Can audio-visual integration strengthen robustness under multimodal attacks?

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

Abstract

In this paper, we propose to make a systematic study on machines' multisensory perception under attacks. We use the audio-visual event recognition task against multimodal adversarial attacks as a proxy to investigate the robustness of audio-visual learning. We attack audio, visual, and both modalities to explore whether audio-visual integration still strengthens perception and how different fusion mechanisms affect the robustness of audio-visual models. For interpreting the multimodal interactions under attacks, we learn a weakly-supervised sound source visual localization model to localize sounding regions in videos. To mitigate multimodal attacks, we propose an audio-visual defense approach based on an audio-visual dissimilarity constraint and external feature memory banks. Extensive experiments demonstrate that audio-visual models are susceptible to multimodal adversarial attacks; audio-visual integration could decrease the model robustness rather than strengthen under multimodal attacks; even a weakly-supervised sound source visual localization model can be successfully fooled; our defense method can improve the invulnerability of audio-visual networks without significantly sacrificing clean model performance. The source code and pre-trained models are released in https://github.com/YapengTian/AV-Robustness-CVPR21.

1. Introduction

Our daily perceptual experiences are specified by multiple cooperated senses with multisensory integration [50]. When we are talking with a person, we can learn her/his spoken words and emotions from the seen lip movements, gestures, facial expressions, and heard speech sounds. Numerous psychological and cognitive studies show that the availability of sensory inputs from several modalities ensures the robustness of the human perception system [66, 29, 75]. However, the robustness highly depends on the reliability of multisensory inputs. For our human perception system, it might fail if certain senses are attacked. For example, the McGurk effect[1] [46] indicates a perceptual illusion, which occurs when a speech sound is paired with the visual component of another sound, leading to the perception of a third speech sound.

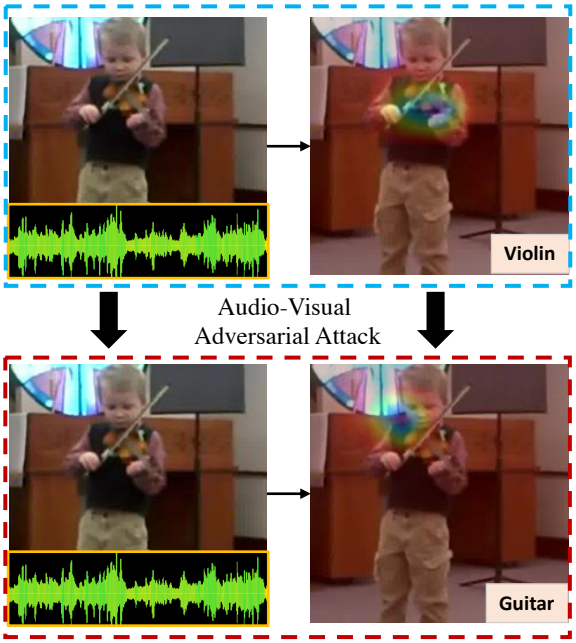


Figure 1: Adding imperceptible perturbations into audio and visual inputs by an audio-visual adversarial attack, our joint perception model predicts a wrong event class: *Guitar* and tend to localize visual regions without the sound source.

For computation models, our community indeed has devoted to develop data-driven approaches in lip reading [15, 58, 14], visually indicated sound separation [20, 25, 53, 87, 86, 81, 22], audio-visual event localization [71, 42, 77, 61, 62], audio-visual video parsing [70], audio-visual embodied navigation [9, 23], and audio-visual action recognition [28, 37, 78] to achieve robust auditory or visual perception by integrating audio and visual information. However, whether these computational perception models still

[1] https://www.youtube.com/watch?v=2k8fHR9jKVM

exhibit robustness under attacks or they are vulnerable to corrupted sensory inputs as in human perception, these have not been systematically evaluated in previous work.

Inspired by the auditory-visual illusion [46] in human perception, we present a systematic study on machines' multisensory integration under attacks. We use the audio-visual event recognition task against multimodal adversarial attacks as a proxy to investigate the robustness of audio-visual learning. Adversarial examples are generated with several different attack methods for audio, visual, and both modalities to evaluate the robustness of our models. In addition, different audio-visual fusion methods are explored to validate the correlation between model robustness and multisensory integration. To visually interpret the audio-visual interactions under attacks, we learn a weakly-supervised sound source visual localization model to localize sounding regions in videos. To mitigate the adversarial multimodal attacks, we propose an audio-visual defense method. It uses external feature memory banks to denoise corrupted features from each modality and learns compact unimodal embeddings by enforcing audio-visual dissimilarity to strengthen invulnerability. For fairly evaluating different defense approaches, we propose a relative improvement (RI) metric that considers results from both clean and attack models and can penalize modality-biased defense models. One audio-visual attack example is illustrated in Fig. 1.

Extensive experiments can validate that our audio-visual models are susceptible to adversarial perturbations, audio-visual integration could weaken model robustness rather than strengthen under multimodal attacks, even a weakly-supervised sound source visual localization model can be successfully fooled, and the proposed audio-visual defense method can improve network invulnerability without significantly sacrificing clean model performance.

The main contributions of our work are: (1) systematically investigating the robustness of audio-visual event recognition models against the adversarial multimodal attack with different attackers and fusion methods; (2) qualitatively interpreting the robustness over multimodal attacks in terms of the sound source spatial localization; (3) proposing a novel audio-visual defense method that uses clean external feature memory banks to denoise adversarial audio and visual features and enforces the multimodal dispersion and unimodal embedding compactness to strengthen invulnerability. (4) finding a shortcut of audio-visual defense originating from the modality bias issue and proposing a new evaluation metric: RI.

2. Related Work

In this section, we discuss some related work on audio-visual learning, adversarial attack, and adversarial defense.

Audio-Visual Learning: Audio and visual modalities in videos can provide synchronized and/or complementary information. The multimodal nature of videos enables a series of new and interesting audio-visual learning problems, such as self-supervised audio-visual representation learning [16, 52, 4, 54, 2, 3, 53, 40, 35], visually indicated sound separation [20, 25, 53, 87, 86, 63, 81, 27, 22, 69], vision-infused audio inpainting [89, 49], sound source spatial localization [34, 38, 64, 71, 3, 53, 59, 36, 1], lip reading [15, 58, 14], audio-visual event localization [71, 42, 77, 61, 62], audio-visual video parsing [70], audio-visual embodied navigation [23, 9], audio-visual action recognition [28, 37, 78, 76], and cross-modal generation and prediction [13, 12, 92, 10, 11, 88, 26, 24, 91, 74, 21, 90, 82]. Although the audio-visual integration with clean data facilitates many audio-visual learning tasks and strengthens model robustness, we do not know whether the robustness still exits when audio and visual modalities are attacked. In this paper, we take audio-visual event recognition as the pretext task to explore audio-visual learning robustness against multimodal adversarial attacks.

Adversarial Attack: Generating adversarial images to attack deep networks have attracted great interests. A pioneer work is proposed by Szegedy *et al*. in [68], which uses a box-constrained L-BFGS-based optimization to predict adversarial perturbations for fooling networks. Following the line of the work, many white-box (network architecture and parameters are known) attack approaches are developed to effectively attack image classifiers, including Fast Gradient Sign Method (FGSM) [30], iterative FGSM [41], DeepFool [48], Projected Gradient Descent (PGD) [45], Jacobian-based Saliency Map Attack (JSMA) [56], Carlini & Wagner's attack [6], Diverse Input Iterative Attack [80], and Momentum-based Iterative Method (MIM) [17]. Building upon research in the visual domain, recent research shows that speech recognition models are also susceptible to adversarial audio examples [5, 65, 85, 7, 60, 18]. But, how adversarial attacks affect universal sound models has not been answered yet.

