Defending Multimodal Fusion Models against Single-Source Adversaries

Karren Yang[1] Wan-Yi Lin[2] Manash Barman[3] Filipe Condessa[2] Zico Kolter[2,3]
[1]Massachusetts Institute of Technology [2]Bosch Center for AI* [3]Carnegie Mellon University
karren@mit.edu , wan-yi.lin@us.bosch.com , mbarman@andrew.cmu.edu
filipe.condessa@us.bosch.com , zkolter@cs.cmu.edu

Abstract

Beyond achieving high performance across many vision tasks, multimodal models are expected to be robust to single-source faults due to the availability of redundant information between modalities. In this paper, we investigate the robustness of multimodal neural networks against worst-case (i.e., *adversarial) perturbations on a single modality. We first show that standard multimodal fusion models are vulnerable to single-source adversaries: an attack on any single modality can overcome the correct information from multiple unperturbed modalities and cause the model to fail. This surprising vulnerability holds across diverse multimodal tasks and necessitates a solution. Motivated by this finding, we propose an adversarially robust fusion strategy that trains the model to compare information coming from all the input sources, detect inconsistencies in the perturbed modality compared to the other modalities, and only allow information from the unperturbed modalities to pass through. Our approach significantly improves on state-of-the-art methods in single-source robustness, achieving gains of 7.8-25.2% on action recognition, 19.7-48.2% on object detection, and 1.6-6.7% on sentiment analysis, without degrading performance on unperturbed* (i.e., *clean) data.*

1. Introduction

Consider a multimodal neural network, illustrated in Figure 1(a), that fuses inputs from k different sources to identify objects for an autonomous driving system. If one of the modalities (*e.g.*, RGB) receives a worst-case or adversarial perturbation, does the model fail to detect the truck in the scene? Or does the model make a robust prediction using the remaining $k-1$ unperturbed modalities (*e.g.*, LIDAR, audio, *etc.*)? This example illustrates the importance of *single-source adversarial robustness* [17] for avoiding catastrophic failures in real-world multimodal systems. In a realistic setting, any single modality may be affected by a worst-case perturbation, whereas multiple modalities usually do not fail simultaneously particularly if the physical sensors are not coupled. Since multimodal models are being increasingly developed for real-world vision tasks [5, 35, 24, 18], it is imperative to investigate whether they are robust to worst-case errors that may affect any single modality and, if they are not, to develop strategies to improve robustness.

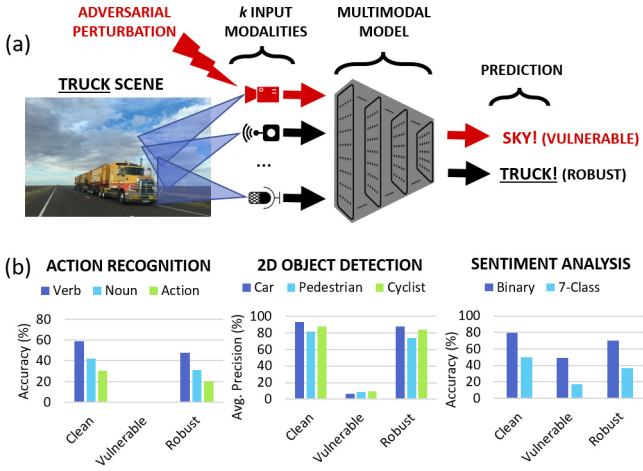

Figure 1. (a) Example of a single source, worst-case (*i.e.*, adversarial) perturbation on a multimodal model. (b) Standard multimodal models are vulnerable to worst-case perturbations on any single modality ("Vulnerable"). Our adversarially robust fusion strategy ("Robust") leverages multimodal consistency to defend against such perturbations without degrading clean performance.

Despite the importance of this problem, we found to the best of our knowledge that empirical studies of single-source adversarial robustness are lacking. Previous empirical works on multimodal robustness have so far only considered single-source corruptions (*e.g.*, dropout, blurring, *etc.*) [16, 15, 17], and although Kim & Ghosh [17] formulate the problem for the adversarial setting, they do not perform an

*KY and MB completed work during an internship at BCAI. Work is sponsored by DARPA (Grant number HR11002020006).

empirical study. In the field of adversarial robustness, most studies have focused on the unimodal setting rather than the multimodal setting [20, 21]. An effective strategy for defending unimodal models against adversaries is *adversarial training* (*i.e.*, end-to-end training of the model on adversarial examples). In principle, adversarial training could be extended to multimodal models as well, but it has several downsides: (1) it is resource-intensive [31] and may not scale well to large, multimodal models that contain many more parameters than their unimodal counterparts; (2) it significantly degrades performance on clean data [21]. For these reasons, end-to-end adversarial training may not be practical for multimodal systems used in real-world tasks.

Contributions. This paper presents, to our knowledge, the first empirical study of single-source adversarial robustness in multimodal systems. Our contributions are two-fold.

(1) We investigate multimodal robustness against single-source adversaries on diverse benchmark tasks with three modalities ($k = 3$): action recognition on EPIC-Kitchens [7], object detection on KITTI [9], and sentiment analysis on CMU-MOSI [40]. We find that standard multimodal fusion practices are vulnerable to single-source adversarial perturbations, even when there are multiple unperturbed modalities that could yield a correct prediction; naive ensembling of features from a perturbed modality with features from clean modalities does not automatically yield robust prediction. As shown in Figure 1(b), a worst-case input at any single modality of a multimodal model can outweigh the other modalities and cause the model to fail. In fact, contrary to expectations, a multimodal model ($k = 3$) under a single-source perturbation does not necessarily outperform a unimodal model ($k = 1$) under the same attack.

