Can audio-visual integration strengthen robustness under multimodal attacks? Appendix

Yapeng Tian Chenliang Xu
University of Rochester
{yapengtian,chenliang.xu}@rochester.edu

Contents

In this appendix, we first provide more experimental details in Sec. A. Then, we show more results in Sec. B.

A. Experimental Details

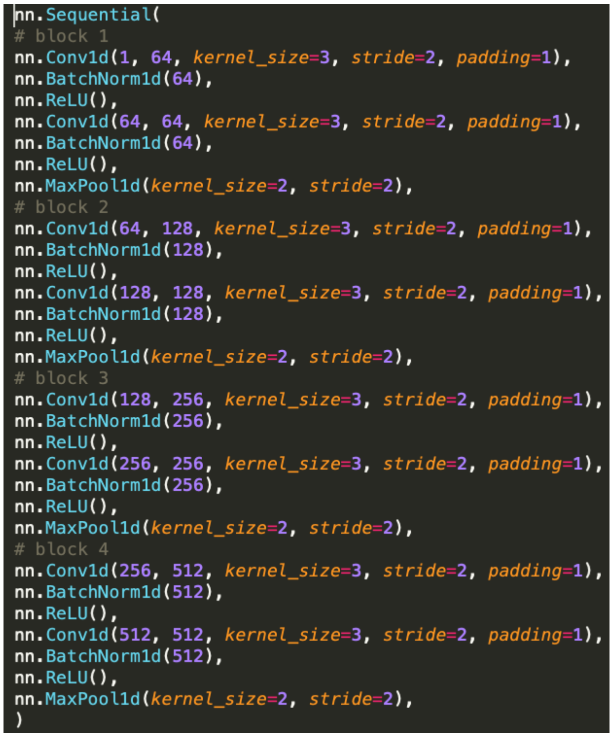
We first introduce an additional dataset: AVE [8] in Sec. A.1. Then, we describe audio and visual data processing in Sec. A.2. In addition, we give more details of audio and visual networks in Sec. A.3. Furthermore, we define the used five different audio-visual fusion functions in Sec. A.4. Finally, more training details are provided in Sec. A.5.

A.1. The AVE Dataset

Besides the MIT-MUSIC and Kinetics-Sounds datasets, we also explore audio-visual model robustness using another popular audio-visual dataset: AVE [8] to further validate our findings. It consists of 4,143 unconstrained videos spanning 28 event categories. As in [8], we divide the data into train/val/test splits of 3,339/402/402 videos, respectively.

A.2. Data Processing

The sampling rates of sounds and video frames are 11025 Hz and 8 fps, respectively. For each video, we sample a 6s audio clip with 1 video frame at the center position of the sound as the inputs of our audio-visual models. We use a pre-trained ResNet18 [3] to extract visual features and a 1-D convolution-based model to extract audio features from input audio waveforms.



```
nn.Sequential(
# block 1
nn.Conv1d(1, 64, kernel_size=3, stride=2, padding=1),
nn.BatchNorm1d(64),
nn.ReLU(),
nn.Conv1d(64, 64, kernel_size=3, stride=2, padding=1),
nn.BatchNorm1d(64),
nn.ReLU(),
nn.MaxPool1d(kernel_size=2, stride=2),
# block 2
nn.Conv1d(64, 128, kernel_size=3, stride=2, padding=1),
nn.BatchNorm1d(128),
nn.ReLU(),
nn.Conv1d(128, 128, kernel_size=3, stride=2, padding=1),
nn.BatchNorm1d(128),
nn.ReLU(),
nn.MaxPool1d(kernel_size=2, stride=2),
# block 3
nn.Conv1d(128, 256, kernel_size=3, stride=2, padding=1),
nn.BatchNorm1d(256),
nn.ReLU(),
nn.Conv1d(256, 256, kernel_size=3, stride=2, padding=1),
nn.BatchNorm1d(256),
nn.ReLU(),
nn.MaxPool1d(kernel_size=2, stride=2),
# block 4
nn.Conv1d(256, 512, kernel_size=3, stride=2, padding=1),
nn.BatchNorm1d(512),
nn.ReLU(),
nn.Conv1d(512, 512, kernel_size=3, stride=2, padding=1),
nn.BatchNorm1d(512),
nn.ReLU(),
nn.MaxPool1d(kernel_size=2, stride=2),
)
```

Figure 1: A Pytorch implemenation of our audio network.

