

# Real-Time Idling Vehicles Detection Using Combined Audio-Visual Deep Learning

Xiwen LI <sup>a,1</sup>, Tristalee MANGIN <sup>b</sup>, Surojit SAHA <sup>a</sup>, Rehman MOHAMMED <sup>a</sup>,  
Evan BLANCHARD <sup>b</sup>, Dillon TANG <sup>b</sup>, Henry POPPE <sup>b</sup>,  
Ouk CHOI <sup>c</sup>, Kerry KELLY <sup>b</sup>, Ross WHITAKER <sup>a</sup>

<sup>a</sup>Scientific Computing and Imaging Institute, University of Utah, USA

<sup>b</sup>Department of Chemical Engineering, University of Utah, USA

<sup>c</sup>Department of Electronics Engineering, Incheon National University, Republic of Korea

**Abstract.** Combustion vehicle emissions contribute to poor air quality and release greenhouse gases into the atmosphere, and vehicle pollution has been associated with numerous adverse health effects. Roadways with extensive waiting and/or passenger drop-off, such as schools and hospital drop-off zones, can result in a high incidence and density of idling vehicles. This can produce micro-climates of increased vehicle pollution. Thus, the detection of idling vehicles can be helpful in monitoring and responding to unnecessary idling and be integrated into real-time or off-line systems to address the resulting pollution. In this paper, we present a real-time, dynamic vehicle idling detection algorithm. The proposed idle detection algorithm and notification rely on an algorithm to detect these idling vehicles. The proposed method relies on a multisensor, audio-visual, machine-learning workflow to detect idling vehicles visually under three conditions: moving, static with the engine on, and static with the engine off. The visual vehicle motion detector is built in the first stage, and then a contrastive-learning-based latent space is trained for classifying static vehicle engine sound. We test our system in real-time at a hospital drop-off point in Salt Lake City. This in situ dataset was collected and annotated, and it includes vehicles of varying models and types. The experiments show that the method can detect engine switching on or off instantly and achieves 71.02 average precision (AP) for idle detection and 91.06 for engine off detection.

**Keywords.** emission mitigation, multimodal ITS, sensing, vision and perception

## 1. Introduction

Poor air quality negatively impacts human health and was globally responsible for 6.5 million deaths and 21 billion in healthcare costs in 2015 [1]. In particular, vehicle pollution has been associated with numerous adverse health effects, such as reduced cognitive function, cancer, and poor reproductive outcomes [2–4]. Likewise, idling vehicles are

<sup>1</sup>Corresponding Author: Xiwen Li, Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, 84112, United States; Email: xiwen.li@utah.edu

significant contributors to greenhouse gas emissions [5]. Roadways where idling vehicles tend to congregate, such as schools and hospital drop-off zones, can produce microclimates of increased vehicle pollution [6]. Populations that are particularly vulnerable to vehicle pollution include children and individuals in wheelchairs since their breathing height is closer to the height of combustion exhaust [7, 8]. Idling vehicles are especially problematic in confined locations such as underground mines [9]. Moreover, idling among fleets, such as long-haul idling trucks at depot/delivery centers, causes excess operational costs due to wasted fuel and engine wear [10, 11].

The detection or monitoring of idling vehicles can impact policies that can subsequently reduce pollution. For instance, research has found that while conventional, static, anti-idling signage and education campaigns have mixed results, anywhere from little effect on driver behavior to improving air quality [6, 12–17], dynamic signage may have a greater impact. For instance, behavioral research has found that dynamic radar-based speed displays are more effective than static signage at reducing vehicle speeds [18–24].

This paper focuses on the design and evaluation of an idling vehicle detection (IVD) system that can be deployed in parking or drop-off areas in order to monitor and respond to driver behaviors. To the best of our knowledge, only one previous work [25] has discussed a method for IVD. That method relies on infrared imaging to detect heat from the engine block of a vehicle, which has several limitations, as described in the next section.

## 2. Related Work

### 2.1. Idling Vehicle Detection

Previous work [25] has proposed to automatically detect IVs via the use of an infrared camera, which monitors the target area. They perform object detection on a heatmap. This method has several disadvantages: (1) The first is high latency because it takes time for heatmap to accumulate and dissipate. Also, many infrared cameras often have relatively low frame rates (e.g., one image every 5 seconds). (2) According to our preliminary experiments, environments with direct sunlight or high ambient temperatures can adversely affect the model's ability to detect a hot engine block. (3) Engine blocks and exhaust pipes (as noted by the authors [25]) are the main high heat area; however, false positives/negatives occur when the engine block faces away from the camera because the engine block is the main heat source. In addition, our preliminary results on detecting heat from vehicle exhaust (rear of car) are inclusive. (4) Infrared imaging cameras are generally more expensive and not easily deployable in a wide variety of settings. The proposed method uses a common RGB webcam and wireless microphone array for video and audio data acquisition. Additionally, the problem is defined as an audio-visual workflow, including vehicle motion detection and audio classification. The method uses real-time video and audio clips and can detect engines switching on or off in less than 1 second. Thus, the proposed audio-visual approach represents an attractive alternative to infrared-based IVD.

### 2.2. Video Understanding

Video understanding is an important area of study within the field of computer vision. Action detection, often the main task of video understanding, typically estimates a 2D

bounding box (region in the image) and label for each action event on video frames. Convolution neural networks (CNNs) have been successfully applied to vehicle motion detection, as in [26–30]. With respect to autonomous driving, works using ego-vehicle cameras such as [31–34] focus on detecting vehicle motion. However, the problem statement in this paper requires surveillance-style camera placement, as in [25, 35], for the purpose of monitoring the target area. The proposed design is vision guided, and we use the SoTA action detection model YOWO [36] as the first stage for locating vehicles in the video frames.