(2) We propose an adversarially robust fusion strategy that can be applied to mid- to late- fusion models to defend against this vulnerability without degrading clean performance. Inspired by recent works that detect correspondence between inputs to defend against image manipulation [13], we hypothesize that a multimodal model can be trained to detect correspondence (or lack thereof) between features from different modalities and use this information to perform a robust feature fusion that defends against the perturbed modality. Our approach extends existing work on adaptive gating strategies [16, 15, 34, 22] with a robust fusion training procedure based on odd-one-out learning [8] to improve single-source adversarial robustness without degrading clean performance. Through extensive experiments, we demonstrate that our approach is effective even against adaptive, white-box attacks with access to the robust fusion strategy. We significantly outperform state-of-the-art methods in single-source robustness [16, 15, 17], achieving gains of 7.8-25.2% on action recognition on EPIC-Kitchens, 19.7-48.2% on 2D object detection on KITTI, and 1.6-6.7% sentiment analysis on CMU-MOSI.

Overall, this paper demonstrates that multimodal models are not inherently robust to single-source adversaries, but that we can improve their robustness without the downsides associated with end-to-end adversarial training in unimodal models. The combination of robust fusion architectures with robust fusion training may be a practical strategy for defending real-world systems against adversarial attacks and establishes a promising direction for future research.

1.1. Related work

Adversarial Robustness. Vision systems based on deep learning models are susceptible to adversarial attacks– additive, worst-case, and imperceptible perturbations on the inputs that cause erroneous predictions [6, 4, 33, 10]. A large number of defense methods against adversarial attacks have been proposed, with the two most effective defenses being end-to-end adversarial training [10, 20, 21], which synthesizes adversarial examples and includes them in training data, and provably robust training, which provides theoretical bounds [37, 25] on the performance. However, these methods have primarily focused on the unimodal setting, in which the input is a single image. In contrast to those works, we consider single-source adversarial perturbations in the multimodal setting and leverage consistent information between modalities to improve the robustness of the model's fusion step. Our training procedure is related to adversarial training in the sense that we also use perturbed inputs, but instead of end-to-end training of model parameters, we focus on designing and training the feature fusion in a robust manner. This strategy brings benefits from adversarial training, while retaining performance on clean data and significantly reducing the number of parameters that need to be trained on perturbed data.

Multimodal Fusion Models. Multimodal neural networks have demonstrated remarkable performance across a variety of vision tasks, such as scene understanding [14], object detection [35, 12], sentiment analysis [40, 3, 39, 19], speech recognition [1], and medical imaging [11]. In terms of fusion methods, several approaches that use gating networks have been proposed to weigh sources adaptively depending on the inputs [22, 34, 23, 2]. These works focus on leveraging multiple modalities to improve clean performance on the task and do not evaluate or extend these approaches to improve single-source robustness, which is our focus.

Single Source Robustness. Several recent works provide important insights into the effects of single-source corruptions such as occlusions, dropout, and Gaussian noise on object detection systems with two modalities ($k = 2$) [16, 15, 17]. In contrast to their work, we consider single-source adversarial perturbations, which explore worst-case failures of multimodal systems due to one perturbed modality. We consider other tasks in addition to object detection and evaluate models with three modalities ($k = 3$), in which

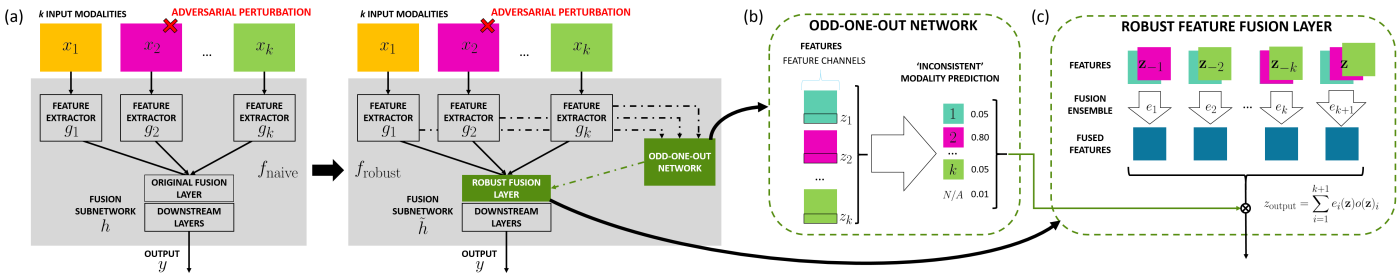

Figure 2. We propose a robust multimodal fusion strategy based on "odd-one-out" learning, an auxiliary self-supervised task in which a model is presented with multiple elements and must predict which one of them is different from the others. A multimodal model augmented with an odd-one-out network can be trained to compare information coming from all the input sources, detect the perturbed modality because it is inconsistent with the other modalities, and only allow information from the unperturbed modalities to pass through.

there are more clean sources than perturbed sources. In terms of defense strategies, robust multimodal fusion methods based on end-to-end robust training [17] and adaptive gating fusion layers [16, 15] have been developed to improve single-source robustness to corruptions. We extend this line of work by developing a robust fusion strategy that leverages correspondence between unperturbed modalities to defend against the perturbed modality, and is effective against more challenging adversarial perturbations.