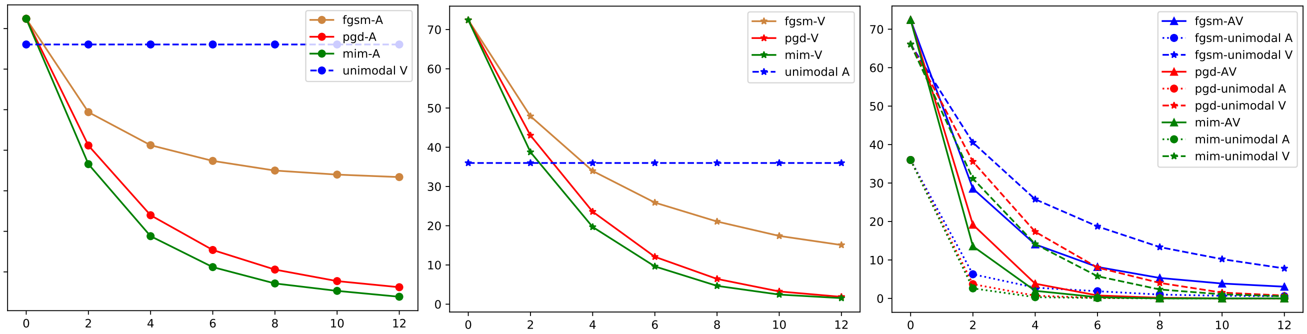
A.3. Architectures

Audio Net: Our audio network takes 6s audio waveforms as inputs and output 512-D audio feature vectors by a global max pooling after the 1-D Convolution-based network as illustrated in Figure 1. The network consists of 8 convolutional layers in 4 building blocks.

Visual Net: We use the ResNet18 [3] removing the final Fully-Connected (FC) layer as our visual network. We also obtain 512-D feature vectors by a global max pooling. But, in the weakly-supervised sound source visual localization, we remove the global max pooling to obtain a 2-D feature

Dataset	Attack Methods	✓AV	✗A	✗V	✗AV	Avg.	Unimodal ✓A	Unimodal ✓V
AVE	FGSM [2]	70.40	40.55	24.88	8.71	24.71	29.85	65.17
	PGD [5]		20.15	11.44	1.99	11.19		
	MIM [1]		15.17	10.20	0.25	8.54		

Table 1: Audio-visual event recognition accuracy on the AVE dataset under different attack methods (ϵ_a, ϵ_v = 0.06). ✗A, ✗V, and ✗AV denote that only audio, only visual, and both audio and visual inputs for our audio-visual network are attacked, respectively. The symbol: ✓ means that inputs are clean. The baselines: Unimodal ✓A and Unimodal ✓V models are two single-modality models.



(a) Audio Attack (b) Visual Attack (c) Audio-Visual Attack

Figure 2: Adversarial robustness against multimodal attacks on the Kinetics-Sounds. The x-axis denotes the attack strength ($\times 10^{-3}$) and we set $\epsilon_a = \epsilon_v$ in the audio-visual attack for a better illustration. For the single-modality attack, the attacked audio-visual models in (a) and (b) still have clean visual and audio information, respectively. When adversarial perturbations become larger, joint perception models with one attacked modality become even worse than the corresponding individual perception models. Thus, an unreliable modality could weaken perception by the other modality in audio-visual models. A similar observation can also be found in the audio-visual attack (*e.g.*, -AV vs. -unimodal V).

map for each frame.