Rather than individual audio and visual adversarial attacks, we investigate audio-visual learning under multimodal attacks, which generate adversarial examples for both audio and visual inputs. Particularly, we explore unconstrained video data from a range of categories (*e.g*., musical instruments and human activities).

Adversarial Defense: The adversarial defense aims to improve the invulnerability of deep models under attacks. To counter adversarial attacks, adversarial training approaches [30, 41, 73, 47] are proposed, which incorporate both clean images and their adversarial counterparts into the training process. Since it is not possible to exploit all different levels of perturbations during adversarial training, the trained models might not be able to generalize to certain unknown attacks. To mitigate adversarial attacks, some ap-

proaches [79, 32, 67] apply different pre-processing steps and transformations on the input image. There are also some defense methods that propose new objective functions [55, 51] to enforce robustness by encouraging compact representations. In the audio domain, there are only a few methods [84, 43] to alleviate adversarial attacks on speech recognition. However, they can only detect adversarial examples and are not able to improve model performance.

Not competing with state-of-the-art defense methods in the image domain, our goal is to investigate how to take the multimodal nature of audio-visual data into consideration for audio-visual defenses and devise unified defense methods, which can alleviate perturbations from both modalities.

3. Method

3.1. Multimodal Adversarial Attack

Let x_v be an input video frame, x_a be an input audio waveform, and y be the corresponding groundtruth label for the multisensory input: $\{x_a, x_v\}$. We denote $\mathcal{F}_\theta$ as our audio-visual network, where θ are the model parameters.

The goal of a multimodal attack is to fool the target multimodal model: $\mathcal{F}_\theta$ by adding human imperceptible perturbations into its inputs from multiple modalities, such as audio: x_a and visual: x_v in our problem. Since there are multiple inputs, we can divide our multimodal attack into two categories: *single-modality attacks* that only generate audio adversarial example x_a^{adv} or visual adversarial example x_v^{adv}, and *audio-visual attacks* that generate both audio and visual adversarial examples: $\{x_a^{adv}, x_v^{adv}\}$.

Adversarial Objective: To force a trained multimodal model $\mathcal{F}_\theta$ to make wrong predictions and the corresponding perturbations be as imperceptible as possible, the objective function for multimodal attacks against $\mathcal{F}_\theta$ with audio and visual inputs is as follows:

$$\begin{aligned} &\underset{x_a^{adv}, x_v^{adv}}{\operatorname{argmax}} \mathcal{L}(x_a^{adv}, x_v^{adv}, y; \theta) \\ &\text{s.t.} \quad ||x_a^{adv} - x_a||_p \leqslant \epsilon_a \\ &\qquad\quad ||x_v^{adv} - x_v||_p \leqslant \epsilon_v, \end{aligned} \tag{1}$$

where $\delta_a = x_a^{adv} - x_a$ is the audio adversarial perturbation, $\delta_v = x_v^{adv} - x_v$ is the visual adversarial perturbation, $\mathcal{L}(\cdot)$ is the loss function to optimize $\mathcal{F}_\theta$, $||\cdot||_p$ is the p-norm, and ϵ_a and ϵ_v are audio and visual perturbation budgets, respectively. With the adversarial objective, the attacker will maximize the loss function by seeking small perturbations within allowed budgets, and try to push the trained model to make incorrect predictions. For single-modality attacks, either ϵ_a or ϵ_v is 0. In this case, our multimodal model can still access clean inputs from the unattacked modality. For audio-visual attacks, both audio and visual inputs will be corrupted. With exploring effects of different single-modality and audio-visual attacks, we can investigate the model robustness under multimodal attacks.

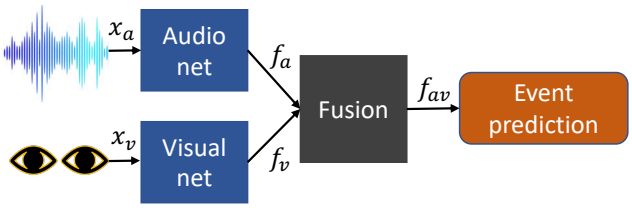


Figure 2: Audio-visual event recognition network. It integrates audio and visual content to predict the event category.

3.2. Audio-Visual Event Recognition

We use audio-visual event recognition task as a proxy to explore the audio-visual model robustness under multimodal attacks. Given an audio waveform: x_a and the corresponding video frame: x_v from a short video clip, the goal of the task is to predict the event category of the video clip. To address the problem, we introduce an audio-visual network[2] as shown in Fig. 2, which can integrate information from the both modalities to infer event labels. It uses an 1D convolution-based audio network to extract an audio feature: $f_a \in \mathcal{R}^d$ from x_a. ResNet [33] is adopted as the visual network to extract a visual feature $f_v \in \mathcal{R}^d$ from x_v. The audio and visual features are integrated by a fusion function outputting a fused feature: f_{av}. In practice, we obtain $f_{av} = [f_a; f_v]$ via concatenating the audio and visual features. Taking the f_{av} as an input, a fully-connected layer with a softmax is used to predict its event class probability p. The cross-entropy objective function: $\mathcal{L}_{CE} = -\sum_{i=1}^{k} y_i log(p_i)$, where k is the category number, is used to force the model to learn discriminative features for each class that be mapped to correct output space.

3.3. Audio-Visual Defense

To defend adversaries and improve the robustness of our audio-visual models, we propose an audio-visual defense method. It includes two parts: learning discriminative and compact unimodal embeddings and external feature memory banks for feature denoising. Next, we will describe the details of our audio-visual defense mechanism.

3.3.1 Learning Discriminative and Compact Features

Our deep models are threatened by adversarial attacks since the attackers, by maximizing the loss function, will force the output across its originally correct decision region. It has been suggested that high intra-class compactness in the feature space can strengthen the adversarial robustness of classifiers since it makes difficulties for the adversarial attackers to find feasible perturbations within its allowed budget and go beyond the correct decision boundary [55, 51].

[2] We include details of the architecture in the supplementary material.

Nevertheless, audio and visual data captured by different senses are essentially distinct. The modality gap in our multimodal task makes the encoded audio, and visual features: f_a and f_v from the same input video different, and thus leads intra-class dispersion in the joint audio-visual feature space. Consequently, our audio-visual model becomes susceptible to adversarial perturbations. To mitigate the intra-class dispersion and strengthen our model robustness, we should learn more compact audio-visual embeddings.

Audio and visual signals that contain synchronized content are ubiquitous, as demonstrated in a wide range of audio-visual tasks [16, 52, 4, 40, 71, 23]. Motivated from the nature synchronization between the two modalities, it is straightforward to alleviate the intra-class dispersion in the multimodal data by enforcing similarities between audio and visual features. Maximizing the audio-visual similarity can force the model to align the features from the two modalities and project them in a similar feature space, which will decrease the intra-class dispersion accompanying the modality gap reduction. However, the synchronization does not mean that the two modalities are identical. One reason for joint modeling is better than individual modeling is that the additional modalities can provide augmented discriminativeness rather than redundant information. Thus, the similarity constraint might weaken the power of our multimodal models since it decreases discriminative information from individual modalities. To further encourage the multimodal dispersion in the synchronized audio and visual signals, instead of maximizing, we minimize the audio-visual similarity. The objective function is formulated as:

$$\mathcal{L}_{Sim} = \frac{f_a \cdot f_v}{max(||f_a||_2 \cdot ||f_v||_2, \eta)}, \tag{2}$$

where we use the cosine similarity as the measurement and $\eta = 1e-8$ is a small scalar to avoid division by zero. Combining the cross-entropy and similarity losses, we can obtain our final objective function:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{Sim}. \tag{3}$$

With the $\mathcal{L}_{Sim}$, the model will tend to learn separated audio and visual embeddings. Meanwhile, the L_{CE} will still urge the features to be discriminative, which will implicitly encourage the both separated unimodal embeddings to be more compact and separable. In this manner, we can simultaneously strengthen the multimodal dispersion and embedding compactness to make our audio-visual model more powerful and robust.

