1 Generating Pseudo Labels for Unlabelled Videos

When training on unlabelled videos, such as SonudNet-Flickr [2] and AVE dataset [15], we need to generate pseudo labels as classification supervision.

First, we use CRNN [13] pretrained on AudioSet [5] and ResNet-18 [6] pretrained on ImageNet [9] to predict classification probabilities on audio and visual message. Next, to reduce noise and assist coarse-grained audiovisual correspondence, we need to organize several general categories as target. Considering AudioSet is annotated with hierarchical ontology, containing four levels of labels from coarse to fine, we choose the first-level labels of 7 classes (human sounds, music, animal, sounds of things, natural sounds, source-ambiguous sounds, and environment) as final classification target. Then we aggregate the predictions from pretrained models into these 7 categories. For audio modality, we directly use the ontology in AudioSet to generate supervision. While for visual modality, inspired by [4, 14], we take similarity of word embeddings [10] and conditional probabilities between labels in ImageNet and AudioSet into consideration to aggregate 1000 classification predictions into 7 as pseudo labels.

2 Experiments on AVE Dataset

2.1 AVE Dataset

AVE dataset [15] contains 4143 10-second video clips covering 28 event categories. This dataset is proper for cross-modality localization since the videos are temporally labelled with audiovisual event boundaries. But annotations are only used for evaluation. In training phase, we feed audiovisual pairs into our model to learn cross-modal alignment in an unsupervised manner. The videos are divided into 3339 for training, 402 for validation and 402 for test. Note that events in testing videos all span less than 10 seconds.

2.2 Cross-Modality Localization

In this task, given a temporal segment of one modality, we aim to accurately localize the temporal position of the synchronized content in the other modality. There are two subtasks, visual localization from audio segments and vice versa, namely A2V and V2A. Following [15], we adopt AVE dataset without labels for training, and only use short-event videos for evaluation.

Concretely, we employ sliding windows to predict the temporal position. Take visual localization from audio (A2V) as an example:

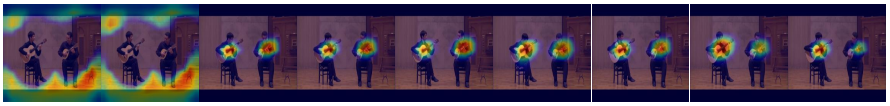
$$t^* = \arg\min_t \sum_{s=1}^{l} f(V_{s+t-1}, \hat{A}_s), \tag{1}$$

where f measures the correspondence score between audio and visual context, $\hat{A}$ represents query l-second audio segment, t^* is the predicted start time when

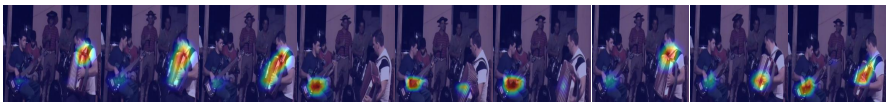
audio and vision synchronize. Strict evaluation metric as [15] is adopted on two subtasks. In Tabel 1, we show our model's results in two different settings, one is only using classification and video-level audiovisual correspondence, the other is to further perform fine-grained alignment. Since it is more challenging to disentangle different events in mixed audio than in video frames, previous methods are poor on V2A. While our method performs much better at capturing temporal information in audio, and outperforms others over a large margin on V2A. Comparing results of our method with different settings, our fine-grained alignment in the second stage further improves performance, but still not the best on A2V. That is because the major target of this task is to distinguish temporal boundaries of audiovisual events, there are few events overlapping at the same time, which restricts the efficacy of our fine-grained alignment.

Table 1. Cross-madality localization accuracy with A2V and V2A subtasks.

Models	DCCA[1]	AVDLN[15]	Ours	Ours w/align
A2V	34.8	44.8	41.5	43.8
V2A	34.1	35.6	43.8	44.3



(a) From background noise to musical instruments.



(b) Duet of accordion and guitar.

(c) Dogs barking interspersed with sound of toy car.

Fig. 1. We visualize the changes of localization maps in videos over time. The frames shown are extracted at 1 fps, the heatmaps show localization responses to corresponding 1-second audio clip. When only with noise, our model mainly focuses on background regions as the first two frames in Fig. 1(a). When there are sounds produced by specific objects, our model can accurately capture the sound makers, e.g., our model can distinguish sounds of guitar and accordion in Fig. 1(b), dog barking and toy-car sound in Fig. 1(c).

We also visualize sound localizetion results on several videos. Fig. 1 vividly shows the changes of sounds on time dimension, which further demonstrates model's capacity of spatio-temporally determining which specific object is making sound.

