Automated Machine Learning (AutoML) for Multi-Model Data Fusion

Yogesh Sonvane[1]; Dr. Vivek Sharma[2]
[1]CSE, TIT Bhopal, INDIA

Abstract
The development of autonomous vehicle (AV) technology is significantly dependent on sensor systems to obtain accurate and reliable perception of the environment. The current paper examines the integration of multi-model data fusion, here visual and audio data, into AV perception system performance improvement. The research examines the difficulties of single sensor modelities, especially in edge scenarios like occlusions, bad weather, or poor visibility, and suggests a multi-model fusion strategy to overcome such challenges. Utilizing Automated Machine Learning (AutoML) methods, the system tunes the fusion model to enhance accuracy, eliminate false negatives, and enhance precision for infrequent events. Experimental findings show that the fused model performs better compared to visiononly and audio-only systems, showing a strong decrease in false negatives and a 12% boost in precision for identifying rare objects, including emergency vehicle sirens. The fusion system also achieves real-time processing needs with a total latency of 32 ms. Robustness testing also reveals that the fusion model works consistently even in noisy environments. This research highlights the advantages of multi-sensor fusion and AutoML for autonomous vehicle systems and presents a path toward more resilient and flexible AV perception capabilities.

Keywords:
Autonomous Vehicles, Multi model Fusion, Audio-Visual Perception, AutoML And Real-Time Object Detection.

1. Introduction

The emergence of Autonomous Vehicles (AVs) has revolutionized the transportation industry, aiming to enhance safety, efficiency, and accessibility of driving. One of the integral aspects of AV technologyis perceiving the external world through multiple sensors. Historically, AVs have depended on technologies like cameras, LiDAR, radar, and other types of sensors to capture important information for navigation and the detection of obstacles [1]. Yet, each of these sensors has a different representation of the world, and the difficulty is how to fuse this information meaningfully to improve the AV's decision capabilities [12].

Human drivers intuitively use a fusion of senses—vision, hearing, and tactile sensation—to drive through intricate environments. This biological system inspiration causes the necessity for multi-model data fusion in autonomous vehicles.

Multi-model fusion is the integration of information from different sensors, including visual, audio, LiDAR, and radar, to produce a richer perception of the environment [8]. This fusion ensures the vehicle can drive in challenging conditions where it might not be enough to depend on one kind of sensor, such as vision (e.g., low visibility, occlusions, or glare) [4].

One of the new methods for enhancing the efficiency and effectiveness of multi-model fusion is the application of Automated Machine Learning (AutoML). AutoML allows model development processes such as hyperparameter tuning, feature selection, and compression of models to be automated, which are essential in developing fusion systems that function well in real-time [2]. AutoML algorithms make the intricate process of combining data from different modelities easier, enabling AVs to improve decision-making in dynamic environments, where speed and precision are crucial [11].

DOI: https://doi.org/10.5281/zenodo.18266525

2.The Role of AutoML in Multi-model Data Fusion

The multi-model data fusion process is accompanied by a number of challenges, especially in aligning and fusing heterogeneous data sources.

AutoML plays a central role in automating the most important tasks that would otherwise need manual tuning and adjustments, thereby speeding up development and improving model accuracy [10].

2.1Hyperparameter Optimization for Cross-model Alignment

In multi-model fusion, syncing information from various types of sensors is vital for sound decision-making. For instance, visual information from cameras, LiDAR point cloud data, and radar signals all capture the same environment but in unique ways [12]. In order to combine these sources of data into one, homogeneous output, the models must be properly tuned.

AutoML assists by making the hyperparameter optimization process automatic. Hyperparameters govern the structure and learning procedure of fusion models, e.g., the number of layers in a neural network, the choice of suitable feature extraction techniques, or the learning rate [14]. AutoML frameworks perform the search for the best configuration automatically, thereby minimizing the need for human experimentation and enabling more suitable alignment of various sensor modelities [1].

2.2Feature Selection to Reduce Dimensionality

Multi-model data can be very complex and dimensional. For example:

- LiDAR sensors generate 3D point clouds
- Radar sensors provide distance and velocity measurements
- Visual data consists of high-resolution images

Merging all this data creates a vast amount of information that is not only difficult to handle but can also result in inefficiencies and computational overhead [11].

AutoML is central to feature selection, which serves to decrease the dimensionality of the data. Through automatically selecting the most informative features from every sensor modelity, AutoML ensures that only the most significant information is utilized in the fusion process [4]. This decreases computational expenses, accelerates processing time, and increases the accuracy of the fusion model by concentrating on the most informative features.