3.3.2 External Feature Memory Bank

When audio and visual inputs are attacked, the features: f_a^{adv} and f_v^{adv} from corresponding audio and visual adversarial examples become noisy and not reliable. To further defend the attackers, we can estimate cleaner audio and visual features: f_a^* and f_v^* to replace f_a^{adv} and f_v^{adv}.

Inspired by conventional sparse representation-based image restoration approaches [19, 83], we propose to adopt external feature memory banks to denoise attacked audio and visual examples at a feature level. Since audio and visual features are reliable in training data, we use them to build audio and visual external feature memory banks: $M_a \in \mathcal{R}^{d\times K}$ and $M_v \in \mathcal{R}^{d\times K}$, respectively, where $M_a[:,k]$ and $M_v[:,k]$ are audio and visual feature vectors from the same video, and we sample totally K samples. To estimate clean features, the adversarial features are first encoded with the external feature memory banks:

$$\begin{aligned} &\min_{\alpha_a} ||f_a^{adv} - M_a\alpha_a||_2^2 + \lambda_a||\alpha_a||_1, \\ &\min_{\alpha_v} ||f_v^{adv} - M_v\alpha_v||_2^2 + \lambda_v||\alpha_v||_1, \end{aligned} \tag{4}$$

where the parameters: λ_a and λ_v balance sparsity of the solutions and fidelity of the approximation, and α_a and α_v are predicted audio and visual coefficients, respectively. Then, the more reliable audio and visual features can be reconstructed by the corresponding encoded coefficients: $f_a^* = M_a\alpha_a$ and $f_v^* = M_v\alpha_v$. We solve the Lasso [72] problems in Eq. 4 using the differentiable Iterative Shrinkage Thresholding Algorithm (ISTA) [31].

With the discriminative, compact, and cleaner audio and visual embeddings, our audio-visual model will be more invulnerable to potential multimodal adversarial attacks.

4. Experiments

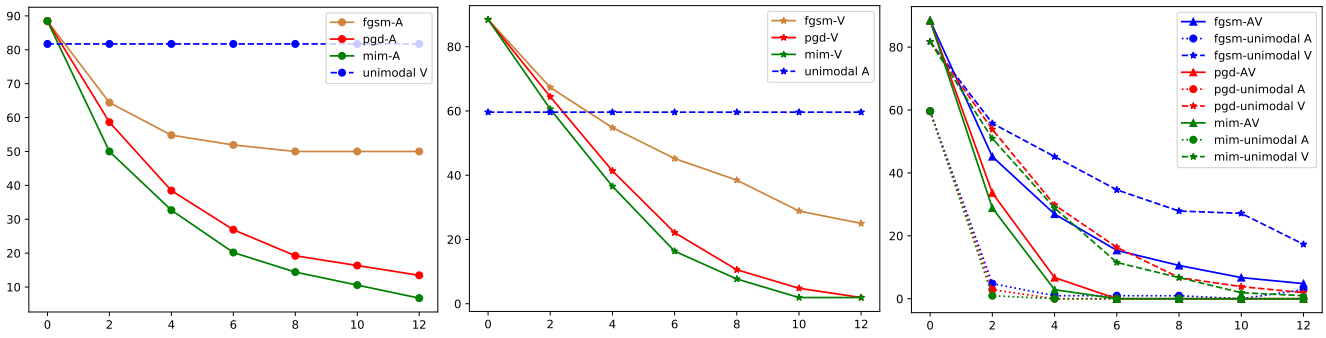
4.1. Datasets

We use two widely used audio-visual datasets: MIT-MUSIC and Kinetics-Sounds for training and evaluation.

MIT-MUSIC: This dataset [87] contains clean audio-visual synchronized musical recordings, which covers 11 instrument categories: accordion, acoustic guitar, cello, clarinet, erhu, flute, saxophone, trumpet, tuba, violin, and xylophone. 520 available videos with solos in the dataset are used to conduct experiments. We randomly divide the data into trian/val/test splits of 312/104/104 videos, respectively.

Kinetics-Sounds: The dataset is a subset of the Kinetics dataset [8], which contains YouTube videos with manually annotated human actions. This subset[3] contains 15516 10 second video clips (9309 training, 3104 validation, 3103 test) in 27 human action categories. Rather than only musical instruments, it includes diverse human activities (*e.g.*, chopping wood, ripping paper, tap dancing, and singing). Besides the diversity of scenes, Kinetics-Sounds is more noisy than the MIT-MUSIC, in which audio and visual content inside some videos might not be completely related.

[3]Kinetics-Sounds is firstly used in [2]. Since some videos in the subset are not available on the Internet, the downloaded dataset is slightly smaller.



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

Figure 3: Adversarial robustness against multimodal attacks on the MIT-MUSIC. 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. But, 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).

4.2. Attack Methods

We evaluate the audio-visual model robustness with l_∞-bounded adversarial perturbations, which is widely used as a standard evaluation metric for adversarial robustness [45]. Three different attack methods are used.

FGSM: The fast gradient sign method (FGSM) [30] computes the gradients of the network to generate adversarial examples x_{adv} by $x^{adv} = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x, y; \theta))$, where x_{adv} is the generated adversarial example, x is the original input, y is the original label, θ refers to model parameters, ϵ is the maximum adversarial perturbation value, and $\mathcal{L}$ is the loss function. For our audio-visual model, we can obtain audio and visual adversarial examples: x^a_{adv} and x^v_{adv} in terms of $x^{adv}_a = x_a + \epsilon_a \cdot \text{sign}(\nabla_{x_a} \mathcal{L}(x_a, x_v, y; \theta))$ and $x^{adv}_v = x_v + \epsilon_v \cdot \text{sign}(\nabla_{x_v} \mathcal{L}(x_a, x_v, y; \theta))$, respectively.

PGD: Projected Gradient Descent (PGD) [45] is an iterative variant of the FGSM. We can perform multi-step attacks based on PGD and generate audio and visual adversarial examples with respect to ϵ_a and ϵ_v, respectively.

MIM: Momentum-based Iterative Method (MIM) [17] integrates a momentum term into the iterative process to further stabilize update directions and mitigate local minima.