2. Single Source Adversarial Perturbations

Let $f : \mathbf{x} \mapsto y$ denote a multimodal model with k input modalities (*i.e.*, $\mathbf{x} = [x_1, \cdots, x_k]$). We aim to understand the extent to which the performance of f is degraded by worst-case perturbations on any single modality $i \in [k]$ (where $[k] = \{1, \cdots, k\}$) while the other $k-1$ modalities remain unperturbed. To this end, we define a *single-source adversarial perturbation* against f on modality i as,

$$\delta^{(i)}(\mathbf{x}, y; f) \;:=\; \arg\max_{||\delta||_p \leq \epsilon} \mathcal{L}(f(x_i + \delta, \mathbf{x_{-i}}), y), \quad (1)$$

where $\mathcal{L}$ is the loss function and $\epsilon > 0$ defines the allowable range of the perturbation $\delta^{(i)}$. If we assume that the multimodal inputs $\mathbf{x}$ and outputs y are sampled from a distribution $\mathcal{D}$, then the *single-source adversarial performance* of f with respect to modality $i \in [k]$ is given by,

$$\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}} \max_{||\delta||_p < \epsilon} \left[\mathcal{L}(f(x_i + \delta, \mathbf{x_{-i}}), y)\right]. \quad (2)$$

The difference between the performance of f on unperturbed data, *i.e.*, $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathcal{L}(f(\mathbf{x}), y)\right]$, and its single-source adversarial performance specified in (2) indicates, on average, the sensitivity of f to its worst-case inputs on modality i. Ideally, a multimodal model that has access to multiple input modalities with redundant information should not be sensitive to perturbations on a single input; it should be able to make a correct prediction by leveraging the remaining $k-1$ unperturbed modalities. However, we find across diverse multimodal benchmark tasks that standard multimodal fusion models are surprisingly vulnerable to these perturbations, even though the clean modalities outnumber the perturbed modality. We discuss the experiments and results in later sections; for now, we emphasize that this vulnerability necessitates a solution.

3. Adversarially Robust Fusion Strategy

Let f_{naive} be a standard multimodal neural network, pretrained to achieve acceptable performance on unperturbed data, *i.e.*, it minimizes $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathcal{L}(f_{\text{naive}}(\mathbf{x}), y)\right]$. Our robust fusion strategy aims to improve the single-source robustness of f_{naive} by leveraging the correspondence between the unperturbed modalities to detect and defend against the perturbed modality. We assume that f_{naive} has a mid- to late- fusion architecture, consisting of the composition of modality-specific feature extractors $g_1, \cdots, g_k$ applied to their respective modalities and a fusion subnetwork h:

$$f_{\text{naive}}(\mathbf{x}) := h(g_1(x_1), g_2(x_2), \cdots, g_k(x_k)), \quad (3)$$

To make f_{naive} robust, we equip it with an auxiliary odd-one-out network and a robust feature fusion layer in place of the default feature fusion operation, as shown in Figure 2(a). Then we perform robust training based on odd-one-out learning [8] and adversarial training [21] that focuses on these new modules. The odd-one-out network o is trained to detect the inconsistent or perturbed modality when presented with feature representations of different modalities (Section 3.1). The robust feature fusion layer ensembles different multimodal fusion operations using the output of the odd-one-out network, ensuring that only the modalities that are consistent with each other are passed to the downstream layers (Section 3.2). We denote the fusion subnetwork h equipped with the robust feature fusion layer as $\tilde{h}$, and we denote the full, augmented multimodal model as f_{robust}, *i.e.*,

$$f_{\text{robust}}(\mathbf{x}) := \tilde{h}(g_1(x_1), g_2(x_2), \cdots, g_k(x_k); o(\{g_i(x_i)\}_{i\in[k]}).$$

Finally, we jointly train the odd-one-out network o and the fusion subnetwork $\tilde{h}$, while keeping the weights and architectures of the feature extractors $g_1, \cdots, g_k$ fixed from f_{naive} (Section 3.3).

3.1. Odd-one-out learning

Odd-one-out learning is a self-supervised task that aims to identify the inconsistent element from a set of otherwise consistent elements [8]. To leverage the shared information between modalities, we propose to augment the multimodal model with an odd-one-out network. Given the set of features $\mathbf{z} = [z_1, \cdots, z_k]$ extracted from the k-modality input, the odd-one-out network predicts whether the multimodal features are consistent with each other (*i.e.*, the inputs are all clean), or whether one modality is inconsistent with the others (*i.e.*, some input has been perturbed). To perform this task, the odd-one-out network must compare the features from different modalities, recognize the shared information between them, and detect any modality that is not consistent with the others. For convenience, we take the features to be the final outputs of the feature extractor networks $g_1, \cdots, g_k$ applied to their respective modalities. In principle, though, these features could also come from any of the intermediate layers of the feature extractors.

Concretely, the odd-one-out network is a neural network o that maps the features $\mathbf{z}$ to a vector of size $k+1$, as shown in Figure 2(b). The i-th entry of this vector indicates the probability that modality i has been perturbed, *i.e.*, z_i is inconsistent with the other features. The $k+1$-th entry of the vector indicates the probability that none of the modalities are perturbed. The odd-one-out network o is trained to perform odd-one-out prediction by minimizing the following cross-entropy loss:

$$-\mathbf{E}_{\substack{(\mathbf{x},y)\sim\mathcal{D} \\ z_i=g_i(x_i)}}\big[\log o(\mathbf{z})_{k+1} + \sum_{i=1}^{k} \log o(z_i^*, \mathbf{z}_{-\mathbf{i}})_i\big], \quad (4)$$

where $z_i^* = g_i(x_i^*)$ is the feature extracted from perturbed input x_i^* that we generate during training.

