AGH University of Krakow

FIELD OF SCIENCE ENGINEERING AND TECHNOLOGY

SCIENTIFIC DISCIPLINE AUTOMATION, ELECTRONIC, ELECTRICAL ENGINEERING AND SPACE TECHNOLOGIES

Summary of accomplishments

Low-level sensor data fusion for Object Detection in an Autonomous Vehicle perception system based on a Machine Learning approach.

Author: Mgr inż. Daniel Dworak

First supervisor: Dr hab. inż. Jerzy Baranowski, prof. AGH
Assisting supervisor: Dr inż. Mateusz Komorkiewicz

Completed in: Faculty of Electrical Engineering, Automatics, Computer Science, and Biomedical Engineering

Kraków, 2023

1. Abstract

Autonomous Driving (AD) is at the forefront of automotive research, promising substantial advancements in safety and user experience. The thesis investigates the pivotal role of sensors, such as cameras, LiDAR (Light Detection And Ranging) and Radar (Radio Detection And Ranging), in perception systems, which are the cornerstone of Autonomous Vehicle (AV). Leveraging Machine Learning (ML) techniques, particularly Convolutional Neural Network (CNN) and Deep Learning (DL), perception systems excel in tasks like Object Detection (OD), crucial for AVs' decision-making processes.

The thesis explores sensor fusion, with a specific focus on Low-Level Fusion (LLF), aiming to amalgamate data from diverse sensors to enhance perception capabilities. The proposed Cross-Domain Spatial Matching (CDSM) method stands out as a novel approach in the domain, addressing the challenge of aligning sensor data from disparate sources and facilitating efficient fusion. By seamlessly integrating CDSM into the Neural Network (NN) architecture, the thesis not only enables end-to-end training but also ensures sufficient inference times during operational deployment, crucial for real-time applications.

To comprehensively evaluate the efficacy of LLF in AV perception systems, the thesis establishes a framework encompassing both single-sensor models and fusion architectures. Utilizing prominent open-source datasets such as KITTI and NuScenes, the performance of these models is rigorously assessed using a range of visual and Key Performance Indicator (KPI) metrics. The results showcase the potential benefits of LLF, with CDSM demonstrating competitive performance compared to State-Of-The-Art (SOTA) fusion methodologies.

Furthermore, the thesis conducts an in-depth analysis of the models' predictions, shedding light on both their strengths and limitations. By highlighting efficiency gains and corner-case scenarios where fusion models diverge from their single-sensor counterparts, the thesis provides valuable insights into the practical implications of LLF in AV perception systems. Additionally, comparisons with existing fusion solutions offer a broader perspective, positioning CDSM among leading techniques in the field.

Ultimately, the findings presented in the thesis contribute to advancing our understanding of LLF in AV perception systems and offer valuable insights into its potential benefits. By bridging the gap between theoretical research and practical implementation, this work contributes to future developments in AD technology, with implications for safety, efficiency, and user experience on the roads of tomorrow.

2. Motivation

Perception forms the backbone of AD systems, enabling critical functions like tracking and planning algorithms. A comprehensive understanding of the environment is essential for AVs to navigate safely. Sensor fusion combines data from multiple sensors, enhancing perception by providing a broader and more reliable view. Utilizing LLF methods for sensor fusion offers additional advantages. LLF methods access unprocessed sensor readings, providing diverse data types and perspectives, strengthening the perception system. Safety is essential for AVs, with perception systems crucial for detecting nearby objects and executing active safety measures. Relying on a single sensor may pose limitations, making sensor fusion vital to integrate multiple sensors and mitigate individual sensor vulnerabilities. LLF-based solutions efficiently validate sensor readings and adjust their impact on outcomes, particularly in scenarios with conflicting data.

Moreover, sensor fusion extends perception capabilities beyond individual sensors' limitations, providing a more accurate representation of the environment. Research focuses on optimizing fusion algorithms, especially LLF solutions based on DL, which uncover cross-sensor dependencies and reveal concealed patterns, enhancing perception insights.

In conclusion, sensor fusion research is driven by perception's critical role in AV systems, safety concerns, and the potential to extend performance beyond individual sensors' capabilities. Advancements in fusion techniques are essential for AD technology, despite posing challenges like synchronization and perspective alignment, which ongoing research aims to address.

3. Research hypothesis

The thesis aims to assess the efficiency of LLF utilizing automotive sensor data for OD in AV perception systems. By integrating multiple sensors, the goal is to enhance system performance and robustness. Through analysis and experimentation, the thesis aims to highlight LLF's potential benefits, including improved perception accuracy, reduced uncertainty, and a more comprehensive understanding of the environment compared to single-sensor approaches.