4.3. Model Robustness under Multimodal Attacks

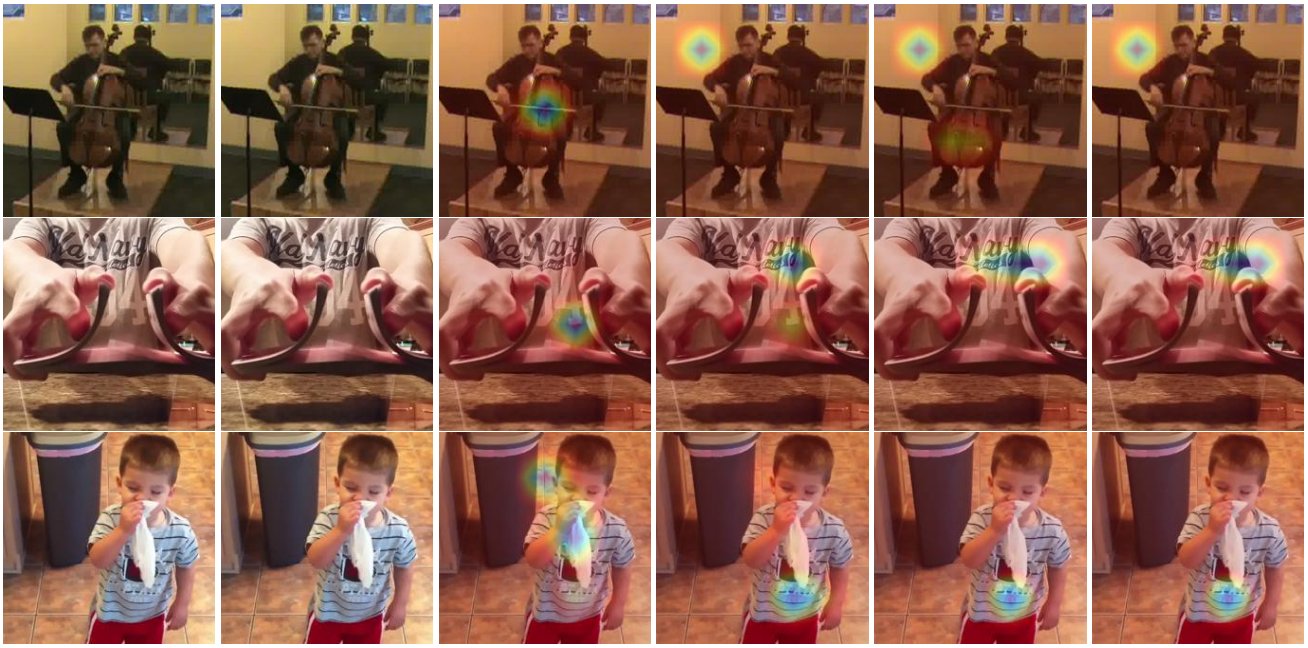
We first investigate the model robustness of audio-visual event recognition under multimodal adversarial attacks. Table 1 shows audio-visual event recognition accuracy on MIT-MUSIC and Kinetics-Sounds datasets under both single-modality and audio-visual attacks with different attackers. To better interpret the multimodal robustness, we also include results from two baselines: Unimodal A and Unimodal V, which are two single-modality models and only use audio and visual modalities, respectively. Clearly, all of the three attack methods: FGSM, PGD, and MIM can significantly decrease recognition results, and the MIM achieves the lowest accuracy under different multimodal attacks. The results show that audio-visual models are susceptible to multimodal adversarial attacks, and the MIM is the most effective attack method among the three attackers.

Dataset	Attack	✓AV	✗A	✗V	✗AV	Avg.	Unimodal ✓A	Unimodal ✓V
MM	FGSM [30]	88.46	50.00	25.00	15.38	30.12	59.62	81.73
	PGD [45]		13.46	**1.92**	**0.00**	5.09		
	MIM [17]		**6.73**	**1.92**	**0.00**	**2.88**		
KS	FGSM [30]	72.42	33.38	15.08	8.18	18.88	35.99	66.08
	PGD [45]		6.22	1.90	0.77	2.96		
	MIM [17]		**3.87**	**1.55**	**0.32**	**1.91**		

Table 1: Audio-visual event recognition accuracy on MIT-MUSIC and Kinetics-Sounds datasets under different attack methods. ✗A, ✗V, and ✗AV denote that only audio, only visual, and both audio and visual inputs for our audio-visual network are attacked, respectively. We set ϵ_a and ϵ_v as 0.12 respectively for ✗A and ✗V, and 0.06 for ✗AV. The symbol: ✓ means that inputs are clean. The baselines: Unimodal ✓A and Unimodal ✓V models are two single-modality models.

From Table 1, we can also see that our clean audio-visual models (✓AV) are better than both clean single-modality A (Unimodal ✓A) and V (Unimodal ✓V) models, which can validate that audio-visual integration can strengthen perception robustness and improve audio-visual event recognition performance when input modalities are clean and reliable. But, the conclusion might not hold if the audio-visual model is attacked. Next, we will analyze it based on multimodal attack results.

Single-Modality Attack: When we use different attackers to perform single-modality attacks on the MIT-MUSIC and Kinetics-Sounds datasets, audio-visual models: ✗A and ✗V are always inferior to Unimodal ✓V and Unimodal ✓A, respectively. For example, the performances drop 91.76% and



Input Frame | Attacked Frame | w/o Attack | Audio Attack | Visual Attack | AV Attack

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.

96.77% on MIT-MUSIC with the MIM attack. Note that ✗A and ✗V have clean visual and audio modalities, respectively. The results can demonstrate that audio-visual integration could weaken event recognition performance, when audio or visual inputs are attacked.

Audio-Visual Attack: Obviously, when inputs from the both modalities are added adversarial perturbations, the audio-visual models: ✗AV obtain even worse performance than the ✗A and ✗V. When we compare it to attacked unimodal models (see Uimodal A and Unimodal V in Table 3), we can see that ✗A of Unimodal A and ✗V of Unimodal V achieve 0.00% and 11.54%, while ✗AV of the audio-visual model is 15.38% under the same FGSM attack on the MIT-MUSIC. Interestingly, the audio-visual model is more invulnerable than the unimodal models against attacks. But when we compare the results from ✗AV of the audio-visual model and ✗V of Unimodal V on the Kinetics-Sounds, joint perception under the audio-visual attack is worse than the visual perception under the single-modality attack. These results validate that one corrupted modality could still help the other modality, but a joint perception is not always better than individual perceptions under audio-visual attacks.

Figure 3 illustrates the adversarial robustness against multimodal attacks with different perturbations. The results can further validate our findings that audio-visual integration may not always strengthen the audio-visual model robustness under multimodal adversarial attacks. The adversarial robustness of the audio-visual models highly depends on the reliability of the multisensory inputs.

Method	✓AV	✗A	✗V	✗AV	Avg.
Sum	88.46	35.58	45.19	3.85	43.27
Concat	88.46	**51.92**	**45.19**	**15.38**	**50.24**
FiLM [57]	83.65	28.85	39.42	3.85	38.95
Gated-Sum [39]	**89.42**	33.65	44.23	4.81	43.03
Gated-Concat [39]	**89.42**	45.19	43.27	13.46	47.84

Table 2: Audio-visual event recognition accuracy with different fusions on the MIT-MUSIC under FGSM attacks.

4.4. Audio-Visual Fusions Against Attacks

Audio-visual fusion strategy is important for the performance of our multimodal model. Here, we are curious about whether different audio-visual fusions would also affect the adversarial robustness. To answer the question, we compare several different audio-visual fusion approaches: Sum, Concatenation (Concat), FiLM [57], Gated Sum (Gated-Sum) [39], and Gated Concatenation (Gated-Concat) [39], where FiLM and Gated-Sum mix updated audio and visual information together as the Sum before the final prediction layer and the Gated-Concat still preserve the individual information as the Concat. Table 2 show audio-visual event recognition results with different fusion methods against FGSM attacks on the MIT-MUSIC dataset.