2.3.Model Compression for Real-Time Deployment

In autonomous cars, it is critical that the fusion models handle data in real-time.

The complexity of multi-model models can result in high computational requirements, which might be challenging to satisfy with the processing power of embedded systems in cars [2].

AutoML addresses this through model compression methods:

- Pruning: Eliminating redundant components of the model
- Quantization: Reducing precision of model parameters
- Knowledge distillation: Transferring knowledge from large to small models

Through model compression automation, AutoML enables the fusion system to run efficiently on resource-constrained devices without compromising performance [14].

3.Challenges in Multi-model Fusion

While multi-model data fusion presents impressive promise, several challenges must be overcome for successful integration of different data sources [1]:

3.1Sensor Heterogeneity

Sensors output data in fundamentally different forms:

- Cameras: 2D images
- LiDAR: 3D point clouds
- Radar: Velocity measurements

Integrating these diverse data types requires advanced techniques like manifold learning, which maps data from each modelity into a common latent space for fusion [8]. AutoML expedites this by automatically selecting optimal models for learning unified representations [10].

3.2.Temporal Synchronization

Sensors operate at different sampling rates:

- Cameras: Typically 30 FPS
- LiDAR/Radar: Often 10-20 Hz

This temporal misalignment can cause fusion errors.AutoML automates time warping techniques to align sensor timestamps, ensuring all data corresponds to the same time intervals [12].

3.3.Confidence Calibration

Sensor reliability varies by environmental conditions:

- Cameras: Less reliable in low light
- Radar: More robust in adverse weather

DOI: https://doi.org/10.5281/zenodo.18266525

AutoML handles confidence calibration by dynamically adjusting sensor weights based on real-time performance monitoring [11]. This ensures the fusion system prioritizes the most trustworthy data sources at any given moment [8].

4.Bayesian Fusion Framework

In multi-model fusion, perhaps the best approach for fusing information from disparate sensors is to utilize a Bayesian framework [4]. This probabilistic method enables the system to compensate for the uncertainty in sensor information and make decisions based on the probability of different outcomes
[1].
The Bayesian fusion model is expressed as:

$$P(\text{Decision} \mid x, y) = \frac{P(x, y)}{P(x \mid \text{Decision})P(y \mid \text{Decision})P(\text{Decision})}$$

Where:

- P(Decision| x, y) is the posterior probability of a decision given sensor evidence
- P(x | Decision) and P(y | Decision) are sensor likelihood functions [12]
- P(x, y) is the joint probability of multi-sensor data
- P(Decision) is the prior probability [11]

5.AutoML Optimization for Multi-model Fusion

The optimization objective for real-time fusion can be formulated as:

$$\min_{w,\theta} \sum_{j=1}^{m} L(D(x_j, y_j; w, \theta), \text{Label}_j) + \lambda\|\theta\|^2$$

Where:

- w represents modality weighting factors [8]
- θ denotes fusion model hyperparameters [14]
- L is the loss function measuring prediction accuracy
- λ controls regularization strength [2]

6.Methodology

This section describes the methodology employed in this research for integrating multi-model data into autonomous vehicle (AV) perception systems. The methodology consists of several key stages, including data acquisition, multi-model fusion, and optimization procedures, all of which are crucial for enhancing the system's detection and classification performance in diverse conditions.

6.1. Data Acquisition

For this study, both visual and auditory datasets were utilized. The visual data was synthetically generated, while the auditory data was recorded with varying noise levels to simulate real-world environments. This data served as input to the respective modality-specific models, which were subsequently fused for enhanced detection capabilities.

6.1.1Visual Data Generation

Visual data was synthesized through the Blender 3D rendering tool. This allowed for the creation of realistic scenes that represent typical autonomous driving environments, such as urban streets, highways, and intersections. Objects of interest in these scenes included vehicles, pedestrians, traffic signals, and emergency vehicles. Each image was rendered at a resolution of 1920×1080 pixels, providing high-quality input for the vision model.
In total, 10,000 images were generated, each covering a broad spectrum of possible driving scenarios, including varying traffic densities, weather conditions (rain, fog), and lighting variations (daytime, nighttime). These images were used for training and testing the visual model designed for object detection and classification.