A.4. Audio-Visual Fusion Methods

We use 5 audio-visual fusion methods to explore how different fusion methods affect audio-visual event recognition against multimodal attacks. Here are formulations of the 5 fusion functions. They use an audio feature: f_a and a visual feature: f_v as inputs and obtain a fused feature: f_{av}.
Sum: It directly sums up the features from the both modalities: $f_{av} = f_a + f_v$.
Concat: The Concat: $f_{av} = [f_a; f_v]$ concatenates the audio and visual features.
FiLM: The FiLM [7] learns to adaptively fuse two different modalities by feature modulations. In our implementation, we use the audio feature as the input of transformation for fusion: $f_{av} = \alpha(f_a) \cdot f_v + \beta(f_a)$, where $\alpha(\cdot)$ is a FC layer and $\beta(\cdot)$ is an identity mapping.
Gated-Sum: The Gated-Sum [4] uses audio and visual features to compute two gates to fuse feature from the other modality, respectively. They can be computed as:

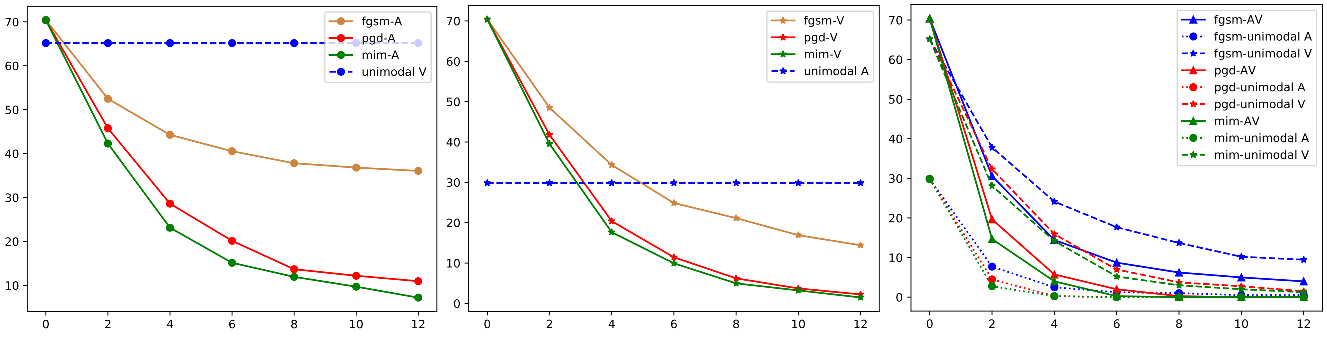
$$f_1 = \sigma(f_a) \cdot f_v, \tag{1}$$

$$f_2 = \sigma(f_v) \cdot f_a, \tag{2}$$

where the $\sigma(\cdot)$ is the Sigmoid function. The fused features: f_1 and f_2 are then combined by the Sum: $f_{av} = f_1 + f_2$.
Gated-Concat: The Gated-Concat is similar to the the Gated-Sum. It also computes f_1 and f_2. But, it fuses by a concatenation: $f_{av} = [f_1; f_2]$.

A.5. Implementation Details

We train our network with the standard SGD using 4 NVIDIA 1080TI GPUs. We set the batch size = 48, the initial learning rate of the audio network $= 1e-4$, the initial learning rate of the visual net $= 1e-3$, the initial learning rate of the fusion network with the final FC layer $= 1e-3$. The epoch numbers are 100, 30, and 100 for the MIT-MUSIC, Kinetics-Sounds, and AVE datasets, respectively. The learning rates drop by multiplying 0.1 after every 30 epochs for the MIT-MUSIC and AVE and every 10 epochs for the Kinetics-Sounds. For PGD [5] and MIM [1], we perform 10-step iterative attacks. The parameters: $\lambda_a = \lambda_v = 0.1$. In addition, when we use our external feature memory banks to defend against attacks, we found averaging the denoised and original features can obtain better performance since there are also optimization errors when computing the audio and visual coefficients.