### 2.3. Audio Classification

A great deal of research has addressed problems related to sound classification such as [37–39] and speech diarization [40]. Devices such as microphone arrays and acoustic cameras [41] can help localize sound sources relative to a visual frame of reference. We have evaluated microphone arrays and beamforming algorithms to separate vehicle engine sound in each direction. Nevertheless, from empirical evaluation, we have observed that the beamforming technique cannot reliably resolve a single vehicle in outdoor setup because of the far-field attenuation and the ambient noise, and the localization accuracy is poor. Therefore, in order to receive a stable and clear audio signal of vehicle engines, we place individual wireless microphones in a row at the roadside, as shown in Fig. 3, and use machine learning models to identify idling engines. Due to access to limited training data, we resort to models built using unsupervised learning techniques, such as contrastive learning [42, 43], on publicly available audio datasets [44–46].

### 2.4. Audio-Visual Learning

Audio-visual learning is an emerging topic in computer vision. Several works, such as [47–50], learn audio-visual, co-occurrence features. However, our IVD problem cannot be resolved by feature co-occurrence because a stationary vehicle has visual presence but lacks audio presence. Furthermore, our experiments show that vehicles are accurately identified in videos, regardless of their engine status. Thus, we have opted for a vision-guided system design, rather than pure, mixed audio-visual recognition.

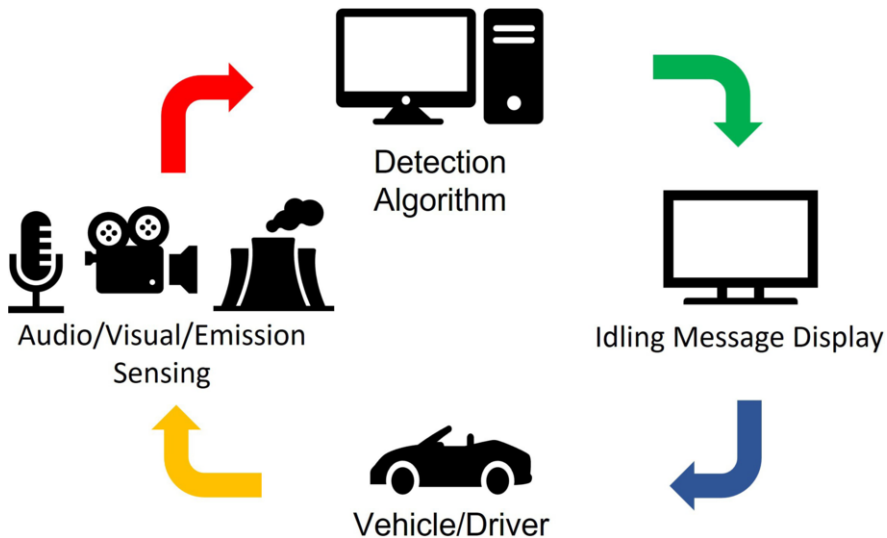
## 3. Method

### 3.1. Problem Definition

We define the problem as localizing and deciding if multiple vehicles in a video clip are idling individually in a drop-off area. We define three vehicle status classes  $Y \in \{Y_{\text{moving}}, Y_{\text{off}}, Y_{\text{idling}}\}$  and show them in Fig. 2(b):

- **Moving.** A vehicle is moving.
- **Engine Off.** A vehicle is stationary with the engine off.
- **Idling.** A vehicle is stationary with the engine on.

By definition, an electric vehicle is either moving or off. We define the problem in an audio-visual manner. To be specific, given a video clip  $V \in R^{D \times H \times W \times C}$  visually con-



**Figure 1.** Proposed System Design. The yellow arrow collects vehicle motion, engine sound, and pollutant concentrations. The red arrow represents data transmission to the computer. The green arrow denotes sending the predicted idling status to the displays. The blue arrow represents the driver receiving the information from the display and potentially making behavior changes.

taining vehicles  $v_1, v_2, \dots, v_n$ , microphone pixel location  $L$  and audio clips  $M \in R^{N_c \times SR}$  at the same moment, our model estimates a bounding box  $BB^{v_i}$  and  $Y^{v_i}$ .  $D$  is the number of video frames.  $H, W$ , and  $C$  are the height, width, and channel number of an RGB frame.  $N_c$  is the number of wireless microphones.  $SR$  represents the microphone sample rate. The observed time-domain audio signal  $M$  is defined as the addition between engine sound  $S$  and environment noise  $N$ :

$$M = S + N(M, S, N \in R^{N_c \times SR}) \tag{1}$$

The model estimates class label  $Y^{v_i}$  of vehicle  $v_i$  given information of bounding box  $BB$ , motion label  $Y_{motion}$  and nearest audio signal  $M^{v_i}$ :

$$P(Y^{v_i} | BB^{v_i}, Y_{motion}^{v_i}, M^{v_i}) \tag{2}$$

To solve this problem, we propose a two-stage visual-guided audio classification algorithm shown in Fig. 2(a). A vehicle detector is trained in the first stage using the SoTA video understanding model, which learns to detect a bounding box  $BB^{v_i}$  and a moving or stationary label  $Y_{motion}^{v_i}$  for each detected vehicle. For each stationary vehicle detected in the frame, we determined (through proximity in the image space) the closest microphone. In the second stage, the model classifies audio acquired from each of the closest microphones into sound with/without engine presence.

