This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3145663, IEEE Transactions on Multimedia

# Multi-modality Sensing and Data Fusion for Multi-vehicle Detection

Debashri Roy, Yuanyuan Li, Tong Jian, Peng Tian, Kaushik Chowdhury, and Stratis Ioannidis

Department of Electrical and Computer Engineering
Northeastern University, Boston, USA
Email: {droy, yuanyuanli, pengtian, jain, krc, ioannidis}@ece.neu.edu

***Abstract*—With the recent surge in autonomous driving vehicles, the need for accurate vehicle detection and tracking is critical now more than ever. Detecting vehicles from visual sensors fails in non-line-of-sight (NLOS) settings. This can be compensated by the inclusion of other modalities in a multi-domain sensing environment. We propose several deep learning based frameworks for fusing different modalities (image, radar, acoustic, seismic) through the exploitation of complementary latent embeddings, incorporating multiple state-of-the-art fusion strategies. Our proposed fusion frameworks considerably outperform unimodal detection. Moreover, fusion between image and non-image modalities improves vehicle tracking and detection under NLOS conditions. We validate our models on the real-world multimodal ESCAPE dataset, showing 33.16% improvement in vehicle detection by fusion (over visual inference alone) over test scenarios with 30-42% NLOS conditions. To demonstrate how well our framework generalizes, we also validate our models on the multimodal NuScene dataset, showing ∼22% improvement over competing methods.**

***Index Terms*—vehicle detection, tracking, multimodal data, fusion, latent embeddings, image, seismic, acoustic, radar.**

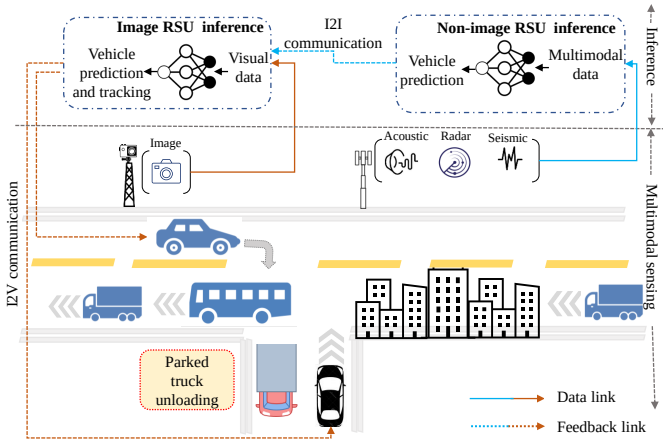



Fig. 1: A typical example of deployment of the proposed fusion architecture in autonomous vehicles in connected vehicular network environment.

## I. Introduction

Vehicle detection and tracking in different road conditions are of vital importance for ensuring safe operability in autonomous driving settings [1]. As humans perceive the driving environment visually, autonomous vehicles rely on visual data, i.e., images and videos. However, in adverse driving conditions, the presence of a vehicle may not be perceived from images/videos only. Adverse driving conditions may arise due to weather (fog, rain etc.) or positional (limited visibility, non-line-of-sight) obstructions. In such cases, camera sensors are increasingly hampered at object detection. These limitations have led the research community to leverage additional modalities beyond just images for safety-critical operations such as vehicle detection and tracking [2].

**Availability and Usability of Multimodal Data:** Given the ongoing revolution in Internet of Things (IoT), pervasive deployment of sensors will result in a multi-trillion dollar market segment within the next few years [3]. These sensors generate data from different modalities, including radar, image, infra-red (IR), seismic, and acoustic, among others. Different modalities capture the situational state of the environment from different perspectives. For example, radar signals are robust to adverse weather conditions (lighting changes, rain, fog) [4], but are vulnerable to clutter and multi-path effects [5], and are insufficient to determine target size [6]. Seismic and acoustic signals can also detect the presence of an object without a direct line-of-sight, but also struggle with the determining object sizes [5]. These non-image modalities are less susceptible to non-line-of-sight (NLOS) conditions, whereas images fare better at tracking and determining object dimensions.

**Motivation of using Multimodal Data Fusion:** The state-of-the-art mainly uses convolutional neural networks (CNNs) based learning frameworks for most accurate object detection with camera images [7], [8], [9], [10]. While great strides have been made in processing and analyzing image source data, currently, we do not have mature frameworks for fusing other sensor modalities with imagery [5]. Furthermore, it is unclear (i) *how* these sensor modalities should be combined/fused together for increased situational awareness, and (ii) *which* subset of these sensors should be fused under varying environmental conditions, particularly under NLOS scenarios.

**Scenario of Interest:** Fig. 1 shows our scenario of interest with two autonomous vehicles wishing to perform mutual detection under NLOS conditions. The black car in the alley cannot view the inbound van (whose view is obstructed by buildings), while the blue car wants to take a turn in the alley,

Approved for Public Release; Distribution Unlimited. This document is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

which is blocked by the bus. The blue car is not aware of the fact that the alley is completely blocked by the parked truck and the stopped black car. Momentary decisions are crucial for ensuring safety in this scenario. Our proposed fusion-based techniques can be deployed in the road-side infrastructure units (RSUs) performing detection and inverence. Even though the autonomous vehicles are equipped with integrated cameras, such cameras often have limited visibility, which is concerning when making safety-critical decisions [2]. Hence, we propose a symbiosis of vehicular and infrastructure communication using multi-domain sensing when making such safety-critical decisions. The RSUs dedicated to collect non-image sensing data can perform fusion-based inference from the multi-domain sensing data to predict vehicles in this scenario. This prediction can be transmitted to the RSU where the visual data is being collected using infrastructure to infrastructure communication (I2I). The second level of fusion-based inference between earlier prediction and visual image happens in this image RSU. The final vehicle detection and tracking (in case of direct LOS) information is then fed back to the vehicles using infrastructure to vehicle communication (I2V). In other words, a vehicle may still use a tracking mechanism from the images, simultaneously receiving more reliable predictions from the deployed infrastructure about presence of vehicles in the surrounding region, impoving decision-making safety.

**Contribution of this Paper:** Our contribution amounts to answering the question of *how* and *when* to fuse different modalities. In that regard, we use the seismic, acoustic, radar and image modalities from multimodal ESCAPE dataset [11], to identify different vehicles in order to validate our proposed fusion frameworks. Our novelty lies in introducing and experimenting with different fusion techniques which can intelligently select the most relevant sensors on a *case-by-case* basis. We list our contributions in detail below:

- We propose fine-tuned models for seismic, acoustic, radar and image modalities to detect and identify each vehicle in a multi-vehicle scenario.
- We provide different fusion frameworks for deep-learning-based fusion between different modalities, which flexibly assign weights on each modality based on relevance. We explore different state-of-the-art fusion strategies to represent the fusion layer from unimodal features.
- We explore fusion from both a *feature-* and *decision-based* perspective. The proposed fusion frameworks work on *ultimate* and *penultimate* (second last) layers of each fine-tuned unimodal models. We also propose a novel *multi-level* fusion framework: feature-based fusion is performed at a first level, and another feature- or decision-based fusion is carried at a second level, using the predictions of the first level.
- The fusion between non-image modalities is performed separately to exploit the correlation in different non-image modalities. Predictions by those *non-image* fusion networks are further fused with the *image* modality to generate the final prediction.
- We evaluate the performance of proposed unimodal and fusion frameworks in two types of settings: (i) *bin-based* and (ii) *scenario-based*. The bin-based setting ensures that each model is trained on part of each scenario, and tested on the remainder. The scenario-based setting validates performance of the proposed frameworks when trained on few scenarios and tested on completely new scenarios, not encountered during training.
- We demonstrate that fusion frameworks qualitatively improve vehicle detection performance for NLOS scenarios. In particular, we apply our proposed fusion frameworks over the ESCAPE [11] and NuScene [12] datasets.

## II. Related Work

Object detection is a widely researched area in the domain of computer vision, and convolutional neural networks [13], [14] have a well-known efficacy in this task. However, the exploitation of cross-modality features along with visual information is still in a nascent stage. It has recently been explored in diverse domains of research involving visual modalities, i.e., video recognition [15], [16], multimedia understanding [17], [18], [19], [20], image and video captioning [21], [22], and person identification [23], to name a few. Additionally, fusion among diverse modalities has its own challenges, such as ensuring diversity among different modalities or the availability of multi-domain sensor data in a time-synchronized manner (see, e.g., Lahat *et al.* [24], [25]). Compared to the other works, we extend the domain of cross-modality feature exploration by fusing both visual and non-image modalities.

Feng *et al.* [26] show that deep learning based fusion techniques can improve performance over single modalities. Other works [27], [28] have also demonstrated improvements in object detection when fusing radar and image modalities. A deep learning approach with Generative Adversarial Networks (GANs) to fuse camera and radar data was proposed by Lekic *et al.* [29]. Chadwick *et al.* [5] propose fusion between low-level representations of radar data and camera images. They use both concatenation or element-wise addition on a fixed layer as fusion strategies. They also use additional extracted features (such as range and the range rate) from the radar as auxiliary image channels. The authors prove the efficiency of these fusion techniques on a self-collected dataset. Similarly, Nobis *et al.* [2], propose a fusion network that is able to learn the optimal layer at which fusion minimizes the training loss. Beyond the fusion between radar and image modalities, light detection and ranging (LiDAR) sensor data has also been explored as a candidate for sensor fusion. Chen *et al.* [30] combine both frontal and birds-eye LiDAR views with camera images for 3D object detection. Moving object classification and tracking using cameras, LiDAR and radar data fusion is presented by Chavez-Garcia *et al.* [31]. Our proposed work differs from the aforementioned state-of-the-art by not only including more modalities during fusion, but also in exploiting novel means for fusing additional modalities within visual data. Most importantly, we depart by explicitly studying detection behavior under NLOS conditions.

In the domain of multi-target detection, fusion has been explored on image and radio frequency data using different machine learning techniques, such as manifold learning [32][33], upstream fusion [34], canonical variation anal-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3145663, IEEE Transactions on Multimedia

ysis [35] and joint sparsity based optimization on data-level fusion [36]. However, these works focus on improving the detection accuracy with direct LOS. There has also been effort to apply fusion between seismic and acoustic data for ground moving vehicle classification by Pan *et al.* [37]. In [38] and [39], Vakil *et al.* present feature level sensor fusion of passive RF and image modalities for vehicle detection and scenario classification. This state-of-the-art motivates us to use fusion of multiple modalities for multi-object detection and tracking; it also demonstrates that image, radar, acoustic, and seismic modalities all are excellent candidates for object detection, as they provide situational awareness from different perspectives. However, designing robust learning frameworks by using different fusion-based approaches among diverse multi-domain data is still open. In this paper, we explicitly bridge that gap by introducing novel fusion techniques among image, radar, acoustic, and seismic modalities.

## III. Problem Statement and Solution Overview

In this section, we present an overview of the problem of multi-vehicle detection using multimodal sensing data, and our proposed approach of using deep learning based fusion techniques as a solution.

### A. Problem Statement

Our main challenge is to detect multiple vehicles when there is no direct LOS. With the advent of Internet of Things (IoT) sensors, there is an abundance of sensors operating on different types of data modalities. We categorize the visual sensing data as *image modality*, and other type of sensing data as *non-image modalities*. Sensors for different modalities can be co-located or separately-located, but they all must be time-synchronized with each other. Our goal is to augment an image detection algorithm from multiple modalities so that we can detect and track vehicles irrespective of LOS and NLOS conditions. The overall idea is to *detect and track* the vehicles using *image* modality under *LOS conditions*, and *detect* the vehicles with help of other *non-image* modalities under *NLOS conditions*. We stress however that our focus is on *improving detection via fusion*, while improving vehicle tracking algorithms out of our scope; we use state-of-the-art vehicle tracking approaches such as [40], [41] for this purpose.

### B. Solution Approach

We propose a deep learning based fusion approach to perform multi-vehicle detection over multimodal sensing data. The overall solution is approached in three steps:

1) **Preprocessing and Unimodal Networks:** The first step is to *preprocess* the time-synchronized raw data and *generate the high-level features* which will be fed to the neural networks for further processing. The generated features of each modality are then passed through fine-tuned unimodal networks and generate predictions. We use the penultimate output of each unimodal network as a measure of data representation of each sensor modality. These *penultimate layers* of unimodal networks extract the low-level latent features for different modalities, whereas the ultimate layers provide us with the initial predictions of the vehicles using those modalities.

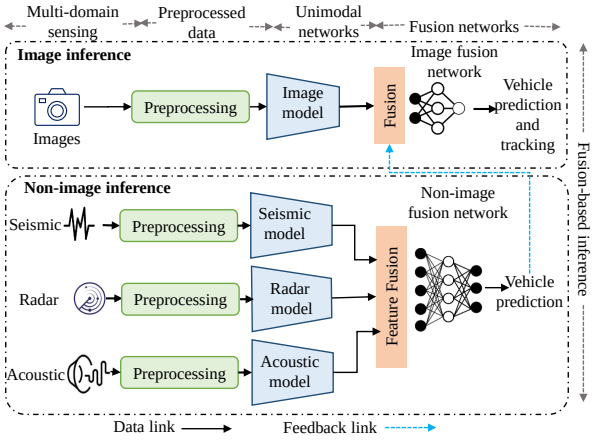



Fig. 2: An overview of our fusion-based non-image and image inference.

2) **Fusion Networks for Non-image Modalities:** The next step of the solution pipeline is to fuse non-image modalities. We explore and compare the performance of different fusion techniques. For example, we explore fusing at the both penultimate and ultimate layers of each non-image unimodal network.
3) **Fusion Networks to Augment Image Modality:** Finally, we propose to fuse the image modality with the ultimate layer of non-image fusion networks. Therefore, the output of non-image fusion network guides the image modality to perform detection even when there is no visible vehicle in the image frame. We again explore several fusion methods. We note that fusion with image data is treated separately so that detection can co-exist with tracking when LOS is available.