(a) Audio Attack (b) Visual Attack (c) Audio-Visual Attack

Figure 3: Adversarial robustness against multimodal attacks on the AVE. The x-axis denotes the attack strength ($\times 10^{-3}$) and we set $\epsilon_a = \epsilon_v$ in the audio-visual attack for a better illustration.

Defense (AVE)	✓AV	✗A	✗V	✗AV	Avg	RI
None	70.40	40.55	24.88	8.71	24.71	0
Unimodal A	29.85	1.24	29.85	1.24	10.77	-54.49
Unimodal V	65.17	65.17	17.66	17.66	33.50	3.56
PCL [6]	61.94	61.69	17.91	17.91	32.50	-0.67
MaxSim	71.64	35.82	25.62	16.42	25.95	2.48
MinSim	70.90	57.21	25.37	21.39	34.66	10.45
ExFMem	71.39	44.78	28.11	10.95	27.94	4.22
MinSim+ExFMem	71.39	58.21	29.35	26.62	38.06	14.34

Table 2: Audio-visual event recognition accuracy on the AVE dataset with different defense methods. Here, we use the FGSM (ϵ_a, ϵ_v = 0.06) to generate audio and visual adversarial examples. Some models (*e.g.*, Unimodal A, Unimodal V, and PCL) highly rely on only one modality, which absolutely makes them more invulnerable to adversarial attacks for another modality. However, they will fail to obtain good performance on clean audio and visual inputs. Top-2 results are highlighted.

B. Experimental Results

To further validate our findings, we show more experimental results on audio-visual model robustness under multimodal attacks in Sec. B.1, sound source localization under attacks in Sec. B.2, and audio-visual defense in Sec. B.3.

B.1. Robustness under Multimodal Attacks

We first show the audio-visual event recognition results on the AVE dataset with different attack methods in the Table 1. Similar to observations on the MIT-MUSIC and Kinetics-Sounds, our audio-visual model can be easily fooled, and the joint perception models: ✗A (with clean visual) and ✗V (with clean audio) are worse than Unimodal ✓V and ✓A, respectively. Thus, audio-visual integration could even weaken event recognition when audio or visual inputs are attacked. We further illustrate results of adversarial robustness against multimodal attacks with different attack strengths on the Kinetics-Sounds and AVE in Figure 2 and Figure 3, respectively. The results can further validate our findings that audio-visual integration may not always strengthen the audio-visual model robustness under multimodal attacks and the adversarial robustness of the audio-visual models highly depends on the reliability of the multisensory inputs.

B.2. Sound Source Localization under Attacks

We show more sound source localization results under multimodal attacks in Figure 4. A large range of events (*e.g.*, chopping wood, playing xylophone, frying, baby crying, running bus) are covered. We can see that the weakly-supervised sound source visual localization model is susceptible to both single-modality and audio-visual attacks.

B.3. Audio-Visual Defense

We show defense results against the FGSM attack on the AVE dataset in Table 2. Similar to results on the Kinetics-Sounds and MIT-MUSIC datasets, the proposed: MinSim and ExFMem can improve audio-visual model robustness against both single-modality and audio-visual attacks and our full model outperforms the compared baselines without the modality bias issue.

Since the MIM is the strongest attacker among the three methods, we provide audio-visual defense results against the MIM attacker on the three different datasets in Table 3 to further demonstrate the effectiveness of our audio-visual defense method. We can see that our method can still improve audio-visual model robustness against the powerful MIM attacker, and it outperforms all of the compared approaches in terms of the RI on the MIT-MUSIC and AVE. The results further demonstrate that our defense method can generalize to different datasets and defend against different attackers. Moreover, we can find that the two models: Unimodal V and PCL, achieve lower performance on the Kinetics-Sounds for clean audio and visual inputs due to the modality bias problem, while they achieve "good" defense results against attacks by the shortcut. The results suggest us to further punish the biased audio-visual models when