From Table 2, we can find that our audio-visual models with Sum, Concat, Gated-Sum, and Gated-Concat fusion mechanisms achieve competitive performance on attack-

free inputs, and FiLM is worse than the other fusion approaches; audio-visual models: ✗A and ✗V with different fusions achieve inferior performance than Unimodal ✓V and Unimodal ✓A, respectively. The results further support that audio-visual integration could decrease event recognition performance when input audio or visual modalities are not reliable. Another interesting observation is that the Concat and Gated-Concat are much better than the Sum, FiLM, and Gated-Sum under audio-visual attacks, and Concat is the most robust fusion among the compared methods. From the results, we can learn that more audio-visual interactions inside the fusion function might weaken the audio-visual model robustness against the audio-visual attacks.

4.5. Visualizing Sound Sources Under Attacks

To visually interpret the audio-visual interactions under multimodal adversarial attacks, we visualize sound sources in video frames. To localize sound sources, we train a weakly-supervised sound source visual localization network. It uses audio-visual event recognition as the pretext task and adopts an audio-guided visual attention mechanism similar to [71, 64] as the localization module. Concretely, we obtain a $N \times N$ visual feature map:$F_v = [f_v^1; ...; f_v^{N^2}] \in \mathcal{R}^{N^2 \times d}$ from an input frame: x_v the ResNet [33]. Given the audio feature vector: f_a and F_v, we compute audio-guided visual attention weights for each spatial position: $w_i = \frac{exp(f_a^T f_v^i)}{\sum_j exp(f_a^T f_v^j)}$ and obtain the attended visual feature $f_v^{att} = \sum_i w_i f_v^i$ to replace f_v in the original audio-visual event recognition network. With optimization, the model will force the attention weights to learn to localize sounding visual regions. Figure 4 illustrates attacked frames and localized sound sources under attacks.

Without attacks, we can see that our localization model can successfully discover the corresponding sounding regions for different events: *playing cello*, *shuffling cards*, and *blowing noise*. From the generated adversarial frames, we can not find perceptible perturbations. But, the model with the attacked frames fails to localize sound sources. Similarly, the model is fooled by the audio and audio-visual attack. The results demonstrate that weakly-supervised sound source localization models can be attacked even without requiring access to any localization losses for an attacker.

4.6. Audio-Visual Defense vs. Multimodal Attacks

Baselines: To validate the effectiveness of the proposed audio-visual defense mechanism, we compare it with several baselines: 1) None: audio-visual network without defense; 2) Unimodal A: audio-only network; 3) Unimodal V: visual-only network; 4) PCL [51]: a recent state-of-the-art adversarial defense approach, which uses a prototype conformity loss to enforce intra-class compactness and an inter-class separation; 5) MaxSim: maximizing audio-visual similarity using the $1-\mathcal{L}_{Sim}$ as a loss term to enforce intra-class compactness of joint audio-visual embeddings; 6) MinSim: the proposed dissimilarity constraint to encourage multimodal dispersion and unimodal compactness; 7) ExFMem: the proposed external feature memory banks; 8) MinSim+ExFMem: our full defense model.

Defense (MUSIC)	✓AV	✗A	✗V	✗AV	Avg	RI
None	88.46	51.92	45.19	15.38	37.50	0.00
Unimodal A	59.62	0.00	59.62	0.00	19.87	-46.47
Unimodal V	81.73	**81.73**	11.54	11.54	34.94	-9.29
PCL [51]	83.65	**81.73**	37.50	36.54	51.91	9.60
MaxSim	89.42	52.88	45.19	31.73	43.27	6.73
MinSim	**91.35**	70.19	46.15	36.54	50.96	16.35
ExFMem	89.42	53.85	50.00	20.19	41.34	4.80
MinSim+ExFMem	90.38	73.08	**53.85**	**42.31**	**56.41**	**20.83**
Defense (Kinetics)	✓AV	✗A	✗V	✗AV	Avg.	RI
None	72.42	36.40	26.35	8.09	23.61	0.00
Unimodal A	35.99	1.87	35.99	1.87	13.24	-46.80
Unimodal V	66.08	**66.08**	18.72	18.72	34.50	4.55
PCL [51]	64.50	63.43	29.28	**28.67**	**40.46**	8.93
MaxSim	71.39	34.95	29.57	21.46	28.66	4.02
MinSim	70.88	52.42	28.12	21.62	34.05	8.99
ExFMem	**72.71**	41.56	29.93	10.44	27.31	3.99
MinSim+ExFMem	71.33	55.96	**30.57**	24.90	37.14	**12.44**

Table 3: Audio-visual event recognition accuracy on the MIT-MUSIC and Kinetics-Sounds with different defense methods. Here, we use the FGSM (ϵ_a, $\epsilon_v = 0.06$) 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. To better evaluate the robustness of our multisensory defense models, we need to consider model performance on both clean and attacked data and the potential modality bias issue. Top-2 results are highlighted.

Evaluation Metrics: To evaluate the performance of different defense methods, we use recognition accuracy as the metric. Results from both the clean model: ✓AV and attacked models: ✗A, ✗V, and ✗AV are computed. Since there are multiple defense results under multimodal attacks for a single model, we also use the averaged accuracy:

$$\text{Avg} = \tfrac{1}{3}(\text{✗A} + \text{✗V} + \text{✗AV}),$$

as an overall metric to evaluate different defenses. However, the metric might not be able to fully reflect the effectiveness of different audio-visual defense methods. For the audio-visual defense, there is a possible shortcut due to the modality bias issue. An audio-visual defense model might mainly make use of information from one dominant modality. If so, the attacks on another modality will not much affect performance, which might make the defense method achieve pretty good results in terms of the Avg. However, the biased audio-visual defense model fails to joint perception and its ✓AV will achieve worse performance. To address the issue, we propose a relative improvement (RI) metric:

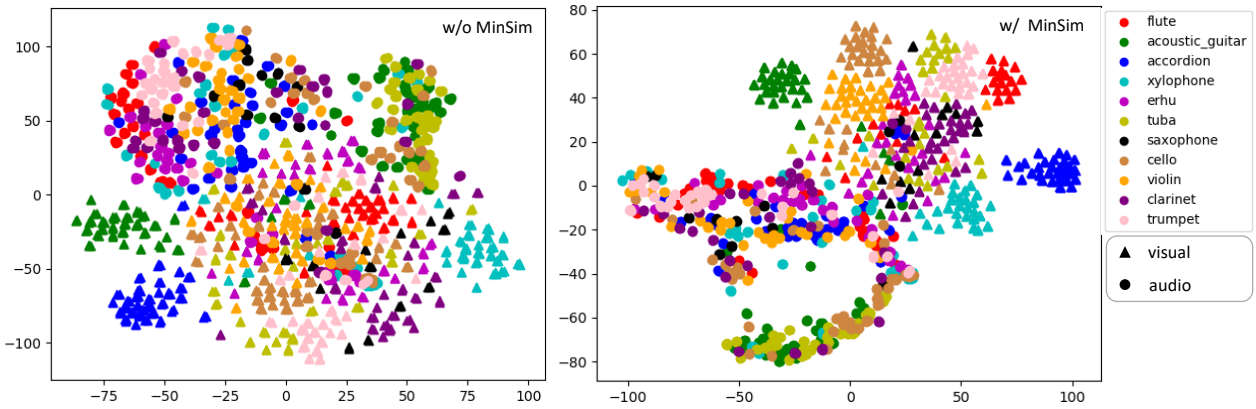


Figure 5: t-SNE visualizations of audio and visual embeddings from w/o MinSim and w/ MinSim models on the MIT-MUSIC. We use symbols: ▲ and ● to denote visual and audio modalities, respectively. Different colors refer to different categories. Our MinSim model can learn more intra-class compact and separable embeddings in separated unimodal spaces.

$$\text{RI} = (\checkmark\text{AV}_m + \text{Avg}_m) - (\checkmark\text{AV}_n + \text{Avg}_n),$$

where we consider results from both clean and attacked models, and the m refers to a defense method and n refers to a base model, which is the baseline: None in our experiments. If a defense method decreases clean model performance, the RI will penalize it accordingly.