3 Comparison with CAM

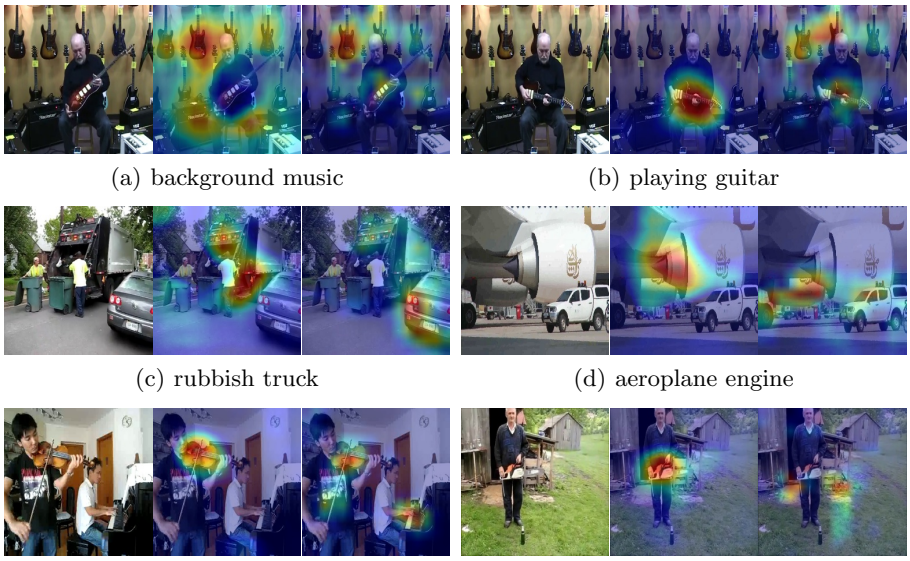
In this section, we compare the localization results between our model and CAM method [16, 11, 3] based on classification. Specifically, our two-stage framework achieves coarse-grained audiovisual correspondence in the category-level at the first stage, and fine-grained sound-object alignment at the second stage. To validate the efficacy of our fine-grained audiovisual alignment in the second stage, we compare our method with category-level CAM output.

Concretely, we adopt the model trained on AVE dataset for comparison, where the classification targets are 7 general categories mentioned above (i.e., human sounds, music, animal, sounds of things, natural sounds, source-ambiguous sounds, and environment). Our model generates localization results following the procedure mentioned in the paper, while for CAM method, we adopt predicted probabilities on audio as prior, and employ CAM to generate class-specific localization maps on visual modality. We visualize some comparison results in Fig. 2. Generally, CAM method cannot distinguish the objects belonging the same category, e.g., aeroplane and car in Fig. 2(d), while our model can precisely localize the specific object making sound in input audio. It is because CAM method performs localization in the category-level, while our model further establishes video- and category-based sound-object association. Additionally, as shown in Fig. 2(a) and Fig. 2(b), in the scene with multiple guitars, with background music sound, our model focuses on the silent guitars hanging on the wall, while with the sound of the man playing guitar, our method precisely localize the guitar held by the man. It is probably because the sound of playing guitar usually coexists with the visual pattern of the interaction between human hands with guitar, while the background music is usually with individually placed music instruments.

Further, we also quantitatively compare the localization results of these two methods on human annotated subset of SoundNet-Flickr dataset [12]. For CAM method, we perform weighted summation on class activation maps over valid categories, where the weights are the normalized predicted probabilities on audio modality. Table 2 shows the results, our two-stage learning framework outperforms CAM method over a large margin, which demonstrates the efficacy of fine-grained sound-object alignment in the second stage.

4 Additional Results

In this section, we present more examples of our localization results in multi-source scenarios. Fig. 3 shows the result in two-source scenes, and the results