Fig. 2 presents a high-level view of our solution pipeline: the multimodal sensing data passes through the proposed feature generation, trained unimodal networks, and fusion networks.

## IV. Dataset, Preprocessing and Unimodal Networks

In this section, we introduce all 4 modalities (image, radar, acoustic, and seismic) of the publicly available multimodal ESCAPE dataset [11], along with implementation details of our preprocessing steps and proposed unimodal networks.

### A. Dataset Description

The **E**xperiments, **S**cenarios, **C**oncept of Oper**a**tions, and **P**rototype **E**ngineering (**ESCAPE**) dataset was collected by the AFRL and Information Directorate to engage the data fusion community in advanced heterogeneous data analytics, design, and understanding. This dataset has time-synchronized electro-optical (EO), infrared (IR), distributed passive radio-frequency (P-RF), radar, acoustic, and seismic sensor data. To provide multimodal data on diverse ground vehicle scenarios, the scenarios in ESCAPE dataset range from single-vehicle to multi-vehicle trajectories. Five different types of ground vehicle targets are present: (a) gas motor gator, (b) diesel motor gator, (c) pickup truck, (d) panel van, and (e) stake rack truck. The dataset contains 13 scenarios. For a detailed description

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3145663, IEEE Transactions on Multimedia

| Scenarios | Date | Duration | Train/test | Vehicles |
|---|---|---|---|---|
| Scenario 1 run 3 | 10/16/2018 | 79s | train | gas gator |
| Scenario 1 run 5 | 10/16/2018 | 35s | train | stake rack |
| Scenario 1 run 7 | 10/16/2018 | 79s | train | pickup |
| Scenario 1 run 8 | 10/19/2018 | 79s | train | stake rack |
| Scenario 1 run 10 | 10/16/2018 | 79s | train | diesel gator |
| Scenario 1 run 11 | 10/16/2018 | 79s | train | pickup |
| Scenario 1 run 12 | 10/19/2018 | 79s | test | stake rack |
| Scenario 2A run 6 | 10/16/2018 | 79s | train | gas gator, diesel gator |
| Scenario 2A run 7 | 10/16/2018 | 79s | train | gas gator, pickup |
| Scenario 2C run 4 | 10/16/2018 | 79s | test | diesel gator, pickup, van |
| Scenario 2C run 5 | 10/16/2018 | 79s | train | diesel gator, pickup, van |
| Scenario 2G run 1 | 10/19/2018 | 59s | train | gas gator, diesel gator |
| Scenario 2H run 1 | 10/19/2018 | 74s | test | gas gator, diesel gator, pickup |

TABLE I: ESCAPE dataset [11] summary. The "Date" column signifies the exact date on which data on that scenario was collected. The "Vehicles" column represents the exact vehicles present in that scenario. The "Train/test" column shows our split on putting a scenario either in train or test set in *scenario-based* evaluations (see Sec. VII).

of the ESCAPE dataset, please refer to [11]; dataset details are summarized in Tab. I.

The dataset comprises both single-vehicle scenarios and multi-vehicle scenarios. The single-vehicle scenarios include all types of ground vehicles, and can be used to train a multimodal representation for each vehicle. Multi-target scenarios (up to 4 vehicles) contain closely spaced targets, opposing targets, passing targets, move-stop-move trajectories, and LOS obstructions. These scenarios pose a challenge for traditional detection approaches. Thus, the ESCAPE dataset provides an excellent platform to explore data fusion approaches leveraging the multimodal representation of each vehicle.

### B. Selected Modalities

We choose *image* (or EO data), *radar*, *acoustic* and *seismic* as the candidate modalities to validate our proposed fusion frameworks. We omit P-RF and IR modalities from further consideration as we observed that they did not add further diversity in the data. The sensors for the IR images were co-located with EO image sensors, thereby adding only redundant information when used along with EO images. Additionally, IR images where collected at a considerably lower resolution. The P-RF data was collected from identical transmitters deployed on vehicles used in a scenario, hence there was no possibility of exploiting vehicle-specific unique transmitter-fingerprinting [42], [43]. Hence, P-RF data were not discriminative for multi-vehicle detection and classification.

### C. Competence of the Selected Modalities

Among additional sensors which have gained popularity for vehicle detection, light detection and ranging (LiDAR) and inertial measurement unit (IMU) sensors [44] are most prominent. The LiDAR sensor gives the 3D representation of an object, hence, it is limited only to external appearance. Vehicles with same make, model and color will have identical representations in the LiDAR modality (for example, gas and diesel gators in ESCAPE dataset). Similarly, IMU gives only information about relative position and velocity. Therefore, the distinguishing factors in same type of vehicles' appearances are not captured in IMU or LiDAR sensors. In contrast, the acoustic and seismic modalities can highlight distinguishing

| Modality | Detects | Does not detect |
|---|---|---|
| **Image ✓** | vehicle appearance | engine type |
| LiDAR ✗ | vehicle appearance | engine type |
| IMU ✗ | vehicle orientation and relative position | vehicle appearance, engine type |
| RF ✓ | vehicle location and path | engine type, vehicle appearance |
| **Radar ✓** | vehicle location and path | engine type, vehicle appearance |
| **Acoustic ✓** | engine type | vehicle appearance |
| **Seismic ✓** | engine type | vehicle appearance |

TABLE II: Analyzing the relevant modalities for multi-vehicle classification: in terms of detecting features in the vehicle. The used modalities are highlighted in bold font. The ticked modalities are included in the ESCAPE dataset.

factors due to vehicle mechanics. The seismic modality captures vibration patterns of different engine types, while acoustical signatures can differentiate among mechanical features across vehicles [11]. However, both these modalities ignore external appearance. Image and radar modalities contribute to determining the area of interest and number of moving objects in the area during object detection and tracking. However, radar is again unable to differentiate between the vehicle's external appearance and mechanical features.

In Tab. II, we summarize the relevance of different modalities with respect to multi-vehicle classification. We choose image, radar, acoustic and seismic modalities as the ESCAPE dataset modalities we use. These choices are consistent with past vehicle detection and tracking methods [35], [38], [39].

### D. Preprocessing and Unimodal Networks

We first describe different preprocessing techniques for generating high-level features across modalities. We then propose different finetuned unimodal architectures for each modality to overcome the issue of overfitting to a particular modality. We present a comprehensive overview of architectures and input sizes of our unimodal network in Tab. III.

*1) Image:* Preprocessing step for image modality mainly consists of designing a finetuned object detection network to generate high level image features. In the ESCAPE dataset, most pixels of an image form the background, hence, directly using a classification model over raw images can diminish classifier prediction accuracy. Therefore, we extract targeted vehicles first, and then classify them. However, existing pre-trained object detection models detect all kinds of objects (e.g., person, umbrella, frisbee), which are irrelevant to our task. Hence, finetuning our object detection network is crucial to our multi-vehicle detection task. We thus perform the three steps to create our image preprocessing pipeline: (a) vehicle detection using a pre-trained object detection network, (b) finetuning the model to our dataset with labeling, (c) extracting the preprocessed high level features from labeled data.

**Object Detection.** We begin with a pre-trained object detection network, R-50-FPN Faster RCNN [14]. Faster RCNN estimates a set of candidate parts of objects that appear within a region, and uses these as inputs to a CNN. The latter extracts features and classifies the scenario as *background* or as containing some real objects. Faster RCNN can detect small objects well, since it has nine anchors in a single grid. This R-50-FPN Faster RCNN model, whose backbone is feature pyramid networks [46], is pretrained on COCO datasets [47].

We present the results of Faster RCNN implementation in Fig. 3(a) for an image in Scenario 2H run 1. The bounding

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3145663, IEEE Transactions on Multimedia


(a) Pre-trained Faster RCNN (b) Finetuned Faster RCNN

Fig. 3: An example output of pre-trained and finetuned Faster RCNN at in Scenario 2H run 1. The results include the bounding boxes of objects, the classification results, and the confidence of the classification. The performance achieves great improvement after the finetuning.


Fig. 4: Bounding box labeling in Scalabel [45].


Fig. 5: An image in Scenario 2A run 6. For this image, the true label is [1, 1, 0, 0, 0], while visible label is [1, 0, 0, 0, 0, 0], where the first bit is a 'non-visible' label.

box indicates an object, and the associated text indicates the object classification and its confidence. However, we get many disparities in detection performance. First, the pre-trained model detects some objects irrelevant to vehicles. Second, many vehicles cannot be detected. Third, even when detected, different types of vehicles cannot be precisely differentiated. Hence, we finetune this model on our image modality data.

**Model Finetuning.** In order to finetune the R-50-FPN Faster RCNN model, we label the images with visible vehicles in our dataset. We use the open source labeling tool Scalabel [45], shown in Fig. 4. Each object in the image is labeled with a rectangle bounding box and its corresponding classification. We have two types of labels: *true labels* and *visible labels*, derived from the metadata of the ESCAPE dataset. For each frame/image in a scenario, true labels represent the vehicles in this scenario, no matter whether the vehicles can be observed in the frame, whereas the visible labels only indicate the observable vehicles in the frame, which are obtained through manual labeling. For example, an image in Scenario 2A run 6 is shown in Fig. 5. The true label for this image is: [1, 1, 0, 0, 0], where each bit represents the presence of gas gator, diesel gator, pickup, van, and stake rack respectively: in this scenario, gas gator and diesel gator are present in the field. The respective visible label for this image however is: [1, 0, 0, 0, 0, 0], where the first bit of this label is a 'non-visible' label, since no vehicles are visible. We introduce this 'non-visible' label to indicate the NLOS settings and will utilize this label in fusion. We label about 15000 images across our 13 scenarios to retrain and test our model.

Having this labeled data, we now finetune the Faster RCNN. Since we have only five types of vehicles, we first trim the last layer from pre-trained Faster RCNN model. Then, we set a very low learning rate to train the model with our labeled data. We improve the precision of the detection dramatically through this finetuning. For example, in Fig. 3(b), when using our finetuned Faster RCNN to detect vehicles, we get the result with 100% precision and high confidence. By construction, no irrelevant objects are detected.

**Feature Extraction.** Though the finetuned Faster RCNN achieves extremely high precision for vehicle detection, its performance on vehicle classification gets increasingly worse due to limited labeled data, low image resolution, small target vehicles, and extremely similar appearance of targeted vehicles. Hence, we extract features from Faster RCNN and further train an additional model on these extracted features specifically for vehicle classification. This has several advantages: (a) we take highly encoded features as inputs focused on vehicle classification, (b) we mitigate overfitting by selecting a smaller architecture than Faster RCNN to perform classification, and (c) the number of computations are reduced during training and, hence, the resulting network is easier to be fused with other modalities. We describe in detail below how we extract and prepare these features.

Faster RCNN consists of four parts: (a) *feature pyramid networks (FPN)*, (b) *region proposal networks (RPN)*, (c) *region of interest (RoI) pooling* and (d) *classification*. Through incorporating both features from FPN and bounding box proposal from RPN, we choose the output of RoI pooling, i.e., *roi_box_features*, as our extracted feature, as these features have fixed size (i.e., 100×1024) and maintain information from both feature network and region proposal.

The images are captured at 14 frames per second. In order to align with other modalities, which have one feature per second, we concatenate those 14 frames within one second as one feature, i.e., the size of feature is 1400×1024, where $d_0^{\mathrm{I}} = 14$, $d_1^{\mathrm{I}} = 1400\times1024$. We take the majority of visible labels to generate a visible label for this concatenated feature. As per the data collection setting, the image feature at the $t^{th}$ second corresponds to features of other modalities at the $(t-1)^{th}$ second. Hence, we discard the image features of first 1 second to synchronize image with other modalities.

**Unimodal Image Network.** Once we have the preprocessed high level features from our finetuned Faster RCNN, we train a ResNet-18 [48] network with those extracted features and visible labels from scratch, explicitly for vehicle classification. The architectural overview of our ResNet-18 model is shown in Fig. 6. Recall that the output of our ResNet-18 is the prediction score for both 5 types of vehicles and a 'non-visible'

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3145663, IEEE Transactions on Multimedia

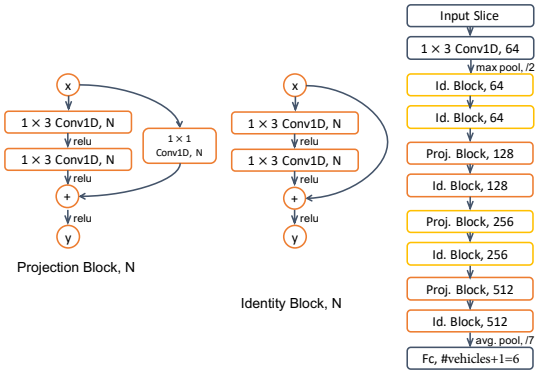
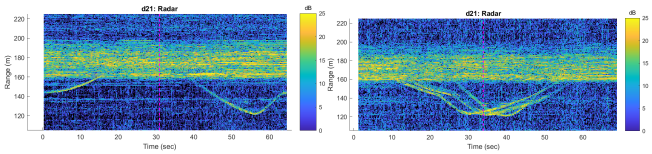
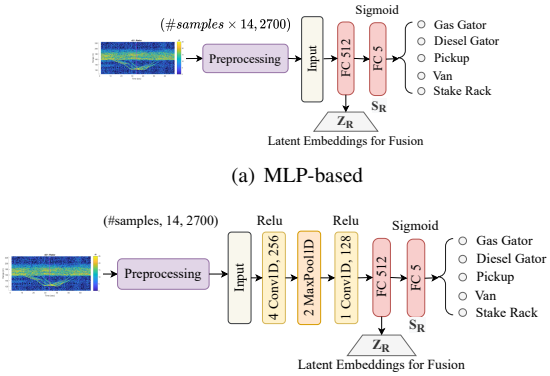



Fig. 6: Architecture of ResNet-18. The last layer of ResNet-18 model is customized to support prediction of 5 vehicles on 1 visible label.




