22-27 MAY 2022
ICASSP 2022
SINGAPORE

On Adversarial Robustness of Large-scale Audio Visual Learning

Juncheng B Li, Shuhui Qu, Xinjian Li, Po-yao Huang,
Florian Metze

junchenl, fmetze@cs.cmu.edu

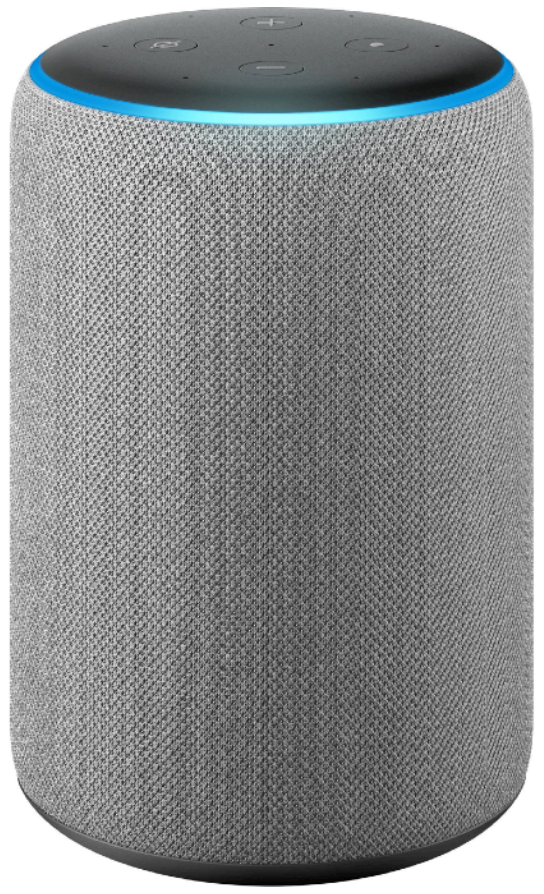
Carnegie Mellon University
Language Technologies Institute

Scan QR code to read our paper:

Audio/Visual Event Recognition in Safety Critical Tasks

Scan QR code to read our paper:

Nest Cams

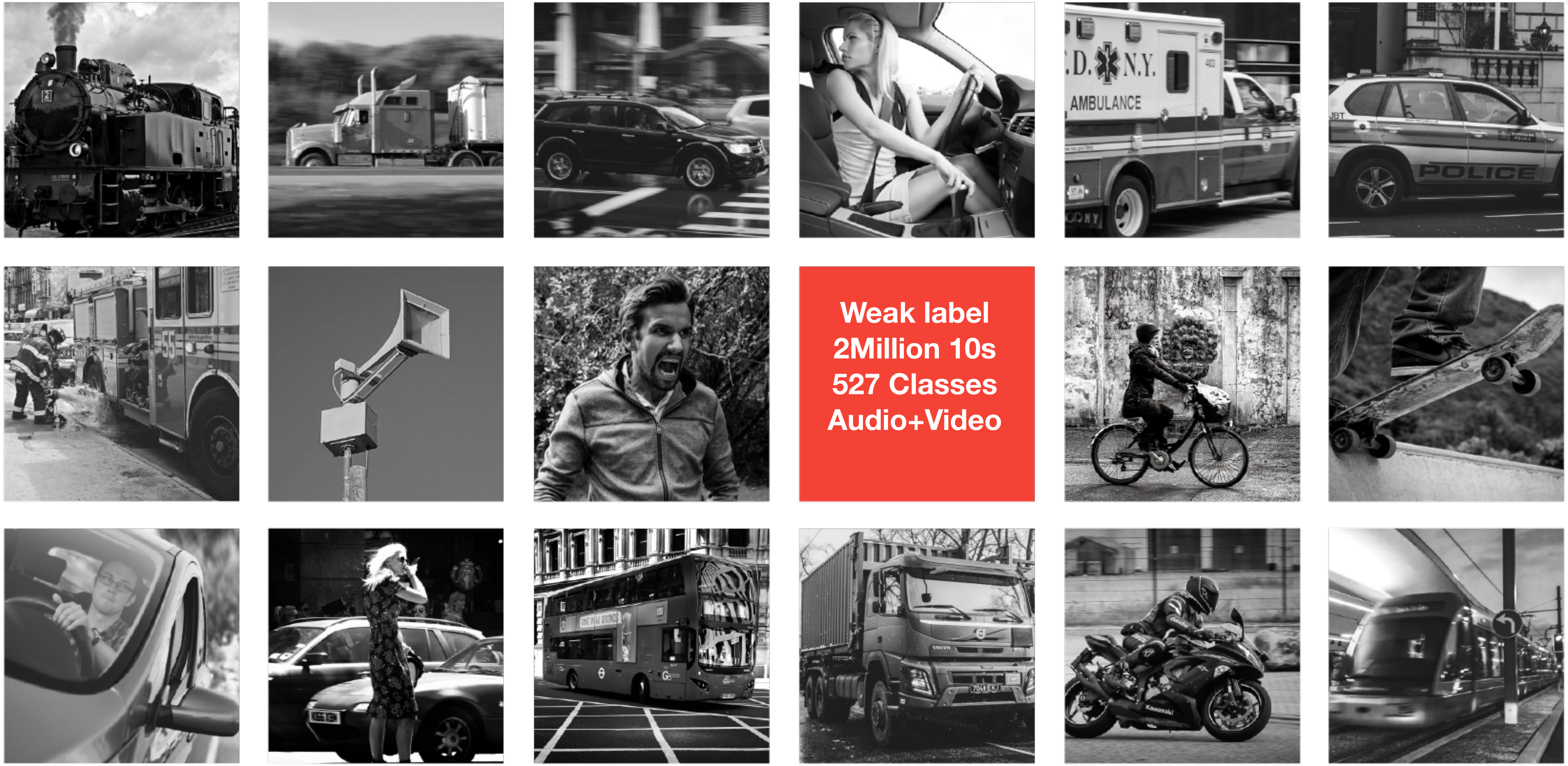


Echo

AI smart speakers

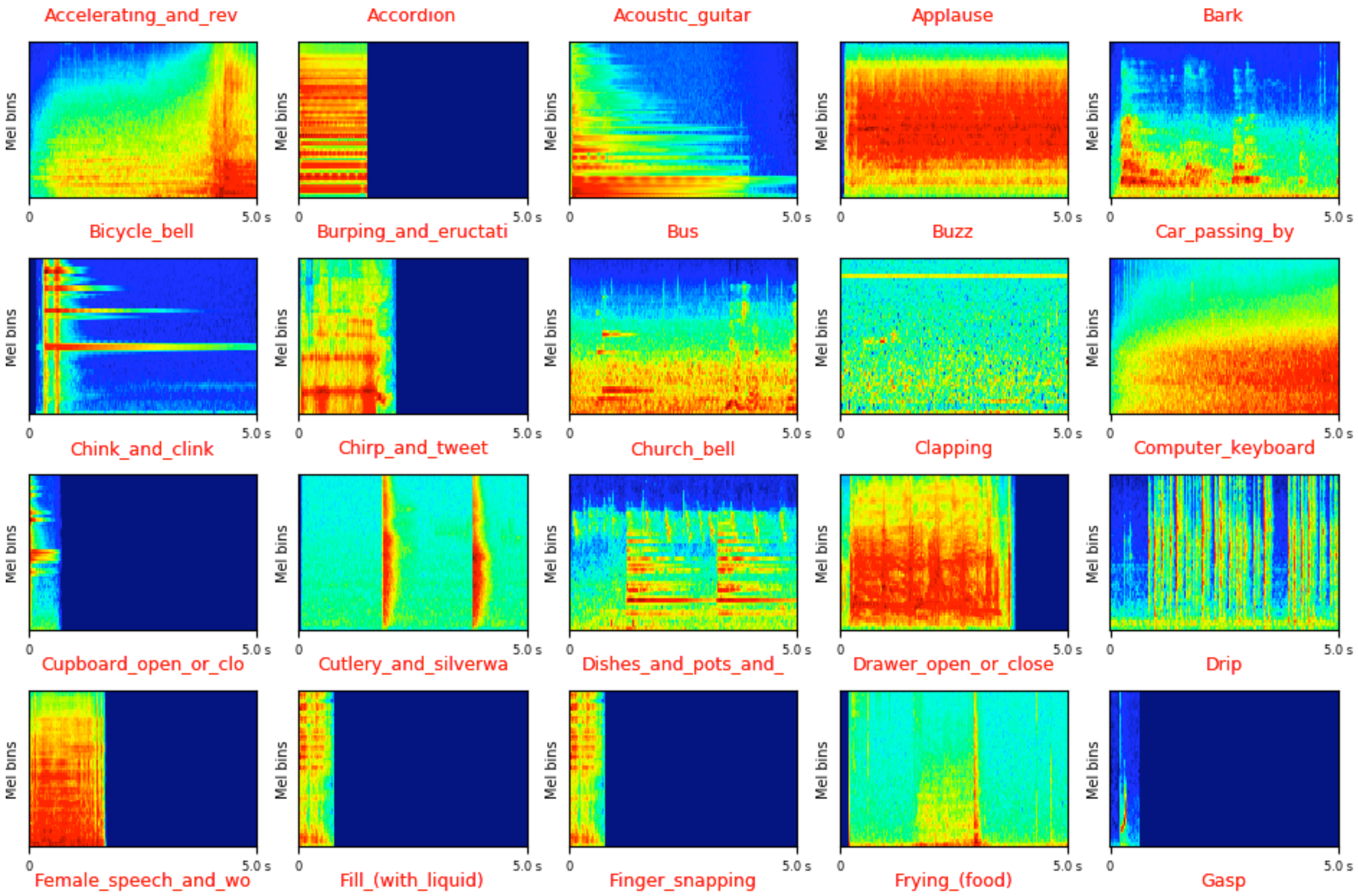
Scan QR code to read our paper:

Dataset: AudioSet, Kinects Sounds



Tasks of audio visual event recognition

- To predict the tag of an audio visual event, such as “Applause” or “Clapping”



LogMel spectrogram of selected audio recordings from AudioSet

Scan QR code to read our paper:

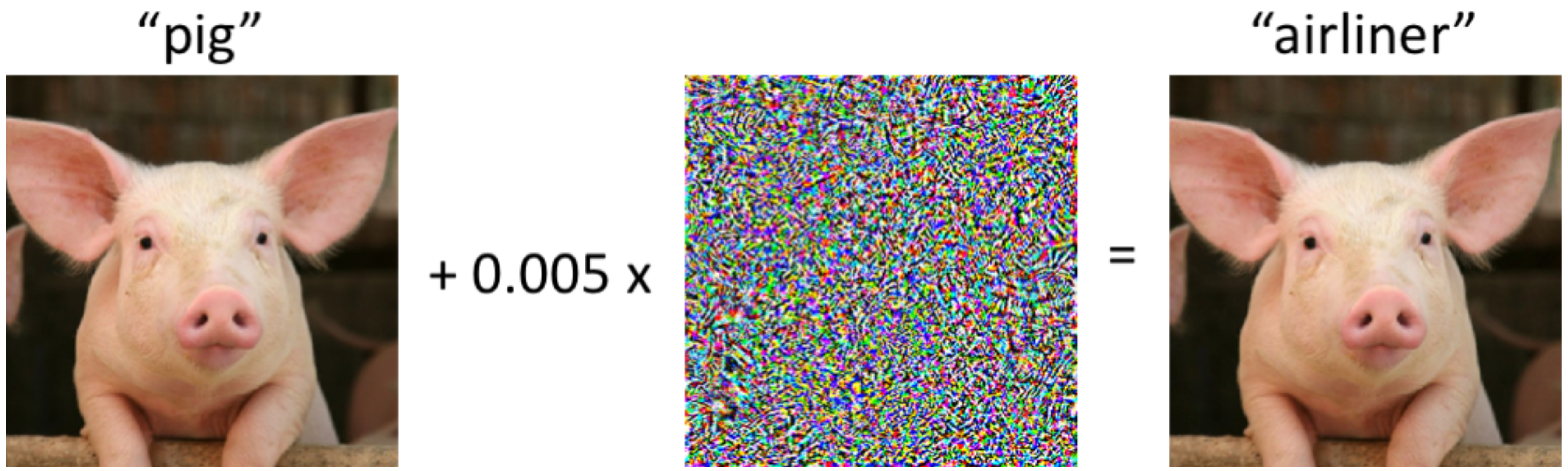
Scan QR code to read our paper:

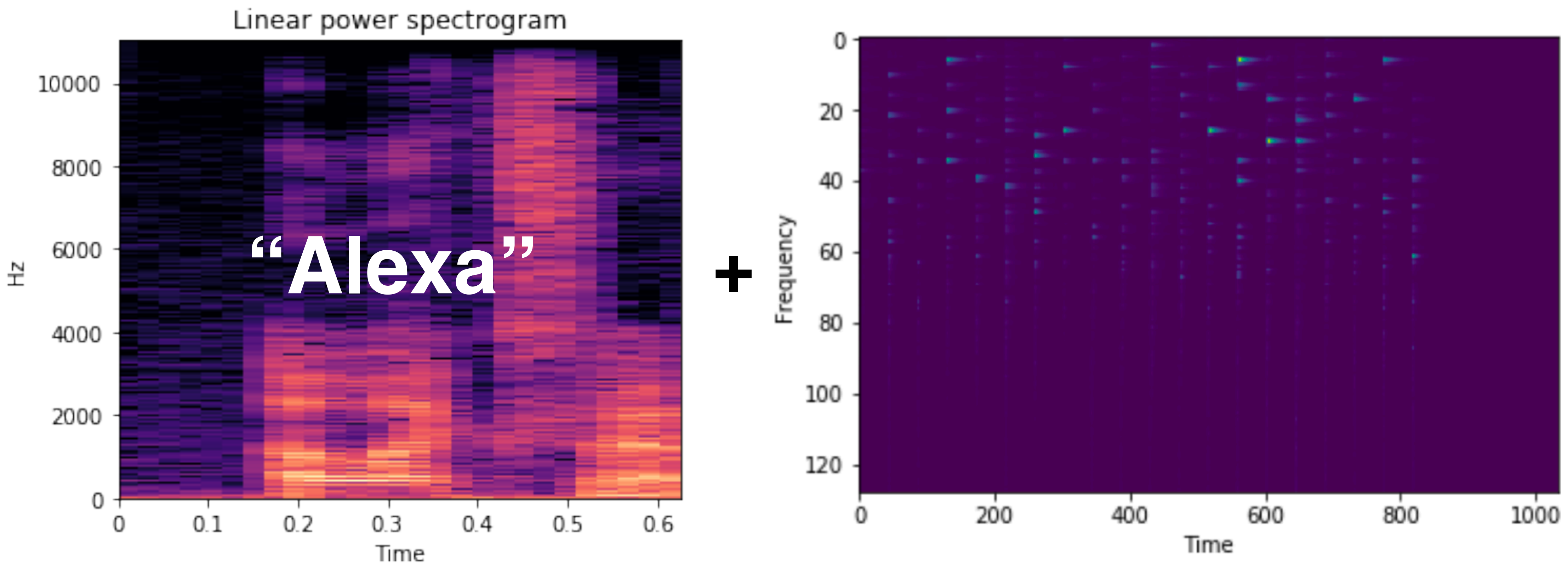
Background: Adversarial Examples

$$\min_{\theta} \sum_{x} \text{loss}(x, \theta)$$

$$\max_{\delta} \text{loss}(x + \delta, \theta)$$

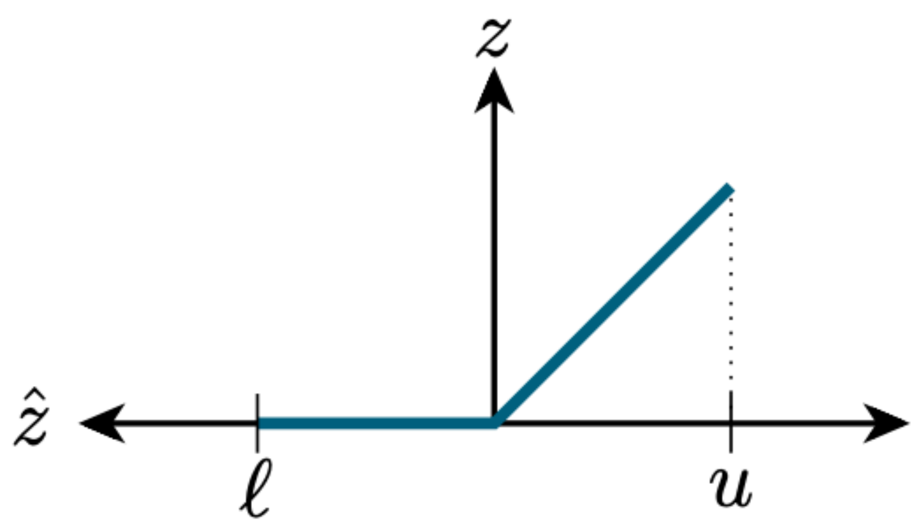
$$\mathbb{E}_{x,y\sim D}\left[\max_{x' \in P(x)} L(f(x'), y)\right]$$



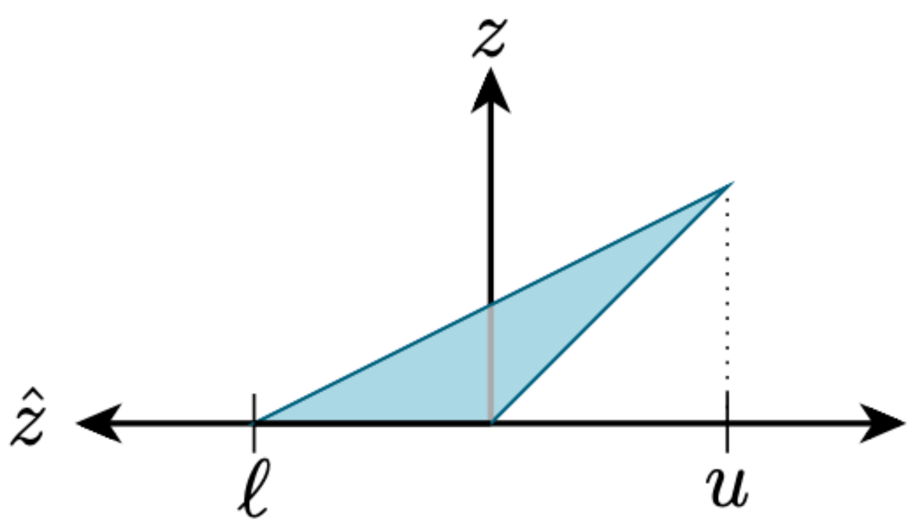


Adversarial Music
Li et al. [2019]