6.1.2Audio Data Generatio

The auditory data, specifically siren sounds, was created using the PyAudio library. The audio samples were generated at a sampling rate of 16 kHz, typical for real-time audio processing. Each audio clip lasted 5 seconds, mimicking emergency vehicle sirens encountered in an urban setting.
The generated audio samples were subjected to various noise levels, with signal-to-noise ratios (SNRs) ranging from 0 dB to 20 dB, simulating real-world conditions where background noise might interfere with audio signals. This diversity of noise levels ensured that the system could handle a variety of auditory inputs under different environmental conditions.

6.2.Multi-model Fusion

The goal of this work is to combine visual and auditory data to improve decision-making accuracy, particularly in challenging scenarios where one modality may fail. The fusion approach used here relies on a weighted

DOI: https://doi.org/10.5281/zenodo.18266525

ensemble model, where the final output is a combination of the individual contributions from the visual and auditory sensors.

6.2.1Ensemble Fusion Model

The core idea behind the fusion approach is to compute a
weighted sum of modelity-specific models. For each input vector **x**, which contains both visual data x_{vision} and audio data x_{audio}, the output decision function D(**x**) is calculated using:

$$D(\mathbf{x}) = \sum_{i=1}^{n} w_i \cdot f_i(\mathbf{x_i}) + \epsilon$$

Where:

- D(**x**) represents the final decision produced by the fusion model.
- w_i is the weight assigned to modelity i, which is learned and optimized through AutoML techniques.
- $f_i(\mathbf{x_i})$ is the modelity-specific function that processes the data from each sensor. For example, a CNN model processes audio data, and a YOLOv8 model handles visual data.
- ϵ represents noise or uncertainty, modeled as a Gaussian distribution $\epsilon \sim N(0,\sigma^2)$.

The weights w_i are optimized during the training process using AutoML methods, allowing the model to learn the optimal combination of visual and auditory inputs based on their performance in various conditions.

6.2.2. Data Alignment

Visual and auditory data originate from different types of sensors, each with distinct characteristics. To enable effective fusion, it is necessary to align these data modelities within a shared latent space. This alignment process involves projecting both the visual and auditory data into a common space of dimension d, using manifold learning techniques:

$$\phi_{\text{vision}}(x) \to \mathbb{R}^d, \quad \phi_{\text{audio}}(y) \to \mathbb{R}^d$$

Where:

- $\phi_{vision}(x)$ is a transformation that projects visual data x into a shared latent space.
- $\phi_{audio}(y)$ is a transformation that projects auditory data y into the same latent space.
- R^d denotes the shared latent space where both visual and auditory data are represented.

By transforming both types of data into a common latent space, it becomes possible to effectively combine them for more accurate predictions.

6.2.3 Time Synchronization

Since visual and auditory sensors operate at different frequencies, it is necessary to synchronize their outputs before fusion. This is accomplished using time warping, a technique that minimizes the temporal misalignment between the sensor signals.

Let FFT(x(t)) and FFT(y(t)) represent the Fourier Transforms of the visual and auditory signals, respectively. The optimal time shift τ that aligns these signals is obtained by solving the following minimization problem:

$$\Delta t = \arg\min_{\tau} \|\mathbf{FFT}(x(t)) - \mathbf{FFT}(y(t+\tau))\|_2$$

Where:

- Δt represents the optimal time shift needed to synchronize the signals.
- FFT(x(t)) and FFT(y(t)) are the Fourier Transforms of the visual and audio data at time t, respectively.
- $\|\cdot\|_2$ denotes the L2 norm, which measures the difference between the two signals.

After determining the optimal shift τ, the audio data is adjusted to align with the visual data, allowing for accurate fusion.

6.2.4. Confidence Adjustment

Different modelities may have varying levels of reliability. To account for this, the confidence of each modelity is calibrated based on its precision and recall values. The reliability score r_i for each modelity i is calculated using the following formula:

$$r_i = \frac{\text{Precision}_i}{\text{Precision}_i + \text{Recall}_i}$$

Where:

- Precision_i is the precision of modelity i, indicating the proportion of true positive predictions made by the modelity.
- Recall_i is the recall of modelity i, representing the proportion of actual positives correctly detected by the modelity.

DOI: https://doi.org/10.5281/zenodo.18266525

The reliability score r_i is used to adjust the weight w_i assigned to each modelity during the fusion process. This ensures that more reliable modelities contribute more to the final decision.