(a) Single vehicle (b) Three vehicles

Fig. 7: Radar plots for single-vehicle and three-vehicle scenarios. We run a matlab scripts to plot these from captured **.imb* files of ESCAPE dataset.




(a) MLP-based

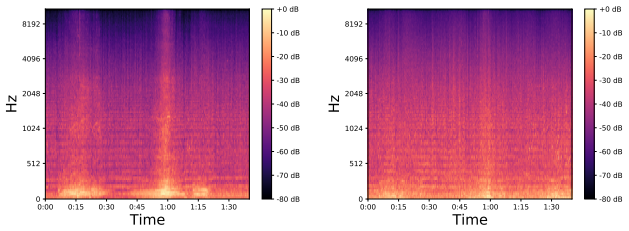
(b) CNN-based

Fig. 8: Proposed neural network architectures for the radar modality.

label, thereby yielding a 6-dimensional vector.

*2) Radar:* The radar modality is used for object detection and ranging the targets. We present a visual representation (Fig. 7) of how the radar data appears differently for single vehicle and multiple vehicle scenarios in the ESCAPE dataset. We extract the I/Q data from the captured radar **.imb* files by designing a matlab script as preprocessing step. The I/Q data of $d_0^{\mathrm{R}}$ samples per second with size of $d_1^{\mathrm{R}}$ has been used as input to the unimodal network for radar, where $d_0^{\mathrm{R}} = 14$ and $d_1^{\mathrm{R}} = 1350$.

**Unimodal Radar Network.** To leverage the raw I/Q data of radar modality, we propose a multi-layer perception (MLP) network (refer to Fig. 8 (a)) with 1 hidden layer, 512 filters, and sigmoid activation to map the raw radar data from $[\mathcal{X}_{min}, \mathcal{X}_{max}]$ range to $[0, 1]$, where $\mathcal{X}_{min}$, and $\mathcal{X}_{max}$ are




(a) Gas gator (b) Pickup

Fig. 9: Spectrogram plot for acoustic sensor data for two different vehicles.

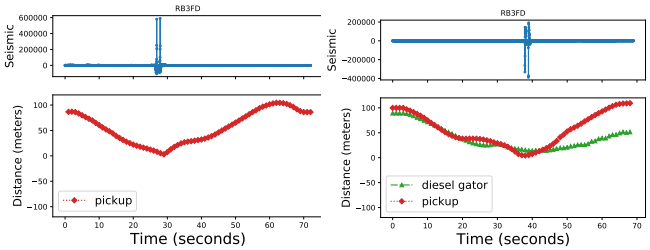



Fig. 10: Seismic plot for single-vehicle and two-vehicle scenarios.

the minimum and maximum values in the raw data. The output layer consists of 5 (number of different vehicles) filters which generate the latent embedding for the fusion network. A sigmoid activation is applied at the output layer.

In addition to the MLP network, we also experimented with a CNN model which captures the spatial correlation in the data with two convolutional layer of kernel sizes 4 and 1, one pooling layer in between them to extract max value of 4 consecutive neurons of the output of first layer. We add two fully connected layers after the convolutional layers as well with the sigmoid activation at the output layer.The proposed CNN-based radar architecture is presented in Fig. 8(b).

*3) Acoustic:* The acoustic modality contains audio signals recorded by 16 microphones. Unlike the aforementioned modalities, we do not directly train a deep model on the raw audio signals. To exploit the intrinsic properties of acoustic signals from different vehicles, we generate spectrograms of recorded audio signals for all the scenarios. From the spectrogram plots, as shown in Fig. 9, we observe that different patterns correspond to different vehicles, which motivates us to classify vehicles via spectrograms. We use $d_0^{\mathrm{A}} = 16$ spectrograms per second with size of $d_1^{\mathrm{A}} \times d_2^{\mathrm{A}}$ as inputs to the acoustic neural network, where $d_1^{\mathrm{A}} = 128$ and $d_2^{\mathrm{A}} = 44$.

**Unimodal Acoustic Network.** As the preprocessed inputs of acoustic network are two dimensional spectrograms and VGG16 [8] has excellent performance on the ImageNet dataset [49], we use a VGG16 network pre-trained on ImageNet, and finetune the weights of the first and last layers.

We also consider VGG-Sound net [50], a version of VGG16 which is pre-trained on different acoustic (e.g., police car siren, guitar, clapping) datasets. We use the pre-trained model which used ResNet18 architecture with NetVLAD aggregation method for audio recognition tasks [51]. We finetune the weights of the first and last layers in this case as well.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3145663, IEEE Transactions on Multimedia

| Modality | Architecture | Input Size |
|---|---|---|
| Image (Detection) | Finetuned Faster-RCNN [14] | $3840 \times 2748$ |
| Image (Classification) | Finetuned Resnet18 [48] | $1400 \times 1024$ |
| Radar | Custom Designed CNN | $14 \times 1350$ |
| Acoustic | VGG16 [8] & VGG-Sound [50] | $16 \times 128 \times 44$ |
| Seismic | Custom Designed Shallow NN | 90 |

TABLE III: The unimodal network architecture and input size for each modality.

*4) Seismic:* Seismic modality is an interesting component to the ESCAPE dataset as different ground targets have uniquely different seismic signature profiles. We plot the seismic data as a function of the distance of the vehicles (from the seismic sensors) in two scenarios in Figure 10. We observe a pulse when the vehicles are nearest to the seismic sensor location. For seismic signals from each sensor, we spit the signal into 1-second bins. Then we further preprocess those and generate 9 features based on the differences between two adjacent bins, e.g., maximum, minimum, average, median, standard deviation etc. As a result, for each second, we obtain $d_0^{\mathtt{S}} = 90$ features from all 10 seismic sensors.

**Unimodal Seismic Network.** Due to the limited dimensionality of seismic modality, we design a shallow neural network to identify vehicle types. Our seismic model consists of a single layer with a ReLU activation.

### E. Fusion Notation

We denote the data matrices after preprocessing for image, radar, acoustic and seismic by $X_{\mathtt{I}} \in \mathbb{R}^{N_t \times d_0^{\mathtt{I}} \times d_1^{\mathtt{I}}}$, $X_{\mathtt{R}} \in \mathbb{R}^{N_t \times d_0^{\mathtt{R}} \times d_1^{\mathtt{R}}}$, $X_{\mathtt{A}} \in \mathbb{R}^{N_t \times d_0^{\mathtt{A}} \times d_1^{\mathtt{A}} \times d_2^{\mathtt{A}}}$, $X_{\mathtt{S}} \in \mathbb{R}^{N_t \times d_0^{\mathtt{S}}}$ respectively, where $N_t$ is the number of training samples. Furthermore, we denote by $(d_0^{\mathtt{I}} \times d_1^{\mathtt{I}})$, $(d_0^{\mathtt{R}} \times d_1^{\mathtt{R}})$, $(d_0^{\mathtt{A}} \times d_1^{\mathtt{A}} \times d_2^{\mathtt{A}})$ and $(d_0^{\mathtt{S}})$ the respective dimensions of the preprocessed image, radar, acoustic, and seismic data. We represent the presence of $\mathcal{V}$ vehicles via binary label matrix $Y \in \{0,1\}^{N_t \times |\mathcal{V}|}$.

*1) Penultimate Layers:* As mentioned in Sec. III-B, the penultimate layer of each unimodal network captures the data representation of each modality. The penultimate layers of the unimodal networks of image, radar, acoustic, and seismic have $d^{\mathtt{I}}$, $d^{\mathtt{R}}$, $d^{\mathtt{A}}$, $d^{\mathtt{S}}$ neurons, respectively. As a result, different sensor modality sample input maps to a vector with that dimension after passing through these unimodal feature extractors. We denote the feature extractor of each modality as $f_{\theta^{\mathtt{I}}}^{\mathtt{I}}$, $f_{\theta^{\mathtt{R}}}^{\mathtt{R}}$, $f_{\theta^{\mathtt{A}}}^{\mathtt{A}}$, and $f_{\theta^{\mathtt{S}}}^{\mathtt{S}}$ for image, radar, acoustic, and seismic data, respectively, each parametrized by weight vectors $\theta^m$, for $m \in \{\mathtt{I}, \mathtt{R}, \mathtt{A}, \mathtt{S}\}$. We refer to the output of these feature extractors as the latent embedding of each modality. Formally,

$$\begin{aligned}
\mathbf{z}_{\mathtt{I}} &= f_{\theta^{\mathtt{I}}}^{\mathtt{I}}(X_{\mathtt{I}}), \qquad f_{\theta^{\mathtt{I}}}^{\mathtt{I}} : \mathbb{R}^{d_0^{\mathtt{I}} \times d_1^{\mathtt{I}}} \mapsto \mathbb{R}^{d^{\mathtt{I}}} \\
\mathbf{z}_{\mathtt{R}} &= f_{\theta^{\mathtt{R}}}^{\mathtt{R}}(X_{\mathtt{R}}), \qquad f_{\theta^{\mathtt{R}}}^{\mathtt{R}} : \mathbb{R}^{d_0^{\mathtt{R}} \times d_1^{\mathtt{R}}} \mapsto \mathbb{R}^{d^{\mathtt{R}}} \\
\mathbf{z}_{\mathtt{A}} &= f_{\theta^{\mathtt{A}}}^{\mathtt{A}}(X_{\mathtt{A}}), \qquad f_{\theta^{\mathtt{A}}}^{\mathtt{A}} : \mathbb{R}^{d_0^{\mathtt{A}} \times d_1^{\mathtt{A}} \times d_2^{\mathtt{A}}} \mapsto \mathbb{R}^{d^{\mathtt{A}}} \\
\mathbf{z}_{\mathtt{S}} &= f_{\theta^{\mathtt{S}}}^{\mathtt{S}}(X_{\mathtt{S}}), \qquad f_{\theta^{\mathtt{S}}}^{\mathtt{S}} : \mathbb{R}^{d_0^{\mathtt{S}}} \mapsto \mathbb{R}^{d^{\mathtt{S}}}
\end{aligned}$$

where $\mathbf{z}_{\mathtt{I}}$, $\mathbf{z}_{\mathtt{R}}$, $\mathbf{z}_{\mathtt{A}}$, and $\mathbf{z}_{\mathtt{S}}$ denote the extracted latent embeddings for input data $X_{\mathtt{I}}$, $X_{\mathtt{R}}$, $X_{\mathtt{A}}$, and $X_{\mathtt{S}}$, respectively.

*2) Ultimate Layers:* We formulate the output of the ultimate layers for each unimodal network in regard to the corresponding latent embeddings. The ultimate layers are just the transformation the penultimate embeddings with suitable activation functions. The outputs of ultimate layers for unimodal networks of image, radar, acoustic, and seismic are denoted by $\mathbf{s}_{\mathtt{I}}$, $\mathbf{s}_{\mathtt{R}}$, $\mathbf{s}_{\mathtt{A}}$, and $\mathbf{s}_{\mathtt{S}}$ respectively, where:

$$\begin{aligned}
\mathbf{s}_{\mathtt{I}} &= \sigma(f_{\theta^{\mathtt{I}}}^{\mathtt{I}}(\mathbf{z}_{\mathtt{I}})), \qquad f_{\theta^{\mathtt{I}}}^{\mathtt{I}} : \mathbb{R}^{d_0^{\mathtt{I}} \times d_1^{\mathtt{I}}} \mapsto \mathbb{R}^{|\mathcal{V}|} \\
\mathbf{s}_{\mathtt{R}} &= \sigma(f_{\theta^{\mathtt{R}}}^{\mathtt{R}}(\mathbf{z}_{\mathtt{R}})), \qquad f_{\theta^{\mathtt{R}}}^{\mathtt{R}} : \mathbb{R}^{d_0^{\mathtt{R}} \times d_1^{\mathtt{R}}} \mapsto \mathbb{R}^{|\mathcal{V}|} \\
\mathbf{s}_{\mathtt{A}} &= \iota(f_{\theta^{\mathtt{A}}}^{\mathtt{A}}(\mathbf{z}_{\mathtt{A}})), \qquad f_{\theta^{\mathtt{A}}}^{\mathtt{A}} : \mathbb{R}^{d_0^{\mathtt{A}} \times d_1^{\mathtt{A}} \times d_2^{\mathtt{A}}} \mapsto \mathbb{R}^{|\mathcal{V}|} \\
\mathbf{s}_{\mathtt{S}} &= \zeta(f_{\theta^{\mathtt{S}}}^{\mathtt{S}}(\mathbf{z}_{\mathtt{S}})), \qquad f_{\theta^{\mathtt{S}}}^{\mathtt{S}} : \mathbb{R}^{d_0^{\mathtt{S}}} \mapsto \mathbb{R}^{|\mathcal{V}|}
\end{aligned}$$

and $\sigma$, $\iota$, $\zeta$ are the sigmoid, linear, and ReLU activation functions, respectively. Next, we formulate different ways to combine different modalities.

### F. Fusion Strategies

We explore several state-of-the-art fusion strategies [52] to combinine different modalities. Even though we formulate these strategies for the three non-image modalities, however, they are extendable to image and other modalities.

*1) Concatenation:* Feature concatenation is an effective strategy for feature-based fusion in machine learning [53]. We perform concatenation in the late fusion approach where the feature representations from the unimodal models are extracted, and then concatenated. The fused feature vector [54] is represented as $\mathbf{z}_{\mathtt{C}} = [\mathbf{z}_{\mathtt{A}}; \mathbf{z}_{\mathtt{R}}; \mathbf{z}_{\mathtt{S}}]$ where $\mathbf{z}_{\mathtt{A}}$, $\mathbf{z}_{\mathtt{R}}$, and $\mathbf{z}_{\mathtt{S}}$ are the latent embeddings for acoustic, radar, and seismic modalities, respectively.