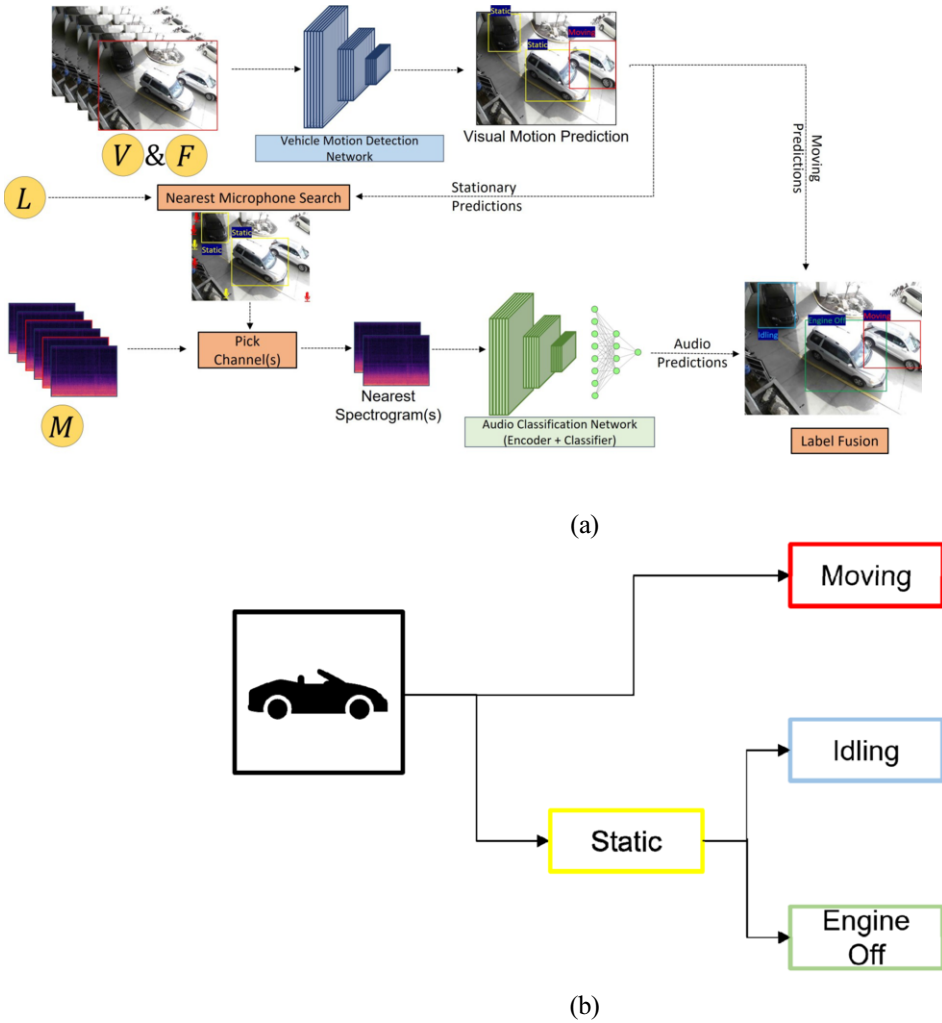


Figure 2. (a) Our IVD Algorithm. (b) Class Definition Hierarchy.

### 3.2. System Setup

Our system setup at the test location is shown in Fig. 3. We set up an RGB camera on a tripod at an elevation of approximately 20 feet, pointed toward the target area. For instance, at the hospital collection area, the RGB camera monitors a two-lane drop-off area, and the wireless microphone transmitters are evenly spaced along the roadside (at 2.6 meters intervals). The microphone transmitters send the acquired signals to the microphone receivers. The microphone receivers are connected to the desktop computer, which collects the data and runs the algorithm. An EMMET C960 webcam was used for the video, and 3 Rode Wireless GO II sets with unidirectional microphones are used for the audio. Each Rode Wireless GO II set has one receiver and two transmitters/microphones, resulting in a total of six microphones.



**Figure 3.** System Setup. Wireless microphones are stuck on the right wall. The camera is mounted on a tripod near rocks.

### 3.3. Vehicle Motion Detection

The vehicle motion detector (as in Fig. 2) estimates, for every detected vehicle, whether it is in motion  $Y_{moving}$  or stationary  $Y_{stationary}$ . For achieving this, we use a deep-learning, video-based, object-detection model, YOWO[36], considered a state-of-the-art benchmark, which is a video-understanding extension of the (static) object detection model YOLO[51]. The neural network extracts 3D clip features using ResNeXt[52], extracts 2D frame features using DarkNet[51], and fuses 3D features and 2D features for action detection. YOWO's input is a video clip  $V \in R^{D \times H \times W \times C}$  and the video clip's last frame  $F \in R^{H \times W \times C}$  of  $V$ . In general, YOWO learns and detects vehicles' motion on  $F$  using  $V$  and  $F$ .

### 3.4. Nearest Microphone Search

Once each vehicle's location and moving/stationary status is found, the algorithm finds the closest microphone to that vehicle. For this, the algorithm relies on image (pixel) locations for each microphone in the video frame, determined interactively by users at the time of system setup. In the setup, we develop a script that accepts a mouse click for each of the six microphone locations in the video frame and produces the pixel-microphone dictionary  $L$  for that deployment. The process takes approximately 1 minute. Having the user input these locations at the beginning of each deployment ensured that the algorithm had the correct microphone pixel location, because the video camera angle can vary due to equipment takedown and installation each deployment. Each predicted  $BB_{v_i}$ 's centroid is computed for the search.  $L$  is stored and can be called during the search for the nearest microphone. The next step uses the audio channel from the nearest microphone for sound classification.

### 3.5. Engine Sound Classification

In the second stage, the algorithm classifies observed sound  $M$  and environment noise  $N$  in Eq. 1. The model assumes that audio  $M = N$  if no idling car shows up near the microphone, and  $M = S + N$  otherwise. Usually, audio where an idling engine is present has power at a certain combination of frequencies relative to audio that consists entirely of background noise (e.g., wind, people talking, distant traffic). We refer the former as foreground signal and the later as background signal.

There are two difficulties when building such a binary classifier in real time. The first difficulty is that there is limited training data because labeling idling/non-idling ground truth is difficult and time-consuming, and would require too many hours of training data, making efficient deployment in new locations infeasible. However, the system requirements indicate that the classifier must generalize to new, previously unheard cars even with limited training data.