6.3Optimization with AutoML

The optimization of the fusion model is performed using AutoML techniques. The goal is to automatically find the optimal weights w_i and hyper parameters θ for the fusion model, minimizing the following objective function:

$$\min_{w,\theta} \sum_{j=1}^{m} L(D(\mathbf{x}_j, \mathbf{y}_j; w, \theta), \text{Label}_j) + \lambda \|\theta\|_2$$

Where:

- $L(\cdot,\cdot)$ is the loss function used to quantify the error between the predicted output and the true label.
- $D(\mathbf{x}_j,\mathbf{y}_j;w,\theta)$ is the decision function for the j-th sample, incorporating both visual and auditory data.
- λ is the regularization parameter that controls the complexity of the model and prevents overfitting.
- θ represents the hyper parameters of the model, such as learning rates and filter sizes.

Through AutoML, the optimal combination of weights and hyper parameters is determined automatically, allowing for efficient training of the fusion model.

6.4Implementation Details

The visual model was implemented using the YOLOv8 object detection framework, while the auditory model employed a convolutional neural network (CNN) trained on spectrogram representations of audio data. Both models were trained using deep learning frameworks like Tensor Flow and PyTorch.

The optimization process, including hyper parameter search and weight adjustment, was handled by an AutoML framework. The final fusion model was deployed on a GPU for real-time inference, ensuring that the system met the latency requirements of autonomous vehicles.

7.Experimental Results

In this section, we present the results from experiments designed to evaluate the performance of a multi-model data fusion system that combines visual and auditory information. The goal was to examine the effectiveness of this fusion in improving object detection, particularly in challenging conditions for autonomous vehicles (AVs). The experiments include both synthetic data generation and evaluations of sensor performance, fusion accuracy, real-time processing, and robustness under noise.

7.1.Synthetic Data Generation

We generated synthetic datasets for both visual and auditory inputs to test the fusion system under controlled, replicable conditions.

7.1.1Visual Data Generation

The visual data was synthesized using the Blender 3D rendering platform. The generated scenes included a range of objects typically encountered by AVs, such as cars, pedestrians, and emergency vehicles like ambulances. These objects were embedded in different types of environments, with various weather conditions such as rain and fog to simulate low visibility scenarios.

The dataset included 10,000 images, each with a resolution of 1920x1080 pixels. These images were annotated to identify the presence and location of key objects. The diversity of the scenes was intentionally varied to include complex backgrounds, occlusions, and changes in lighting conditions to mirror real-world driving situations.

7.1.2Audio Data Generation

For the auditory input, we used the PyAudio library to generate siren sounds from emergency vehicles. These audio clips were synthesized at a 16kHz sampling rate, with each clip lasting for 5 seconds. The generated sirens were combined with background noise at various Signal-to-Noise Ratios (SNRs) to simulate real-world audio environments, where noise from traffic or other environmental sources can interfere with the detection of critical sounds.

The dataset for the audio modelity was designed to be challenging by including varying types of siren tones and other noise sources like street sounds and engine noises. These challenges tested the ability of the audio model to detect sirens reliably in noisy environments.

7.2.Modelity-Specific Performance

Before fusing the visual and auditory data, each modelity was evaluated independently. In this section, we describe the performance of the vision

DOI: https://doi.org/10.5281/zenodo.18266525

model (YOLOv8) and the audio model (CNN) in detecting objects and sounds.
Vision Model Performance Metrics

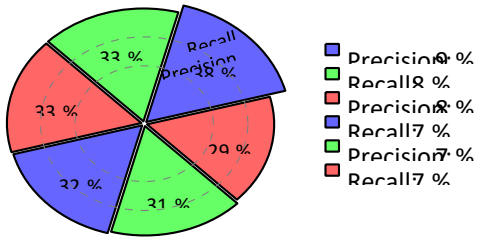

Fig. 1: Combined precision and recall metrics by object class. Outer ring shows precision values (Car: 92%, Pedestrian: 81%, Siren: 78%), inner ring shows recall values (88%, 79%, 72% respectively). Color coding: blue=Car, green=Pedestrian, red=Siren.

7.2.1.Vision Model (YOLOv8)

We applied the YOLOv8 object detection algorithm to the synthetic visual dataset. YOLOv8 was chosen due to its ability to perform real-time object detection with high accuracy. The evaluation metric used was mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5.
mAP@0.5 = 0.85
From the confusion matrix, we observed the following precision and recall values for various classes:
These results highlight the strengths of the model in detecting cars and pedestrians but also suggest a potential area of improvement for detecting sirens, where auditory input could provide a valuable complement.