*2) Multiplicative Interactions (MI):* Multiplicative interactions generalize tensor products to include few learnable parameters so that rank and structures of those parameters can be constrained on the type of the output [55]. We formulate the multiplicative interactions between three modalities in the form of bilinear products:

$$\mathbf{z}_{\mathtt{MI}} = f_{\mathtt{MI}}^1(\mathbf{z}_{\mathtt{M}_1}, \mathbf{z}_{\mathtt{M}_2}), \qquad f_{\mathtt{MI}}^1 : (\mathbf{z}_{\mathtt{M}_1}\mathbf{W}^1\mathbf{z}_{\mathtt{M}_2} + \mathbf{z}_{\mathtt{M}_1}^T\mathbf{U}^1 + \mathbf{V}^1\mathbf{z}_{\mathtt{M}_2} + \mathbf{b}^1)$$

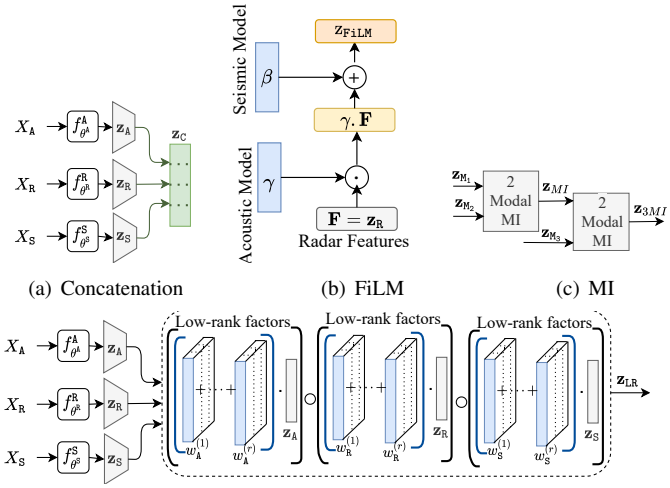
$$\mathbf{z}_{\mathtt{3MI}} = f_{\mathtt{MI}}^2(\mathbf{z}_{\mathtt{MI}}, \mathbf{z}_{\mathtt{M}_3}), \qquad f_{\mathtt{MI}}^2 : (\mathbf{z}_{\mathtt{MI}}\mathbf{W}^2\mathbf{z}_{\mathtt{M}_3} + \mathbf{z}_{\mathtt{MI}}^T\mathbf{U}^2 + \mathbf{V}^2\mathbf{z}_{\mathtt{M}_3} + \mathbf{b}^2)$$

where $\mathbf{z}_{\mathtt{M}_1}$, $\mathbf{z}_{\mathtt{M}_2}$, and $\mathbf{z}_{\mathtt{M}_3}$ are the extracted unimodal features from the modalities $\mathtt{M}_1$, $\mathtt{M}_2$, and $\mathtt{M}_3$, respectively, representing the acoustic, radar, and seismic modalities in different permutations. Matrices $\mathbf{W^1}, \mathbf{W^2}, \mathbf{U^1}, \mathbf{U^1}, \mathbf{V^1}, \mathbf{V^2}$, and vectors $\mathbf{b^1}$, $\mathbf{b^2}$ are the trainable parameters. The fused features from two and three modalities are represented by $\mathbf{z}_{\mathtt{MI}}$ and $\mathbf{z}_{\mathtt{3MI}}$, respectively. The outcome of $\mathbf{z}_{\mathtt{MI}}$ and $\mathbf{z}_{\mathtt{3MI}}$ will vary depending on the sequence of the participating modalities.

*3) Feature-wise Linear Modulation (FiLM) Layer:* FiLM layers [56] can be used to perform simple computations on network's intermediate features, where specific parameters $\gamma$ and $\beta$ are set to modulate a neural (base modality) network's features $\mathbf{F}$. The latter form a feature-wise affine transformation based on conditioning information. FiLM layers can be exploited to fuse multiple modalities in which transformed features per modality are conditioned on the relevance of each modality. Hence, our formulated FiLM layer for fusion is described as:

$$\mathbf{z}_{\mathtt{FiLM}}(\mathbf{F}|\gamma, \beta) = \gamma\mathbf{F} + \beta,$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3145663, IEEE Transactions on Multimedia




(a) Concatenation (b) FiLM (c) MI

(d) Low rank multimodal fusion layer

Fig. 11: Representation of different fusion strategies used as fusion layer in the proposed fusion networks. The '·' in (b) signifies a Hadamard product. The '∘' in (d) indicates the element-wise product over a sequence of tensor. The shown graphics are considering fusion among radar, acoustic and seismic modality as an example. However, all these strategies can be extended to image and more number of modalities.

where $\gamma$ is $f^{\mathtt{A}}_{\theta^{\mathtt{A}}}$ and $\beta$ is $f^{\mathtt{S}}_{\theta^{\mathtt{S}}}$ and the base modal features $\mathbf{F}$ is $\mathbf{z}_{\mathtt{R}}$. We denote the fused features by $\mathbf{z}_{\mathtt{FiLM}}$.

*4) Low Rank Multimodal Fusion:* The idea of low rank multimodal fusion is to decompose the weight tensor into M sets of modality-specific factors [57]. Parameter $r$ is specified by the rank of the tensor which makes the decomposition valid. As per [57], the generated fused features are defined as:

$$\mathbf{z}_{\mathtt{LR}} = \sum_{i=1}^{r} \left[ \Lambda_{m=1}^{M} \left[ \mathbf{w}_m^{(1)}, \mathbf{w}_m^{(2)}, \cdots, \mathbf{w}_m^{(r)} \right] \cdot X_m \right]_{i,:},$$

where $\Lambda$ is the element-wise product over a sequence of tensors, '·' is the Hadamard product, $\mathbf{w}_m$ is the weight matrix for modality $m \in (\mathtt{A}, \mathtt{R}, \mathtt{S})$, $r$ is the rank, and $X_m$ is the input.

An overview of the fusion strategies we use is presented in Fig. 11. Finally, we denote the fusion layer by $\mathcal{F}(\cdot) = \mathbf{z}$, where $\mathbf{z} \in (\mathbf{z}_{\mathtt{C}}, \mathbf{z}_{\mathtt{3MI}}, \mathbf{z}_{\mathtt{FiLM}}, \mathbf{z}_{\mathtt{LR}})$.

## V. Proposed Fusion Frameworks for Non-Image Modalities

Non-image unimodal networks can predict vehicle classes but image networks can additionally be used for tracking the vehicle location from the image modality. Hence, we first introduce different novel fusion frameworks among non-image unimodal networks, then we discuss about the image fusion in the next section.

Our fusion frameworks work on the basic principles of *feature-based* and *decision-based* fusion. In feature-based fusion, deep learning approach automatically assigns higher weights on the more relevant modalities than the other during the training. However, in decision-based fusion, we set specific rules so that the more relevant modalities get prioritized over others. Most of the proposed fusions work with the feature-based fusion principle, unless explained otherwise.

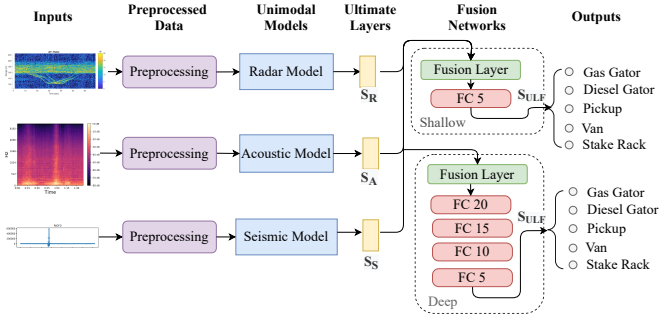



Fig. 12: Fusion at ultimate layers. The top block of fusion network represents the proposed shallow fusion among all modalities after fusion layer. The bottom block represents the proposed deep fusion network with 4 dense layers after fusion layer. Training is performed on each block separately.

### A. Fusion at Ultimate Layer of Non-image Modalities

One general strategy to attack the feature-based fusion is to combine the output of ultimate layers of different modalities. The ultimate layer represents the high-level representation of each unimodal model. However, the penultimate layer captures the intrinsic low-level representation. We propose a feature-based fusion on the ultimate layer output of the unimodal models and added both deep and shallow fusion networks. The details of the proposed networks are presented in Fig. 12. The selection of the network layers and architecture has been determined depending on the performance on the training data.

To define this fusion network, we take earlier defined ultimate layers of each modality: $\mathbf{s}_{\mathtt{R}}$, $\mathbf{s}_{\mathtt{A}}$ and $\mathbf{s}_{\mathtt{S}} \in \mathbb{R}^{|\mathcal{V}|}$, the fused representation matrix $\mathbf{s}_{\mathtt{U}}$ is generated as: $\mathbf{s}_{\mathtt{U}} = \mathcal{F}(\mathbf{s}_{\mathtt{R}}, \mathbf{s}_{\mathtt{A}}, \mathbf{s}_{\mathtt{S}}) \in \mathbb{R}^{3\times|\mathcal{V}|}$, where $\mathcal{F}(\cdot)$ represents one of the fusion strategies mentioned in Sec. IV-F.

We denote the ultimate layer fusion network as $f^{\mathtt{ULF}}_{\theta^{\mathtt{ULF}}}(\cdot)$:

$$\mathbf{s}_{\mathtt{ULF}} = \sigma(f^{\mathtt{ULF}}_{\theta^{\mathtt{ULF}}}(\mathbf{s}_{\mathtt{U}})), \qquad f^{\mathtt{ULF}}_{\theta^{\mathtt{ULF}}} : \mathbb{R}^{3\times|\mathcal{V}|} \mapsto \mathbb{R}^{|\mathcal{V}|},$$

where $\sigma$ is the sigmoid activation function.

### B. Fusion at Penultimate Layer of Non-image Modalities

Given the latent embeddings at penultimate layers of all modalities, we also design two types of fusion frameworks: aggregated fusion and incremental fusion, by fusing the latent feature embedding from each unimodal network.

*1) Aggregated Fusion Framework:* Given $\mathbf{z}_{\mathtt{R}} \in \mathbb{R}^{d^{\mathtt{R}}}$, $\mathbf{z}_{\mathtt{A}} \in \mathbb{R}^{d^{\mathtt{A}}}$ and $\mathbf{z}_{\mathtt{S}} \in \mathbb{R}^{d^{\mathtt{S}}}$, we fuse them in aggregated manner and generate the combined latent feature matrix $\mathbf{z}$ as:

$$\mathbf{z} = \mathcal{F}(\mathbf{z}_{\mathtt{R}}, \mathbf{z}_{\mathtt{A}}, \mathbf{z}_{\mathtt{S}}) \in \mathbb{R}^{d^{\mathtt{R}}+d^{\mathtt{A}}+d^{\mathtt{S}}}.$$

We denote the aggregated fusion network as $f^{\mathtt{AGF}}_{\theta^{\mathtt{AGF}}}(\cdot)$:

$$\mathbf{s}_{\mathtt{AGF}} = \sigma(f^{\mathtt{AGF}}_{\theta^{\mathtt{AGF}}}(\mathbf{z})), \qquad f^{\mathtt{AGF}}_{\theta^{\mathtt{AGF}}} : \mathbb{R}^{d^{\mathtt{R}}+d^{\mathtt{A}}+d^{\mathtt{S}}} \mapsto \mathbb{R}^{|\mathcal{V}|}$$

where $\sigma$ is the sigmoid. The aggregated fusion network, shown in Fig. 13, fuses the latent embeddings of the penultimate layers from three non-image unimodal networks, and adds two convolutional module and three fully connected (FC) layer afterwards. First convolutional module consists of two convolutional layer of kernel sizes 4 and 1, one pooling maximum

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3145663, IEEE Transactions on Multimedia

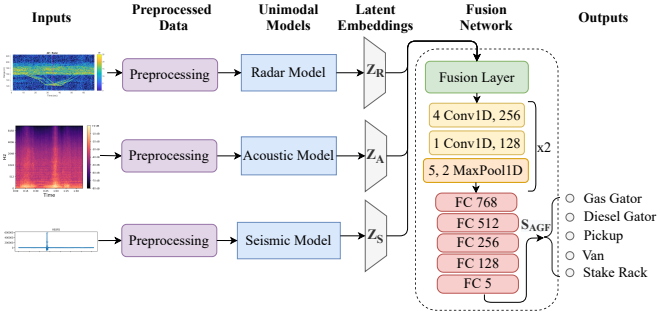



Fig. 13: Aggregate fusion at penultimate layers. The latent embeddings of each non-image modality contribute to the overall fusion network. The output of the fusion network gives prediction of each vehicle.

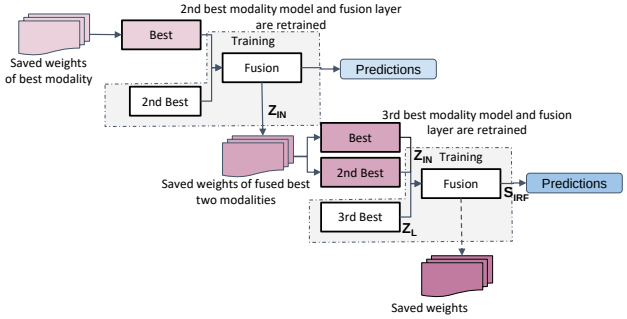



Fig. 14: Incremental fusion at penultimate layers.

value of 5 consecutive neurons. The second convolutional module has similar settings with the only difference of pooling maximum value from 2 consecutive neurons.

*2) Incremental Fusion Framework:* For incremental fusion model, we first sort unimodal networks according to their performance. We fuse the embedding of penultimate layers of the best unimodal networks and the second best one. In this step, we freeze the best model and only retrain the second best model and the fusion model. This forces the second best model and the fusion model to learn different information than the best model. Similarly, when incorporating the third best model, we only retrain the third and fusion layers. A conceptual overview of the incremental fusion framework is illustrated in Fig. 14.