Background: Point-wise Robustness, Adversarial Training



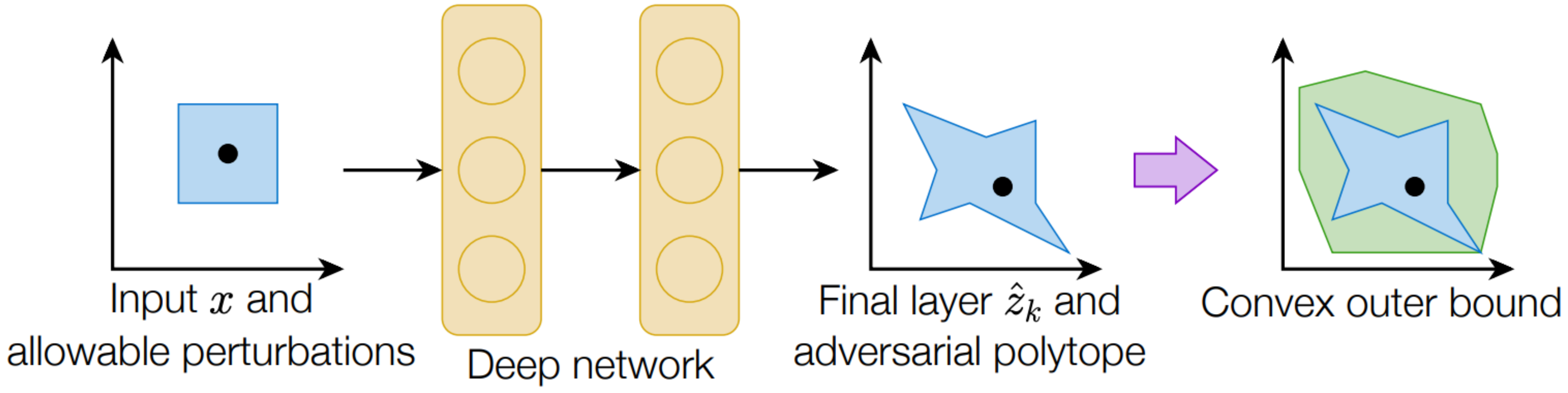
Bounded ReLU set

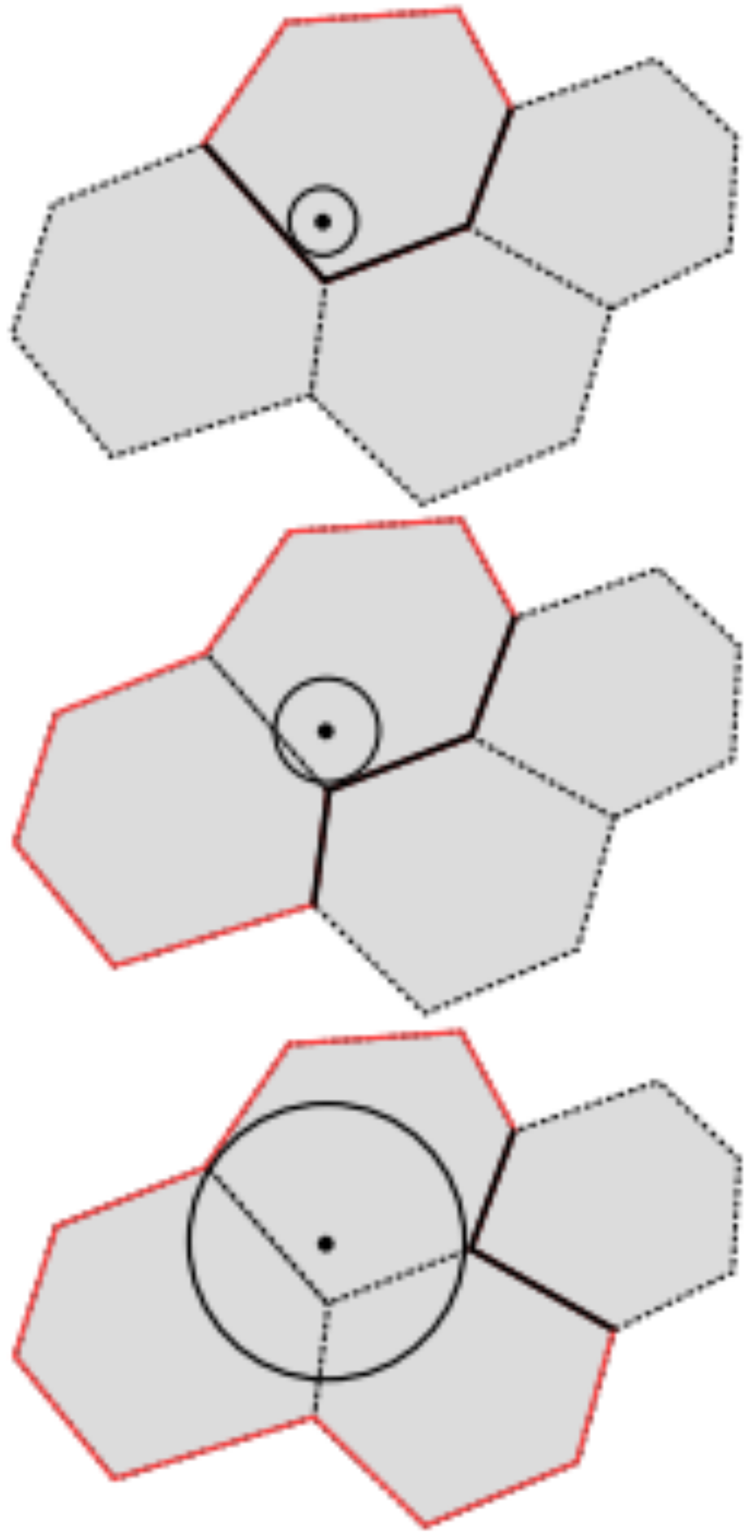


Convex relaxation

Provable Defense

Wong , Kolter [2018]





Centered Chebyshev Ball

Jordan et al. [2019]

ImageNet L2-robust accuracy

	ε-train	
ε-test	0.0	3.0
0.0	**76.13% / -**	57.90% / -
0.5	3.35% / 2.98%	**54.42% / 54.42%**
1.0	0.44% / 0.37%	**50.67% / 50.67%**
2.0	0.16% / 0.14%	**43.04% / 43.02%**
3.0	0.13% / 0.12%	**35.16% / 35.09%**

Adversarial Training

Ilyas, Madry et al. [2021]

Main Questions and Answers

Q1) *Are multi-modal models necessarily more robust than uni-modal models?*
Answer: Not Necessarily. *see Theorem 1.*

Q2) *How to efficiently measure the robustness of multi-modal learning?*
Answer: Previous works only focused on point-wise robustness, we should also look into class-wise robustness.