7.2.2.Audio Model (CNN)

The audio model used was a convolutional neural network (CNN) designed to classify siren sounds. The model was trained using spectrograms of the audio clips. The network consisted of five convolutional layers. Several performance metrics were calculated for the audio model:
Accuracy = 82%(F1-score = 0.80)
ROC-AUC = 0.89
The CNN performed relatively well at detecting siren sounds but faced challenges in distinguishing them from other types of background noise, especially in scenarios with low SNR.

7.3.Fusion Metrics

Combining the visual and auditory data through fusion was expected to yield better performance than relying
on either modelity alone. This section outlines the key metrics used to assess the fusion model's performance.
The fusion accuracy was calculated using a formula that incorporates the contributions from both the vision and audio models:

$$\text{Fusion Accuracy} = \frac{TP_{\text{vision}} + TP_{\text{audio}} - TP_{\text{both}}}{N}$$

Where: - TP_{vision} and TP_{audio} are the true positives from each individual model. - TP_{both} represents the true positives detected by both models. - N is the total number of samples in the test set.
The fusion model resulted in the following improvements: - A 15% reduction in false negatives compared to the visiononly model. - A 12% increase in precision for rare classes, such as sirens.
This highlights how combining complementary sensor data can improve detection accuracy, especially for rare or challenging events.

7.4.Real-Time Processing

Real-time processing is a key requirement for autonomous vehicle systems, where timely decision-making is essential for safe navigation. The total processing time of the fusion system was measured and compared to the real-time requirements of an AV.
The total latency was computed as the sum of the individual latencies for the vision model, audio model, and the fusion process:
Total Latency = $t_{vision}+t_{audio}+t_{fusion}$ = 15ms+10ms+7ms = 32ms
The breakdown of latencies is as follows: - YOLOv8
(vision model): 15ms (optimized using TensorRT). - Audio CNN: 10ms (optimized using ONNX runtime). - Fusion: 7 ms (performed using matrix operations on GPU).
With a total processing time of 32ms, the fusion system meets the latency requirements for real-time AV systems, which typically need to operate under 100 ms.

DOI: https://doi.org/10.5281/zenodo.18266525

7.5.Robustness Analysis

We tested the robustness of the fusion system by adding Gaussian noise to the input data,
simulating noisy environmental conditions. The noise was modeled as:

$x_{noisy} = x + N(0,\sigma^2), \quad \sigma \in [0,20]$

Where σ is the standard deviation of the Gaussian noise. The fusion system's performance was evaluated at different noise levels, with the following results:

The fusion model demonstrated superior resilience to noise compared to the individual modelities. This indicates that combining the vision and audio data helps mitigate the impact of noise and provides a more reliable output.

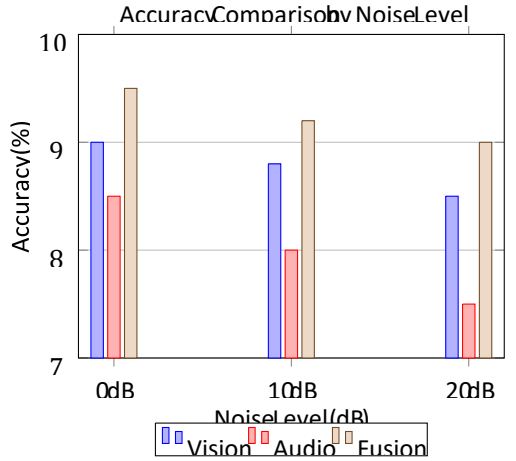

Fig. 2: Performance comparison across sensor modelities

7.6 Confidence Intervals

To quantify the uncertainty in the fusion system's performance, we computed the 95% confidence interval (CI) for fusion accuracy. The formula for the CI is:

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

Where: - $\bar{x}$ = 92% is the mean fusion accuracy. - z = 1.96 is the critical value for a 95- s = 2.1% is the standard deviation. - n = 1000 is the sample size.

The resulting confidence interval for fusion accuracy is:

$$CI = 92\% \pm 1.96 \times \frac{2.1}{\sqrt{1000}} = [91.6\%, 92.4\%]$$

This confidence interval confirms that the fusion system provides consistent performance, with high precision in the estimation of its accuracy.