The next difficulty is that the data has multiple outliers because the real-world deployment environment must deal with practical limitations on the placement of microphones, microphone cutoff, and environmental interference. Some microphones are not placed in the optimum location for engine noise due to the presence of sidewalks, wheelchair access, or existing infrastructure. The placement of the microphones also impacts the signal quality that the transmitter has with the receiver. Additionally, sound events such as helicopters flying overhead or people talking right by the microphone can confuse the classifier. Because of these issues, we have found that a simple frequency-selection-based power threshold is not reliable, and we have developed a machine-learning-based method applied to time-frequency data to differentiate  $S$  and  $N$ . Moreover, in practice, it is time-consuming to annotate (video) frame-by-frame idle ground truth, because an annotator must watch the video and refer to the audio back and forth to achieve frame-level accuracy.

The problem of building classifiers with limited training data is an important and active area of machine-learning research, with some promising preliminary results. One effective strategy is to develop latent spaces using large amounts of unlabeled or weakly labeled training data, and then to leverage this latent space to the target problem with limited data. In this light, we have developed a supervised contrastive learning approach for subsequent audio classification. Supervised contrastive learning (SCL) (first proposed in the computer vision literature [53]) delineates interclass audio samples while alienating intraclass samples. Following architecture and SCL loss functions from [53], the architecture for the proposed system relies on a ResNet50 feature encoder, a fully connected projector, and a linear/nonlinear classifier. First, the encoder and projector are pretrained using SCL loss on a large public audio dataset ESC-50. Next, the pretrained encoder is frozen and fed a small amount of the labeled foreground/background data from the new site/deployment. We found that the on-site data in the SCL latent spaces is distributed in an easily separable fashion. Thus, we can use simpler classifier (with fewer degrees of freedom) to differentiate between foreground and background data.

For processing by the neural network, the time-domain, audio signal is converted to a 2D spectrogram through a short-time Fourier transform (STFT). Compared to a regular Fourier transform, STFT reflects the local frequency domain over time, computes magnitude, and concatenates them vertically into a spectrogram with shape  $T \times F$ , where  $T$  is the number time steps, and  $F$  is the number of frequency bins. The spectrograms are

encoded using a ResNet50, which is pretrained on ImageNet [53], and then fine tuned on ESC-50 [54] using the SCL approach described above. Similar to [39] and [38], the model maps the input spectrogram to a normalized 2048 dimension vector and further projects it to a normalized 64 dimension latent vector on a hypersphere. A classifier is trained to classify 2048 dimension latent vectors.

In the last step, we keep  $Y$  and  $BB$  for moving vehicles. We replace  $Y$  as either  $Y_{idling}$  or  $Y_{eof}$  for stationary vehicles according to audio classifier's output.

## 4. Experiments

The system was deployed at the main entrance of the hospital test site for 14 test days. On average, about 10 vehicles are picking up and dropping off patients every hour, with some typically busier times (e.g., morning and mid afternoon). The vehicle motion detector, engine sound classifier, and entire IVD pipeline performance are evaluated qualitatively and quantitatively on a single day of held-out data.

### 4.1. Dataset

#### 4.1.1. Audio-Visual IVD Dataset

We performed field tests at the test site for 14 days. To the best of our knowledge, there is no existing IVD dataset matching our setup. Because video and audio sample amounts differ significantly (foreground audio samples are very limited since drivers' idling time is much shorter than entire recording), they are collected separately on different days. Sampled every 1 second from the first 3-days 10-hour recordings, our video training set consists of 33015 clips, and our video validation set consists of 8252 clips. Our video test set has 13271 clips sampled from a random one test day recording from the remaining 11 days. All video data is annotated with bounding boxes and motion status. The camera fps is 25. The audio training set has 8721 foreground (positive) samples and 30618 background (negative) samples collected every 1 second from a random five days' first 80% of recording. To evaluate the audio model's generalization ability over new vehicles, the audio validation set has 2491 positive samples and 8245 negative samples sampled from those 5 days' last 20% of recording. The sample rate of the microphone is 48000Hz. All audio samples are 5 seconds long and centered at the video clip's last frame. They are input pairs for the algorithm. The data acquisition script synchronizes the video and audio. Notes are taken on-site for annotating the ground truth per frame.

#### 4.1.2. Urban Sound Classification

Our audio model is pretrained on ESC-50. ESC-50 [54] is a public environment sound event classification dataset that contains 2000 audio recordings evenly distributed over 50 classes, including engine idling and common background noise sources similar to that observed at the test-site field deployments, such as wind, speech, and helicopters. In the ESC-50 dataset, each audio sample is 5 seconds long, and the samples are recordings from the public project Freesound.



## 4.2. Experimental Setup

The vehicle motion model is trained using an Adam optimizer with a learning rate of 0.0001 with batch size 20 on a Nvidia RTX4090 24GB GPU. Following [36], five anchors are precomputed on the training dataset. The audio model is trained using the same GPU and learning rate 0.0001 with batch size 128. The audio latent space dimension is 2048. Each input audio sample is 5 seconds long and is normalized. Since audio positive and negative samples are imbalanced, we manually balance them in each mini-batch for stochastic gradient descent.

## 4.3. Metrics

We evaluate our algorithm by computer vision’s common object detection metrics: average precision (AP), mean average precision (mAP), and accuracy. These metrics also compute intersection over union (IoU), precision, and recall curve.

### 4.3.1. Intersection Over Union (IoU)

IoU measures how predicted bounding box  $BB_p$  overlaps with ground truth box  $BB_{gt}$ . It is a common metric used for object detection in computer vision.