3.2. Robust Feature Fusion Layer

To integrate the output of the odd-one-out network o into the multimodal model, we propose a feature fusion layer inspired by the mixture-of-experts layer [32]. This layer consists of an ensemble of $k+1$ feature fusion operations $e_1, \cdots, e_{k+1}$, each of which is specialized to exclude one modality, as illustrated in Figure 2(c). Formally, each fusion operation takes the multimodal features $\mathbf{z}$ as input and performs a fusion of a subset of the features as follows:

$$e_i(\mathbf{z}) = \text{NN}\big(\oplus\, \mathbf{z}_{-i}\big)\ \forall i \in [k], \quad e_{k+1}(\mathbf{z}) = \text{NN}\big(\oplus\, \mathbf{z}\big),$$

where $\oplus$ denotes the concatenation operation and NN stands for a shallow neural network. By definition, e_i is responsible for performing a fusion of features from all the modalities *except* for i, and only e_{k+1} fuses features from all the modalities. If feature z_i is not consistent with features from the other $k-1$ modalities because it results from a perturbed input, then e_i receives more weight than the other fusion operations based on the output of the odd-one-out network:

$$z_{\text{output}} = \sum_{i=1}^{k+1} e_i(\mathbf{z}) o(\mathbf{z})_i, \quad (5)$$

We form a robust fusion subnetwork $\tilde{h}$ by equipping the fusion subnetwork h with this robust feature fusion layer. Then $\tilde{h}$ and o are trained to optimize clean performance,

$$\mathbf{E}_{\substack{(\mathbf{x},y)\sim\mathcal{D} \\ z_i=g_i(x_i)}}\big[\mathcal{L}\big(\tilde{h}(\mathbf{z}; o(\mathbf{z})), y\big)\big], \quad (6)$$

as well as the single-source robust performance,

$$\mathbf{E}_{\substack{(\mathbf{x},y)\sim\mathcal{D} \\ z_i=g_i(x_i)}}\big[\mathcal{L}\big(\tilde{h}(z_i^*, \mathbf{z}_{-\mathbf{i}}; o(z_i^*, \mathbf{z}_{-\mathbf{i}})), y\big)\big], \quad (7)$$

with respect to each modality, where $z_i^* = g_i(x_i^*)$ is the feature extracted from perturbed input x_i^* that we generate during training. Note that one of the arguments into the fusion network $\tilde{h}$ is now the output of o.

Spatiotemporal Dimensions. Our formulations assume that $z_1, \cdots, z_k$ are one-dimensional feature representations, in which case the odd-one-out network o and fusion operations $e_1, \cdots, e_{k+1}$ can be implemented as shallow fully-connected networks (*e.g.*, two fully-connected layers). In many multimodal models, the features also have spatiotemporal dimensions that are aligned between different modalities, *i.e.*, $z_i \in \mathbb{R}^{c_i \times N_1 \times \cdots \times N_d}$, where c_i is the number of feature channels and $N_1 \times \cdots \times N_d$ are the shared spatiotemporal dimensions (*e.g.*, audio and visual features extracted from a video are aligned along the temporal axis, features

Algorithm 1 Robust Training Strategy.

procedure GRADIENTUPDATE
  $\ell_{\text{odd}} \leftarrow 0$
  $\ell_{\text{task}} \leftarrow 0$
  Sample $\mathbf{x} = [x_1, \cdots, x_k], y$ from $\mathcal{D}$
  $\mathbf{z} = [z_1, \cdots, z_k] \leftarrow [g_1(x_1), \cdots, g_k(x_k)]$
  $\ell_{\text{odd}} \leftarrow \ell_{\text{odd}} - \log o(\mathbf{z})_{k+1}$
  $\ell_{\text{task}} \leftarrow \ell_{\text{task}} + \mathcal{L}\big(h(\mathbf{z}, o(\mathbf{z})), y\big)$
  for $i \in [k]$ **do**
    $\delta^{(i)} \leftarrow \delta^{(i)}(\mathbf{x}, y; f_{\text{robust}})$ (Eqn. 1)
    $z_i^* \leftarrow g_i(x_i + \delta^{(i)})$
    $\ell_{\text{odd}} \leftarrow \ell_{\text{odd}} - \log o(z_i^*, \mathbf{z}_{-\mathbf{i}})_i$
    $\ell_{\text{task}} \leftarrow \ell_{\text{task}} + \mathcal{L}\big(h(z_i^*, \mathbf{z}_{-\mathbf{i}}, o(z_i^*, \mathbf{z}_{-\mathbf{i}})), y\big)$
  $\ell \leftarrow \ell_{\text{odd}} + \ell_{\text{task}}$
  Update o, h based on $\nabla \ell$

Dataset	Tasks	Input Modalities	Model	Adversarial Perturbation	Evaluation Metrics
EPIC-Kitchens [7]	Action recognition	Visual frames; Motion frames (flow); Audio (spectrogram)	Feature extractors: BNInception [14] (all); Fusion: feed-forward network + temporal pooling; Odd-one-out network: feed-forward network	PGD (10-step): ϵ = 8/256 (vision) ϵ = 8/256 (motion) ϵ = 0.8 (audio)	Top-1, top-5 accuracy: Verbs, nouns, actions
KITTI [9]	2D object detection	Visual frame; Depth map (Velodyne); Depth map (stereo image)	Feature extractors: Darknet19 [26] (all); Fusion: 1×1 conv layer + YOLO [27]; Odd-one-out network: 1×1 conv net;	PGD (10-step): ϵ = 16/256 (all)	Average precision: Cars (>0.7 IoU), Pedestrians (>0.5 IoU), Cyclists (>0.5 IoU)
MOSI [40]	Sentiment analysis	Visual frame; Audio (mel ceptron); Text	Feature extractors: FaceNet[29] +LSTM (vision), MFCC+LSTM (audio), transformer [] (text); Fusion: feed-forward network Odd-one-out network: feed-forward network	PGD (10-step): ϵ = 8/256 (vision) ϵ = 0.8 (audio) word replacement [28], 1-word per sentence (text)	Binary accuracy 7-class accuracy

Table 1. A summary table of our experimental setups.

extracted from different visual modalities are aligned along the spatial axes). In those cases, our odd-one-out network and fusion operations are more efficiently implemented as convolutional neural networks with $1 \times \cdots \times 1$ filters. This enables us to compute the losses in Equations (4) and (5) in parallel over the spatiotemporal dimensions.