Results: Table 3 shows defense results of different methods on the MIT-MUSIC and Kinetics-Sound. Although the single-modality model: Unimodal A is not affected by the visual attack, it achieves worse results on the ✓AV and ✗A. We can obtain a similar observation from another modality-biased defense model: Unimodal V. The both defense methods fail to improve robustness on the MIT-MUSIC dataset.

Interesting results are from the recent defense method: PCL. We can find that the PCL is almost invulnerable to audio attacks (see ✓AV vs. ✗A and ✗V vs. ✗AV) and can also improve the model robustness under visual and audio-visual attacks. From the observation, we can learn that the PCL is a visual-biased defense model. Although the PCL can achieve good results in terms of Avg and even RI, it fails to learn an effective multimodal model. The results further remind us to consider both the modality issue and defense results when we evaluate audio-visual defense methods.

The MaxSim can achieve better performance against audio-visual attacks, however, it is limited in handling single-modality attacks. The results validate that the MaxSim fails to learn compact and powerful unimodal audio and visual embeddings. Compared to the MaxSim, our MinSim is overall more robust against both single-modality and audio-visual attacks. Adding the external feature memory bank, the performance of our defense model is further improved. From the results, we can see that our full defense model outperforms all the compared methods on the RI and can achieve comparable or even better clean model performance than the base model.

To further validate our MinSim defense, we show t-SNE [44] visualizations of learned audio and visual embeddings from w/o MinSim and w/ MinSim in Fig. 5. We can see that our MinSim model learns more intra-class compact and inter-class separable embeddings (especially for the visual) in separated unimodal feature spaces.

5. Conclusion and Future Work

In this paper, we investigate the audio-visual model robustness under multimodal attacks. We cast multimodal attacks into two different categories: single-modality attacks and audio-visual attacks. Using the audio-visual event recognition task as a proxy with different fusion and attack methods, we find that audio-visual integration does not always strengthen the perception robustness under multimodal attacks, and it could even decrease performance when the input modalities are not reliable.

We use the human perception system as a guidance to help us develop computational models. However, there are indeed gaps between AV models and the real perception system and our research is limited by existing learning tools. Humans can perceive events from single modalities when the other modalities are missing. However, our study shows that AV models are susceptible to attacks since they try to exploit information from both modalities fully. Considering the observation and our results, a promising future direction is to design robust AV models that can perform attacked modality-aware predictions.

Acknowledgement: We would like to thank the anonymous reviewers for the constructive comments. This work was supported in part by NSF 1741472, 1813709, and 1909912. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *ECCV*, 2020. 2

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Int. Conf. Comput. Vis.*, pages 609–617, 2017. 2, 4

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Eur. Conf. Comput. Vis.*, pages 435–451, 2018. 2

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900, 2016. 2, 4

[5] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 513–530, 2016. 2

[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 2

[7] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE, 2018. 2

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6299–6308, 2017. 4

[9] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc, Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Eur. Conf. Comput. Vis.*, 2020. 1, 2

[10] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018. 2

[11] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019. 2

[12] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357, 2017. 2

[13] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017. 2

[14] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3444–3453. IEEE, 2017. 1, 2

[15] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016. 1, 2

[16] Virginia R de Sa. Learning classification with unlabeled data. In *NIPS*, pages 112–119, 1994. 2, 4

[17] 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, 5, 13, 14

[18] Tom Dörr, Karla Markert, Nicolas M Müller, and Konstantin Böttinger. Towards resistant audio adversarial examples. In *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, pages 3–10, 2020. 2

[19] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006. 4

[20] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *TOG*, 2018. 1, 2

[21] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *Eur. Conf. Comput. Vis.*, 2020. 2

[22] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020. 1, 2

[23] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707. IEEE, 2020. 1, 2, 4

[24] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. *arXiv preprint arXiv:2005.01616*, 2020. 2

[25] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 1, 2

[26] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 2

[27] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019. 2

[28] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. *arXiv preprint arXiv:1912.04487*, 2019. 1, 2

[29] Bryan Gick and Donald Derrick. Aero-tactile integration in speech perception. *Nature*, 462(7272):502–504, 2009. 1

[30] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 5, 13

[31] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning*, pages 399–406, 2010. 4

[32] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017. 3

[33] 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. 3, 7, 12

[34] John R Hershey and Javier R Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Advances in neural information processing systems*, pages 813–819, 2000. 2

[35] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9248–9257, 2019. 2

[36] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. In *Advances in Neural Information Processing Systems*, 2020. 2

[37] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5492–5501, 2019. 1, 2

[38] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 88–95. IEEE, 2005. 2

[39] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *AAAI*, 2018. 6, 12

[40] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, pages 7763–7774, 2018. 2, 4

[41] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 2

[42] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2002–2006. IEEE, 2019. 1, 2

[43] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Detecting adversarial attacks on audio-visual speech recognition. *arXiv preprint arXiv:1912.08639*, 2019. 3

[44] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8

[45] 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. 2, 5, 13

[46] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976. 1, 2

[47] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training. *arXiv preprint arXiv:1507.00677*, 2015. 2

[48] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 2

[49] Giovanni Morrone, Daniel Michelsanti, Zheng-Hua Tan, and Jesper Jensen. Audio-visual speech inpainting with deep learning. *arXiv preprint arXiv:2010.04556*, 2020. 2

[50] Micah M Murray and Mark T Wallace. *The neural bases of multisensory processes*. CRC Press, 2011. 1

[51] Aamir Mustafa, Salman Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. Adversarial defense by restricting the hidden space of deep neural networks. In *Int. Conf. Comput. Vis.*, pages 3385–3394, 2019. 3, 7, 13, 14

[52] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011. 2, 4

[53] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. *European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[54] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Eur. Conf. Comput. Vis.*, pages 801–816. Springer, 2016. 2

[55] Tianyu Pang, Kun Xu, Yinpeng Dong, Chao Du, Ning Chen, and Jun Zhu. Rethinking softmax cross-entropy loss for adversarial robustness. *arXiv preprint arXiv:1905.10626*, 2019. 3

[56] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016. 2

[57] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 6, 12

[58] Stavros Petridis, Yujiang Wang, Zuwei Li, and Maja Pantic. End-to-end multi-view lipreading. *arXiv preprint arXiv:1709.00443*, 2017. 1, 2

[59] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[60] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International Conference on Machine Learning*, pages 5231–5240. PMLR, 2019. 2

[61] Janani Ramaswamy. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In *ICASSP*, pages 4372–4376. IEEE, 2020. 1, 2

[62] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2970–2979, 2020. 1, 2

[63] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *ICASSP 2019-2019 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2357–2361. IEEE, 2019. 2

[64] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 2, 7

[65] Liwei Song and Prateek Mittal. Poster: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 2583–2585, 2017. 2

[66] William H Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954. 1

[67] Bo Sun, Nian-hsuan Tsai, Fangchen Liu, Ronald Yu, and Hao Su. Adversarial defense by stratified convolutional sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11447–11456, 2019. 3