Following the same notation as aggregated fusion, the combined latent feature matrix $\mathbf{z}$ can be represented as:

$$\mathbf{z} = \mathcal{F}(\mathbf{z}_{\mathtt{IN}}, \mathbf{z}_{\mathtt{L}}) \in \mathbb{R}^{d^{\mathtt{R}}+d^{\mathtt{A}}+d^{\mathtt{S}}},$$

where $\mathbf{z}_{\mathtt{IN}} \in \{\mathcal{F}(\mathbf{z}_{\mathtt{R}}, \mathbf{z}_{\mathtt{A}}), \mathcal{F}(\mathbf{z}_{\mathtt{R}}, \mathbf{z}_{\mathtt{S}}), \mathcal{F}(\mathbf{z}_{\mathtt{A}}, \mathbf{z}_{\mathtt{S}})\}$, and $\mathbf{z}_{\mathtt{L}} \in \{\mathbf{z}_{\mathtt{R}}, \mathbf{z}_{\mathtt{A}}, \mathbf{z}_{\mathtt{S}}\}$, $\mathbf{z}_{\mathtt{IN}}$ refers to the fused latent embeddings of the best two modalities, and $\mathbf{z}_{\mathtt{L}}$ is the latent embedding of the third best modality. The notations for incremental fusion is parameterized over $\mathtt{IRF}$. We denote the incremental fusion network as $f^{\mathtt{IRF}}_{\theta^{\mathtt{IRF}}}(\cdot)$ as:

$$\mathbf{s}_{\mathtt{IRF}} = \sigma(f^{\mathtt{IRF}}_{\theta^{\mathtt{IRF}}}(\mathbf{z})), \qquad f^{\mathtt{IRF}}_{\theta^{\mathtt{IRF}}} : \mathbb{R}^{d^{\mathtt{R}}+d^{\mathtt{A}}+d^{\mathtt{S}}} \mapsto \mathbb{R}^{|\mathcal{V}|}$$

### C. Multi-level Fusion of Non-Image Modalities

So far the proposed feature-based fusion networks exploit the correlation in either the ultimate or penultimate layers of the unimodal networks. However, such feature-based fusion networks can further be exploited along with the unimodal networks for an improved decision at the second level. The output of the ultimate layers of each unimodal or fusion network represent the sensitivity of that particular model for detecting each vehicles. Hence, we propose few multi-level fusion techniques which involve: (i) unimodal radar network, (ii) unimodal acoustic network, (ii) unimodal seismic network, and (iv) penultimate feature fusion network (Sec. V-B).

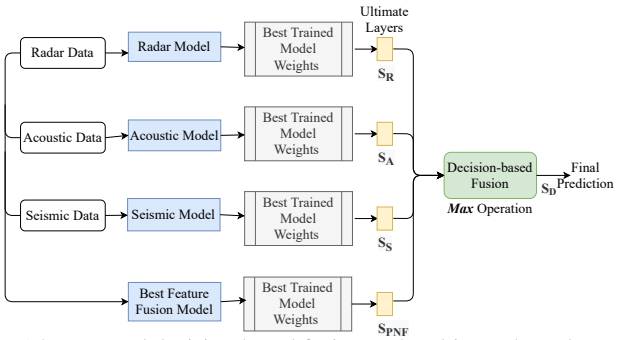



Fig. 15: Proposed decision-based fusion at the ultimate layer between all modalities and penultimate fusion model. The preprocessing step in between the raw data and proposed neural network models is not shown in this block diagram.

*1) Multi-level Decision-based Fusion Framework:* The first multi-level fusion uses penultimate layer feature-based fusion (Sec. V-B) at first level, and decision-based fusion of the ultimate layers at second level. To define the decision-based fusion at the second level, we use the ultimate layers of unimodal and penultimate fusion networks: $\mathbf{s}_{\mathtt{R}}$, $\mathbf{s}_{\mathtt{A}}$, $\mathbf{s}_{\mathtt{S}}$, and $\mathbf{s}_{\mathtt{PNF}} \in \mathbb{R}^{|\mathcal{V}|}$, where $\mathbf{s}_{\mathtt{PNF}} \in (\mathbf{s}_{\mathtt{AGF}}, \mathbf{s}_{\mathtt{IRF}})$. Finally, the decision at second level is defined as: $\mathbf{s}_{\mathtt{D}} = \max(\mathbf{s}_{\mathtt{R}}, \mathbf{s}_{\mathtt{A}}, \mathbf{s}_{\mathtt{R}}, \mathbf{s}_{\mathtt{PNF}})$.

The graphical illustration of multi-level decision based fusion approach is presented in Fig. 15. The main idea behind using decision-based fusion at the second level is to selectively use either one of the proposed unimodal model or feature-based fusion model on the fly for each vehicle. Here, we use $\max$ operator as selection criteria among the ultimate layers. This concept can be explained with one example. Suppose, one vehicle is best detected by acoustic model, however another vehicle is best detected from the aggregated fusion network. This decision based fusion will dynamically select the output of acoustic model for first vehicle, and aggregated fusion network for the second vehicle. However, this idea works on the assumption that best performing model will generate the maximum value at the ultimate layer for each vehicle. This might not be true for all the cases, as it is possible that some model falsely generates a high valued output for a vehicle. As a result, the decision-based fusion may propagate model mispredictions from first level to the decision at the second level. Hence, we move forward adding fully connected layers at the second-level, as a measure of assigning weights to the output from the first-level.

*2) Multi-level Quadratic Shallow Fusion Framework:* In our second attempt of adding another level of fusion at the ultimate layer, we generate the quadratic features from the fused output of all the models of first level. We use the fusion strategies mentioned in Sec. IV-F for fusing the ultimate

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3145663, IEEE Transactions on Multimedia

layer outputs of each models. The relevance of each model of first-level is pronounced through the quadratic features and further dynamically assigned weights by the one fully connected output layer. The overall fusion approach is depicted in Fig. 16(a).

To define the quadratic-shallow fusion at the second level, we use the previously defined unimodal and fusion networks. The representation of combined matrix $\mathbf{s}_{\mathtt{SF}}$ is:

$$\mathbf{s}_{\mathtt{SF}} = \mathbf{q}(\mathcal{F}(\mathbf{s}_{\mathtt{R}}, \mathbf{s}_{\mathtt{A}}, \mathbf{s}_{\mathtt{S}}, \mathbf{s}_{\mathtt{PNF}})) \in \mathbb{R}^{|\mathbf{q}(4\times|\mathcal{V}|)|},$$

where $\mathbf{q}(\cdot)$ extracts the *quadratic features* from the ultimate layer at first level. The quadratic features represents the features generated by applying a quadratic function on the inputs. The quadratic function outputs all possible combinations of the input elements. The $\mathbf{q}(\cdot)$ on an input $[x_1, \cdots, x_n]$ is defined as: $\mathbf{q}(x_1, \cdots, x_n) = \{x_i x_j | 1 \leq i \leq n,\ 1 \leq j \leq n\}$, where $|\mathbf{q}(\cdot)| = n^2/2$. We denote the multi-level quadratic shallow fusion network as $f^{\mathtt{SF}}_{\theta^{\mathtt{SF}}}(\cdot)$, where:

$$\mathbf{s}_{\mathtt{SF}} = \sigma(f^{\mathtt{SF}}_{\theta^{\mathtt{SF}}}(\mathbf{z})), \qquad f^{\mathtt{SF}}_{\theta^{\mathtt{SF}}} : \mathbb{R}^{|\mathbf{q}(4\times|\mathcal{V}|)|} \mapsto \mathbb{R}^{|\mathcal{V}|},$$

and $\sigma$ is the sigmoid activation.

*3) Multi-level Deep Fusion Framework:* We also propose a deep neural network at the second-level of fusion, as shown in Fig. 16 (b). The deep fusion network on the output of ultimate layer will intelligently assigns higher weights to the outputs of the more relevant models of first-level. We use 4 fully connected layers as the deep fusion architecture of second-level. This network architecture is presented in Fig. 16 (b).

Similar to other multi-level fusion frameworks, the deep fusion at the second level can also be defined using the ultimate layers of unimodal and fusion networks. In this case, the representation of combined matrix $\mathbf{s}_{\mathtt{DF}}$ is:

$$\mathbf{s}_{\mathtt{DF}} = \mathcal{F}(\mathbf{s}_{\mathtt{R}}, \mathbf{s}_{\mathtt{A}}, \mathbf{s}_{\mathtt{S}}, \mathbf{s}_{\mathtt{PNF}}) \in \mathbb{R}^{4\times|\mathcal{V}|}.$$

We denote the multi-level deep fusion network as $f^{\mathtt{DF}}_{\theta^{\mathtt{DF}}}(\cdot)$:

$$\mathbf{s}_{\mathtt{DF}} = \sigma(f^{\mathtt{DF}}_{\theta^{\mathtt{DF}}}(\mathbf{z})), \qquad f^{\mathtt{DF}}_{\theta^{\mathtt{DF}}} : \mathbb{R}^{4\times|\mathcal{V}|} \mapsto \mathbb{R}^{|\mathcal{V}|}.$$

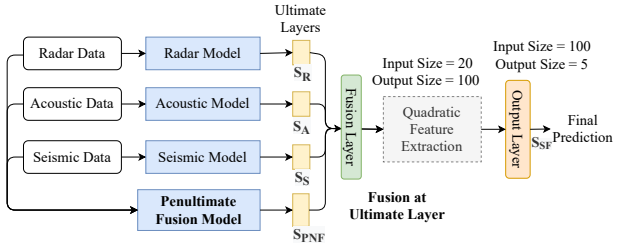
where $\sigma$ is the sigmoid function.

## VI. Proposed Fusion Frameworks to Augment Image Modality

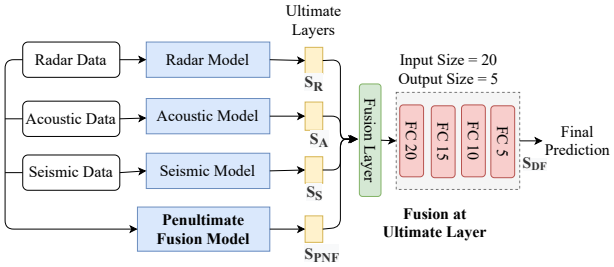
After exploring different types of fusion networks with non-image modalities, we now augment those with the image modality. The image model can be fused in multiple ways with the non-image unimodal/fused models. We present three such novel image fusion techniques where ultimate layer of unimodal image network is fused either in unified way or with one of the non-image fusion networks.

### A. Image Unified Fusion Framework

We first approach the integration of image modality in unified way as the other non-image modalities. Hence, the proposed architecture is designed to use either ultimate layers or extracted latent embeddings from all the modalities and combine those in an unified way. The fused features are later passed through the fusion networks similar to Fig. 12 and

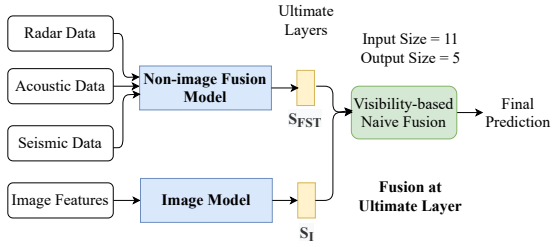



(a) Quadratic shallow fusion

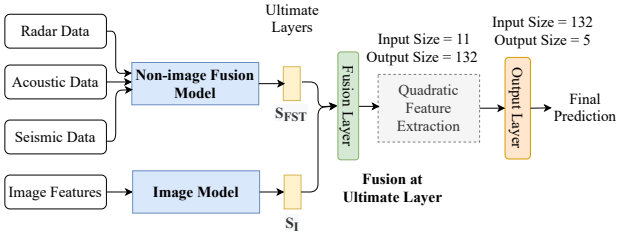



(b) Deep fusion

Fig. 16: Proposed multi-level fusion framework at ultimate layers. The preprocessing step is not shown in this block diagram to highlight the proposed fusion more.




(a) Image naïve fusion




(b) Image quadratic fusion

Fig. 17: Proposed image fusion frameworks at ultimate layers. The preprocessing step is not shown in this block diagram to highlight the proposed fusion more.

13, where image modality is also integrated to the end-to-end architecture, and one more fully connected layer is added to deal with the added dimensionality of the image feature than the 3 modality version.

*1) Fusion at Ultimate Layer:* To define the unified fusion at the ultimate layer, the scores of each modality: $\mathbf{s}_{\mathtt{R}}$, $\mathbf{s}_{\mathtt{A}}$, $\mathbf{s}_{\mathtt{S}}$, and $\mathbf{s}_{\mathtt{I}} \in \mathbb{R}^{|\mathcal{V}|}$, are combined to generate the representation matrix from $\mathcal{F}(\mathbf{s}_{\mathtt{R}}, \mathbf{s}_{\mathtt{A}}, \mathbf{s}_{\mathtt{S}}, \mathbf{s}_{\mathtt{I}}) \in \mathbb{R}^{4\times|\mathcal{V}|}$. We denote this unified fusion network by $f^{\mathtt{UUNF}}_{\theta^{\mathtt{UUNF}}}(\cdot)$:

$$\mathbf{s}_{\mathtt{UUNF}} = \sigma(f^{\mathtt{UUNF}}_{\theta^{\mathtt{UUNF}}}(\mathcal{F}(\mathbf{s}_{\mathtt{R}}, \mathbf{s}_{\mathtt{A}}, \mathbf{s}_{\mathtt{S}}, \mathbf{s}_{\mathtt{I}}))), \qquad f^{\mathtt{UUNF}}_{\theta^{\mathtt{UUNF}}} : \mathbb{R}^{4\times|\mathcal{V}|} \mapsto \mathbb{R}^{|\mathcal{V}|},$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3145663, IEEE Transactions on Multimedia

where $\sigma$ is the sigmoid activation function.

