, the number of group tokens used in GroupViT is a hyper-parameter, and they must tune it carefully across each grouping stage.

2) **Audio-Visual Grouping.** We propose the audio-visual grouping module for extracting individual semantics with category-aware information from the mixture spectrogram and image. However, GroupViT used the grouping mechanism on only visual patches without explicit category-aware tokens involved. Therefore, GroupViT can not be directly applied to a sound spectrogram for solving sound localization problems from mixtures. Moreover, they utilized multiple grouping stages during training and the number of grouping stages is a hyper-parameter. In our module, only one audio-visual grouping stage with disentangled audio-visual category tokens is enough to learn disentangled audio-visual representations in the multi-modal semantic space.

3) **Audio-Visual Class as Weak Supervision.** We leverage the audio-visual category as the weak supervision to address the sound localization problem from the mixtures, while GroupViT used a trivial contrastive loss to match the global visual representations with text embeddings. In this case, GroupViT required a large batch size for self-supervised training on large-scale visual-language pairs. In contrast, we do not need unsupervised learning on the large-scale simulated mixture data with extensive training costs.

2. Depth of Transformer Layers and Grouping Strategies

The depth of transformer layers and grouping strategies used in the proposed AVG affect the extracted and grouped representations for source localization from the mixtures. To explore such effects more comprehensively, we varied the depth of transformer layers from $\{1, 3, 6, 12\}$ and ablated the grouping strategy using Softmax and Hard-Softmax. During training, the Gumbel-Softmax [1, 2] was applied as the alternative to Hard-Softmax to make it differentiable.

We report the comparison results of source localization performance in Table 1. When the depth of transformer layers is 3 and using Softmax in AVG, we achieve the best localization performance in terms of all metrics. With the increase of the depth from 1 to 3, the proposed AVGN consistently raises results as better disentangled audio-visual representations are extracted from encoder embeddings of the raw mixture and image. However, increasing the depth from 3 to 12 will not continually improve the result since 3 transformer layers might be enough to extract the learned category-aware embeddings for audio-visual grouping with only one grouping stage. Furthermore, replacing Softmax with Hard-Softmax significantly deteriorates the localization performance, which shows the importance of the proposed AVG in extracting disentangled audio-visual representations with category-aware semantics from the audio

*Corresponding author.

Depth	AVG	MUSIC-Solo			MUSIC-Duet			
		AP(%)	IoU@0.5(%)	AUC(%)	CAP(%)	PIAP(%)	CIoU@0.3(%)	AUC(%)
1	Softmax	75.8	53.6	45.7	47.6	53.1	28.5	23.2
3	Softmax	**77.2**	**58.1**	**48.5**	**50.6**	**57.2**	**32.5**	**24.6**
6	Softmax	76.7	57.5	47.9	50.1	56.9	32.1	24.3
12	Softmax	76.3	57.3	47.6	49.8	56.7	31.7	24.1
3	Hard-Softmax	73.2	47.6	43.1	42.5	49.2	24.8	21.5

Table 1. Exploration studies on the depth of self-attention transformer layers and grouping strategies in Audio-Visual Grouping (AVG) module.

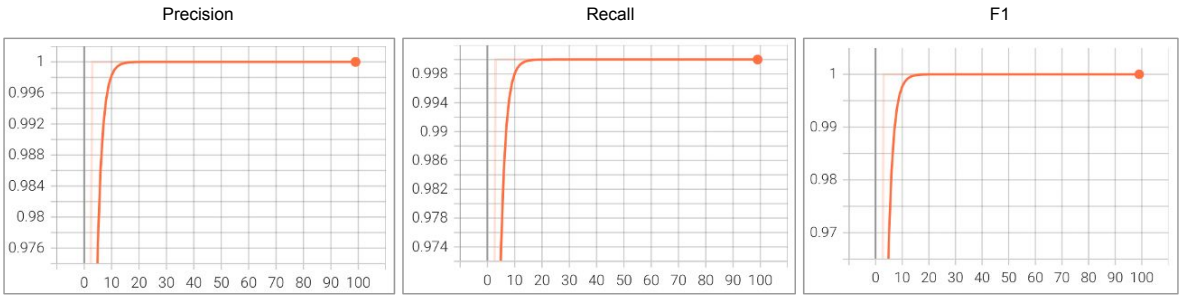

Figure 1. Quantitative results (Precision, Recall, and F1 score) of learned audio-visual class tokens.