(a) background music (b) playing guitar

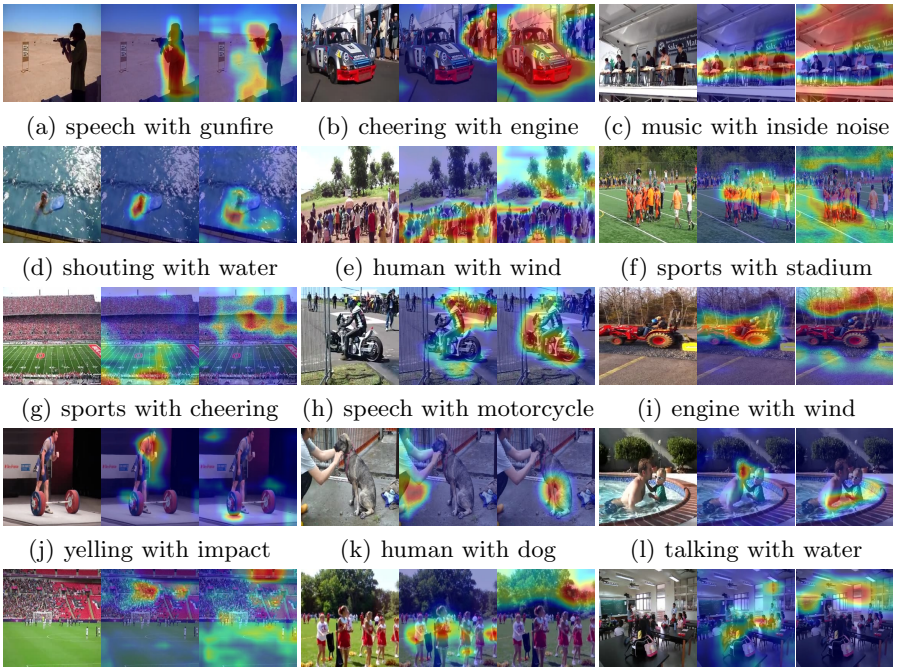
(c) rubbish truck (d) aeroplane engine

(e) playing violin (f) wood sawing

Fig. 2. We show some comparison between our model and CAM method. The images in each subfigure are listed as: original image, localization result of our model, result of CAM method. It is clear that CAM method cannot distinguish the objects belonging to the same category, e.g., violin and piano in Fig. 2(e), but our model can precisely localize the object that makes sound in input audio.

Table 2. Quantitative localization results on SoundNet-Flickr subset, cIoU and AUC are reported.

Methods	cIoU@0.5	AUC
Random	7.2	30.7
Attention[12]	43.6	44.9
DMC AudioSet[7]	41.6	45.2
CAVL AudioSet[8]	50.0	49.2
Ours Stage-one	44.2	48.1
Ours Stage-two	**52.2**	**49.6**



(a) speech with gunfire (b) cheering with engine (c) music with inside noise

(d) shouting with water (e) human with wind (f) sports with stadium

(g) sports with cheering (h) speech with motorcycle (i) engine with wind

(j) yelling with impact (k) human with dog (l) talking with water

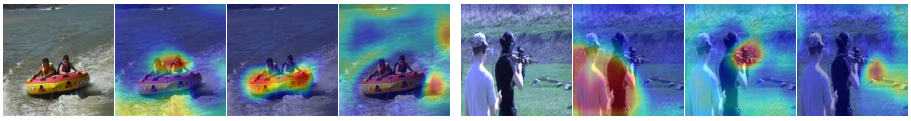
(m) screaming with stadium (n) yelling with wind (o) speech with classroom

Fig. 3. We visualize the localization maps corresponding to different elements contained in the mixed sounds of two sources. The results qualitatively demonstrate our model's performance in multi-source sound localization.

generally demonstrate our model's capacity of distinguishing different sound sources.

We also show some localization results under three-source scenes in Fig. 4. In Fig. 4(a), it is interesting that the boat is being towed by something off-screen, and the engine sound actually comes from the unseen object, while our model associates them as a sound-object pair. This is probably because the visual pattern of boats usually coexist with engine sound, and these two are of the same category, eventually they become highly correlated.

We present cross-modal retrieval results based on the aligned audiovisual features in Fig. 5. Concretely, we use an image or a clip of audio as query, and treat other audio or images in the dataset as gallery. We calculate the distance between query and gallery features, and take the top-5 nearest examples shown in Fig. 5.



(a) shouting, engine and water (b) speaking, gunfire and wind

Fig. 4. We show the localization results of three-source scenes, and each localization map corresponds to one potential sound source.

(a) sound of baby as query and top-5 retrieved images

(b) sound of helicopter engine as query and top-5 retrieved images

(c) image of crowd people as query and top-5 retrieved audio

Fig. 5. Cross-modal retrieval results, with an image/sound as query and retrieve top-5 most similar audio/images.

References

1. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International conference on machine learning. pp. 1247–1255 (2013)
2. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29, pp. 892–900. Curran Associates, Inc. (2016), http://papers.nips.cc/paper/6146-soundnet-learning-sound-representations-from-unlabeled-video.pdf
3. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 839–847 (March 2018). https://doi.org/10.1109/WACV.2018.00097
4. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5177–5186 (2019)
5. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780 (March 2017). https://doi.org/10.1109/ICASSP.2017.7952261
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
7. Hu, D., Nie, F., Li, X.: Deep multimodal clustering for unsupervised audiovisual learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
8. Hu, D., Wang, Z., Xiong, H., Wang, D., Nie, F., Dou, D.: Curriculum audiovisual learning. arXiv preprint arXiv:2001.09414 (2020)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012), http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
10. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014), http://www.aclweb.org/anthology/D14-1162
11. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
12. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., So Kweon, I.: Learning to localize sound source in visual scenes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
13. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(11), 2298–2304 (Nov 2017). https://doi.org/10.1109/TPAMI.2016.2646371
14. Sun, Y., Ghaffarzadegan, S.: An ontology-aware framework for audio event classification. ArXiv **abs/2001.10048** (2020)

15. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: The European Conference on Computer Vision (ECCV) (September 2018)
16. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)