*2) Fusion at Penultimate Layer:* Given $\mathbf{z}_{\mathtt{R}} \in \mathbb{R}^{d^{\mathtt{R}}}$, $\mathbf{z}_{\mathtt{A}} \in \mathbb{R}^{d^{\mathtt{A}}}$, $\mathbf{z}_{\mathtt{S}} \in \mathbb{R}^{d^{\mathtt{S}}}$, and $\mathbf{z}_{\mathtt{I}} \in \mathbb{R}^{d^{\mathtt{I}}}$ we fuse them in unified way and generate the combined latent feature matrix $\mathbf{z}$ as:

$$\mathbf{z} = \mathcal{F}(\mathbf{z}_{\mathtt{R}}, \mathbf{z}_{\mathtt{A}}, \mathbf{z}_{\mathtt{S}}, \mathbf{z}_{\mathtt{I}}) \in \mathbb{R}^{d^{\mathtt{R}}+d^{\mathtt{A}}+d^{\mathtt{S}}+d^{\mathtt{I}}}.$$

We denote the unified fusion at penultimate layer as $f^{\mathtt{PUNF}}_{\theta^{\mathtt{PUNF}}}(\cdot)$:

$$\mathbf{s}_{\mathtt{PUNF}} = \sigma(f^{\mathtt{PUNF}}_{\theta^{\mathtt{PUNF}}}(\mathbf{z})), \qquad f^{\mathtt{PUNF}}_{\theta^{\mathtt{PUNF}}} : \mathbb{R}^{d^{\mathtt{R}}+d^{\mathtt{A}}+d^{\mathtt{S}}+d^{\mathtt{I}}} \mapsto \mathbb{R}^{|\mathcal{V}|},$$

where $\sigma$ is the sigmoid. The unified way of fusing all four modalities is restricted by assigning equal importance to all four modalities before fusion. However, we want to enforce higher importance to the image modalities than the non-image ones, hence we propose two more fusion frameworks which take images separately than the other non-image modalities, and eventually gives higher importance to the image modality.

### B. Image Naïve Fusion Framework

In our experimental validation, we notice that if vehicles are visible, the image modality provides accurate prediction. Recall that besides 5 types of vehicles, we define a 'non-visible' class. We take advantage of the accuracy of image through this 'non-visible' class, and propose a image fusion framework shown in Fig. 17 (a). If image feature predicts that vehicles are visible, then we use image prediction as final prediction. Otherwise, if image feature predicts that vehicles are not visible, we use prediction from other modalities as final prediction. Hence, we call this specific fusion as naïve image fusion. This fusion works with the principle of *decision-based* fusion, as we design the logic of taking the decision of using the prediction from either image model or non-image fusion model.

To define this fusion, we use the ultimate layers of unimodal image and earlier described non-image fusion (Sec. V-B and V-C) networks: $\mathbf{s}_{\mathtt{I}}$, $\mathbf{s}_{\mathtt{FST}}$, where $\mathbf{s}_{\mathtt{FST}} \in (\mathbf{s}_{\mathtt{PNF}}, \mathbf{s}_{\mathtt{DF}}, \mathbf{s}_{\mathtt{SF}})$. Finally, this fusion is defined as:

$$\mathbf{s}_{\mathtt{INF}} = \begin{cases} \mathbf{s}_{\mathtt{I}}, & \text{if } \mathtt{nonvisible}(\mathbf{s}_{\mathtt{I}}) = 0, \\ \mathbf{s}_{\mathtt{FST}}, & \text{otherwise.} \end{cases}$$

### C. Image Quadratic Fusion Framework

We further propose a trainable approach for image fusion, shown in Fig. 17 (b). To explore the relationship between each output before fusing with images, we generate the quadratic features from both non-image and image modalities. The quadratic features are defined in the same way as in Sec. V-C2. This quadratic features include not only each output (dimension $2 \times |\mathcal{V}| + 1$), but also the multiplication of each output (dimension $\mathbf{q}(2 \times |\mathcal{V}| + 1)$, which is 121 considering $|\mathcal{V}|$ = 5). After that, we employ a fully connected layer to map our result into dimension $|\mathcal{V}|$.

The representation of combined matrix $\mathbf{s}_{\mathtt{IQF}}$ is:

$$\mathbf{s}_{\mathtt{IQF}} = \mathbf{q}(\mathcal{F}(\mathbf{s}_{\mathtt{I}}, \mathbf{s}_{\mathtt{FST}})) \in \mathbb{R}^{|\mathbf{q}(2\times|\mathcal{V}|+1)|}.$$

The $\mathbf{q}(\cdot)$ extracts the quadratic features from the ultimate layer of image model and one of the non-image fusion model.

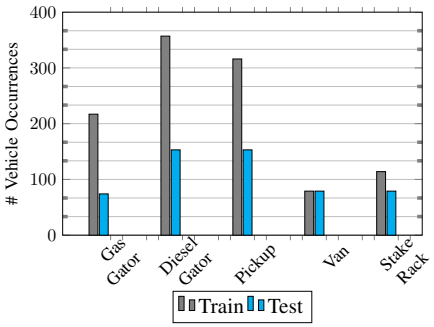



Fig. 18: Vehicle occurrences in the train and test set for scenario-based split.

We denote the multi-level quadratic image fusion network as $f^{\mathtt{IQF}}_{\theta^{\mathtt{IQF}}}(\cdot)$, so that:

$$\mathbf{s}_{\mathtt{IQF}} = \sigma(f^{\mathtt{IQF}}_{\theta^{\mathtt{IQF}}}(\mathbf{z})), \qquad f^{\mathtt{IQF}}_{\theta^{\mathtt{IQF}}} : \mathbb{R}^{|\mathbf{q}(2\times|\mathcal{V}|+1)|} \mapsto \mathbb{R}^{|\mathcal{V}|},$$

where $\sigma$ is the sigmoid activation function.

## VII. Experiments

In this section, we discuss about the setup of different experiments and evaluation metrics for analyzing the performance of the proposed unimodal and fusion networks.

### A. Experimental Setup

We split the dataset of 13 scenarios into 958 non-overlapped 1-second bins. We consider two types of train-test splits: *bin-based* and *scenario-based* splits. In *bin-based* split: we randomly select $80\%$, and $20\%$ of bins for each scenario into training and testing sets, respectively, and combine all scenarios to form the training and testing sets for the entire dataset. In *scenario-based* split: we split 13 scenarios into 10 training scenarios and 3 testing scenarios. The distribution of different vehicles in test and train set for scenario-based split is shown in Fig. 18. Scenario-based splitting uses data trained at different runs to predict performance at other runs, while bin-based allows testing on data across all scenarios.

For all modalities except RGB imagery, we consider the labels of 1-second bins directly from the existing vehicles in the scenario. For the image modality, we consider two types labels: true labels and visible labels as described in Sec. IV-D. For each frame/image in a scenario, true labels represent the vehicles in this scenario, even though the vehicles cannot be observed in the frame; while the visible labels only indicates the observable vehicles in the frame. Besides, we use a 'non-visible' class to indicate no vehicles situation. Scenario-wise 'non-visible' vs 'visible' labels has been depicted in Fig. 19.

The details of used optimizers and learning rates for training of the proposed models are presented in Tab. IV.

### B. Evaluation Metrics

For each 1-second bin, we binarize the labels to existence vs non-existence for each type of vehicle. We use Area under the receiver operating characteristic (ROC) curve (AUC), and average precision (AP) as our evaluation metrics for vehicle

| Proposed Models | Learning Rates | Optimizer |
|---|---|---|
| Radar Unimodal Network | $10^{-4}$ | Adam |
| Acoustic Unimodal Network | $10^{-4}$ | Adam |
| Seismic Unimodal Network | $10^{-3}$ | SGD |
| Image Unimodal Network | $10^{-4}$ | Adam |
| Non-image Fusion Networks | $10^{-4}$ | Adam |
| Image Fusion Networks | $10^{-2}$ | Adam |

TABLE IV: Optimizers and learning rates used for training different proposed models.

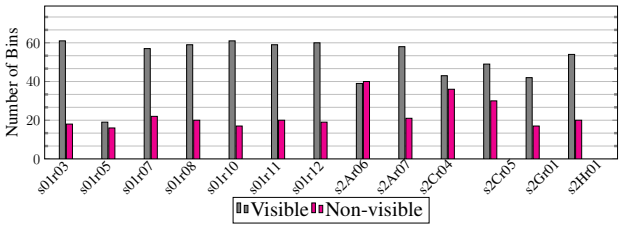



Fig. 19: Scenario-wise 'visible' and 'non-visible' labels per bin. The 'non-visible' bin refers to the fact that no vehicle was visible from image for that bin, even though the vehicles were present as per the non-image modalities.

classification. The ROC curve is a performance measurement for the classification problems at various threshold settings. The AUC tells how much the model is capable of distinguishing between classes. On the other hand, AP represents the area under the precision-recall curve. Higher the AUC and AP are, the better the classification model is.

Additionally, the considered metrics to evaluate the object detection task of the image preprocessing step are mean average precision (mAP) and mean average recall (mAR).

## VIII. Performance Analysis

In this section, we provide discussion regarding the performance of different unimodal and fusion networks presented in Sec. IV-D, V, and VI. We use *python* and *pytorch* to implement the proposed models.

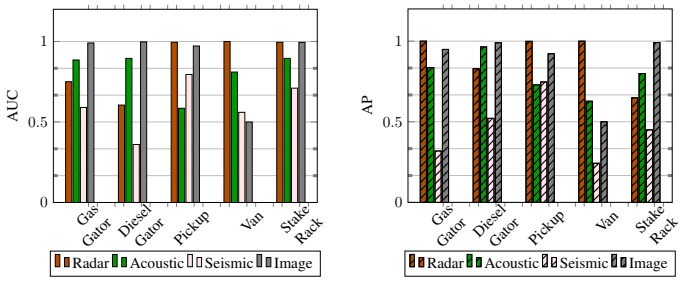
### A. Performance of Proposed Unimodal Networks

The average AUCs and APs of all the vehicles for different settings of the unimodal networks are presented in Tab. V. As mentioned in Sec. IV-D, we propose one MLP and one CNN based models for radar, and specific models for other modalities. We also explore the impacts for using the normalized or unnormalized inputs. Since the normalized features gave better performance than the unnormalized ones, we use normalized features only, for our experiments. It is also evident from the results (presented in Tab. V) that VGGSound outperforms VGG16 as it is pre-trained on different types of acoustic data. However, VGG16 gives better performance when fused with other modalities. From our preliminary experiments, we observed that the acoustic model with pre-trained VGG16 gives 20-30% better performance in terms of both evaluation metrics (AUC and AP) when fused with other modalities. Hence, we consider VGG16 as the best performing acousitc unimodal network for fusion.

The vehicle-wise AUC and AP comparison for the best settings of each unimodal network is presented in Fig. 20. We use the best settings (highlighted in bold font in Tab V) of each unimodal model for all the fusion networks in the rest of the performance evaluations. Finally, the image modality is trained with visible labels. We evaluate the result through both true label and visible label. In true labels, although vehicles are not observable, they provide positive labels. Thus, evaluation performance with respect to true labels is significantly poorer than results with visible labels.

| Settings | Average AUC | | Average AP | |
|---|---|---|---|---|
| | sc-based | bin-based | sc-based | bin-based |
| Radar-MLP | 0.766 | 0.712 | 0.818 | 0.724 |
| **Radar-CNN** | 0.742 | 0.996 | 0.791 | 0.999 |
| **Acoustic (VGG16)** | 0.752 | 0.876 | 0.726 | 0.856 |
| Acoustic (VGGSound) | 0.776 | 0.906 | 0.823 | 0.817 |
| **Seismic** | 0.564 | 0.642 | 0.385 | 0.526 |
| Image (true labels) | 0.698 | 0.703 | 0.757 | 0.649 |
| **Image (visible labels)** | 0.986 | 0.796 | 0.956 | 0.823 |

TABLE V: Area Under the ROC (AUC) and Average Precision (AP) for each modality over the ESCAPE dataset (higher is better, 1 is best). We use the best settings (highlighted in bold font) of each unimodal model for the fusion networks for the rest of the performance evaluations.




(a) AUC (b) AP

Fig. 20: Vehicle-wise AUCs and APs for best settings of proposed unimodal networks. The evaluation metrics are averaged over scenario-based and bin-based settings for each vehicle.

| Metric | Gas gator | Diesel gator | Pickup | Van | Stake rack |
|---|---|---|---|---|---|
| mAP | 0.736 | 0.828 | 0.836 | 0.531 | 0.837 |
| mAR | 0.799 | 0.869 | 0.870 | 0.533 | 0.864 |

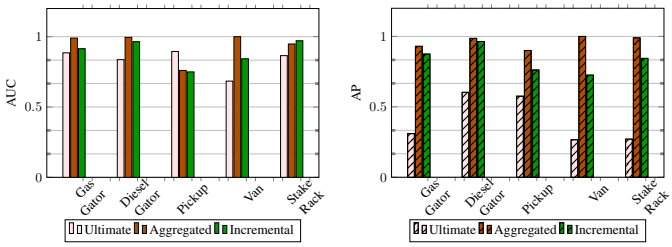
TABLE VI: The mAP and mAR for IoU = (0.50:0.05:0.95) in object detection. We train and evaluate the object detection only in *bin-based* settings.