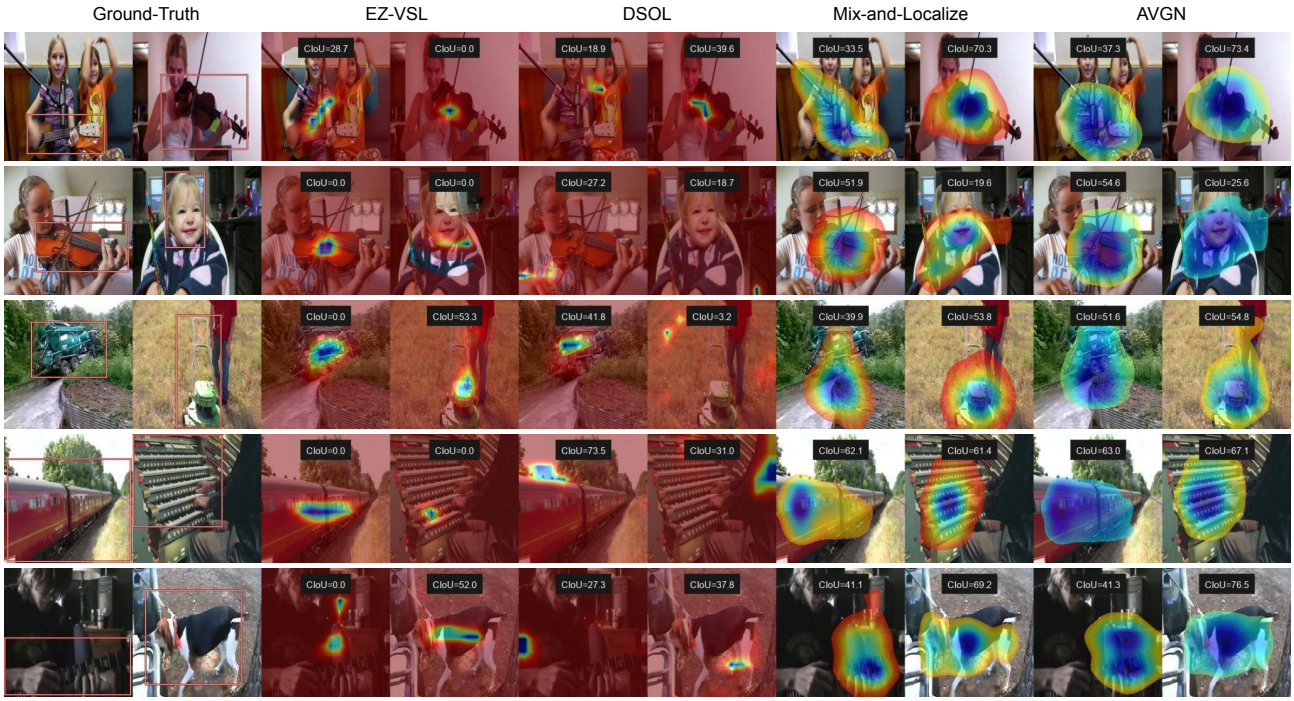

Figure 2. Qualitative comparisons with single-source and multi-source baselines on multi-source localization. Note that blue refers to high attention values and red for low attention values. The proposed AVGN produces more accurate and high-quality localization maps for each source.

mixture and image for localizing each sounding source.

3. Quantitative Validation on Audio-Visual Category Tokens

Learnable Audio-Visual Category Tokens are essential to aggregate audio-visual representations with category-aware

semantics from the sound mixture. To quantitatively validate the rationality of learned audio-visual category token embeddings, we compute the Precision, Recall, and F1 scores of audio-visual classification using these embeddings across training iterations. The quantitative results are reported in Figure 1. As can be seen, all metrics rise to 1 at epoch 20, which indicates that each learned audio-visual category token has disentangled information with category-aware semantics. These quantitative results further demonstrate the effectiveness of audio-visual category tokens in the audio-visual grouping for extracting disentangled audio-visual representations from images and audio mixtures for localizing each source more accurately.

4. Qualitative Visualization on Source Localization

To qualitatively demonstrate the effectiveness of our method, we report more visualization results in Figure 2. We can observe that the proposed AVGN achieves decent localization performance in terms of more accurate and high-quality localization maps for each sound source.

References

[1] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017. 1

[2] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017. 1

[3] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. *arXiv preprint arXiv:2202.11094*, 2022. 1