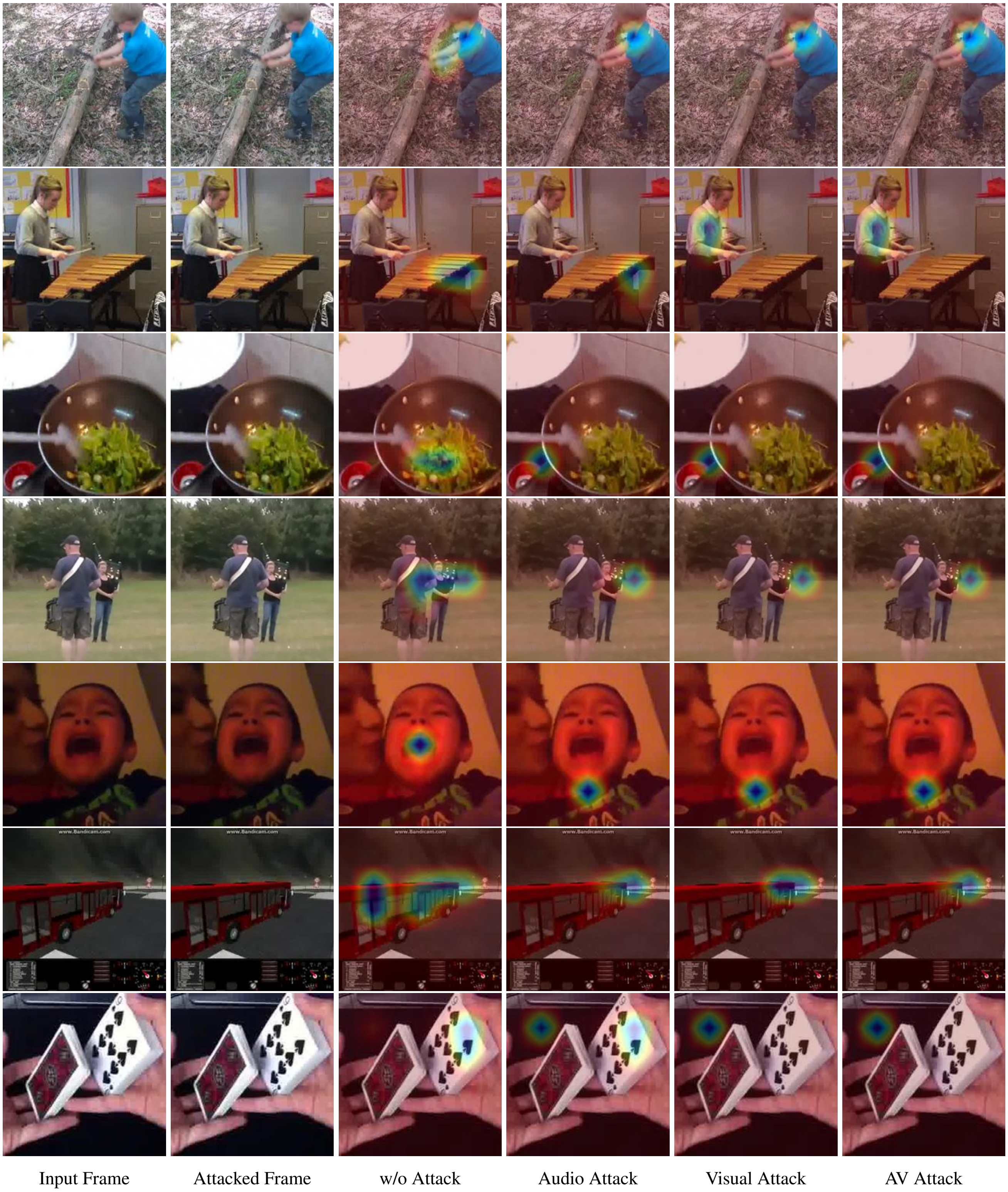


Figure 4: Visualizing sound sources under multimodal attacks. The adversarial perturbations in attacked video frames are almost imperceptible. Both single-modality and audio-visual attacks can successfully fool the weakly supervised sound source visual localization model without using sounding object location supervision.