3.3. Robust Training Procedure

The multimodal model f_{robust}, which is equipped with an odd-one-out network o and fusion subnetwork $\tilde{h}$, contains a mechanism to compare information coming from all the input sources, detect that the perturbed modality is inconsistent with the other unperturbed modalities, and only allow information from the unperturbed modalities to pass through. During training, we generate perturbed inputs x_i^* using the single-source adversarial perturbations from Equation 1, *i.e.*, we let

$$x_i^* = x_i + \delta^{(i)}(\mathbf{x}, y, f_{\text{robust}}).$$

Note that this adversarial perturbation is generated against f_{robust}. In other words, our approach performs adversarial training of the fusion network and also leverages the adversarial examples to provide self-supervised labels for odd-one-out learning. We optimize the parameters of the odd-one-out network o and the fusion subnetwork $\tilde{h}$ with respect to the losses in Equations (4), (6), and (7), as shown in Algorithm 1. We do not retrain the feature extractors $g_1, \cdots, g_k$, which are already pretrained on clean data.

4. Experiments

We evaluate the single-source adversarial robustness of multimodal models on three benchmark tasks: action recognition on EPIC-Kitchens, 2D object detection on KITTI, and sentiment analysis on MOSI. Existing benchmark tasks for studying single source corruptions in multimodal models have primarily focused on the object detection task with two modalities [17, 16, 15]. The benchmarks that we consider involve three input modalities and span a larger variety of tasks and data sources, ensuring generality of the conclusions drawn. A summary can be found in Table 1.

4.1. Multimodal Benchmark Tasks

Action recognition on EPIC-Kitchens. EPIC-Kitchens is the largest egocentric video dataset consisting of 39,596 video clips [7]. The objective is to predict the action taking place in the video, which is composed of one verb and one noun out of 126 and 331 classes respectively. Three modalities are available from the original dataset: visual information (RGB frames), motion information (optical flow), and audio information.

Object Detection on KITTI. KITTI is an autonomous driving dataset [9] that contains stereo camera and LIDAR information for 2D object detection, where the objective is to draw bounding boxes around objects of interest from predefined classes, *i.e.*, car, pedestrian, cyclist, etc. Existing works use different combinations and processed versions of the available data modalities for object detection. For the proposed benchmark, we consider the following three modalities based on common use in the literature: (1) RGB frames, which are used by the majority of detection methods, (2) LIDAR points projected to a sparse depth map and (3) a depth map estimated from the stereo views [36].

Sentiment Analysis on CMU-MOSI. Multimodal Opinion-level Sentiment Intensity Corpus (CMU-MOSI) [40] is a multimodal dataset for sentiment analysis consisting of 93 video clips of movie reviews, each of which are divided into an average of 23.2 segments. Each segment is labeled with a continuous sentiment intensity between $[-3, 3]$. The objective is to predict the sentiment on a binary scale (*i.e.*, negative v. positive) or 7-class scale (*i.e.*, rounding to the nearest integer). MOSI contains three modalities: text, video and audio.

4.2. Implementation Details

Model Architecture and Training. For each task, we consider mid- to late- multimodal models that use the architectures summarized in column 4 of Table 1. We first train baseline multimodal models for each task on clean data to obtain f_{naive}. We then augment these models with the odd-one-out network and robust feature fusion layer as described in Section 3 to obtain f_{robust}, and perform robust training ac-

Fusion	Clean			Visual Perturbation			Motion Perturbation			Audio Perturbation		
	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
Oracle (Upper Bound)	-	-	-	55.8	31.4	21.9	50.0	37.2	23.8	53.9	39.2	25.6
Concat Fusion	59.0	42.1	30.2	0.1	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0
Mean Fusion	56.8	40.4	27.6	0.3	0.8	0.0	0.3	0.3	0.0	0.4	0.3	0.0
LEL+Robust [17]	61.2	**43.1**	30.5	22.3	11.6	6.6	25.4	24.6	12.0	20.4	17.7	8.0
Gating+Robust [16, 15]	60.9	43.0	30.6	26.0	10.9	6.2	35.9	26.9	14.3	21.3	16.2	7.0
Ours	**61.5**	42.5	**31.4**	**48.0**	**24.2**	**16.8**	**48.5**	**35.6**	**22.1**	**46.5**	**33.3**	**22.1**
Δ-Clean	**2.5**	**0.3**	**1.2**	**47.7**	**23.4**	**16.8**	**48.2**	**35.3**	**22.1**	**46.1**	**33.0**	**22.1**
Δ-Robust	**0.3**	-0.6	**0.8**	**22.0**	**13.3**	**10.2**	**12.6**	**8.7**	**7.8**	**25.2**	**15.6**	**14.1**

Table 2. Top-1 classification accuracy results on EPIC-Kitchens dataset under clean data and single-source adversarial perturbations on each modality. Higher is better. Due to space constraints, we defer Top-5 accuracy to Supplementary Materials.