Moreover, the thesis investigates the end-to-end fusion of sensor data within a NN architecture, employing DL methodologies. This exploration seeks to uncover hidden patterns in complex, high-dimensional data spaces, potentially amplifying LLF's effectiveness in AV perception systems.

Finally, recognizing the growing importance of explainability in AD systems, the thesis also explores Explainable AI (XAI) techniques to enhance interpretability in perception NN models. This endeavour aims to shed light on the underlying reasoning behind model predictions, critical for understanding decision-making processes in AV perception systems.

4. Methods and results

The primary focus of the thesis revolves around the fusion of low-level sensor data. It introduces a novel methodology named CDSM, specifically designed to overcome challenges associated with merging data from sensors belonging to different domains. The detailed explanation of this methodology serves as the cornerstone of the thesis, including its theoretical foundations: alignment, aggregation, and fusion processes. Following this theoretical groundwork, the thesis proceeds to present and analyze the obtained results in depth. Through careful examination, it highlights the advantages of the fusion approach and provides empirical evidence to support the stated hypotheses. This comprehensive analysis underscores the significance and effectiveness of the proposed fusion methodology in enhancing perception systems. In addition to the core research on sensor data fusion, the thesis also tackles the realm of XAI. It explores the adaptation of the popular Gradient-weighted Class Activation Maps (Grad-CAM) method to a novel context, demonstrating its applicability in elucidating complex decision-making processes within LiDAR and Radar pointclouds. This exploration further enriches the thesis by broadening its scope to encompass diverse data domains and methodologies.

4.1 CDSM fusion method

The proposed fusion approach adopts a multi-view setup, utilizing separate network architectures to process camera images and 3D pointcloud data. In this setup, the camera input undergoes processing within the 2D image domain, while the pointcloud data is handled in an enhanced Bird's Eye View (BEV). Both networks generate predictions in their respective domains, with feature maps from both models passed to a common block for fusion and final 3D predictions. This approach falls under the late LLF category, as sensor information is fused during the processing phase, and preprocessed feature maps enter the fusion block, rather than raw sensor readings.

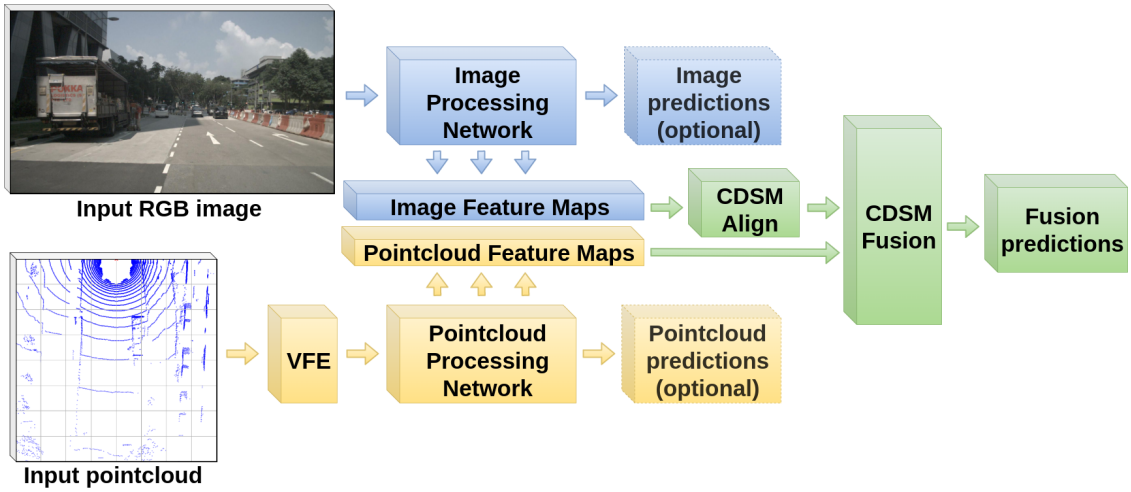

Figure 4.1: CDSM solution architecture overview for image and pointcloud fusion.

To enable this late fusion, a novel component called CDSM is introduced, as presented in Figure 4.1. Its purpose is to fuse feature maps from intermediate layers of the camera and Radar networks, creating a unified internal representation for object predictions in 3D space. A critical challenge lies in aligning feature maps from disparate domains, which is accomplished through the CDSM align block, ensuring optimal fusion of information from both sources.