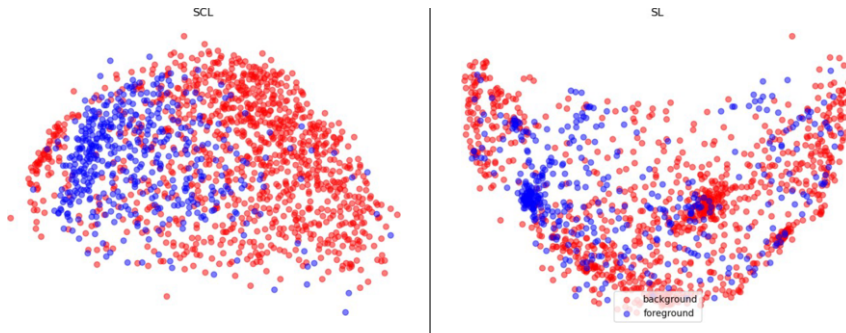
$$IoU = \frac{BB_p \cap BB_{gt}}{BB_p \cup BB_{gt}}$$

### 4.3.2. Metrics for Quantitative Analysis

In object detection, a true positive is defined as a detection with bounding-box IoU (relative to ground-truth bounding boxes) greater than a threshold, whereas a false positive is defined as a detection/output from the neural network with IoU smaller than a given threshold. A false negative is defined as a ground truth without a corresponding, overlapping, bounding box from the NN. For object detection tasks, typical choices for these thresholds are 0.5 and 0.75. With precision and recall computed from these, a Precision-Recall (PR) curve can be plotted to evaluate the performance of an object detector on a particular class. In audio classification, precision and recall are defined in the conventional manner for binary classifiers. AP is the interpolated area under PR curve for each class. mAP is the average AP value over classes.

## 4.4. Vehicle Motion Detection Performance

We evaluate the performance of vehicle motion detection using mAP and AP. For test-site deployment, the evaluation focuses on the motion detector’s generalization ability across test data. Because the tripod is redeployed each day, the camera pose varies slightly from day to day. The lighting conditions, traffic conditions, and types of vehicles also differ. For example, the variability of light conditions between sunny and cloudy days is significant. The motion detector was trained and validated on the first 3 train/validation days. Sampled every second, the first 3-day data is divided into 80% training clips and 20% validation clips, resulting in 33015 training clips and 8252 validation clips. Additionally, 13271 test clips are sampled in the same way for another random single day. Table 1 summarizes the performance comparison between validation set and test set. APs of



**Figure 4.** MDS 2D Visualization on Audio Encoder Latent Space. Due to the huge amount of validation samples, we feed part of our validation data into trained SCL and SL’s encoder. The left side shows projected latent space of SCL in 2D dimension. The right side shows SL’s latent space. Red dots are background samples. Blue dots are foreground samples.

stationary and moving are comparable for validation and test clips. For the same IoU threshold, the test data’s stationary AP is about 15% lower than validation data. Moving AP is even 9% higher because, we believe, validation samples have fewer vehicles than test samples. However, from these results we can conclude that the trained vehicle motion detector is capable of localizing vehicle motion on a different single day’s test data with 3 days of training data, even with different pose and lighting conditions.

**Table 1.** Vehicle Motion Detection Performance

	data type	mAP (%)	AP Stationary (%)	AP Moving (%)
IoU@0.5	validation	84.49	95.83	73.15
IoU@0.75	validation	60.94	80.85	41.03
IoU@0.5	test	87.04	91.14	82.94
IoU@0.75	test	48.22	64.50	31.94

#### 4.5. Engine Sound Classifier Performance

The sound classifier is evaluated using precision, recall, and F-score. Groundtruth static vehicles are selected, and the respective ground truth bounding boxes are fixed to find the closest microphone to cut audio samples from for training and testing. The effectiveness of supervised contrastive learning (SCL) and supervised learning (SL) is compared by visualizing validation data latent vectors in normalized 2048 dimension space, as seen in Fig. 4. This visualization shows that in the SCL latent space positive and negative vectors are clustered and potentially separable. To simulate nonlinearity, we chose a two-layer perceptron with 1024 internal nodes with a relu activation as the classifier on the latent space. After training, the classifier’s performance on validation data was computed and is shown in Table 2. Although both have similar recalls, SCL’s F-score is generally better than SL by 0.15. Along with latent visualization, we believe clustered SCL latent space can adopt test data better than SL from this observation.

Several factors affect the classifier’s performance. Unidirectional microphones can concentrate on only upfront vehicle sound. However, since our model does not separate mixed audio, it can pick up surrounding loud engine sounds and predict them as a false

**Table 2.** Audio Classification Accuracy

	Precision	Recall	F-score
SL (ResNet50)	0.5121	0.8691	0.6444
SCL (ResNet50)	<b>0.5913</b>	<b>0.8687</b>	<b>0.7036</b>

positive even with a directional microphone. These cases cause low precision. Also, we find audio cutout from the wireless microphones affects the classifier’s performance. The microphones exhibit intermittent cutout (signal loss) throughout the outdoor recordings. Such signal loss was found to be a common hardware issue for outdoor wireless microphones, even with the most advanced affordable wireless microphone set on the market. Aurally, the cutoff signal has no sound. Digitally, the acquired signal value bounces between specific values, destroying the semantic meaning of the audio samples.

#### 4.6. IVD Performance

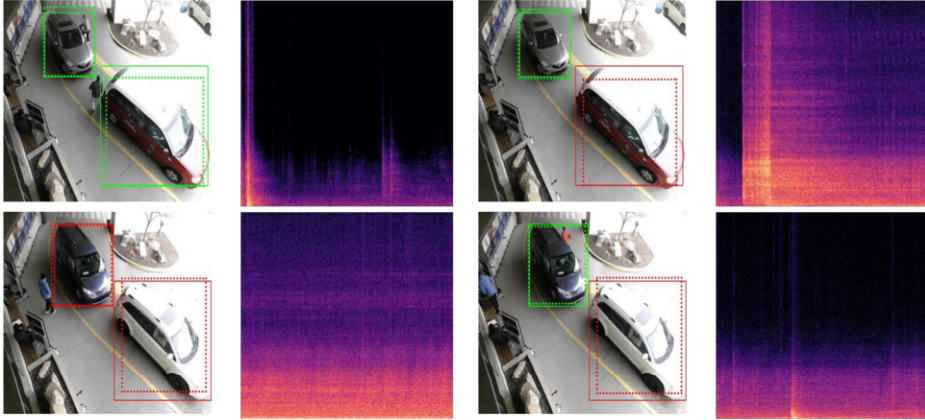
We perform audio-visual combined evaluation based on our audio validation set. Since our audio validation set is sampled per 1 second (25 frames), we expand combined evaluation set to per entire second (enlarged by 25 times). Each input pair has a 16-frame clip and a 5-second audio sample.