*1) Object Detection Performance for Image Modality:* To quantify the performance of the object detection step of finetuned Faster RCNN, presented in Sec. IV-D1, the corresponding evaluation metrics are presented in Tab. VI. Here, the mean Average Precision (mAP) corresponds to the average precision for Intersection Over Union (IoU) from 0.5 to 0.95 with a step size of 0.05. Similarly, mean Average Recall (mAR) corresponds to the average recall for the IoU of same settings. Note that as object detection is considered as a preprocessing step in our proposed approach, we only train and evaluate our object detection model in *bin-based* settings.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3145663, IEEE Transactions on Multimedia

| Settings | Average AUC | | Average AP | |
|---|---|---|---|---|
| | sc-based | bin-based | sc-based | bin-based |
| Ultimate Layer Fusion (Shallow), $S_1$ | 0.531 | 0.916 | 0.664 | 0.657 |
| Ultimate Layer Fusion (Shallow), $S_2$ | 0.425 | 0.525 | 0.527 | 0.401 |
| Ultimate Layer Fusion (Shallow), $S_3$ | 0.674 | 0.646 | 0.611 | 0.597 |
| **Ultimate Layer Fusion (Deep),** $S_1$ | 0.767 | 0.898 | 0.462 | 0.348 |
| Ultimate Layer Fusion (Deep), $S_2$ | 0.446 | 0.509 | 0.584 | 0.357 |
| Ultimate Layer Fusion (Deep), $S_3$ | 0.661 | 0.877 | 0.459 | 0.342 |
| **Aggregated Penultimate Layer Fusion,** $S_1$ | 0.886 | 1 | 0.924 | 0.998 |
| Aggregated Penultimate Layer Fusion, $S_2$ | 0.788 | 0.887 | 0.728 | 0.824 |
| Aggregated Penultimate Layer Fusion, $S_3$ | 0.757 | 0.923 | 0.684 | 0.851 |
| Incremental Ultimate Layer Fusion | 0.544 | 0.517 | 0.444 | 0.382 |
| **Incremental Penultimate Layer Fusion** | 0.841 | 0.936 | 0.805 | 0.863 |

TABLE VII: Average AUCs and APs for proposed feature-based fusion networks over the ESCAPE dataset (higher is better, 1 is best). Three settings are used: (i) $S_1$ means training the whole model from scratch; (ii) $S_2$ means loading the weights from pre-trained unimodal networks and training the fusion layers only; (iii) $S_3$ means loading the weights from pre-trained unimodal networks and retraining the whole model. Each column specify same meaning as of the earlier table.




(a) AUC (b) AP

Fig. 21: Vehicle-wise AUCs and APs for best settings of proposed fusion networks on non-image modalities. The evaluation metrics are averaged over scenario-based and bin-based settings for each vehicle.

### *B. Performance of Ultimate and Penultimate Layer Fusion on Non-image Modalities*

As presented in Sec. V-A and V-B, we first evaluate the performance of feature-based fusion at ultimate and penultimate layer on non-image modalities. The average AUCs and APs for all the vehicles for fusion at ultimate, penultimate and successive (incremental) stages with different settings are presented in Tab. VII. Vehicle-wise AUC and AP comparison for the best settings (highlighted in bold font in Tab. VII) of each fusion network is presented in Fig. 21. From the results it is clear that aggregated penultimate layer fusion, while training from scratch, gives the best performance, hence we use this model as the penultimate fusion network for the multi-level fusion, presented in Sec. V-C. Next, we analyze the performance of different fusion strategies.

### *C. Performance of Different Fusion Strategies*

In this set of experiments, we perform different fusion strategies presented in Sec. IV-F on the fusion networks considered in Sec. VIII-B. Without loss of generality, we choose to present the best performing one from Tab. VII, which is the aggregated fusion at penultimate layer trained from scratch. However, we observe a similar trend in the performance for other fusion frameworks as well. The extensive results on the AUC evaluation metric over different used strategies are presented in Tab. VIII. We observe that results vary in wide range for both *bin-based* and *scenario-based* evaluations. Even though these state-of-the-art fusion strategies dynamically extract different learnable parameters from the features, we observe the concatenation or late fusion outperforms the rest for the aggregated fusion framework. We observe the similar performance for other proposed fusion frameworks as well. Hence, we choose concatenation as the selected fusion strategy in the proposed fusion frameworks for the rest of our experiments.

### *D. Performance of Proposed Multi-level Fusion Frameworks on Non-image Modalities*

Once we derive the best performing feature-based fusion model and best fusion strategy, we next evaluate the proposed multi-level fusion frameworks (discussed in Sec. V-C). The average AUCs for all the vehicles for multi-level decision based, quadratic shallow, and deep fusion frameworks are presented in Tab. IX. The best performing multi-level fusion for each type is highlighted in bold font in Tab. IX.

Note that both evaluation metrics follow the same trend when averaged over vehicles, as presented in Tab. V, VII, and IX. Hence, we focus on one evaluation metric, AUCs, for the detailed vehicle-wise results for the remainder of the performance evaluation.

### *E. Performance of Fusion with Image Modality*

Next we move forward analyzing the performance of two types image fusion with non-image fusion networks, introduced in Sec. VI. In the first set of experiments involving images, we uniformly fuse all four modalities in different ways of unified fusion (Sec. VI-A). We present the results in Tab. X. Next, for the naïve (Sec. VI-B) and quadratic (Sec. VI-C) image fusion frameworks, 4 different non-image fusion frameworks with top performance are taken into account, listed as: (a) aggregated penultimate layer fusion (trained from scratch), (b) multi-level deep fusion (baseline) (c) multi-level deep fusion (4 layers), and (d) multi-level quadratic shallow fusion. In general, quadratic image fusion framework achieves better classification performance than naïve image fusion. This can be justified by the fact that image naïve fusion is based on manual and experiential selection, whereas in quadratic image fusion, the selection is learnt by hidden layer. Image naïve fusion with non-image multi-level deep fusion (4 layers), and image quadratic fusion with non-image multi-level quadratic shallow fusion achieve the best average performance. The detailed results of all the best performing unimodal, non-image fusion, and image fusion networks are presented in Tab. X.

### *F. Dealing with 'non-visible' Vehicles*

After carefully observing each frames of the scenarios reported in Tab. I, we present a visual representation of the scenario-wise 'visible' and 'non-visible' bins in Fig. 19. For the scenario-based setting, we have 3 scenarios in the test set, where there are no visible vehicles in 69 bins out of the total 163 bins. In the bin-based setting, there are 113 bins without

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3145663, IEEE Transactions on Multimedia

| Fusion Strategies | Gas gator | | Diesel gator | | Pickup | | Van | | Stake rack | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | sc-based | bin-based | sc-based | bin-based | sc-based | bin-based | sc-based | bin-based | sc-based | bin-based | sc-based | bin-based |
| **Concatenation** | 0.940 | 1 | 0.970 | 1 | 0.540 | 1 | 1 | 1 | 0.980 | 1 | 0.886 | 1 |
| Low Rank Tensor Fusion [57] | 0.980 | 0.430 | 0.520 | 0.570 | 0.500 | 0.450 | 0.920 | 0.580 | 0.720 | 0.550 | 0.493 | 0.518 |
| Feature-wise Linear Modulation [56] | 0.908 | 0.996 | 0.965 | 1 | 0.825 | 0.998 | 1 | 1 | 0.919 | 0.999 | 0.876 | 0.999 |
| Multiplicative Interaction [55] | 0.540 | 0.500 | 0.460 | 0.480 | 0.550 | 0.570 | 0.550 | 0.510 | 0.420 | 0.500 | 0.510 | 0.519 |

TABLE VIII: AUCs for different fusion strategies for aggregated fusion framework over the ESCAPE dataset (higher is better, 1 is the best). Vehicle-wise detection performance of the best performing fusion strategy is highlighted.

| Settings | Average AUC | | Average AP | |
|---|---|---|---|---|
| | sc-based | bin-based | sc-based | bin-based |
| Multi-level Decision-based Fusion | 0.860 | 0.998 | 0.811 | 0.999 |
| Multi-level Quadratic Shallow Fusion | 0.870 | 0.974 | 0.883 | 0.980 |
| **Multi-level Deep Fusion (baseline)** | 0.906 | 0.988 | 0.949 | 0.997 |
| Multi-level Deep Fusion (4 layers) | 0.900 | 0.976 | 0.727 | 0.989 |

TABLE IX: Average AUCs and APs for proposed multi-level fusion frameworks over the ESCAPE dataset (higher is better, 1 is best). The baseline of multi-level deep fusion refers to using a single dense layer (than the 4 layers) at the second level of fusion.

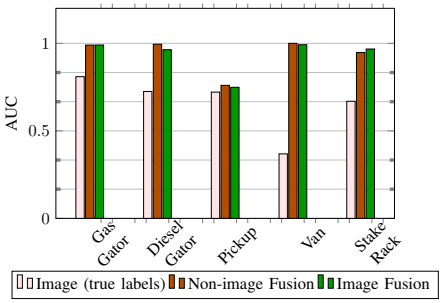



Fig. 22: Vehicle-wise AUCs for best settings of proposed fusion networks on image and non-image modalities. The AUCs are averaged over scenario-based and bin-based settings for each vehicle.

visible vehicles out of the total 370 bins in the test set. This proves that in 42.33% and 30.54% of the times, in the scenario-based and bin-based settings, respectively, all the vehicles are in NLOS. We plot the average AUCs (over all the vehicles) achieved for image modality and other fusion techniques in Fig. 22, to show that the NLOS situations result in lower AUCs for image modality, which later got improved by fusion with non-image modalities. In Fig. 22, the 'non-image fusion' bar represents the results from the best non-image fusion framework which is the multi-level deep fusion (baseline), and the 'image fusion' bar represents the top-performing image fusion framework which is the image quadratic fusion, from Tab. X. Here, we observe a 33.16% improvement in average AUC (vehicle and settings wise) after fusion with non-image modalities than when we use only the image modality with true labels. It is to be noted here that the visible labeling was one of the steps of preprocessing, hence, we compare our fusion result with ground truth of true labeling of image modality.

Thus, we establish that when non-image modalities are combined with the image modality, multi-vehicle detection can improve in NLOS situations, where image modality fails to detect the vehicle. Through prudent exploitation of non-image modalities (such as acoustic, radar, seismic, etc.), it is possible to detect the vehicle even for NLOS situations.

### *G. Performance of Proposed Fusion Frameworks on Other Multimodal Datasets*

Finally, we use existing multimodal datasets to evaluate and compare the performance of our proposed fusion frameworks with the current state-of-the-art. Since our proposed work encompasses the detection and classification of vehicles, we consider the available multimodal datasets for autonomous driving, such as NuScene [12], KITTI [58], [59], and OLIMP [60], for this evaluation. Among these datasets, we choose NuScene as the most suitable candidate for the performance comparison, due to the availability of image and radar data, which matches with the used sensor modalities in our proposed fusion frameworks. Additionally, the fact that NuScene has multiple vehicles present in some of the frames, further reinforces our decision. We use 850 scenes, and 2 sensor modalities (image and radar) of the NuScene data, having 4 vehicles (construction vehicles, motorcycle, trailer, and truck) for detection and classification. Without the loss of generality, we choose only the front camera and front radar sensor data in NuScene for this evaluation purpose, the results of which are presented in Tab. XI.

Note that NuScene is mainly curated for detection, hence the state-of-the-art on this dataset focuses on detection performance, e.g., [61], which uses 6 different sensors per modality. In contrast, our proposed method uses only 1 camera and 1 radar sensor, making it more data efficient. For a fair comparison, we implement existing fusion strategies (low rank multimodal fusion [57] and multiplicative interactions [55]), and compare with the proposed aggregated and incremental fusion frameworks, for classification task on this dataset. We use the evaluation metrics mentioned in Sec. VII-B.

From Tab. XI, we observe that the proposed aggregated fusion framework outperforms the AUC and AP scores of unimodal image network by 9.39% and 7.66%. These improvements have significant impact on various safety critical issues involving vehicle classification. Also, our proposed fusion framework shows a performance improvement of 22.73% in AUC and 18.06% in AP, and 21.93% in AUC and 15.18% in AP over the low rank multimodal fusion [57] and multiplicative interaction [55] fusion techniques, respectively, for vehicle classification.

## IX. Conclusions

We introduced different types of fusion frameworks to broadly leverage intrinsic situational information in different types of data modalities. These deep-learning based fusion frameworks intelligently assign variable weights per modality at different level of fusion. We evaluate our proposed model on real world multimodal ESCAPE dataset and NuScene dataset.