Q3) *How to fuse different modalities to achieve a more robust multi-modal model?*
Answer: *We propose multimodal mixup as a cheap alternative to adversarial training.*

Scan QR code to read our paper:

Multimodal Adversarial Perturbation

Scan QR code to read our paper:

Multimodal Loss:

$$L_{multi} = L(f(g(x_{m_1}) \oplus h(x_{m_2}) \oplus \cdots \oplus z(x_{m_k})), y),$$

Our Goal:

$$\underset{\delta_A \in C(x_A), \delta_V \in C(x_V)}{\text{Maximize}} [\mathbf{E}_{x_A, y \sim \mathcal{D}_\mathcal{A}; x_V, y \sim \mathcal{D}_\mathcal{V}}, [L(f(x'), y)]]$$

$$\text{subject to } C(x) = \{x \in \mathbb{R}^d : ||x||_p \leq \epsilon\}.$$

Audio Perturbation:

$$\delta_A = \mathcal{P}_\epsilon \left(\delta_A - \alpha \frac{\nabla_{\delta_A} L(f(g(x_A + \delta_A) \oplus h(x_V)), y)}{\|\nabla_{\delta_A} L(f(g(x_A + \delta_A) \oplus h(x_V)), y)\|_p} \right)$$

Video Perturbation:

$$\delta_V = \mathcal{P}_\epsilon \left(\delta_V - \alpha \frac{\nabla_{\delta_V} L(f(h(x_V + \delta_V) \oplus g(x_A)), y)}{\|\nabla_{\delta_V} L(f(h(x_V + \delta_V) \oplus g(x_A)), y)\|_p} \right)$$

Multimodal Perturbation:

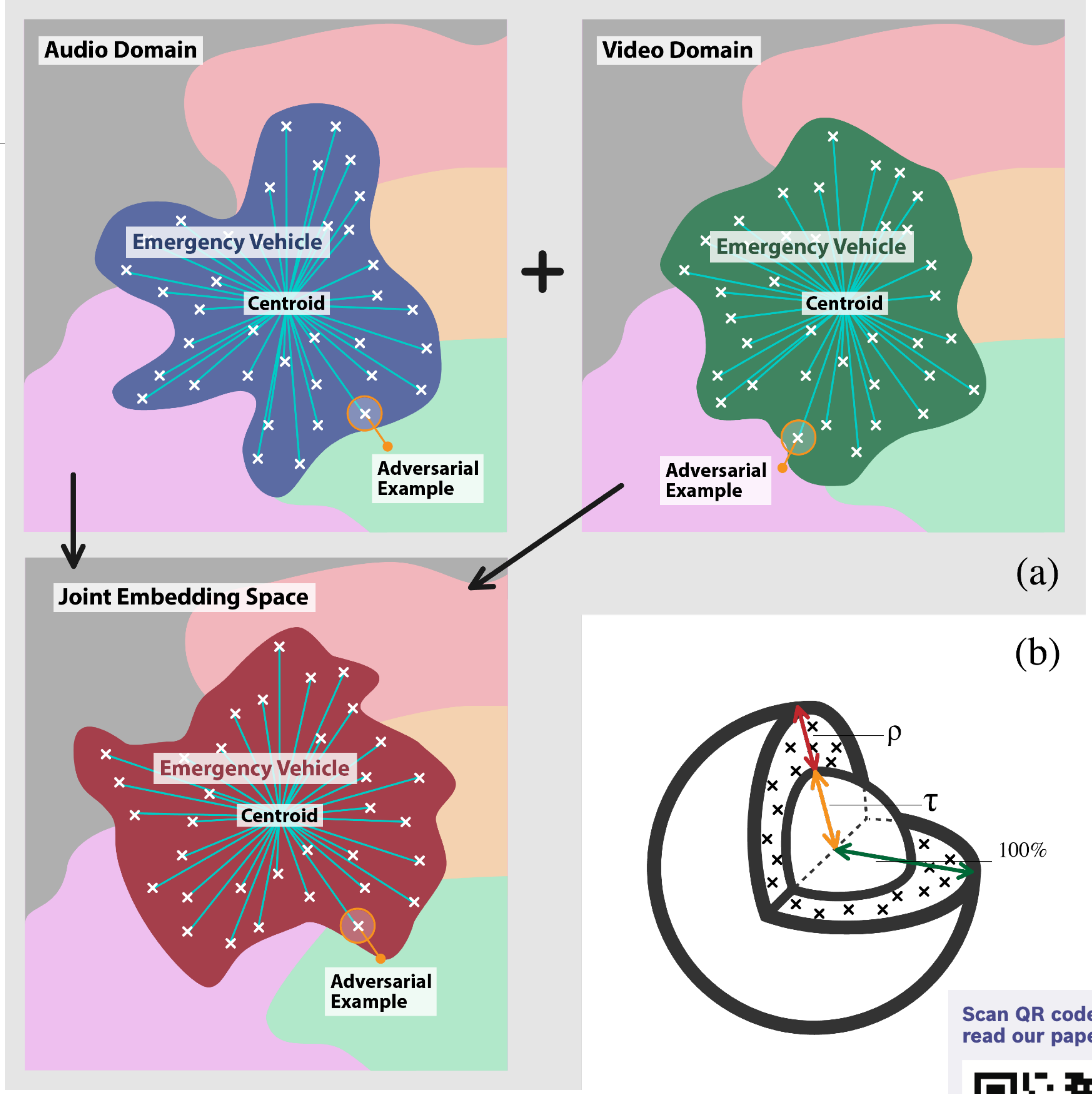
$$\delta_A, \delta_V := \mathcal{P}_\epsilon(\delta_{(V,A)} - \alpha \frac{\nabla_{\delta_{(V,A)}} L(f(h(x_V + \delta_V) \oplus g(x_A + \delta_A)), y)}{\|\nabla_{\delta_{(V,A)}} L(f(h(x_V + \delta_V) \oplus g(x_A + \delta_A)), y)\|_p})$$

Our Approach

Theorem 1 *There exists a sample $x_i \in \mathcal{D}$, and a unimodal sample-wise attack $\exists||\delta_{A,i}||_p \leq \epsilon_A$ or $\exists||\delta_{V,i}||_p \leq \epsilon_V$ that can break a multimodal fusion network $f((x_{V,i} \oplus x_{A,i}), y_i)$, changing its prediction label y_i.*

Here, *D* is the dataset, and ε_A and ε_V are the point-wise robustness threshold for each uni-modal of sample x_i. Therefore, as a conjecture, a unimodal attack can break a multimodal model, which we empirically verified the existence of such cases in our experiments.

The proof of Theorem 1 can be found in the appendix page.



Scan QR code to read our paper:

Scan QR code to read our paper:

Convolutional Self-Attention Network (CSN)

Audio Encoding Network

- 10 Stacked Convolutions and Pooling Layers. 5 pooling layers are insert after every 2 convolution layers.
- The outputs of the convolution encoder are fed into 2 transformer blocks to further model the global interaction among frames.

Video Encoding Network (3D-CNN)

• R(2+1)D block which decomposes the 3D (spatial-temporal) CNN into a spatial 2D convolution followed by a temporal 1D convolution.

