AVNet: Multimodal Emergency Vehicle Classification via Audio-Visual Transformers and Knowledge Distillation

Amar Dabaja, PhD Electrical Engineering Student, College of Engineering
Vijay John, PhD, Associate Professor, College of Arts and Sciences
Lawrence Technological University

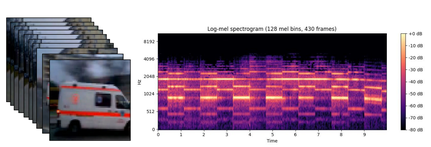
BACKGROUND INFORMATION

- As autonomous driving advances, vehicles must reliably detect and respond to critical road events to operate safely.
- **Recognizing emergency vehicles** is essential so the vehicle can yield promptly as required by traffic laws and to protect first responders and the public.
- Current methods exist to detect emergency vehicles based on either visual data or audio recordings. Existing bimodal sensor fusion models rely on the presence of both modalities and can't be used for unimodal perception
- Robust detection requires **fusing both** audio and visual cues to handle challenging conditions such as occlusions, poor lighting, or elevated noise levels that may obscure sirens or line of sight.

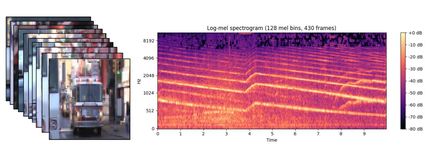
PROJECT OBJECTIVES AND PROPOSAL

- We propose **AVNet**, a novel multimodal deep learning model that classifies emergency vehicles into **four categories (police car, ambulance, fire truck, and road/non-emergency)** using **either audio data, video data, or their fusion**.
- AVNet is designed to handle three different input subsets (audio-alone, video-alone, or audio-video multimodal data). The system remains robust even in the event of missing modalities by using **null embeddings for missing modalities**.
- The model integrates an **Audio Spectrogram Transformer (AST) branch and a Vision Transformer** (ViT) **branch.**
- A **student-teacher knowledge distillation strategy** is adopted using unimodal teacher models to train a multimodal student network.

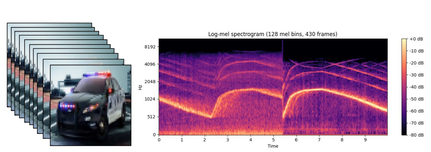
DATASET AND PREPROCESSING



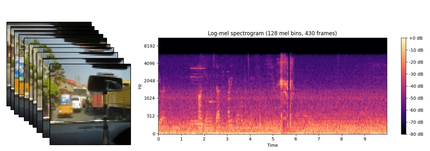
Ambulance Class Example



Fire Engine Class Example



Police Car Class Example

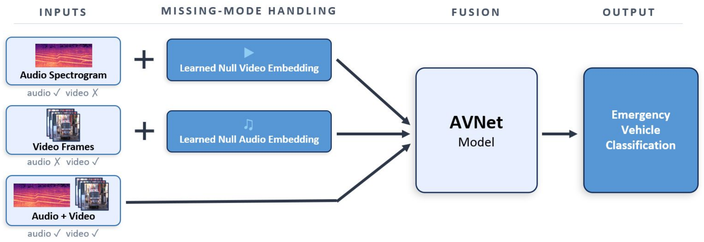


Road Traffic Class Example

MODEL ARCHITECTURE

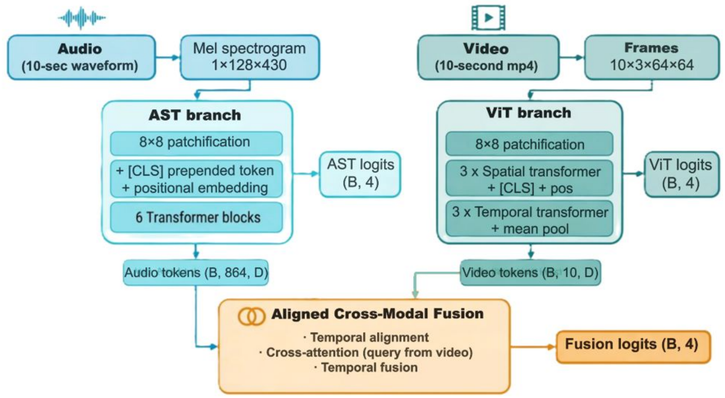
OVERVIEW

The AVNet model supports audio-only input, visual-only input, or combined audio-visual input. Missing modalities are handled via an added null embedding.