[68] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[69] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[70] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: weakly-supervised audio-visual video parsing. In *Eur. Conf. Comput. Vis.*, 2020. 1, 2

[71] 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. 1, 2, 4, 7, 12

[72] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. 4

[73] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 2

[74] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Semantic object prediction and spatial sound super-resolution with binaural sounds. *arXiv preprint arXiv:2003.04210*, 2020. 2

[75] Katharina von Kriegstein. A multisensory perspective on human auditory communication. *The neural bases of multisensory processes*, pages 683–702, 2012. 1

[76] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12695–12705, 2020. 2

[77] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[78] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 1, 2

[79] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017. 3

[80] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2730–2739, 2019. 2

[81] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 882–891, 2019. 1, 2

[82] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[83] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010. 4

[84] Zhuolin Yang, Bo Li, Pin-Yu Chen, and Dawn Song. Characterizing audio adversarial examples using temporal dependency. *arXiv preprint arXiv:1809.10875*, 2018. 3

[85] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 103–117, 2017. 2

[86] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1735–1744, 2019. 1, 2

[87] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018. 1, 2, 4

[88] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9299–9306, 2019. 2

[89] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In *Int. Conf. Comput. Vis.*, pages 283–292, 2019. 2

[90] Hang Zhou, Yasheng Sun, Wu Wayne, Chen Change Loy, Xiaogang Wang, and Liu Ziwei. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[91] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *European Conference on Computer Vision*, pages 52–69. Springer, 2020. 2

[92] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3550–3558, 2018. 2

Appendix

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 [71] 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 [71] to further validate our findings. It consists of 4,143 unconstrained videos spanning 28 event categories. As in [71], 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 [33] to extract visual features and a 1-D convolution-based model to extract audio features from input audio waveforms.

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 6. The network consists of 8 convolutional layers in 4 building blocks.