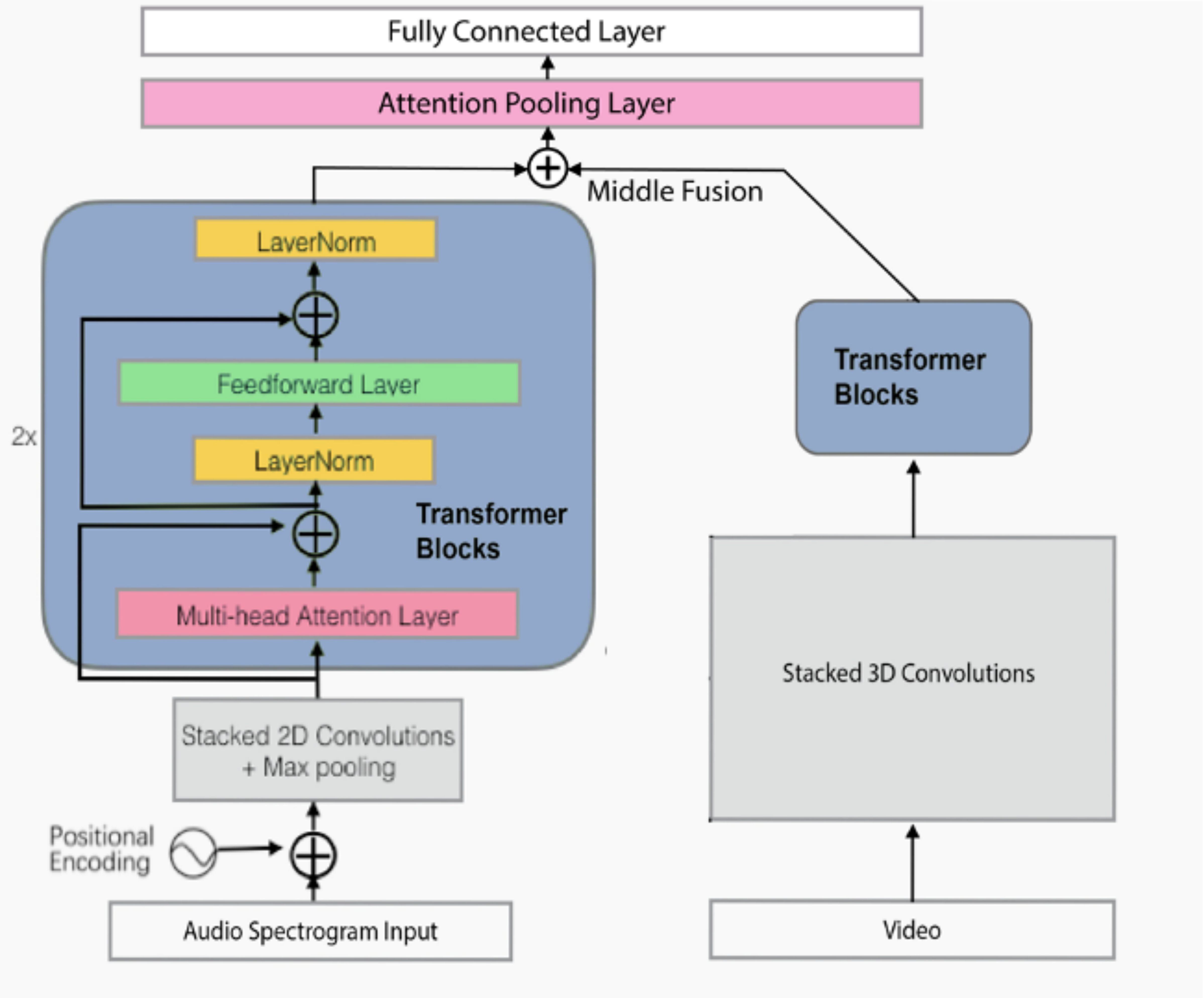


Fig1. The overall architecture of the network studied (left) audio branch (right) video branch

Main Questions and Answers

Q1) Are multi-modal models necessarily more robust than uni-modal models?
Answer: *Not Necessarily. see Theorem 1.*

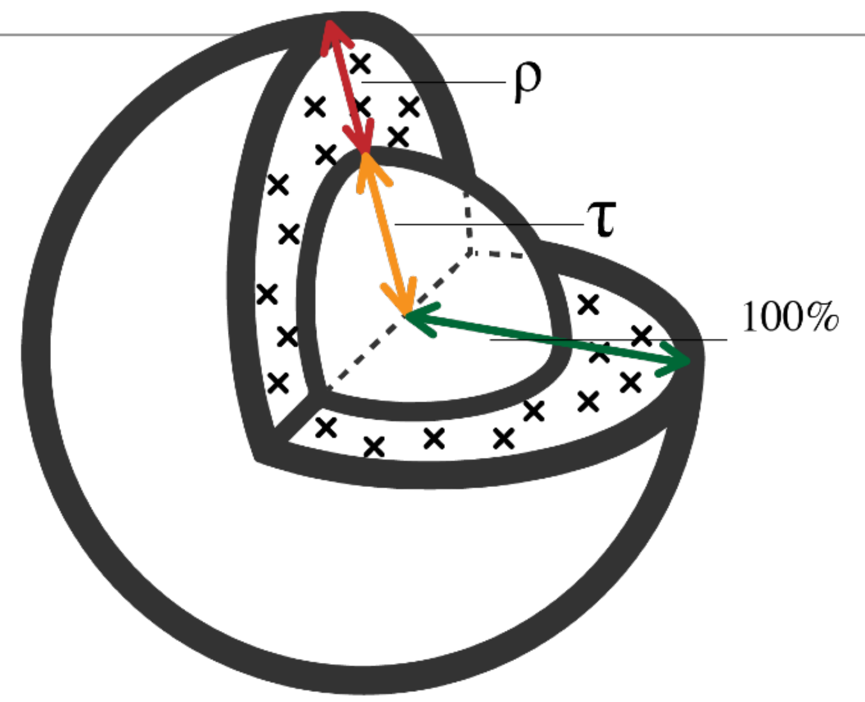
Q2) How to efficiently measure the robustness of multi-modal learning?
Answer: Previous works only focused on point-wise robustness, we should also look into class-wise robustness.

Q3) How to fuse different modalities to achieve a more robust multi-modal model?
Answer: *We propose multimodal mixup as a cheap alternative to adversarial training.*

Scan QR code to read our paper:

Class-wise Robustness Metric

(b)



Centroid-based Density Metric :

$$\rho_c^{R_\tau,p,c} = \frac{n_c - n_{\tau,c}}{log(V_d^p(R_{p,c})) - log(V_d^p(R_{\tau,p,c}))}$$

In the equation, the numerator is the number of class samples whose l_p distance to centroid larger than τ quantile of samples in class c;

$R_{\tau,p,c}$ is the τ quantile of all class sample's l_p distance to the class's centroid.