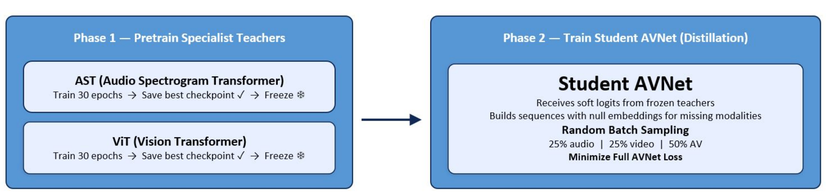


AVNET COMPONENTS

The AVNet model consists of an Audio Spectrogram Transformer path for audio input, a Vision Transformer path for video input, and a fusion branch that aligns the two temporally.



MODEL TRAINING: KNOWLEDGE DISTILLATION



Optimization involves two loss functions:

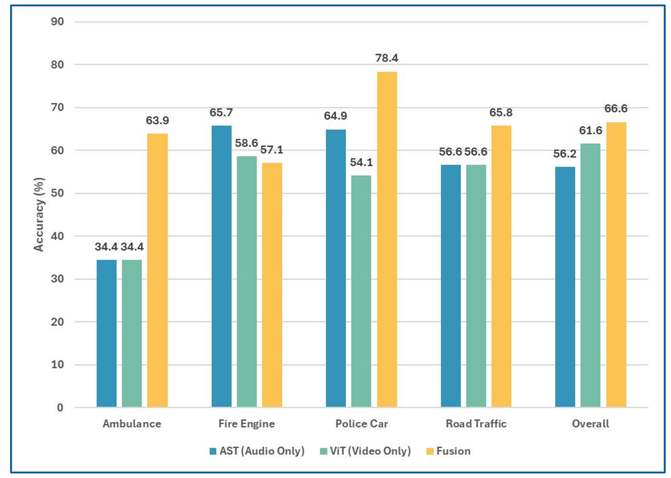
- KL Divergence Loss measures how different the student's soft predictions are from the teacher's soft predictions.

$$\mathcal{L}_{\text{KD}}(s,t;T) = T^2 \cdot \text{KL}\left(\sigma\left(\tfrac{t}{T}\right) \parallel \sigma\left(\tfrac{s}{T}\right)\right)$$

- Cross-Entropy with Label Smoothing compares the true label to the student's predictions, with label smoothing to prevent overconfidence.

$$\mathcal{L}_{\text{CE}}(s,y) = -\sum_{c} \tilde{y}_c \log \sigma(s_c)$$

RESULTS AND IMPLICATIONS



Model Strengths:

- Multi-modal fusion improves accuracy
- Handling of missing modalities
- Modular design allows flexibility
- Soft labels from teachers stabilize learning

Model Weaknesses:

- Small dataset (~2800 clips)
- Limited GPU capacity
- Class-specific confusion (e.g. acoustic overlap, visual resemblance)
- No pretrained weights

FUTURE WORK

- Complete comparative study to assess model performance against existing classification models that utilize audio and visual data.
- Complete ablation study to optimize model parameters.
- Incorporate additional datasets for improved accuracy and generalization.
- Utilize pretrained models for improved initialization.
- Increase model capacity on a larger GPU.
- Expand to include additional classes, such as train horns and air medical transportation.
- Add a real-time processing interface for embedded systems applications.

REFERENCES

1. G. Hinton, O. Vinyals, J. Dean, Distilling the Knowledge in a Neural Network, NIPS Deep Learning Workshop, 2015.
2. Y. Gong, Y-A. Chung, J. Glass, AST: Audio Spectrogram Transformer, Interspeech, 2021.
3. A. Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR, 2021.
4. J. F. Gemmeke et al., Audio Set: An Ontology and Human-Labeled Dataset for Audio Events, ICASSP, 2017.
5. A. Vaswani et al., Attention Is All You Need, NeurIPS, 2017.
6. A. Romero et al., FitNets: Hints for Thin Deep Nets, ICLR, 2015.
7. N. Samson et al., Detection and Classification of Emergency Vehicles from Audio and Video Inputs using Deep Learning Techniques, JSCDM, 2025.