##### 4.6.1. Qualitative Evaluation

Fig. 5 shows the combined audio-visual IVD detection. The trained vehicle motion detector localizes each vehicle very well. As a result, the method is able to find the correct nearest microphone. Additionally, the acquired audio signal is distinct between foreground and background sound. The first row shows two vehicles with the bottom one switching on. The highlighted spectrogram indicates engine ignition has stronger power on frequency bins. The second row shows two vehicles with the upper one switching off, as the corresponding spectrogram turns darker. It also turns out our system is able to handle multiple vehicles with the help of unidirectional microphones. Two per-frame predictions during 10-minute interval are computed and shown in Fig. 6. By comparing the color and shape of trajectories, we believe our model is generally capable of capturing correct vehicle position and engine status over a long period of time.

##### 4.6.2. Quantitative Evaluation

IVD performance is also evaluated using AP and mAP as three IVD classes. Compared to Table 1, AP Engine Off is similar to AP Stationary. AP Idling is about 10% drop compared to AP Stationary. We believe this is a reasonable drop by combining audio and visual errors. With similar foreground and background accuracy in the previous section, AP Idling is about 20% lower than AP Engine Off. Thus, our system struggles more with detecting idling vehicles than with engine-off vehicles, which we believe is due to fewer foreground training and validation samples. Also, our current training set is not big enough to cover a variety of engine sounds, which can lower audio classifier’s performance on a different vehicle. However, when threshold IoU is 0.5 and 0.75, our model still has comparable values with models solving other video understanding tasks.



**Figure 5.** IVD Visual Performance. Each row includes detected results and groundtruth annotations along with the corresponding spectrogram. Red, green, and blue bounding boxes are idle, non-idle, and moving labels respectively. Dotted and solid rectangles are prediction and ground truth.

**Table 3.** IVD Performance on Validation Data

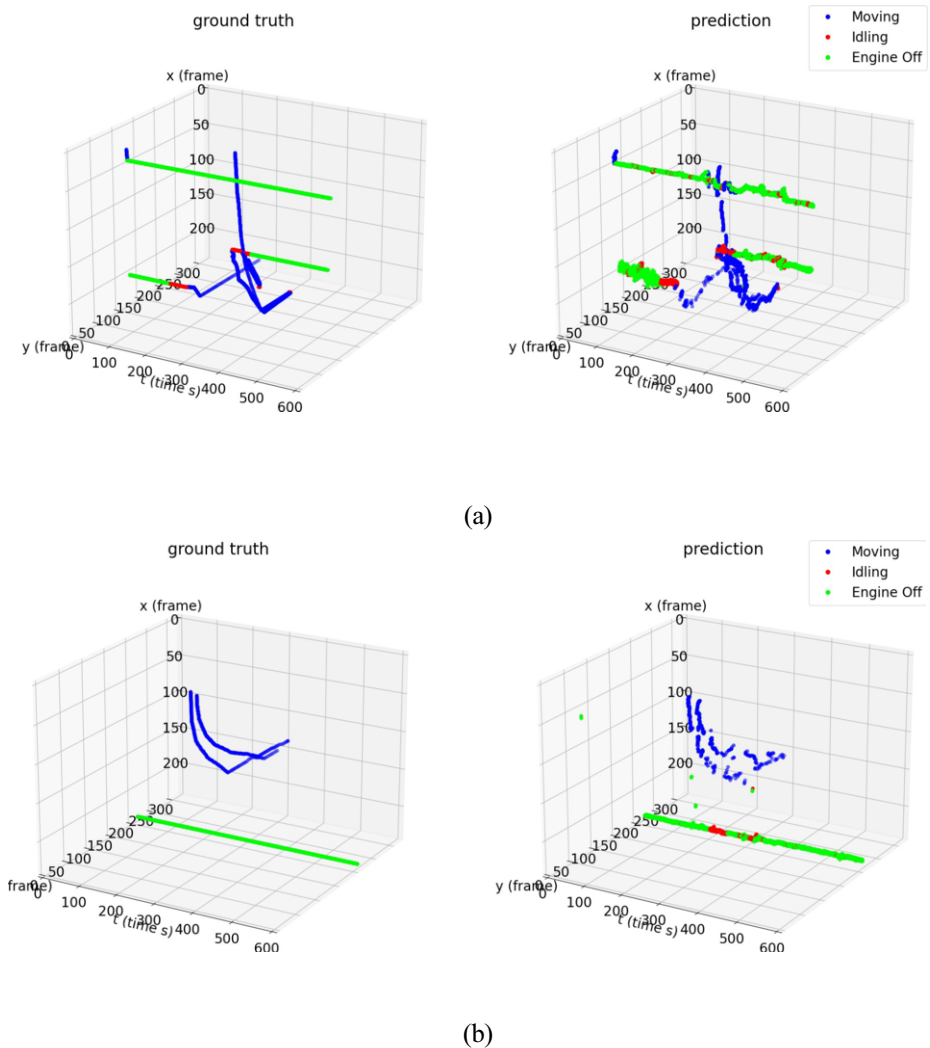
	mAP	AP Moving	AP Idling	AP Engine Off
IoU@0.5	80.12	78.27	71.02	91.06
IoU@0.75	30.39	25.75	17.61	47.82

## 5. CONCLUSIONS

In this work, we create a camera and microphone setup and formulate a new problem for IVD. We build an audio-visual algorithm to solve the problem. By deploying the system in real-time for 11 days, it detects IVs in most circumstances and displays smart messages to drivers. We believe this system can be refined and adapted to more real-world scenarios.

