Supplementary Material

Liuyuan Wen
lywen@mail.ustc.edu.cn

School of Physical Sciences
USTC
Hefei, China

In this supplementary material, we present a detailed introduction to the datasets used in Section 1. Furthermore, additional ablation studies on the negative loss terms and threshold tuning, along with t-SNE visualizations, are unveiled in Section 2 and Section 3, respectively.

1 Datasets

In this study, we conduct experiments on three widely used datasets: VGGSound [2], UCF [10], and ActivityNet [4].

VGGSound: The VGGSound dataset is a large-scale audio-visual dataset consisting of short clips of audio sounds extracted from videos uploaded to YouTube. It is designed to be a comprehensive resource for studying the correspondence between audio and visual elements in diverse and challenging acoustic environments. We selected 276 classes that are clearly labeled in our experiment, resulting in 93,752 videos.

UCF: The UCF101 dataset is widely recognized for action recognition in videos. As an extension of the UCF50 dataset, it is notable for its diversity and complexity, making it one of the most challenging datasets for video-based human action recognition. We included only the 51 classes that contain audio information, resulting in 6,816 videos.

ActivityNet: The ActivityNet dataset is a large-scale video benchmark designed for human activity understanding. It contains 200 different types of activities across 20,348 video clips collected from YouTube. This dataset is significant for its size and diversity, making it a challenging and comprehensive benchmark for temporal activity detection.

2 Additional Ablation Study

2.1 Effects of Negative Loss Terms of Unseen Classifier

As explained in Section 3.4, prior research often focuses solely on losses associated with positive samples [5, 6, 7, 8, 11], while neglecting losses from negative samples. To foster more robust training, we accord them equal importance and compute all three types of losses for both positive and negative samples. We present an ablation study of different loss terms in Table 1. $\mathbf{UC_{acc}}$ denotes the accuracy of the unseen classifier, while **HM** scores are provided with other parts of the model weights fixed for a fair comparison. It is evident that the inclusion of $\mathcal{L}^-_{trip}$ alone can remarkably enhance scores. Taking VGGSound as an example, $\mathcal{L}^+ + \mathcal{L}^-_{trip}$ yields 8.35% and 10.36% for $\mathbf{UC_{acc}}$ and **HM**, respectively, surpassing $\mathcal{L}^+$ alone with scores of 7.83% and 9.92%. On the other hand, while the addition of either $\mathcal{L}^-_{rec}$ or $\mathcal{L}^-_{reg}$

Table 1: Effects of negative loss terms of unseen classifier

Loss	VGGSound		UCF		ActivityNet	
	$\mathbf{UC_{acc}}$	**HM**	$\mathbf{UC_{acc}}$	**HM**	$\mathbf{UC_{acc}}$	**HM**
$\mathcal{L}^+$	7.83	9.92	20.69	29.86	9.9	12.83
$\mathcal{L}^+ + \mathcal{L}^-_{trip}$	8.35	10.36	25.31	34.9	9.81	12.9
$\mathcal{L}^+ + \mathcal{L}^-_{rec}$	7.36	9.62	18.77	27.19	9.42	12.72
$\mathcal{L}^+ + \mathcal{L}^-_{reg}$	7.33	9.43	21.81	30.99	8.88	11.97
$\mathcal{L}^+ + \mathcal{L}^-$(ours)	**8.78**	**11.16**	**28.21**	**37.89**	**11.49**	**14.38**

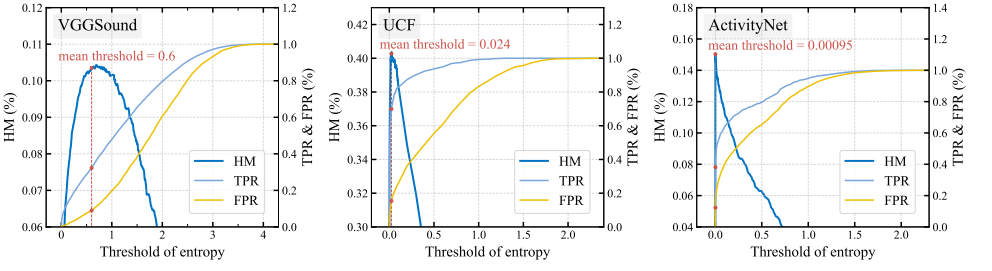

Figure 1: Effects of OOD-entropy thresholds on **HM**, **TPR**, and **FPR** across three datasets. In each figure, the left vertical axes correspond to **HM** lines, while the right axes correspond to **TPR** and **FPR** lines. Red dots and words represent the average entropy of seen classes from the training data, which evidently aligns closely with the value maximizing **HM**.

alone does not seem particularly effective, the combination of all three negative loss terms performs admirably. Indeed, $\mathcal{L}^+ + \mathcal{L}^-$ outperforms $\mathcal{L}^+$ in all aspects, achieving the highest scores of 8.78% and 11.16% on VGGSound.

