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 [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.

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, 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...
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 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 exists 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 \(F_\theta\) as our audio-visual network, where \(\theta\) are the model parameters.

The goal of a multimodal attack is to fool the target multimodal model: \(F_\theta\) 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 \(F_\theta\) to make wrong predictions and the corresponding perturbations be as imperceptible as possible, the objective function for multimodal attacks against \(F_\theta\) with audio and visual inputs is as follows:

\[
\arg\max_{x_a^{adv}, x_v^{adv}} L(x_a^{adv}, x_v^{adv}, y; \theta)
\]

\[
\text{s.t. } ||x_a^{adv} - x_a||_p \leq \epsilon_a
\]

\[
||x_v^{adv} - x_v||_p \leq \epsilon_v,
\]

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, \(L()\) is the loss function to optimize \(F_\theta\), \(||\cdot||_p\) is the \(p\)-norm, and \(\epsilon_a\) and \(\epsilon_v\) 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 \(\epsilon_a\) or \(\epsilon_v\) 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-modal attacks, we can investigate the model robustness under multimodal attacks.

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 \(x_a\), and visual features: \(f_v \in \mathcal{R}^d\) 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: \(L_{CE} = -\sum_{i=1}^{k} y_i \log(p_i)\), 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 in-class dispersion in the joint audio-visual feature space. Consequently, our audio-visual model becomes susceptible to adversarial perturbations. To mitigate the in-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 in-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 in-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:

\[
\mathcal{L}_{\text{Sim}} = \frac{f_a \cdot f_v}{\max(||f_a||_2 \cdot ||f_v||_2, \eta)},
\]

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:

\[
\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{\text{Sim}}.
\]

With the \( \mathcal{L}_{\text{Sim}} \), the model will tend to learn separated audio and visual embeddings. Meanwhile, the \( \mathcal{L}_{CE} \) 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: \( f^\alpha_a \) and \( f^\alpha_v \) 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: \( M_a \in \mathbb{R}^{d \times K} \) and \( M_v \in \mathbb{R}^{d \times K} \), 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:

\[
\min_{\alpha_a} ||f^a_{adv} - M_a \alpha_a||^2 + \lambda_a ||\alpha_a||_1, \\
\min_{\alpha_v} ||f^v_{adv} - M_v \alpha_v||^2 + \lambda_v ||\alpha_v||_1,
\]

where the parameters: \( \lambda_a \) and \( \lambda_v \) balance sparsity of the solutions and fidelity of the approximation, and \( \alpha_a \) and \( \alpha_v \) are predicted audio and visual coefficients, respectively. Then, the more reliable audio and visual features can be reconstructed by the corresponding encoded coefficients: \( f^\alpha_a = M_a \alpha_a \) and \( f^\alpha_v = M_v \alpha_v \). 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 train/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.
4.2. Attack Methods

We evaluate the audio-visual model robustness with $l_\infty$-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 $x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x, y; \theta))$, where $x_{adv}$ is the generated adversarial example, $x$ is the original input, $y$ is the original label, $\theta$ refers to model parameters, $\epsilon$ 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: $x_{adv}^{a}$ and $x_{adv}^{v}$ 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 $\epsilon_{a}$ and $\epsilon_{v}$, 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.

### Table 1: Audio-visual event recognition accuracy on MIT-MUSIC and Kinetics-Sounds datasets under different attack methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Attack</th>
<th>$\mathbf{\mathcal{X}A}$</th>
<th>$\mathbf{\mathcal{XV}}$</th>
<th>$\mathbf{\mathcal{XAV}}$</th>
<th>Avg. Unimodal $\mathbf{\mathcal{A}}$</th>
<th>Unimodal $\mathbf{\mathcal{V}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM</td>
<td>FGSM [30]</td>
<td>15.38</td>
<td>8.18</td>
<td>91.76</td>
<td>44.36</td>
<td>72.42</td>
</tr>
<tr>
<td></td>
<td>PGD [45]</td>
<td>72.42</td>
<td>91.76</td>
<td>50.00</td>
<td>15.38</td>
<td>8.18</td>
</tr>
<tr>
<td></td>
<td>MIM [17]</td>
<td>22.22</td>
<td>19.00</td>
<td>72.42</td>
<td>15.38</td>
<td>8.18</td>
</tr>
</tbody>
</table>