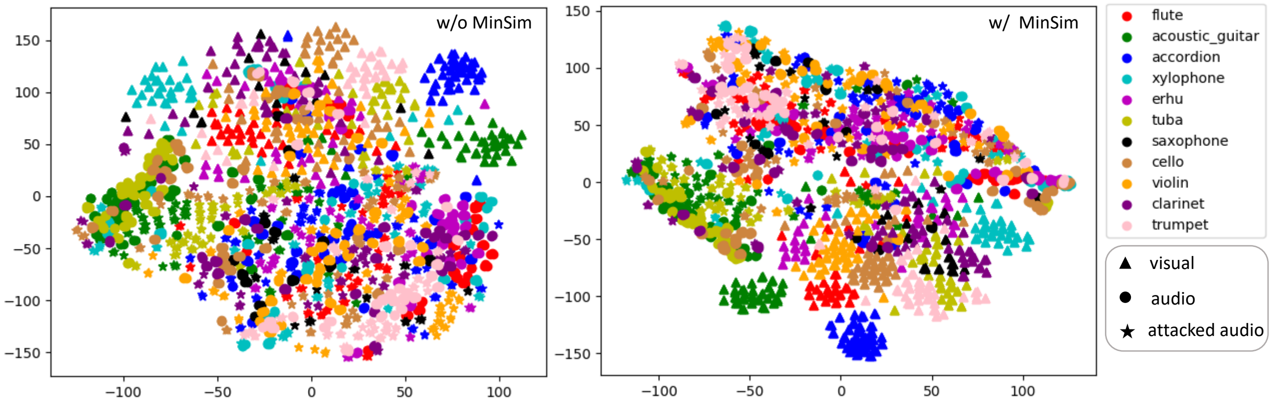


Figure 5: t-SNE visualizations of audio (clean and attacked) and visual embeddings from w/o MinSim and w/ MinSim models on the MIT-MUSIC. We use symbols: ▲, ●, and ⋆ to denote visual, audio, attacked audio modalities, respectively. Different colors refer to different categories. Our MinSim model can learn more intra-class compact and separable embeddings in separated unimodal spaces. Thus, the attacked audio samples generated by w/ MinSim are much closer to clean samples in the same categories (*e.g.*, violin, tuba, flute) than the adversarial audio examples obtained by the w/o MinSim.