The proposed fusion NN architecture operates at the feature level, employing the CDSM alignment and fusion methodology. It requires preprocessed sensor data in the form of extracted feature maps, integrating SOTA methods from single-sensor perception algorithms. Inspiration for those submodels' design is drawn

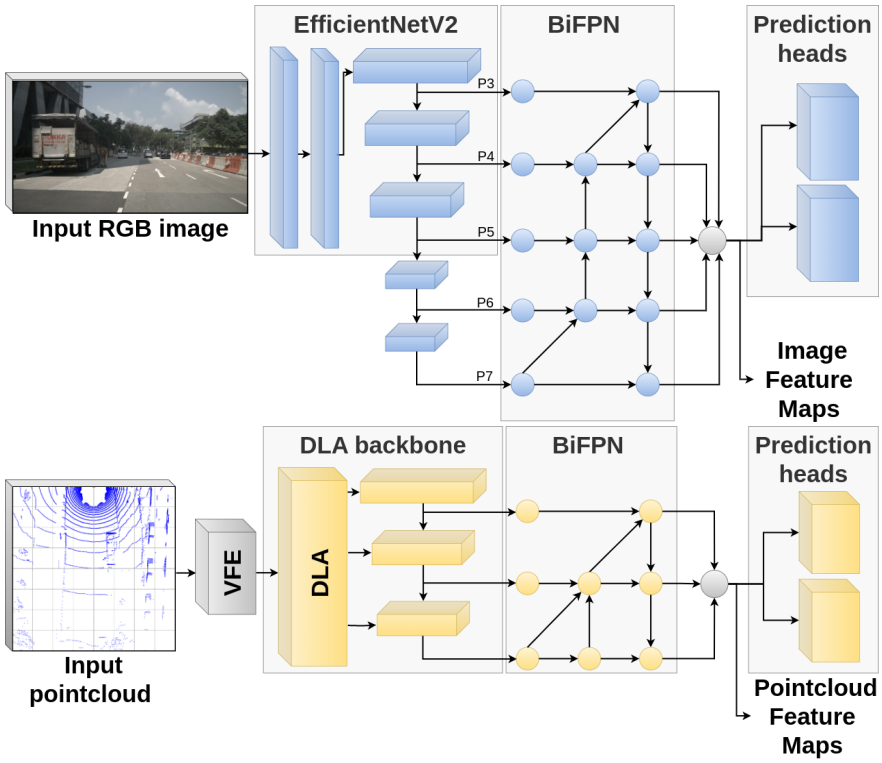

Figure 4.2: Single-sensor features extraction sub-models.

from SOTA in a 2D image and a 3D pointcloud OD domains, allowing for the utilization of optimized network structures validated in their respective contexts. Those single-sensor feature extraction submodels are presented in Figure 4.2. Additionally, incorporating single-sensor feature extraction networks enables comparison of intermediate results, aiding in assessing sensor contribution and fusion gain.

The CDSM block aims to spatially align information extracted from 2D camera images and 3D pointcloud data. Initially, feature maps from the intermediate layers of each network are misaligned (Figure 4.3). The CDSM method incorporates Domain Alignment and Fusion, with the former addressing misalignment issues. To facilitate alignment, a unified space called the Vehicle Coordinate System (VCS) is introduced, providing a standardized framework for representation and alignment. The VCS is a Cartesian coordinate system centred on the car's front axle, with axes defined accordingly. By considering the VCS, camera images and pointcloud data can be aligned consistently. The alignment process involves a custom CDSM rotation layer, which extracts indexes from the original tensor and calculates a rotation matrix using quaternion rotations. The rotation matrix is then applied to the indexes to achieve spatial alignment. Additionally, carefully designed rotation parameters ensure proper orientation and alignment of camera feature maps with the VCS. The specific rotation operations are crucial for aligning tensors within the VCS, ensuring not only proper orientation but also a unified reference point. This level of alignment is significant for seamless fusion between different sensor modalities. The chosen rotation operations address the unique spatial requirements of the fusion process, ensuring coherent positioning within the VCS.