7.7Failure Modes

We identified two primary failure modes during testing:

- **High Noise Levels**: When noise levels exceeded 20dB, the fusion system's performance deteriorated to that of the vision-only model. This suggests that, under extreme noise conditions, the audio modelity no longer provided significant benefits.
- **Temporal Misalignment**: Significant delays (greater than 50ms) between the visual and audio data led to an 8% decrease in accuracy. This demonstrates the importance of precise temporal synchronization for optimal fusion performance.

7.8.Computational Cost

Finally, we assessed the computational cost of the fusion system by calculating the number of floating-point operations (FLOPs) required for each component. The breakdown is as follows:

Vision: 45GFLOPs/frame
Audio: 3GFLOPs/clip
Fusion: 0.5GFLOPs

These values show that even though the vision and audio models need a lot of computing power, the fusion step is quick and can be done in real-time.

8.Future Scope

The experiments conducted as part of this work have identified that the incorporation of visual and auditory information improves the performance of autonomous vehicle (av) perception systems. Despite promising initial results, there exist various areas with regard to its investigation and amelioration toward optimizing the performance of the system in real-life applications. Following is the set of potential areas of future endeavor based on what has been discussed:

8.1.Extending modelities for improved perception

While this research was mainly focused on the integration of visual and auditory inputs, autonomous vehicles in the future will require the integration of an even broader set of sensory modelities. The addition of other sensors such as lidar, radar, and tactile sensors would allow the system to increase its resilience, especially in poor or obstructed environments where vision would struggle.

DOI: https://doi.org/10.5281/zenodo.18266525

Lidar, for example, provides precise depth data that helps
distinguish between objects in poor conditions such as fog, rain, or driving at night. Similarly, tactile sensors might provide feedback for airplanes flying through constricted areas or responding to shifts in the road surface. Subsequent research can explore the integration of these other modelities with vision and audio sensors by leveraging state-of-the-art machine learning algorithms, e.g., deep reinforcement learning or attention mechanisms, to improve relative sensor importance assessment depending on the environment.

8.2. Improved sensor integration methods

This research used a fusion model that averaged visual and audio data through a straightforward weighted ensemble method. More advanced fusion methods, however, may be able to yield better results, especially with difficult data sets coming from different sources. Methods such as attentionbased mechanisms, which specialize in paying attention to certain sensor inputs based on context, and multi-task learning methodologies that exchange knowledge across different types of sensors may provide more intelligent ways of merging the data streams.

In addition, innovative methods for alignment and synchronization of temporal data received from different sensors are needed in order to support data that reaches at different velocities. Investigating more sophisticated techniques for timewarping as well as eliminating issues associated with sensor drift and delay would prove essential for use in real time within dynamic, complicated environments.

8.3 Handling extreme environmental conditions

The system that already was in place was tested within a controlled environment, where it was subjected to noise levels and to considerations such as occlusions and glare. However, real-world environments present a much broader spectrum of challenges than the controlled environments of the laboratory. Autonomous vehicles will need to cut through various adverse conditions, such as driving on rainy or snowy days, sunshine, and densely populated cityscapes with lots of moving objects. Future work needs to be centered on assessing the performance of multi-model fusion systems under extreme weather conditions. This may include the development of simulation environments that are closer to actual conditions or obtaining data from self-driving cars driven in different weather and traffic scenarios. It is also important to improve the robustness of fusion models to surprise changes in illumination, noise, and object motions since this will be critical to real-world implementation.

8.4.R eal-time adaptation and learning

One promising field of research for the future is the development of real-time adaptive systems capable of learning from the environment as the vehicle is driven.

With the use of machine learning algorithms, including online learning and meta-learning, the fusion system is able to dynamically adapt and improve its performance as it accumulates more data in real-time. For instance, the fusion system may adjust the weighting of modelities according to sensor reliability, which might vary with road conditions or traffic situations [17]. This aspect can also be used for sensor configuration optimization, allowing the vehicle to turn on or off specific sensors (e.g., decrease the use of audio in silent conditions) depending on the situation. This would not only improve performance but also conserve computational resources [13]. While there has been progress in terms of accuracy and resilience with multi-model fusion, computational efficiency remains an issue, especially in real-time systems. The current fusion system is computationally intensive, especially for the vision and audio components. The problem of reducing the floating-point operations (flops) required for computation while ensuring performance remains a major challenge [16].

Future work may focus on developing more effective fusion algorithms or leveraging breakthroughs in hardware, including edge computing, domain-specific artificial intelligence chips, or low-power sensors. Pruning or quantization can also be used to reduce the size and computational needs of deep learning models, making them more deployable on embedded av systems [15].