Table 1: Audio-visual event recognition accuracy on MIT-MUSIC and Kinetics-Sounds datasets under different attack methods. $\mathbf{\mathcal{X}A}$, $\mathbf{\mathcal{XV}}$, and $\mathbf{\mathcal{XAV}}$ denote that only audio, only visual, and both audio and visual inputs for our audio-visual network are attacked, respectively. We set $\epsilon_{a}$ and $\epsilon_{v}$ as 0.12 respectively for $\mathbf{\mathcal{X}A}$ and $\mathbf{\mathcal{XV}}$, and 0.06 for $\mathbf{\mathcal{XAV}}$. The symbol: $\checkmark$ means that inputs are clean. The baselines: Unimodal $\mathbf{\mathcal{A}}$ and Unimodal $\mathbf{\mathcal{V}}$ models are two single-modality models.

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.

From Table 1, we can also see that our clean audio-visual models ($\mathbf{\mathcal{XAV}}$) are better than both clean single-modality A (Unimodal $\mathbf{\mathcal{A}}$) and V (Unimodal $\mathbf{\mathcal{V}}$) 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: $\mathbf{\mathcal{X}A}$ and $\mathbf{\mathcal{XV}}$ are always inferior to Unimodal $\mathbf{\mathcal{V}}$ and Unimodal $\mathbf{\mathcal{A}}$, respectively. For example, the performances drop 91.76% and
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 $\mathcal{X}_A$ and $\mathcal{X}_V$ 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: $\mathcal{XAV}$ obtain even worse performance than the $\mathcal{XA}$ and $\mathcal{XV}$. When we compare it to attacked unimodal models (see Unimodal A and Unimodal V in Table 3), we can see that $\mathcal{XA}$ of Unimodal A and $\mathcal{XV}$ of Unimodal V achieve 0.00% and 11.54%, while $\mathcal{XAV}$ 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 $\mathcal{XAV}$ of the audio-visual model and $\mathcal{XV}$ 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.

### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\mathcal{XAV}$</th>
<th>$\mathcal{XA}$</th>
<th>$\mathcal{XV}$</th>
<th>$\mathcal{XAV}$ Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td>88.46</td>
<td>35.58</td>
<td>45.19</td>
<td>3.85</td>
</tr>
<tr>
<td>Concat</td>
<td>88.46</td>
<td><strong>51.92</strong></td>
<td><strong>45.19</strong></td>
<td><strong>15.38</strong></td>
</tr>
<tr>
<td>FiLM [57]</td>
<td>83.65</td>
<td>28.85</td>
<td>39.42</td>
<td>3.85</td>
</tr>
<tr>
<td>Gated-Sum [39]</td>
<td><strong>89.42</strong></td>
<td>33.65</td>
<td>44.23</td>
<td>4.81</td>
</tr>
<tr>
<td>Gated-Concat [39]</td>
<td><strong>89.42</strong></td>
<td>45.19</td>
<td>43.27</td>
<td>13.46</td>
</tr>
</tbody>
</table>

Table 2: Audio-visual event recognition accuracy with different fusions on the MIT-MUSIC under FGSM attacks.
free inputs, and FiLM is worse than the other fusion approaches; audio-visual models: $\mathbf{X} \mathbf{A}$ and $\mathbf{X} \mathbf{V}$ with different fusions achieve inferior performance than Unimodal $\mathbf{V}$ and Unimodal $\mathbf{A}$, 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, 63] as the localization module. Concretely, we obtain a $N \times N$ visual feature map: $F_v = [f_v^1; \ldots; f_v^N] \in \mathcal{R}^{N^2 \times d}$ 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:

$$w_i = \frac{\exp(f_a^i f_v^i)}{\sum_j \exp(f_a^j f_v^j)}$$

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 sound-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 sound 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.