## ACKNOWLEDGMENT

We thank Intermountain Healthcare and LDS Hospital in Salt Lake City, Utah, United States for providing the field test location for this study. This material is based upon work supported by the National Science Foundation under Grant No. 1952008 SCC-IRG TRACK 2: SMART AIR: Informing driving behavior through dynamic air-quality sensing and smart messaging.



**Figure 6.** Two Vehicle Trajectory Reconstruction Examples During Ten-minute Intervals.  $x$  and  $y$  are centroids of bounding box video frames and  $t$  is the time axis. In examples (a) and (b), the shape of the trajectory is the vehicle path through the video sequence. The color of the trajectory indicates the vehicle status, moving, idling, or engine off, of the ground truth and the prediction. In case (a), two vehicles are parked in the target area at the beginning. One of them ignited and drove away. The third vehicle came, parked, and stopped the engine later. Case (b) also includes three vehicles. One was off in the target area, while the other two drove through.

## References

- [1] OECD. The Economic Consequences of Outdoor Air Pollution; 2016. Available from: <https://www.oecd-ilibrary.org/content/publication/9789264257474-en>.
- [2] Annavarapu RN, Kathi S. Cognitive disorders in children associated with urban vehicular emissions. Environmental pollution. 2016;208 Pt A:74-8.

- [3] Rice MB, Rifas-Shiman SL, Litonjua AA, Oken E, Gillman MW, Kloog I, et al. Lifetime Exposure to Ambient Pollution and Lung Function in Children. *American journal of respiratory and critical care medicine*. 2016;193 8:881-8.
- [4] Lewtas J. Air pollution combustion emissions: characterization of causative agents and mechanisms associated with cancer, reproductive, and cardiovascular effects. *Mutation research*. 2007;636 1-3:95-133.
- [5] Carbon Pollution from Transportation;. <https://www.epa.gov/transportation-air-pollution-and-climate-change/carbon-pollution-transportation>.
- [6] Ryan PH, Reponen T, Simmons MJH, Yermakov M, Sharkey K, Garland-Porter D, et al. The impact of an anti-idling campaign on outdoor air quality at four urban schools. *Environmental science Processes & impacts*. 2013;15 11:2030-7.
- [7] Sharma A, Kumar P. A review of factors surrounding the air pollution exposure to in-pram babies and mitigation strategies. *Environment international*. 2018;120:262-78.
- [8] Kenagy HS, Lin C, Wu H, Heal MR. Greater nitrogen dioxide concentrations at child versus adult breathing heights close to urban main road kerbside. *Air Quality, Atmosphere, & Health*. 2015;9:589 595.
- [9] Mischler SE, Colinet JF. Controlling And Monitoring Diesel Emissions In Underground Mines In The United States; 2010.
- [10] Widla J. A complete guide to fleet idling: Understand, detect and stop true idling; 2022. Accessed: 2023-05-15. "<https://www.geotab.com/blog/detect-stop-true-fleet-idling/>".
- [11] Scales M. The hidden impact of idling engines; 2022. Accessed: 2023-05-15. "<https://www.canadianminingjournal.com/featured-article/the-hidden-impact-of-idling-engines/>".
- [12] Mendoza DL, Benney TM, Bares R, Fasoli B, Anderson C, Gonzales SA, et al. Air Quality and Behavioral Impacts of Anti-Idling Campaigns in School Drop-Off Zones. *Atmosphere*. 2022.
- [13] Eghbalian C, Sharkey K, Garland-Porter D, Alam M, Crumpton M, Jones C, et al. A community-based participatory research partnership to reduce vehicle idling near public schools. *Journal of environmental health*. 2013;75 9:14-9.
- [14] Meleady R, Abrams D, de Vyver JV, Hopthrow T, Mahmood L, Player A, et al. Surveillance or Self-Surveillance? Behavioral Cues Can Increase the Rate of Drivers' Pro-Environmental Behavior at a Long Wait Stop. *Environment and Behavior*. 2017;49:1156 1172.
- [15] de Vyver JV, Abrams D, Hopthrow T, Purewal K, de Moura GR, Meleady R. Motivating the selfish to stop idling: self-interest cues can improve environmentally relevant driver behaviour. *Transportation Research Part F-traffic Psychology and Behaviour*. 2018;54:79-85.
- [16] Dowds J, Sullivan JL, Aultman-Hall L. Comparisons of Discretionary Passenger Vehicle Idling Behavior by Season and Trip Stage with Global Positioning System and Onboard Diagnostic Devices. *Transportation Research Record*. 2013;2341:76 82.
- [17] Carrico AR, Padgett PO, Vandenbergh MP, Gilligan JM, Wallston KA. Costly myths: An analysis of idling beliefs and behavior in personal motor vehicles \$. *Energy Policy*. 2009;37:2881-8.