Fusion	Clean			Visual (RGB) Perturbation			Depth (Velo) Perturbation			Depth (Stereo) Perturbation		
	Car	Pedest.	Cyclist	Car	Pedest.	Cyclist	Car	Pedest.	Cyclist	Car	Pedest.	Cyclist
Oracle (Upper Bound)	-	-	-	90.4	80.1	86.4	93.2	79.3	85.3	92.8	80.5	87.4
Concat Fusion	93.5	**81.5**	**87.7**	14.3	10.7	12.3	1.58	11.1	8.82	3.57	4.64	7.23
Mean Fusion	**93.6**	77.7	86.7	12.6	15.2	10.5	3.16	12.9	7.88	3.08	8.03	7.77
LEL+Robust [17]	71.4	64.2	80.0	3.95	15.4	13.9	6.83	20.6	24.8	9.39	24.2	24.7
Gating+Robust [16, 15]	89.4	74.7	84.6	57.2	54.2	56.0	46.5	45.7	45.6	41.6	47.4	48.8
Ours	90.6	79.9	85.4	**85.1**	**73.9**	**82.3**	**87.8**	**71.1**	**85.8**	**89.8**	**76.8**	**84.7**
Δ-Clean	-3.0	-1.6	-2.3	**70.8**	**58.7**	**70.0**	**74.6**	**58.2**	**77.0**	**86.2**	**68.8**	**76.9**
Δ-Robust	**1.2**	**5.2**	**0.8**	**27.9**	**19.7**	**26.3**	**41.3**	**25.4**	**40.2**	**48.2**	**29.4**	**35.9**

Table 3. Evaluation of Average Precision for 2D object detection on the KITTI dataset under clean data and single-source adversarial perturbations on each modality. Higher is better. Due to space constraints, only performance at medium difficulty is shown. See Supplementary Materials for full table with easy/medium/hard difficulties.

cording to Algorithm 1. Additional details are deferred to Supplementary Materials.

Adversarial Attacks. The adversarial perturbations for each task are summarized in column 5 of Table 1. We attack individual modalities using projected gradient descent (PGD) [21], except text, for which we use word replacement [28]. Note that these perturbations are white-box adaptive attack, *i.e.*, attacks are generated with full knowledge of f_{robust}. In the Supplementary Materials, we also describe and show results for other types of attacks, such as transfer attacks, targeted attacks, and feature-level attacks [38].

Evaluation Metric. The metrics used for each task are summarized in column 6 of Table 1. For the action recognition, we consider classification accuracy of verbs, nouns, and actions. For object detection, we consider the average precision of car, pedestrian, and cyclist detection at intersection-over-union (IoU) thresholds shown in the table, and at three difficulty levels following the KITTI evaluation server [9]. For sentiment analysis, we consider binary and 7-class prediction accuracy. For each metric, we consider clean performance as well as performance under single-source attacks.

4.3. Baselines

In addition to our approach, we evaluate two types of methods: standard multimodal models trained with clean data (standard training), and state-of-the-art robust multimodal models [17, 16, 15] with robust training.

Concatenation Fusion with Standard Training ("Concat Fusion"). We use multimodal models with the same feature extractors and concatenate features before the final layers, which is a standard method for fusing features.

Mean Fusion with Standard Training ("Mean Fusion"). For each modality, we train a unimodal model with the same feature extractor and final layers as the multimodal model on clean data. Then we fuse the unimodal model outputs by taking their mean, *i.e.*, $z_{\text{output}} = \sum_{i \in [k]} z_i$. For action recognition and sentiment analysis, we perform mean fusion on the logits layer. For object detection, we perform the fusion prior to the YOLO layer. Mean fusion is a common fusion practice used in late fusion models, and in the context of defenses against perturbations, it is equivalent to a soft voting strategy between the different modalities.

Latent Ensembling Layer with Robust Training ("LEL+Robust" [17]). This approach involves (1) training on clean data and data with each single-source corruption in an alternating fashion, and (2) ensembling the multimodal features using concatenation fusion followed by a linear network. We adapt their strategy to our model by training our multimodal models with their fusion layer on data augmented with single-source perturbations.

Information-Gated Fusion with Robust Training ("Gating+Robust"[16, 15]). This approach applies a multiplicative gating function to features from different modalities before ensembling them. The adaptive gating function is trained on clean data and data with single-source corruptions. We adapt their robustness strategy to our models by training our multimodal models with their gated

Fusion	Clean		Audio Perturbation		Video Perturbation		Text Perturbation	
	2-class	7-class	2-class	7-class	2-class	7-class	2-class	7-class
Oracle (Upper Bound)	-	-	78.64	49.10	73.36	47.84	69.82	40.28
Concat Fusion	79.82	49.69	56.92	21.38	51.23	19.75	39.50	9.97
Mean Fusion	78.09	46.14	52.63	20.75	49.37	17.02	35.50	8.88
LEL+Robust [17]	79.09	49.92	69.21	39.51	63.15	35.17	58.14	21.23
Gating+Robust [16, 15]	78.82	46.37	69.31	38.26	64.23	31.88	59.39	25.14
Ours	**82.03**	**50.89**	**73.18**	**42.06**	**69.94**	**38.20**	**66.13**	**30.20**
Δ-Clean	**2.21**	**1.20**	**16.26**	**20.68**	**18.71**	**18.45**	**26.53**	**20.23**
Δ-Robust	**1.94**	**0.97**	**3.87**	**2.55**	**5.71**	**3.03**	**6.74**	**5.06**

Table 4. Binary and seven-class classification results (%) on MOSI. Higher is better. Random guess is 50% for binary and 14.3% for seven class classification.

feature fusion layer on data augmented with single-source adversarial perturbations.