| Frameworks | Gas gator | | Diesel gator | | Pickup | | Van | | Stake rack | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | sc-based | bin-based | sc-based | bin-based | sc-based | bin-based | sc-based | bin-based | sc-based | bin-based | sc-based | bin-based |
| Radar | 0.500 | 1 | 0.210 | 1 | 1 | 0.990 | 1 | 1 | 1 | 0.990 | 0.742 | 0.996 |
| Acoustic (VGG16) | 0.930 | 0.840 | 0.880 | 0.910 | 0.360 | 0.810 | 0.750 | 0.870 | 0.840 | 0.950 | 0.752 | 0.876 |
| Seismic | 0.470 | 0.710 | 0.240 | 0.480 | 0.840 | 0.750 | 0.510 | 0.610 | 0.760 | 0.660 | 0.564 | 0.642 |
| Image (visible labels) | 0.995 | 0.987 | 0.995 | 1 | 0.944 | 1 | 1 | 0 | 0.996 | 0.993 | 0.986 | 0.796 |
| Image (true labels) | 0.887 | 0.732 | 0.718 | 0.732 | 0.695 | 0.748 | 0.296 | 0.441 | 0.897 | 0.861 | 0.697 | 0.703 |
| Aggregated Fusion | 0.940 | 1 | 0.970 | 1 | 0.540 | 1 | 1 | 1 | 0.980 | 1 | 0.886 | 1 |
| Multi-level Decision-based Fusion | 0.980 | 1 | 1 | 0.990 | 0.600 | 1 | 0.830 | 1 | 0.890 | 1 | 0.860 | 0.998 |
| Multi-level Quadratic Shallow Fusion | 1 | 0.990 | 1 | 0.970 | 0.550 | 0.930 | 0.970 | 0.990 | 0.830 | 0.990 | 0.870 | 0.974 |
| **Multi-level Deep Fusion (baseline)** | 1 | 0.970 | 0.980 | 0.990 | 0.560 | 0.990 | 1 | 0.990 | 0.990 | 1 | 0.906 | 0.988 |
| Multi-level Deep Fusion (4 layers) | 0.990 | 0.990 | 1 | 0.960 | 0.510 | 0.970 | 1 | 0.970 | 1 | 0.990 | 0.900 | 0.976 |
| Image Unified Fusion (ultimate layer-shallow) | 0.891 | 0.930 | 0.939 | 0.989 | 0.410 | 0.869 | 0.814 | 0.897 | 0.930 | 0.957 | 0.797 | 0.929 |
| Image Unified Fusion (ultimate layer-deep) | 0.605 | 0.893 | 0.975 | 0.949 | 0.511 | 0.912 | 0.535 | 0.874 | 0.984 | 0.999 | 0.722 | 0.926 |
| Image Unified Fusion (penultimate layer) | 0.953 | 0.990 | 0.964 | 0.996 | 0.527 | 0.999 | 1 | 1 | 0.993 | 0.997 | 0.888 | 0.997 |
| Naïve Fusion: Image and Aggregated Fusion | 0.959 | 0.900 | 0.958 | 0.891 | 0.817 | 0.936 | 0.584 | 0.566 | 0.990 | 0.996 | 0.862 | 0.858 |
| Naïve Fusion: Image and Deep Fusion (baseline) | 0.995 | 0.892 | 0.972 | 0.879 | 0.813 | 0.928 | 0.581 | 0.561 | 0.995 | 0.982 | 0.871 | 0.848 |
| **Naïve Fusion: Image and Deep Fusion (4 layers)** | 0.995 | 0.896 | 0.995 | 0.889 | 0.806 | 0.920 | 0.584 | 0.565 | 0.996 | 0.993 | 0.875 | 0.853 |
| Naïve Fusion: Image and Quadratic Shallow Fusion | 0.988 | 0.895 | 0.952 | 0.889 | 0.815 | 0.912 | 0.576 | 0.566 | 0.967 | 0.987 | 0.859 | 0.850 |
| Quadratic Fusion: Image and Aggregated Fusion | 0.902 | 0.996 | 0.896 | 1 | 0.641 | 1 | 0.916 | 1 | 0.966 | 0.999 | 0.864 | 0.999 |
| Quadratic Fusion: Image and Deep Fusion (baseline) | 1 | 0.980 | 0.955 | 0.961 | 0.529 | 0.989 | 0.706 | 0.934 | 0.991 | 0.993 | 0.836 | 0.971 |
| Quadratic Fusion: Image and Deep Fusion (4 layers) | 0.989 | 0.996 | 1 | 0.973 | 0.495 | 0.992 | 0.828 | 0.968 | 0.991 | 0.986 | 0.861 | 0.983 |
| **Quadratic Fusion: Image and Quadratic Shallow Fusion** | 0.988 | 0.993 | 0.941 | 0.986 | 0.589 | 0.999 | 0.987 | 0.997 | 0.961 | 0.974 | 0.893 | 0.999 |

TABLE X: AUCs for each best performing unimodal, non-image, and image fusion networks over the ESCAPE dataset (higher is better, 1 is best). Vehicle-wise detection performance of only using image true labels and final fusion network are highlighted.

| Frameworks | Construction Vehicles | | Motor Cycle | | Trailer | | Truck | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | AP | AUC | AP | AUC | AP | AUC | AP | AUC | AP |
| Radar | 0.400 | 0.314 | 0.766 | 0.637 | 0.380 | 0.224 | 0.814 | 0.951 | 0.590 | 0.531 |
| Image (true label) | 0.624 | 0.408 | 0.637 | 0.550 | 0.679 | 0.539 | 0.785 | 0.954 | 0.681 | 0.613 |
| Low Rank Tensor Fusion [57] | 0.547 | 0.291 | 0.669 | 0.542 | 0.664 | 0.509 | 0.548 | 0.894 | 0.607 | 0.559 |
| Multiplicative Interactions [55] | 0.564 | 0.348 | 0.648 | 0.507 | 0.771 | 0.554 | 0.519 | 0.884 | 0.611 | 0.573 |
| **Aggregated Fusion*** | 0.689 | 0.478 | 0.746 | 0.627 | 0.768 | 0.575 | 0.780 | 0.960 | 0.745 | 0.660 |
| Incremental Fusion* | 0.663 | 0.450 | 0.739 | 0.605 | 0.764 | 0.584 | 0.775 | 0.957 | 0.735 | 0.649 |

TABLE XI: Comparison of AP and AUC for subset of NuScene dataset for radar and image modality with competing methods. The methods marked with '*' are proposed in this paper with the concatenation as fusion strategy. The best performing one is highlighted.

The results show that the use of fusion improves in vehicle detection and identification performance than using the single modalities, especially in the limited visibility states. Additionally, the proposed fusion frameworks also outperform the state-of-the-art competitors. Moving forward, vehicle detection can be further enhanced by designing fusion frameworks which exploits the orthogonality within the single modalities.

## ACKNOWLEDGMENTS

This work was partly supported by the Air Force Research Laboratory, and the US National Science Foundation grant CNS-2112471. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the United States Air Force.

## REFERENCES

[1] A. Petrovskaya and S. Thrun, "Model based vehicle detection and tracking for autonomous urban driving," *Autonomous Robots*, vol. 26, p. 123–139, 2009.

[2] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pp. 1–7, 2019.

[3] PwC, "Sensing the future of the Internet of Things," 2014.

[4] F. De Ponte Müller, "Survey on ranging sensors and cooperative techniques for relative positioning of vehicles," *Sensors*, vol. 17, no. 2, 2017.

[5] S. Chadwick, W. Maddern, and P. Newman, "Distant vehicle detection using radar and vision," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8311–8317, 2019.

[6] T. Stanislas, LeoandPeynot, "Characterisation of the delphi electronically scanning radar for robotics applications," *Proceedings of the Australasian Conference on Robotics and Automation*, pp. 1–10, 2015.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, pp. 1097–1105, 2012.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[11] P. Zulch, M. Distasio, T. Cushman, B. Wilson, B. Hart, and E. Blasch, "ESCAPE Data Collection for Multi-Modal Data Fusion Research," in *2019 IEEE Aerospace Conference*, pp. 1–10, 2019.

[12] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[13] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 2018.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," 2016.

[15] Y. Fu, L. Zhang, J. Wang, Y. Fu, and Y.-G. Jiang, "Depth guided adaptive meta-fusion network for few-shot video recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, p. 1142–1151, 2020.

[16] W. Zhao, H. Lu, and D. Wang, "Multisensor image fusion and enhancement in spectral total variation domain," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 866–879, 2018.

[17] D. Semedo and J. a. Magalhães, "Adaptive temporal triplet-loss for cross-modal embedding learning," in *Proceedings of the 28th ACM International Conference on Multimedia*, p. 1152–1161, 2020.

[18] Q. Lin, B. Yan, J. Li, and W. Tan, "Mmfl: Multimodal fusion learning for text-guided image inpainting," in *Proceedings of the 28th ACM International Conference on Multimedia*, p. 1094–1102, 2020.

[19] Y. Yi, H. Wang, and Q. Li, "Affective video content analysis with adaptive fusion recurrent network," *IEEE Transactions on Multimedia*, vol. 22, no. 9, pp. 2454–2466, 2020.

[20] X. Zhang, X. Gao, W. Lu, L. He, and J. Li, "Beyond vision: A multimodal recurrent attention convolutional neural network for unified image aesthetic prediction tasks," *IEEE Transactions on Multimedia*, vol. 23, pp. 611–623, 2021.

[21] B. Zhang, L. Li, L. Su, S. Wang, J. Deng, Z.-J. Zha, and Q. Huang, "Structural semantic adversarial active learning for image captioning," in *Proceedings of the 28th ACM International Conference on Multimedia*, p. 1112–1121, 2020.

[22] Y. Yuan, L. Ma, J. Wang, and W. Zhu, "Controllable video captioning with an exemplar sentence," in *Proceedings of the 28th ACM International Conference on Multimedia*, p. 1085–1093, 2020.

[23] Y. Liu, W. Zhou, M. Xi, S. Shen, and H. Li, "Vision meets wireless positioning: Effective person re-identification with recurrent context propagation," in *Proceedings of the 28th ACM International Conference on Multimedia*, p. 1103–1111, 2020.

[24] D. Lahat, T. Adalý, and C. Jutten, "Challenges in multimodal data fusion," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, pp. 101–105, 2014.

[25] D. Lahat, T. Adalı, and C. Jutten, "Multimodal Data Fusion: An Overview of Methods, Challenges and Prospects," *Proceedings of the IEEE*, vol. 103, pp. 1449–1477, Aug. 2015.

[26] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2021.

[27] K. Kim, C. Lee, D. Pae, and M. Lim, "Sensor fusion for vehicle tracking with camera and radar sensor," in *2017 17th International Conference on Control, Automation and Systems (ICCAS)*, pp. 1075–1077, 2017.

[28] H. Jha, V. Lodhi, and D. Chakravarty, "Object detection and identification using vision and radar data fusion system for ground-based navigation," in *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 590–593, 2019.

[29] V. Lekic and Z. Babic, "Automotive radar and camera fusion using generative adversarial networks," *Elsevier Computer Vision and Image Understanding*, vol. 184, pp. 1–8, 2019.

[30] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[31] R. O. Chavez-Garcia and O. Aycard, "Multiple sensor fusion and classification for moving object detection and tracking," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 525–534, 2016.

[32] D. Shen, P. Zulch, M. Disasio, E. Blasch, G. Chen, Z. Wang, J. Lu, and R. Niu, "Manifold learning algorithms for sensor fusion of image and radio-frequency data," in *2018 IEEE Aerospace Conference*, pp. 1–9, 2018.

[33] D. Shen, E. Blasch, P. Zulch, M. Distasio, R. Niu, J. Lu, Z. Wang, and G. Chen, "A joint manifold leaning-based framework for heterogeneous upstream data fusion," *Journal of Algorithms & Computational Technology*, vol. 12, no. 4, pp. 311–332, 2018.

[34] D. Garagić, J. Peskoe, F. Liu, M. S. Claffey, P. Bendich, J. Hineman, N. Borggren, J. Harer, P. Zulch, and B. J. Rhodes, "Upstream fusion of multiple sensing modalities using machine learning and topological analysis: An initial exploration," in *2018 IEEE Aerospace Conference*, pp. 1–8, 2018.

[35] E. Blasch, A. Vakil, J. Li, and R. Ewing, "Multimodal Data Fusion Using Canonical Variates Analysis Confusion Matrix Fusion," in *IEEE Aerospace Conference*, pp. 1–10, 2021.

[36] R. Niu, P. Zulch, M. Distasio, E. Blasch, G. Chen, D. Shen, Z. Wang, and J. Lu, "Joint sparsity based heterogeneous data-level fusion for multitarget discovery," in *2018 IEEE Aerospace Conference*, pp. 1–8, 2018.

[37] Q. Pan, J. Wei, H. Cao, N. Li, and H. Liu, "Improved ds acoustic–seismic modality fusion for ground-moving target classification in wireless sensor networks," *Pattern Recognition Letters*, vol. 28, no. 16, pp. 2419–2426, 2007.

[38] A. Vakil, J. Liu, P. Zulch, E. Blasch, R. Ewing, and J. Li, "Feature Level Sensor Fusion for Passive RF and EO Information Integration," in *IEEE Aerospace Conference*, pp. 1–9, 2020.

[39] A. Vakil, J. Liu, P. Zulch, E. Blasch, R. Ewing, and J. Li, "A Survey of Multimodal Sensor Fusion for Passive RF and EO Information Integration," *IEEE Aerospace and Electronic Systems Magazine*, vol. 36, no. 7, pp. 44–61, 2021.

[40] H. Qiu, X. Liu, S. Rallapalli, A. J. Bency, K. Chan, R. Urgaonkar, B. Manjunath, and R. Govindan, "Kestrel: Video analytics for augmented multi-camera vehicle tracking," in *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pp. 48–59, 2018.

[41] H. Chen, B. Guo, Z. Yu, and Q. Han, "Crowdtracking: Real-time vehicle tracking through mobile crowdsensing," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7570–7583, 2019.

[42] D. Roy, T. Mukherjee, M. Chatterjee, and E. Pasiliao, "RF Transmitter Fingerprinting Exploiting Spatio-Temporal Properties in Raw Signal Data," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 89–96, 2019.

[43] D. Roy, T. Mukherjee, M. Chatterjee, E. Blasch, and E. Pasiliao, "RFAL: Adversarial Learning for RF Transmitter Identification and Classification," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 2, pp. 783–801, 2020.

[44] Y. Li and J. Ibanez-Guzman, "Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems," *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 50–61, 2020.

[45] Scalabel Project, "A scalable open-source web annotation tool," 2021.

[46] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255, IEEE, 2009.

[50] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGG-Sound: A Large-scale Audio-Visual Dataset," in *IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, 2020.

[51] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1437–1451, 2018.

[52] P. P. L. et al., "Multibench: Multiscale benchmarks for multimodal representation learning," in *NeurIPS 2021 Datasets and Benchmarks Track*, 2021.

[53] J.-M. Perez-Rua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "MFAS: Multimodal Fusion Architecture Search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[54] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[55] S. M. Jayakumar, W. M. Czarnecki, J. Menick, J. Schwarz, J. Rae, S. Osindero, Y. W. Teh, T. Harley, and R. Pascanu, "Multiplicative interactions and where to find them," in *International Conference on Learning Representations*, 2020.

[56] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[57] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2247–2256, 2018.

[58] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[59] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[60] A. Mimouna, I. Alouani, A. Ben Khalifa, Y. El Hillali, A. Taleb-Ahmed, A. Menhaj, A. Ouahabi, and N. E. Ben Amara, "Olimp: A heterogeneous multimodal dataset for advanced environment perception," *Electronics*, vol. 9, no. 4, 2020.

[61] R. Nabati and H. Qi, "Centerfusion: Center-based radar and camera fusion for 3d object detection," *arXiv preprint arXiv:2011.04841*, 2020.