- [18] Winnett MA, Wheeler A. VEHICLE-ACTIVATED SIGNS - A LARGE SCALE EVALUATION; 2003.
- [19] Ullman GL, Rose ER. Evaluation of Dynamic Speed Display Signs. *Transportation Research Record*. 2005;1918(1):92-7. Available from: <https://doi.org/10.1177/0361198105191800112>.
- [20] Lee CK, Lee S, Choi B, Lee Oh Y. Effectiveness of Speed-Monitoring Displays in Speed Reduction in School Zones. *Transportation Research Record*. 2006;1973:27-35.
- [21] Cruzado I, Donnell ET. Evaluating Effectiveness of Dynamic Speed Display Signs in Transition Zones of Two-Lane, Rural Highways in Pennsylvania. *Transportation Research Record*. 2009;2122:1-8.
- [22] Stern J, Schlag B, Schulze C, Degener S, Butterwegge P, Gehlert T. Evaluation des Dialog-Displays - Berliner Studien / Evaluation of dynamic speed display signs - The Berlin studies; 2010.
- [23] Ando R, Noda K, Mimura Y, Yamazaki M, Yang J, Ogino H, et al. Long-term effect analysis of dynamic speed display sign in streets. 2017 4th International Conference on Transportation Information and Safety (ICTIS). 2017:522-9.
- [24] Gehlert T, Schulze C, Schlag B. Evaluation of different types of dynamic speed display signs. *Transportation Research Part F-traffic Psychology and Behaviour*. 2012;15:667-75.
- [25] Bastan M, Yap KH, Chau LP. Idling Car Detection with ConvNets in Infrared Image Sequences. 2018 IEEE International Symposium on Circuits and Systems (ISCAS). 2018:1-5.
- [26] Bertasius G, Torresani L, Shi J. Object Detection in Video with Spatiotemporal Sampling Networks. 2018.
- [27] Feichtenhofer C, Fan H, Malik J, He K. SlowFast Networks for Video Recognition. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2018:6201-10.
- [28] Ramaswamy A, Seemakurthy K, Gubbi J, Purushothaman B. Spatio-temporal action detection and localization using a hierarchical LSTM. 2020.
- [29] Chen S, Sun P, Xie E, Ge C, Wu J, Ma L, et al. Watch Only Once: An End-to-End Video Action Detection Framework. 2021.
- [30] Sánchez-Caballero A, de López-Diz S, Fuentes-Jimenez D, Losada-Gutiérrez C, Marrón-Romera M, Casillas-Pérez D, et al. 3DFCNN: real-time action recognition using 3D deep neural networks with raw depth information. *Multimedia Tools and Applications*. 2022:24119–24143.
- [31] Siam M, Mahgoub H, Zahran M, Yogamani S, Jagersand M, El-Sallab A. MODNet: Motion and Appearance based Moving Object Detection Network for Autonomous Driving. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC); 2018. p. 2859-64.
- [32] Siam M, Mahgoub H, Zahran M, Yogamani SK, Jägersand M, Sallab AE. MODNet: Motion and Appearance based Moving Object Detection Network for Autonomous Driving. 2018 21st International Conference on Intelligent Transportation Systems (ITSC). 2018:2859-64.

- [33] Song R, Kasun LLC, Peng X, Lin Z, Cao Q, Huang G. Motion Embedding for On-road Motion Object Detection for Intelligent Vehicle Systems. 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). 2022:3354-9.
- [34] Zhao W, Yin J, Wang X, Hu J, Qi B, Runge T. Real-Time Vehicle Motion Detection and Motion Altering for Connected Vehicle: Algorithm Design and Practical Applications. *Sensors (Basel, Switzerland)*. 2019;19.
- [35] Lopez M, Griffin T, Ellis K, Enem A, Duhan C. Parking Lot Occupancy Tracking Through Image Processing. In: *International Conference on Computers and Their Applications*; 2019.
- [36] Köpüklü O, Wei X, Rigoll G. You Only Watch Once: A Unified CNN Architecture for Real-Time Spatiotemporal Action Localization. *ArXiv*. 2019;abs/1911.06644.
- [37] Kong Q, Cao Y, Iqbal T, Wang Y, Wang W, Plumbley MD. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2019;28:2880-94.
- [38] Saeed A, Grangier D, Zeghidour N. Contrastive Learning of General-Purpose Audio Representations. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020:3875-9.
- [39] Nasiri A, Hu J. SoundCLR: Contrastive Learning of Representations For Improved Environmental Sound Classification. *ArXiv*. 2021;abs/2103.01929.
- [40] Kotti M, Moschou V, Kotropoulos C. Speaker segmentation and clustering. *Signal Processing*. 2008;88:1091-124.
- [41] Seeing the Sound: A New Multimodal Imaging Device for Computer Vision. 2015 IEEE International Conference on Computer Vision Workshop (ICCVW). 2015:693-701.
- [42] Chen T, Kornblith S, Norouzi M, Hinton G. A Simple Framework for Contrastive Learning of Visual Representations. In: *International Conference on Machine Learning*; 2020.
- [43] He K, Fan H, Wu Y, Xie S, Girshick R. Momentum Contrast for Unsupervised Visual Representation Learning. In: *IEEE Conference on Computer Vision and Pattern Recognition*; 2020.
- [44] Salamon J, Jacoby C, Bello JP. A Dataset and Taxonomy for Urban Sound Research. In: *Proceedings of the 22nd ACM international conference on Multimedia*; 2014.
- [45] Piczak KJ. ESC: Dataset for Environmental Sound Classification. In: *Proceedings of the 23rd ACM international conference on Multimedia*; 2015.
- [46] Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Lawrence W, Moore RC, et al. AUDIO SET: AN ONTOLOGY AND HUMAN-LABELED DATASET FOR AUDIO EVENTS. In: *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*; 2017.
- [47] Afouras T, Asano YM, Fagan F, Vedaldi A, Metze F. Self-supervised object detection from audio-visual correspondence. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021:10565-76.
- [48] Owens A, Efros AA. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. In: *European Conference on Computer Vision*; 2018.
- [49] Badamdorj T, Rochan M, Wang Y, Cheng L. Joint Visual and Audio Learning for Video Highlight Detection. In: *International Conference on Computer Vision*; 2021.



- [50] Parekh S, Essid S, Ozerov A, Duong NQ, Perez P, Richard G. Weakly supervised representation learning for unsynchronized audio-visual events. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2018.
- [51] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:6517-25.
- [52] Xie S, Girshick RB, Dollár P, Tu Z, He K. Aggregated Residual Transformations for Deep Neural Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:5987-95.
- [53] Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised Contrastive Learning. *ArXiv*. 2020;abs/2004.11362.
- [54] Piczak KJ. ESC: Dataset for Environmental Sound Classification. Proceedings of the 23rd ACM international conference on Multimedia. 2015.