Upper Bound ("Oracle (Upper Bound)"). To obtain an empirical upper bound for robust performance under attacks against each modality, we train and evaluate 2-modal models that exclude the perturbed modality. We refer to this model as the oracle because it assumes perfect knowledge of which modality is attacked (*i.e.*, a perfect odd-one-out network), which is not available in practice.

5. Results

How robust are standard multimodal models to single-source perturbations? A key motivation for building multimodal models that fuse features from several modalities, beyond improving model performance, is to improve the robustness of the model to perturbations at any given modality. To this end, we first ask how well multimodal models trained on clean data that use concatenation fusion (a standard mid-fusion approach) or mean fusion (a standard late fusion approach) fare against a worst-case perturbation on any single modality. Since these models utilize features from three input modalities ($k = 3$), we hypothesized that ensembling the perturbed features from one modality with the clean features from two other modalities could boost the robust performance of the model, at least compared to a unimodal model that receives the same attack.

Our empirical results suggest that standard multimodal models trained on clean data that use concatenation fusion ("Concat Fusion") or mean fusion ("Mean Fusion") are surprisingly vulnerable against single-source adversarial perturbations. Across the benchmark tasks, we observe drastic drops in performance of both types of models when any one of the modalities receives a worst-case perturbation (see Rows 2-3 of Table 2, 3, 4). For the more challenging tasks with larger output spaces, such as action recognition on EPIC-Kitchens and object detection on KITTI, the multimodal models are not significantly better than a unimodal model under the same attack, and their performance is often close to zero. In the Supplementary Materials, we show that similar results hold when we use other types of adversarial perturbations, such as transfer attacks, targeted attacks, and feature-level attacks [38]. An alarming conclusion drawn from the unimodal transfer attacks is that an attacker can successfully perturb a single modality of the multimodal model even without knowledge of how the different modalities are fused or what the inputs from unperturbed modalities are. Overall, these results show that standard multimodal fusion practices are not sufficiently robust against worst-case perturbations on a single modality and demonstrate the need for robust strategies.

How effective is the proposed approach at defending against single-source adversaries? Our proposed approach focuses on designing and training a robust feature fusion, using odd-one-out learning to leverage the correspondence between unperturbed modalities to detect and defend against any perturbed modality. We achieve significant gains in single-source adversarial robustness across all benchmarks and tasks under *white-box adaptive attacks*, *i.e.*, perturbations were generated with full knowledge of f_{robust}. The main results are shown in Tables 2, 3, 4. Across the board, our method significantly improves the robustness of the standard models (see **"Δ-Clean"**). Our approach also *significantly outperforms* the state-of-the-art robust fusion methods, including the "Gating+Fusion" method which combines an adaptive gating function with robust training (see **"Δ-Robust"**). In the Supplementary Materials, we show that our fusion strategy is also more robust against other single-source perturbations such as transfer attacks, targeted attacks, and feature-level attacks [38]. Comparing our method to the "Oracle (Upper Bound)" row, which always fuses the modalities that are unperturbed, we can see that our method is within -20% of this empirical upper bound. We conclude that our approach outperforms the state-of-the-art in robust multimodal fusion and is close to empirical upper bound that can be achieved without learning robust features using end-to-end adversarial training.

Detection accuracy of odd-one-out network. We hypothesize that our approach is effective because odd-one-out learning enables the model to compare features from the different modalities, recognize consistent information between the unperturbed modalities, and exclude any per-

Action Recognition on EPIC-Kitchens				
Odd-one-out network	Clean	Visual Perturb	Motion Perturb	Audio Perturb
Unaligned features	**66.8**	**73.4**	**88.6**	**84.7**
Aligned Features	55.9	54.7	41.3	52.8
Object Detection on KITTI				
Odd-one-out network	Clean	RGB Perturb	Velo Perturb	Stereo Perturb
Unaligned features	**96.2**	**93.5**	**98.2**	**98.0**
Aligned Features	91.9	86.8	94.4	90.4
Sentiment Analysis on MOSI				
Odd-one-out network	Clean	Audio Perturb	Video Perturb	Text Perturb
Unaligned features	**94.8**	**95.3**	**91.2**	**86.4**
Aligned Features	80.3	90.4	87.3	78.5

Table 5. Detection rate (%) of odd-one-out networks that use unaligned vs. aligned representations of features from each modality. Higher is better. Random guess is 25%.

turbed modality that is inconsistent with the others. To determine if this is the case, we ask how well the odd-one-out network performs in detecting adversarial perturbations from each modality. The results in Table 5 (see **"Unaligned features"**) suggest that the odd-one-out network is highly effective at filtering out features from the perturbed modalities, and performs better when there is more redundant information between the two unperturbed modalities.

In relative comparison between the three tasks, we observe that the odd-one-out network is more successful at filtering out perturbed modalities on the KITTI and MOSI benchmarks than on the EPIC-Kitchens benchmark. This is consistent with our observation that the three modalities in the KITTI benchmark are highly redundant, which enables the odd-one-out network to detect the perturbed modality more easily. For the MOSI benchmark, the text modality contains the most information for the task, and audio and vision provide partly redundant information with text, which is also reflected in the results. In contrast, the three modalities in the EPIC-Kitchens benchmark (*e.g.*, vision, motion, audio) contain a relatively higher degree of complementary (*i.e.*, non-redundant) information, which increases the difficulty of odd-one-out learning. For example, if there is less shared information between the motion and audio inputs for a particular sample, then it may be harder to detect that the visual input is inconsistent.