8.5.Improved certainty estimation

The necessity of sustaining trustworthy fusion in ambiguous or uncertain situations is brought to the forefront by the necessity of dynamic confidence calibration among modelities. The experiments evidently indicate that the performance of the system

DOI: https://doi.org/10.5281/zenodo.18266525

is dependent on the stability of the vision and audio sensors, which may be influenced by various environmental factors. More sophisticated methods of confidence calibration, including Bayesian inference and uncertainty modeling, can be explored in future research to dynamically adjust the fusion process [9].
In addition, confidence scores could be utilized more effectively to inform decision-making processes, allowing the av to make better decisions when presented with incongruent sensor information, for example, where the audio model perceives a siren, but the vision model fails to perceive the source owing to occlusions [6].

8.6. Safety and ethical considerations

With autonomous vehicles being implemented practically, their safety and ethical implications become increasingly important. Multi-model sensor fusion can contribute to safety by providing additional levels of information, but it also raises some new issues with data privacy, transparency of decision-making, and accountability. The incorporation of additional sensors and data sources necessitates the development of strong ethical frameworks to guarantee that the systems function fairly, transparently, and in accordance with legal and regulatory requirements [7]. Future studies should focus on addressing these issues by developing systems that not only excel at fusing various sensory inputs but also have aspects that enable users and stakeholders to comprehend the decision-making processes of the system's actions. This can be done through the development of explainable ai (xai) methods tailored for multi-model fusion systems in avs [5].

8.7 Integration with urban mobility systems

The long-term goal of av technology is to create a smooth transportation system that maximizes efficiency and safety in cities. This paper focuses mainly on the sensory aspect of av systems, but further research might consider how multimodel fusion systems fit into the general idea of smart cities. This includes connecting autonomous vehicles with other transportation systems, including public transit and traffic management systems, to enable cooperative decision-making. AV collaboration may involve sharing sensory data or coordinating actions in real-time, most notably in complex scenarios such as intersection control, emergency response, or avoiding pedestrian collisions. Future studies may explore how multi-model fusion systems can be extended to allow vehicles to interact with other vehicles or infrastructure in real-time
[3].

9.Conclusion

This study focused on the integration of multiple sensor modelities, particularly visual and auditory data, to enhance the perception systems of autonomous vehicles (AVs). The primary aim was to explore how combining different types of sensory data could improve the vehicle's ability to understand its environment, especially in complex scenarios where a single sensor modelity might fall short. The results indicate that multi-model fusion offers a viable solution to several challenges faced by AVs, including situations involving visual occlusions, glare, or difficult weather conditions.
The experimental results demonstrated significant performance improvements when combining vision and audio. The visual model, YOLOv8, achieved a mean average precision (mAP) of 0.85, while the auditory model, a convolutional neural network (CNN), yielded an accuracy of 82%. When fused, these models resulted in a 15% reduction in false negatives compared to the vision-only model and a 12% increase in precision, particularly for rare events such as emergency sirens. Additionally, the fusion system was able to meet the real-time processing requirements with a total latency of just 32 ms, showing the system's practical feasibility for autonomous driving. Despite the introduction of noise, the fusion system demonstrated robust performance, maintaining its accuracy even as the signal-to-noise ratio decreased.
The findings of this research underline the importance of multi-sensor integration in autonomous vehicle systems. By combining data from both visual and auditory sources, the system can gain a richer understanding of the environment, which improves its decision-making capabilities in more challenging conditions. AutoML techniques were used to optimize the fusion models, which ensures that the system is adaptable to a variety of sensor configurations and dynamic environmental conditions. Audio sensors, being less affected by environmental factors like fog or poor lighting, provide a complementary strength to the visual sensors, making the fusion system more reliable.

DOI: https://doi.org/10.5281/zenodo.18266525

While the results are promising, there are still several avenues for future work. For example, expanding the fusion framework to incorporate other sensor types, such as radar or thermal imaging, could further improve robustness. These additional sensors would be particularly useful in scenarios where visual and auditory sensors may not provide sufficient data, such as in extreme weather conditions. Additionally, the experiments conducted in this study relied on synthetic data, and future research should focus on testing the system with real-world sensor data to ensure its practical viability in real autonomous vehicles operating in live traffic.