2.2 Parameter Tuning of Thresholds of OOD-entropy

All three bias reduction methods explored in Section 4.3.1 require parameter tuning of thresholds, outputs exceeding or falling below which are categorized as seen or unseen accordingly. Figure 1 illustrates the curves of **HM**, **TPR**, and **FPR** as they change with entropy thresholds. Unlike *Calibrated Stacking* and *OOD-binary*, whose thresholds necessitate tuning through iterative processes such as "for" loops, the thresholds of *OOD-entropy* can be readily determined using the average entropy of seen classes from the training data (indicated by red dots and the descriptor *mean threshold* in Figure 1). It's evident that all *mean threshold*s closely align with the value that maximizes **HM**. Additionally, we observe significant variation in thresholds corresponding to the points with the highest **HM** across different datasets, namely 0.6, 0.024, and 0.00095 for VGGSound, UCF, and ActivityNet respectively. This variability underscores the inefficiency of tuning entropy thresholds through iterative loops.

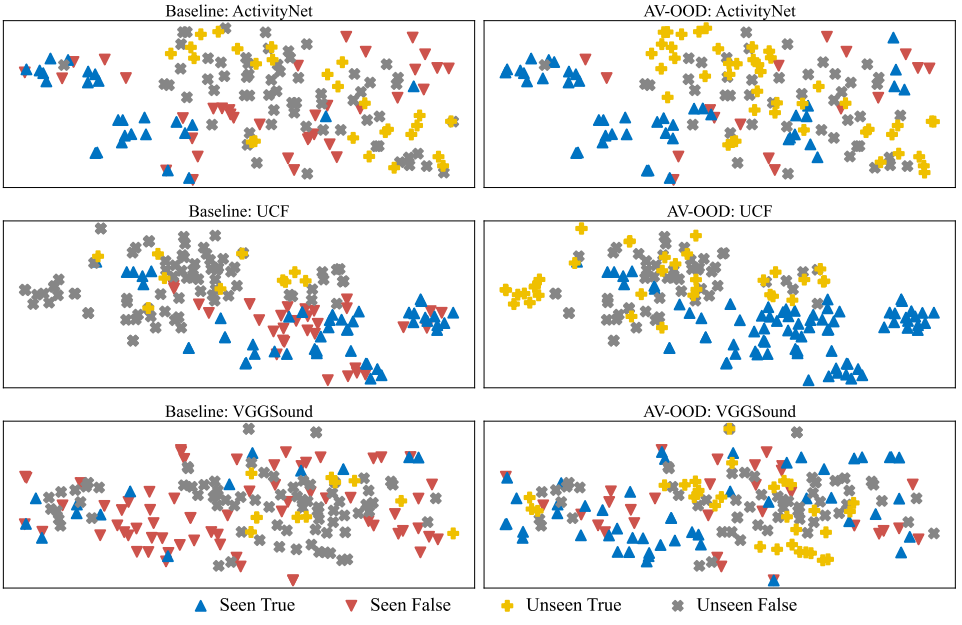

Figure 2: Comparison between Baseline and our AV-OOD in t-SNE visualizations from test classes across three datasets. In each subplot, we plot 6 seen classes and 6 unseen classes, with 15 samples randomly selected from each category. As to legends, *Ture* and *False* denote correctly or incorrectly classified samples respectively. Thus, as an example, *Seen False* refers to incorrectly classified samples from seen classes. (They can be misclassified as either a seen class or an unseen class. This is not reflected in the figure.)

3 Qualitative Results

In Figure 2, we present t-SNE visualizations for comparison between the *Baseline* AVCA [9] and our proposed *AV-OOD*. Feature embeddings are obtained using attributes $a \oplus v$, and all of them are grouped into four categories: *Seen True*, *Seen False*, *Unseen True*, and *Unseen False*. We can clearly see that *AV-OOD* outperforms *Baseline* in terms of both seen and unseen classes. Although *Baseline* adopts calibrated stacking to reduce bias towards seen classes, this method comes at the cost of sacrificing the accuracy of seen classes. This highlights the superiority of our framework, where the OOD detector and separate expert classifier cooperate, resulting in better overall performance.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, volume 70, pages 214–223, 2017.

[2] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pages 721–725, May 2020.

[3] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *NeurIPS*, volume 30, 2017.

[4] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[5] Jingwei Hong, Zulqarnain Hayder, Junting Han, Ping Fang, Mehrtash Harandi, , and Lars Petersson. Hyperbolic audio-visual zero-shot learning. In *ICCV*, pages 7873–7883, 2023.

[6] Wenrui Li, Zhengyu Ma, Liang-Jian Deng, Hengyu Man, and Xiaopeng Fan. Modality-fusion spiking transformer network for audio-visual zero-shot learning. In *ICME*, pages 426–431, 2023.

[7] Wenrui Li, Xi-Le Zhao, Zhengyu Ma, Xingtao Wang, Xiaopeng Fan, , and Yonghong Tian. Motion-decoupled spiking transformer for audio-visual zero-shot learning. In *ACM MM*, pages 3994–4002, 2023.

[8] Yapeng Li, Yong Luo, and Bo Du. Audio-visual generalized zero-shot learning based on variational information bottleneck. In *ICME*, pages 450–455, 2023.

[9] O. B. Mercea, L. Riesch, A. Koepke, and Z. Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *CVPR*, pages 10553–10563, 2022.

[10] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint*, 2012.

[11] Q. Zheng, J. Hong, and M. Farazi. A generative approach to audio-visual generalized zero-shot learning: Combining contrastive and discriminative techniques. In *IJCNN*, pages 1–8, 2023.