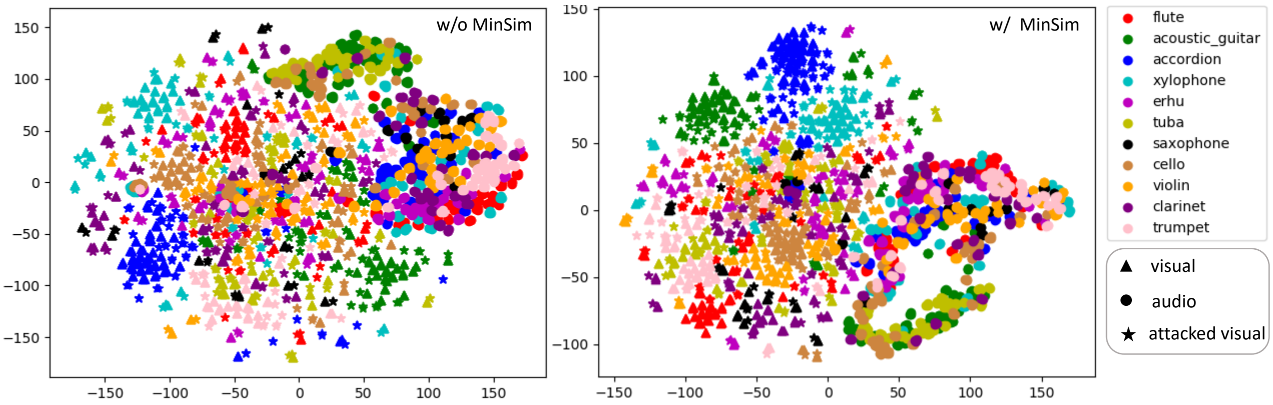


Figure 6: t-SNE visualizations of audio and visual (clean and attacked) embeddings from w/o MinSim and w/ MinSim models on the MIT-MUSIC. We use symbols: ▲, ●, and ⋆ to denote visual, audio, attacked visual modalities, respectively. Different colors refer to different categories. The attacked visual samples generated by w/ MinSim are much closer to clean samples in the same categories (*e.g.*, accordion, xylophone, and flute) than the adversarial visual examples obtained by the w/o MinSim.

we evaluate audio-visual defense methods.

We show t-SNE visualizations of both attacked audio and attacked visual embeddings from w/o MinSim and w/ MinSim in Figure 5 and Figure 6, respectively. We can see that the attacked samples generated by w/ MinSim are closer to clean samples in the same categories than the attacked sampled produced by w/o MinSim, especially for the attacked audio embedding in Figure 5, since the w/ MinSim can force our audio-visual models to strengthen multimodal dispersion and unimodal compactness. The results can further validate the effectiveness of the proposed MinSim defense mechanism.

References

[1] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9185–9193, 2018.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[4] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *AAAI*, 2018.

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[6] Aamir Mustafa, Salman Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. Adversarial defense

Defense (MUSIC)	✓AV	✗A	✗V	✗AV	Avg	RI
None	88.46	20.19	16.35	0.00	12.18	0.00
Unimodal A	59.62	0.00	59.62	0.00	19.87	-21.15
Unimodal V	81.73	81.73	11.54	11.54	34.94	16.03
PCL [6]	83.65	79.81	17.31	17.31	38.14	21.15
MaxSim	89.42	25.00	31.73	15.38	24.04	12.84
Ours (Full)	90.38	64.42	27.88	18.27	36.86	26.60
Defense (Kinetics)	✓AV	✗A	✗V	✗AV	Avg	RI
None	72.42	11.18	9.63	0.32	7.04	0.00
Unimodal A	35.99	0.10	35.99	0.10	12.06	-31.41
Unimodal V	66.08	66.08	5.77	5.77	25.87	12.49
PCL [6]	64.50	62.98	20.26	19.97	34.40	19.44
MaxSim	71.39	17.65	18.78	13.89	16.77	8.70
Ours (Full)	71.33	44.72	16.04	9.99	23.58	15.45
Defense (AVE)	✓AV	✗A	✗V	✗AV	Avg	RI
None	70.40	15.17	10.20	0.25	8.54	0.00
Unimodal A	29.85	0.00	29.85	0.00	9.95	-39.14
Unimodal V	65.17	65.17	5.22	5.22	25.20	11.43
PCL [6]	61.94	61.44	7.21	6.97	25.21	8.21
MaxSim	71.64	15.17	12.94	8.96	12.36	5.06
Ours (Full)	71.39	52.99	14.43	11.44	26.29	18.74

Table 3: Audio-visual defense against the MIM [1] attack on the MIT-MUSIC, Kinetics-Sounds, AVE datasets. Here, we use the MIM with ϵ_a, ϵ_v = 0.06 to generate audio and visual adversarial examples. Our full defense method combines the MinSim and ExFMem. Our audio-visual defense method can successfully defend against strong MIM attacks without the modality bias problem. Top-2 results are highlighted.

by restricting the hidden space of deep neural networks. In *Int. Conf. Comput. Vis.*, pages 3385–3394, 2019.

[7] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[8] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.