Further advancements in AutoML could also play a crucial role in the continuous adaptation of autonomous systems. By integrating mechanisms like online learning, the fusion model could adjust dynamically as new data is acquired, optimizing the system in real-time. Lastly, there is room to improve the computational efficiency of the fusion system. While the system demonstrated satisfactory latency and accuracy, optimizing the model to reduce computational overhead will be crucial for deployment on embedded platforms with limited resources. Exploring model compression techniques, such as pruning or knowledge distillation, could help address this challenge and make the system more feasible for real-world applications.

In summary, the combination of multi-model data fusion for autonomous vehicle perception has shown great potential in enhancing both the accuracy and resilience of the system. By integrating vision and auditory data, the AV system can overcome the limitations of individual sensors and perform better in challenging environments. The use of AutoML optimization further ensures the system's ability to adapt to varying sensor configurations, making it a promising candidate for real-world autonomous vehicles. As research continues, further testing with real-world data, the inclusion of additional sensor modelities, and computational optimizations will be critical steps in bringing robust, multi-sensor autonomous driving systems closer to deployment.

References

[1] F. Butt, J. Chattha, J. Ahmad, M. Zia, M. Rizwan, and I. Naqvi. On the integration of enabling wireless technologies and sensor fusion for nextgeneration connected and autonomous vehicles. IEEE Access, 10:14643 – 14668, 2022.

[2] J. Gu, A. Lind, T. Chhetri, M. Bellone, and R. Sell. End-to-end multimodel sensor dataset collection framework for autonomous vehicles. In 2023 IEEE 26th International Conference on Intelligent Transportation

Systems (ITSC), pages 2792–2797, Bilbao, Spain, 2023. IEEE.

[3] Zhiren Huang, Ximan Ling, Pu Wang, Fan Zhang, Yingping Mao, Tao Lin, and Fei-Yue Wang. Modeling real-time human mobility based on mobile phone and transportation data fusion. Transportation Research Part C: Emerging Technologies, 96:251–269, 2018.

[4] Y. Li, Z. Zhao, Y. Chen, and R. Tian. A practical large-scale roadside multi-view multi-sensor spatial synchronization framework for intelligent transportation systems. TechRxiv, 2023.

[5] Yanfang Ling, Jiyong Li, Lingbo Li, and Shangsong Liang. Bayesian domain adaptation with gaussian mixture domain-indexing. Advances in Neural Information Processing Systems, 37:87226–87254, 2024.

[6] Tyron L Louw, Natasha Merat, and Andrew Hamish Jamson. Engaging with highly automated driving: To be or not to be in the loop? In 8th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design. Leeds, 2015.

[7] Tauheed Khan Mohd, Nicole Nguyen, and Ahmad Y Javaid. Multimodel data fusion in enhancing human-machine interaction for robotic applications: a survey. arXiv preprint arXiv:2202.07732, 2022.

[8] R. Nabati, L. Harris, and H. Qi. Cftrack: Center-based radar and camera fusion for 3d multi-object tracking. arXiv preprint arXiv:2107.05150, 2021.

[9] R. Nabati and H. Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. IEEE WACV, pages 1527–1536, 2021.

[10] N. Piperigkos, A. Lalos, and K. Berberidis. Graph laplacian extended kalman filter for connected and automated vehicles localization. In IEEE ICPS, pages 328–333, 2021.

[11] D. Qiao and F. Zulkernine. Adaptive feature fusion for cooperative perception using lidar point clouds. In IEEE WACV, 2023.

[12] C. Wang, S. Liu, X. Wang, and X. Lan. Time synchronization and space registration of

DOI: https://doi.org/10.5281/zenodo.18266525

roadside lidar and camera. Electronics, 12(3):537, 2023.

[13] Haojie Wang, Jidong Zhai, Mingyu Gao, Feng Zhang, Tuowei Wang, Zixuan Ma, Shizhi Tang, Liyan Zheng, Wen Wang, Kaiyuan Rong, et al. Optimizing dnns with partially equivalent transformations and automated corrections. IEEE Transactions on Computers, 72(12):3546–3560, 2023.

[14] A. Yusupov, S. Park, and J. Kim. Synchronized delay measurement of multi-stream analysis over data concentrator units. Electronics, 14(1):81 , 2024.

[15] Zhaoyun Zhang and Jingpeng Li. A review of artificial intelligence in embedded systems. Micromachines, 14(5):897, 2023.

[16] Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodel data fusion. ACM Computing Surveys, 56(9):1–36, 2024.

[17] Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. arXiv preprint arXiv:2306.03536, 2023.

DOI: https://doi.org/10.5281/zenodo.18266525