Intuitively, the density in the outer crust of a ball as is shown in Fig. 1(b) above.

Generally, the higher the density of the crust, the more robust the samples within/below the crust are.

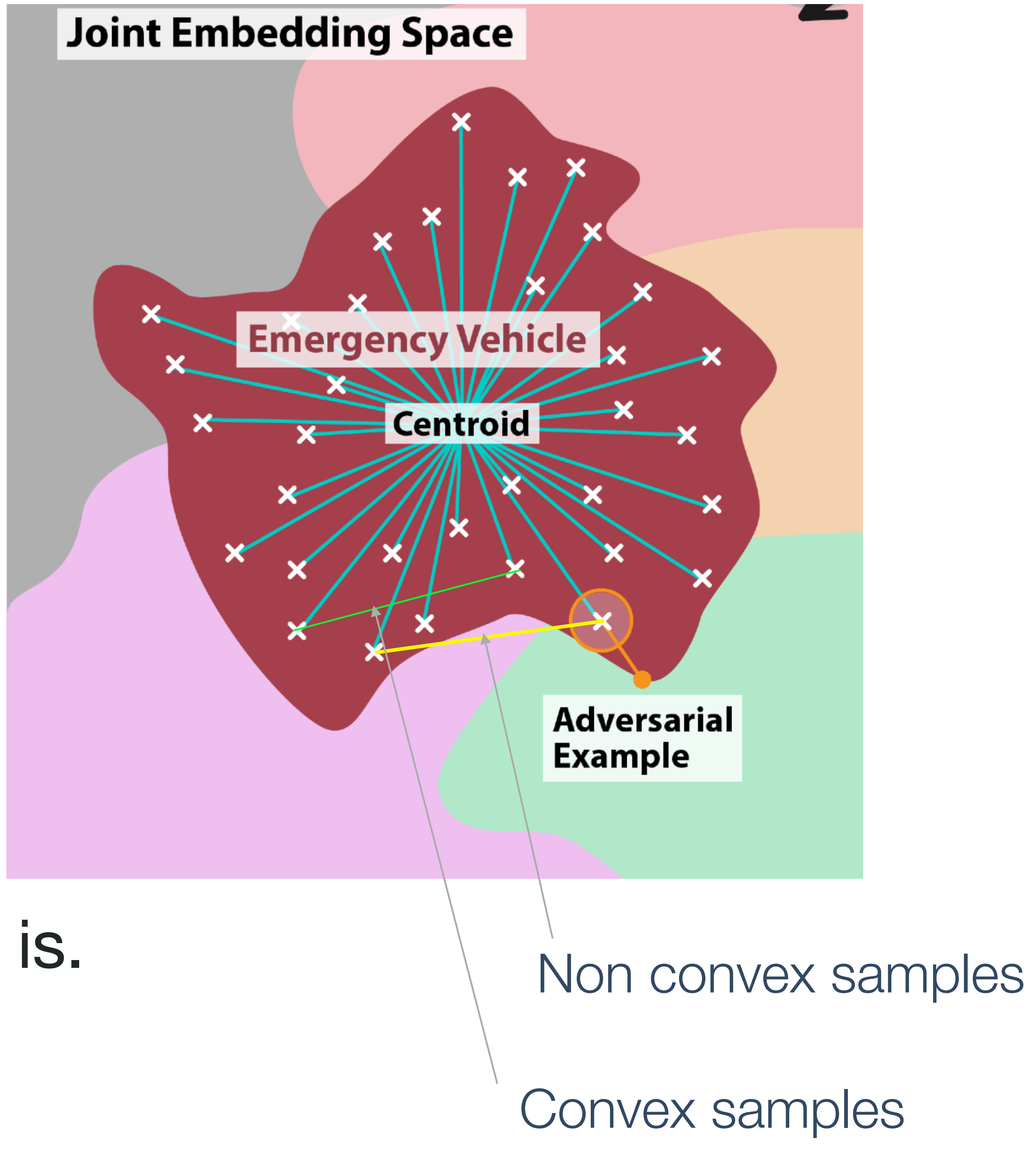
Scan QR code to read our paper:

Class-wise Robustness Metric

Convexity-based Metric :

$$\kappa_c = \frac{\sum_{i=1}^{n} \mathbb{1}\{f(\hat{x}_i) = y_c\}}{n}$$

For each class C in the dataset, we construct the convex set of $S = \{x_s | x_s = \theta x_1 + (1 - \theta) x_2,\ \theta \sim U[0,1],\ \forall x_1, x_2 \in C\}$, and sample n points from it $\{x_1, \ldots, x_n | x_i \in S\}$, we set $n = 2000$, where y_c is the class label.



The higher the κ_c is, the more convex the decision boundary of class C is.