Visual Net: We use the ResNet18 [33] 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 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: $f_{av} = f_a + f_v$.

```
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 6: A Pytorch implemenation of our audio network.

Concat: The Concat: $f_{av} = [f_a; f_v]$ concatenates the audio and visual features.
FiLM: The FiLM [57] 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 [39] uses audio and visual features to compute two gates to fuse feature from the other modality, respectively. They can be computed as:

$$f_1 = \sigma(f_a) \cdot f_v, \tag{5}$$
$$f_2 = \sigma(f_v) \cdot f_a, \tag{6}$$

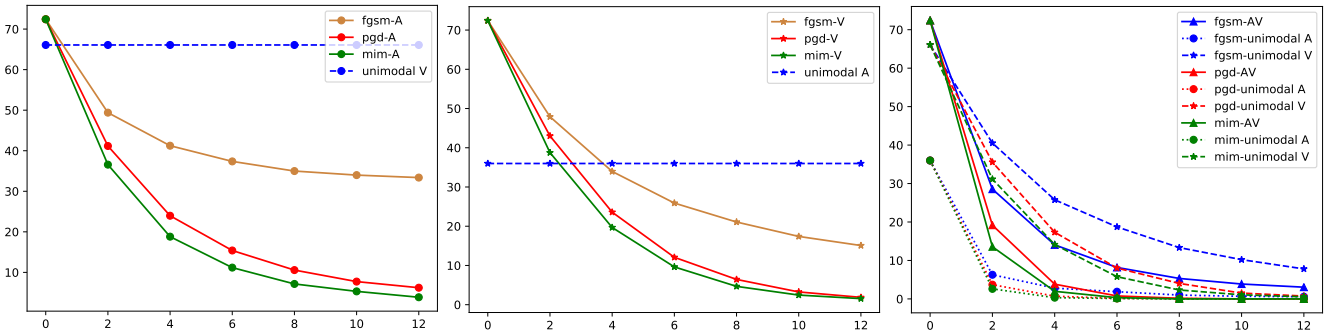
where the $\sigma(\cdot)$ is the Sigmoid function. The fused features: f_1 and f_2 are then combined by the Sum: $f_{av} = f_1 + f_2$.
Gated-Concat: The Gated-Concat is similar to the the Gated-Sum. It also computes f_1 and f_2. But, it fuses by a concatenation: $f_{av} = [f_1; f_2]$.

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

Dataset	Attack Methods	✓AV	✗A	✗V	✗AV	Avg.	Unimodal ✓A	Unimodal ✓V
AVE	FGSM [30]	70.40	40.55	24.88	8.71	24.71	29.85	65.17
	PGD [45]		20.15	11.44	1.99	11.19		
	MIM [17]		15.17	10.20	0.25	8.54		

Table 4: 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 7: 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).

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 [45] and MIM [17], 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.

B. Experimental Results

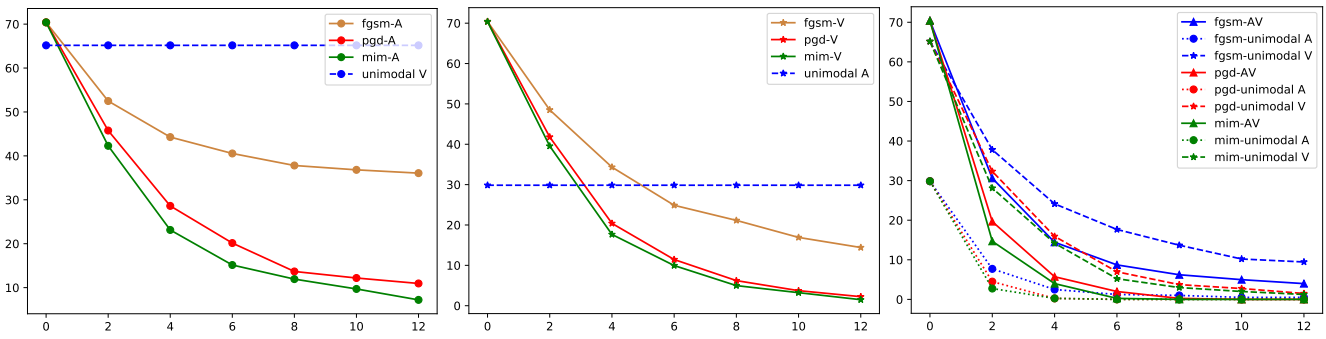
To further validate our findings, we show more experimental results on audio-visual model robustness under multimodal attacks in Sec. B.1, sound source localization under attacks in Sec. B.2, and audio-visual defense in Sec. B.3.

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 4. 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 at-

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 [51]	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 5: 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.



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

Figure 8: 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.

tack strengths on the Kinetics-Sounds and AVE in Figure 7 and Figure 8, 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 9. 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 5. 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 6 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

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 [51]	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 [51]	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 [51]	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 6: Audio-visual defense against the MIM [17] 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.

results against attacks by the shortcut. The results suggest us to further punish the biased audio-visual models when 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 10 and Figure 11, 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 10, since the w/ Min-

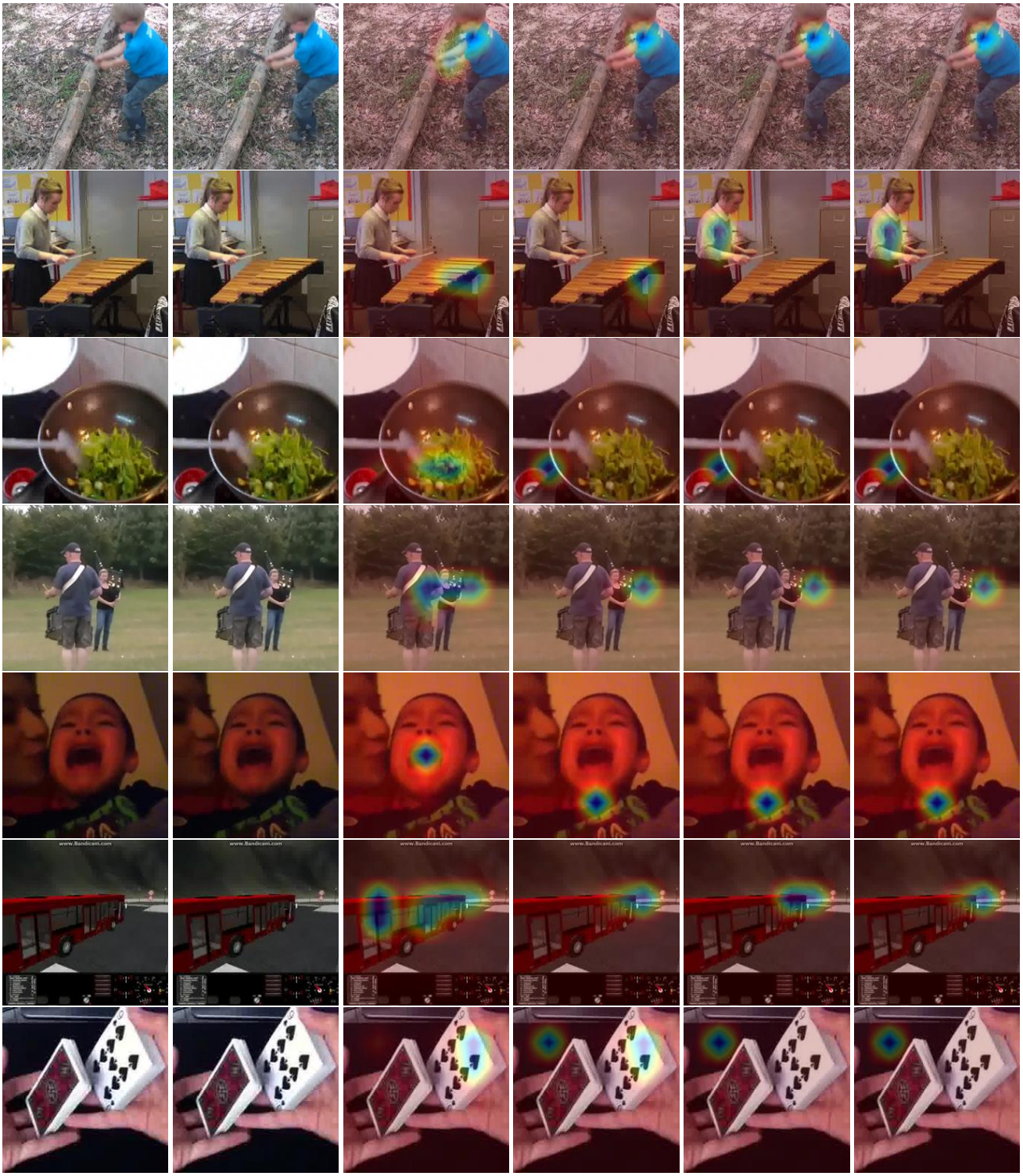


Figure 9: Visualizing sound sources under multimodal attacks. The adversarial perturbations in attacked video frames are almost imperceptible. Both single-modality and audio-visual attacks can successfully fool the weakly supervised sound source visual localization model without using sounding object location supervision.

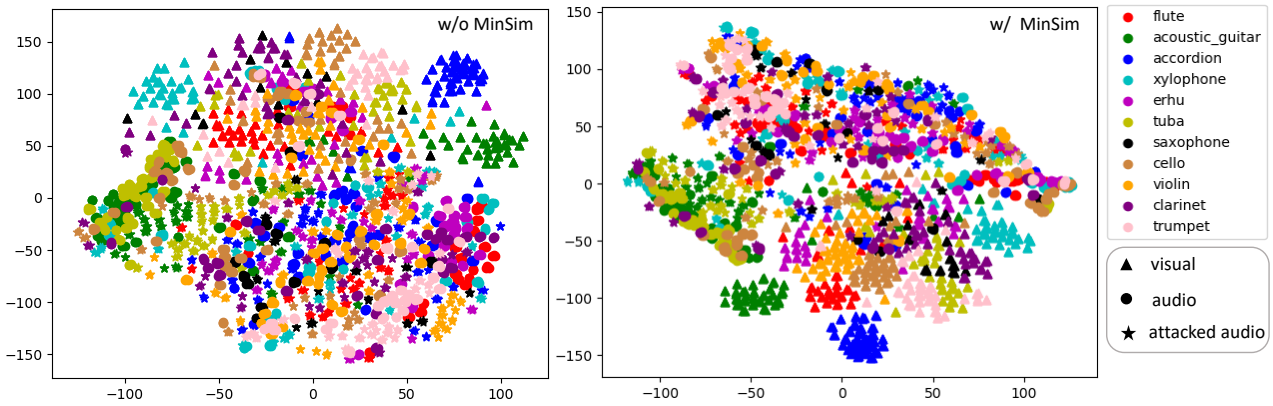


Figure 10: t-SNE visualizations of audio (clean and attacked) and visual embeddings from w/o MinSim and w/ MinSim models on the MIT-MUSIC. We use symbols: ▲, ●, and ⋆ to denote visual, audio, attacked audio modalities, respectively. Different colors refer to different categories. Our MinSim model can 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.

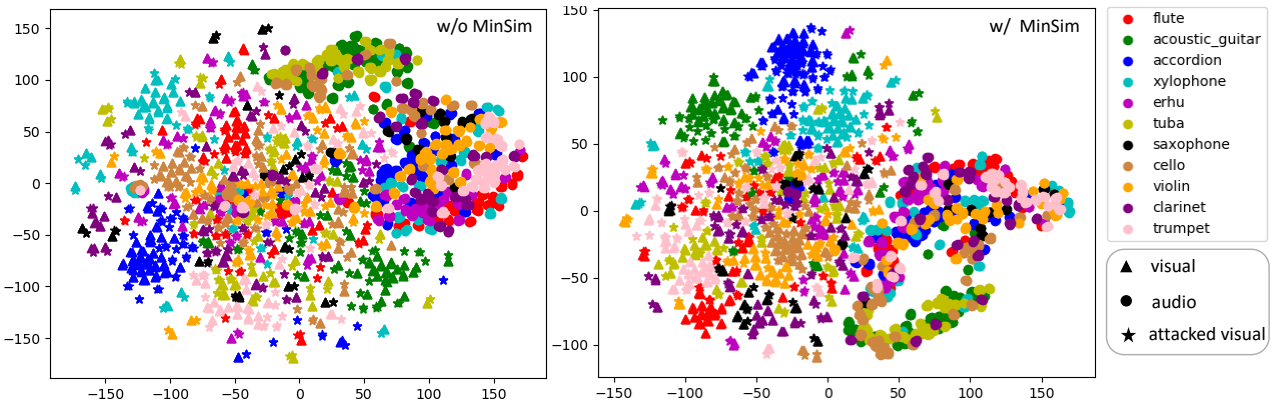


Figure 11: t-SNE visualizations of audio and visual (clean and attacked) embeddings from w/o MinSim and w/ MinSim models on the MIT-MUSIC. We use symbols: ▲, ●, and ⋆ to denote visual, audio, attacked visual modalities, respectively. Different colors refer to different categories. The attacked visual 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.

Sim 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.