Since the performance of the odd-one-out network in Table 5 translates to the robust performance of the full model in Tables 2, 3, 4, we asked if we could improve the detection accuracy using features that are already aligned, rather than the unaligned representations from individual feature extractors. For example, for action recognition and sentiment analysis, one can compute differences between logits output by unimodal models; similarly, for object detection, one can compute differences between bounding box coordinates and object confidences output by unimodal models. We found that such strategies (Table 5, **"Aligned features"**) were generally less effective than using the unaligned features. This suggests that the context information available in the unaligned features from the feature extractors may be helpful for detecting the perturbed modality. We defer the full robust performance of our models based on different odd-one-out networks (including a random baseline) to the Supplementary Material.

	# Parameters (Approx in Millions)	
Task	**Feature Extractors (Not Trained)**	**Fusion Network (Trained)**
EPIC-Kitchens	30.8	57.9
KITTI	201.1	6.8
CMU-MOSI	253.4	12.3

Table 6. Number of parameters (in millions) in the feature extractors and fusion networks of our multimodal models.

What are the advantages over learning robust features for each modality (*i.e.*, end-to-end adversarial training)? *Clean performance:* Robust features that are trained on perturbed data are known to perform *significantly* worse on unperturbed data [21]. In contrast, our approach is built on top of feature extractors pretrained on clean data and does not notably degrade clean performance. Across all three benchmarks, our performance on clean (unperturbed) data is comparable with fusion models with standard training (see the "Clean" column of the "Δ-Clean" rows in Tables 2, 3, 4). *Resource utilization:* End-to-end adversarial training on perturbed data is resource-intensive [30] and may not be feasible for large multimodal models. In contrast, our Algorithm 1 for robust fusion only requires training the parameters in the odd-one-out network and the fusion network and not the parameters in the feature extractors. This can drastically reduce the number of parameters that need to be trained on perturbed data. Table 6 shows the number of parameters in the feature extractors vs. the fusion network for our different models. Note that on KITTI, our approach achieves single-source robustness by robustly training only ∼3.3% of the model parameters.

6. Conclusion

This paper presents, to our knowledge, the first empirical study of multimodal robustness under single-source worst-case (adversarial) perturbations. We show across multiple benchmarks that standard multimodal fusion models are surprisingly vulnerable to single-source adversaries and provide an effective solution based on a robust feature fusion that leverages multimodal consistency through odd-one-out learning. The methods and experiments are focused on digital attacks on a single modality in the case where $k \geq 3$. Important directions for future work include physically-realisable attacks, as well as attacks on multiple modalities at once, which would require scaling up our robust fusion approach and training strategy. We believe that this first work on single-source adversarial attack and defense can serve as a basis for these future directions.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 2

[2] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017. 2

[3] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. 2

[4] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013. 2

[5] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 1

[6] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, et al. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108. ACM, 2004. 2

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 5

[8] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017. 2, 3, 4

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2, 5, 6

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

[11] Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(2):162–169, 2019. 2

[12] Ehtesham Hassan, Yasser Khalil, and Imtiaz Ahmad. Learning feature fusion in deep learning-based object detector. *Journal of Engineering*, 2020, 2020. 2

[13] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 101–117, 2018. 2

[14] 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. 2, 5

[15] Jaekyum Kim, Jaehyung Choi, Yechol Kim, Junho Koh, Chung Choo Chung, and Jun Won Choi. Robust camera lidar sensor fusion via deep gated information fusion network. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1620–1625. IEEE, 2018. 1, 2, 3, 5, 6, 7

[16] Jaekyum Kim, Junho Koh, Yecheol Kim, Jaehyung Choi, Youngbae Hwang, and Jun Won Choi. Robust deep multimodal learning based on gated information fusion network. In *Asian Conference on Computer Vision*, pages 90–106. Springer, 2018. 1, 2, 3, 5, 6, 7

[17] Taewan Kim and Joydeep Ghosh. On single source robustness in deep fusion models. In *Advances in Neural Information Processing Systems*, pages 4814–4825, 2019. 1, 2, 3, 5, 6, 7

[18] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018. 1

[19] Ayush Kumar and Jithendra Vepa. Gated mechanism for attention based multi modal sentiment analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4477–4481. IEEE, 2020. 2

[20] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale, 2016. 2

[21] 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, 3, 6, 8

[22] Oier Mees, Andreas Eitel, and Wolfram Burgard. Choosing smartly: Adaptive multimodal fusion for object detection in changing environments. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 151–156. IEEE, 2016. 2

[23] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017. 2

[24] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018. 1

[25] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples, 2018. 2

[26] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger, 2016. 5

[27] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 5

[28] Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1085–1097, 2019. 5, 6

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 5

[30] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! 2019. https://github.com/ashafahi/free_adv_train. 8

[31] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems*, pages 3358–3369, 2019. 2

[32] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 4

[33] 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

[34] Abhinav Valada, Johan Vertens, Ankit Dhall, and Wolfram Burgard. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4644–4651. IEEE, 2017. 2

[35] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *ESANN*, 2016. 1, 2

[36] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 5

[37] Eric Wong and J Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv preprint arXiv:1711.00851*, 2017. 2

[38] Qiuling Xu, Guanhong Tao, Siyuan Cheng, Lin Tan, and Xiangyu Zhang. Towards feature space adversarial attack. *arXiv preprint arXiv:2004.12385*, 2020. 6, 7

[39] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. *arXiv preprint arXiv:1802.00927*, 2018. 2

[40] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016. 2, 5