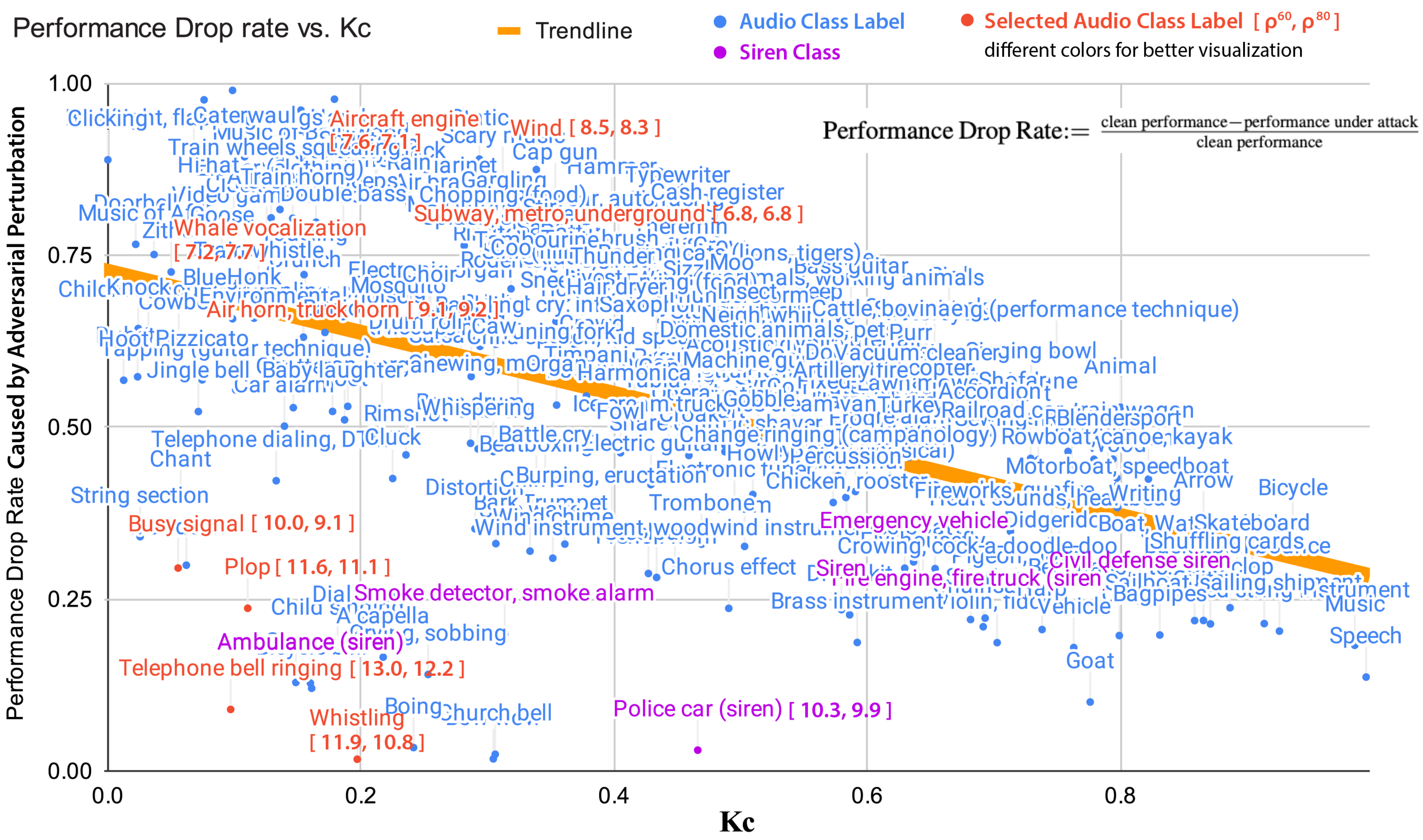
Scan QR code to read our paper:

Main Questions and Answers

Q3) How to fuse different modalities to achieve a more robust multi-modal model?

Answer: We propose **multimodal mixup** as a cheap alternative to adversarial training. We desire to augment the **less convex** classes of training data with more samples from the "**denser**" samples which are closer to the center of its feature space.

We tune mixup temperature between audio and video samples according to empirical threshold of the above-mentioned Density metric ρ and the Convexity metric κ_C


Performance Drop rate vs. Kc
Trendline
Audio Class Label
Siren Class
Selected Audio Class Label [ρ60, ρ80]
different colors for better visualization
Performance Drop Rate:= clean performance − performance under attack / clean performance
Performance Drop Rate Caused by Adversarial Perturbation
Kc
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
Aircraft engine [7.6, 7.1]
Wind [8.5, 8.3]
Subway, metro, underground [6.8, 6.8]
Whale vocalization [7.2, 7.7]
Air horn, truck horn [9.1, 9.2]
Busy signal [10.0, 9.1]
Plop [11.6, 11.1]
Smoke detector, smoke alarm
Ambulance (siren)
Telephone bell ringing [13.0, 12.2]
Whistling [11.9, 10.8]
Police car (siren) [10.3, 9.9]
Emergency vehicle
Civil defense siren
Siren
Fire engine, fire truck (siren)
Cap gun
Typewriter
Cash register
Train wheels squealing
Double bass
Chopping (food)
Tambourine
Hair dryer
Bass guitar
Cattle, bovinae
Domestic animals, pets
Purr
Vacuum cleaner
Animal
Harmonica
Accordion
Gobble
Blender
Rowboat, canoe, kayak
Motorboat, speedboat
Arrow
Bicycle
Writing
Skateboard
Shuffling cards
Bagpipes
Music
Speech
Goat
Hoot
Pizzicato
Jingle bell
Baby laughter
Car alarm
Whispering
Telephone dialing, DTMF
Cluck
Chant
String section
Burping, eructation
Trombone
Chorus effect
Crowing, cock-a-doodle-doo
Brass instrument
Chicken, rooster
Fireworks
Percussion
Battle cry
A capella
Church bell
Boing
Wind instrument, woodwind instrument
Mosquito
Environmental noise
Artillery fire
Railroad car, train wagon
Change ringing (campanology)
Insect
Saxophone
Tuning fork
Timpani
Goose
Clickin
Caterwaul
Music of Asia
BlueHonk
Knock
Cowbell
Tapping (guitar technique)
Caw
Drum roll
Crowd
Fowl
Chewing, mastication
Crying, sobbing
Didgeridoo
Vehicle
Sailboat, sailing ship
Ice cream truck
Hammer
Scary music
Clarinet
Gargling
Thunder
Moo
Scratching (performance technique)
Neigh, whinny
Singing bowl
Distortion
Trumpet
Wind chime
Electric guitar
Electronic tuner
Rimshot

Scan QR code to read our paper:

Results:

Table 1. *Performance of our best performing model on AudioSet, and their performance against the adversarial perturbation, using the overall architecture shown in Fig 2.*

Here, mAP is the mean average precision, AUC is the area under the false positive rate and true positive rate.

The d-prime can be calculated from AUC [1].

AT denotes adversarial training.
*A **red** text color indicates the **most potent** perturbation against the model.*

Models	Attack	**mAP**	**AUC**	**d-prime**
Audio UniModal (PANNS) [23]	No	0.383	0.963	2.521
Audio UniModal	Yes	0.183	0.895	1.770
Mid Fusion (G-blend) [14]	No	**0.427**	0.971	2.686
Mid Fusion	Yes A+V	0.182	0.889	1.836
Mid Fusion	Yes V-only	0.339	0.954	2.441
Mid Fusion	Yes A-only	0.310	0.940	2.276
Mid Fusion mixup	No	0.424	0.972	2.711
Mid Fusion mixup	Yes A+V	0.234	0.891	1.983
Mid Fusion *AT*	No	0.397	0.964	2.530
Mid Fusion *AT*	Yes A+V	0.199	0.900	1.861

Carnegie Mellon University
Language Technologies Institute

22-27 MAY 2022
ICASSP 2022
SINGAPORE

Scan QR code to read our paper:

IEEE Signal Processing Society

IEEE

Conclusion

1. Multimodal Networks are not always more robust than their unimodal counterparts.
2. Our density and convexity metric could effectively measure robustness of models in large-scale.
3. We propose multimodal mixup as an alternative to adversarial training.

Thank you!

Carnegie Mellon University
Language Technologies Institute