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

$$\text{Avg} = \frac{1}{3} (\mathbf{X} \mathbf{A} + \mathbf{X} \mathbf{V} + \mathbf{X} \mathbf{AV}),$$

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 $\checkmark$AV will achieve worse performance. To address the issue, we propose a relative improvement (RI) metric:

<table>
<thead>
<tr>
<th>Defense (MUSIC)</th>
<th>$\checkmark$AV</th>
<th>$\mathbf{X} \mathbf{A}$</th>
<th>$\mathbf{X} \mathbf{V}$</th>
<th>$\mathbf{X} \mathbf{AV}$</th>
<th>Avg</th>
<th>RI</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>88.46</td>
<td>51.92</td>
<td>45.19</td>
<td>15.38</td>
<td>37.50</td>
<td>0.00</td>
</tr>
<tr>
<td>Unimodal A</td>
<td>59.62</td>
<td>0.00</td>
<td>59.62</td>
<td>0.00</td>
<td>19.87</td>
<td>-46.47</td>
</tr>
<tr>
<td>Unimodal V</td>
<td>81.73</td>
<td>81.73</td>
<td>11.54</td>
<td>11.54</td>
<td>34.94</td>
<td>-9.29</td>
</tr>
<tr>
<td>PCL [51]</td>
<td>83.65</td>
<td>83.65</td>
<td>37.50</td>
<td>36.54</td>
<td>51.91</td>
<td>9.60</td>
</tr>
<tr>
<td>MaxSim</td>
<td>89.42</td>
<td>52.88</td>
<td>45.19</td>
<td>31.73</td>
<td>43.27</td>
<td>6.73</td>
</tr>
<tr>
<td>MinSim</td>
<td>91.35</td>
<td>70.19</td>
<td>46.15</td>
<td>36.54</td>
<td>50.96</td>
<td>16.35</td>
</tr>
<tr>
<td>ExFMem</td>
<td>89.42</td>
<td>53.85</td>
<td>50.00</td>
<td>20.19</td>
<td>41.34</td>
<td>4.80</td>
</tr>
<tr>
<td>MinSim+ExFMem</td>
<td>90.38</td>
<td>73.08</td>
<td>53.85</td>
<td>42.31</td>
<td>56.41</td>
<td>20.83</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Defense (Kinetics)</th>
<th>$\checkmark$AV</th>
<th>$\mathbf{X} \mathbf{A}$</th>
<th>$\mathbf{X} \mathbf{V}$</th>
<th>$\mathbf{X} \mathbf{AV}$</th>
<th>Avg</th>
<th>RI</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>72.42</td>
<td>36.40</td>
<td>26.35</td>
<td>8.09</td>
<td>23.61</td>
<td>0.00</td>
</tr>
<tr>
<td>Unimodal A</td>
<td>35.99</td>
<td>1.87</td>
<td>35.99</td>
<td>1.87</td>
<td>13.24</td>
<td>-46.80</td>
</tr>
<tr>
<td>Unimodal V</td>
<td>66.08</td>
<td>66.08</td>
<td>18.72</td>
<td>18.72</td>
<td>34.50</td>
<td>4.55</td>
</tr>
<tr>
<td>PCL [51]</td>
<td>64.50</td>
<td>63.43</td>
<td>29.28</td>
<td>28.67</td>
<td>40.46</td>
<td>8.93</td>
</tr>
<tr>
<td>MaxSim</td>
<td>71.39</td>
<td>34.95</td>
<td>29.57</td>
<td>21.46</td>
<td>28.66</td>
<td>4.02</td>
</tr>
<tr>
<td>MinSim</td>
<td>70.88</td>
<td>52.42</td>
<td>28.12</td>
<td>21.62</td>
<td>34.05</td>
<td>8.99</td>
</tr>
<tr>
<td>ExFMem</td>
<td>72.71</td>
<td>41.56</td>
<td>29.93</td>
<td>10.44</td>
<td>27.31</td>
<td>3.99</td>
</tr>
<tr>
<td>MinSim+ExFMem</td>
<td>71.33</td>
<td>55.96</td>
<td>30.57</td>
<td>24.90</td>
<td>37.14</td>
<td>12.44</td>
</tr>
</tbody>
</table>
RI = (√AV_m + Avg_m) − (√AV_n + 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 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


[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