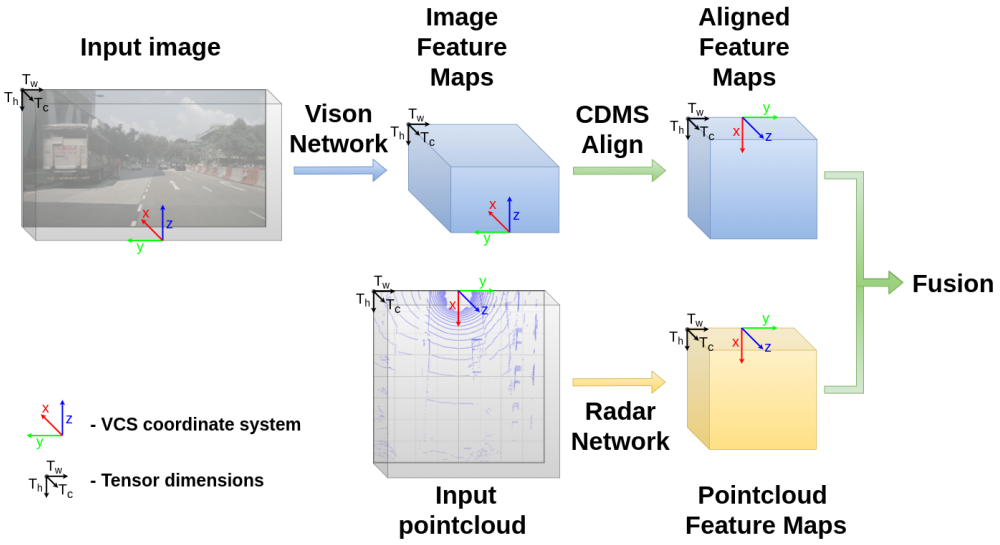

Figure 4.3: CDSM domain alignment method visualization.

As presented in the architecture overview, fundamental components necessary for sensor data fusion have been already established in the proposed pipeline. Single-sensor submodels process inputs, yielding feature maps representing data from each sensor. CDSM domain alignment transfers image features to a unified VCS in BEV, enabling fusion between camera images and LiDAR or Radar pointcloud data. Three fusion methods are introduced: one-to-one concatenation, feature-wise aggregation, and range-based aggregation. One-to-one concatenation merges feature maps directly, while feature-wise aggregation combines

vision features before concatenation with pointcloud features. Range-based aggregation integrates distance-specific information into the fusion process, distributing camera feature maps based on specific distance ranges as illustrated in Figure 4.4.

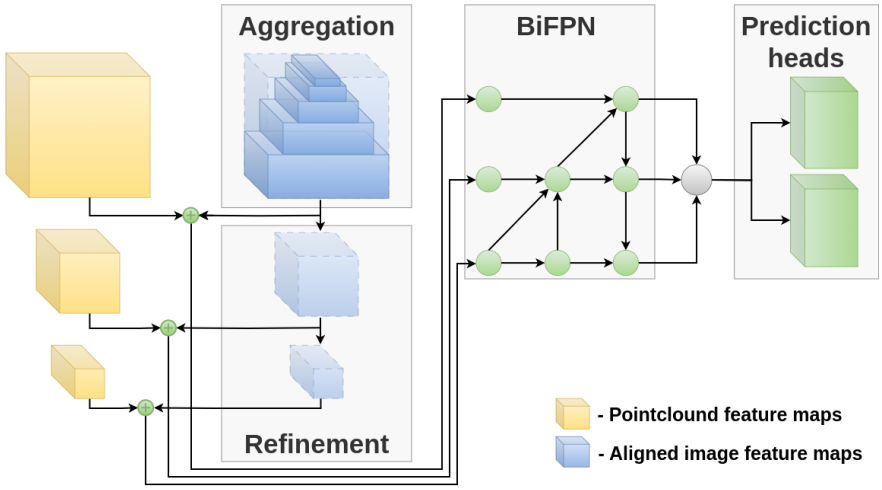

Figure 4.4: CDSM range-based aggregation fusion method.

For feature-wise aggregation, camera feature maps are concatenated on a unified BEV grid, then refined through convolutional layers to establish spatial correlations and generate refined grids for fusion with pointcloud features. In range-based aggregation, camera feature maps are distributed based on distance ranges, accommodating range-specific characteristics. Features are adjusted to maintain consistency with the camera's Field-of-View (FoV), ensuring relevant information is incorporated while excluding irrelevant details. Finally, refined feature grids are fused with pointcloud features to generate 3D predictions.

4.2 CDSM fusion results

The accuracy and reliability of perception system components are crucial for AV applications. Effective evaluation methods are necessary to ensure their functionality, especially in challenging automotive environments. Evaluation processes should not only confirm functionality but also provide quantitative measures of performance. In the field of OD, well-established metrics like precision, recall, F1 score, and Mean Average Precision (mAP) are commonly used to assess NN model performance. Furthermore, considering the additional features in 3D OD compared to 2D, a set of supplementary metrics is introduced, including Mean Average Translation Error (mATE), Mean Average Size Error (mASE), and Mean Average Orientation Error (mAOE), alongside the combination of those metrics in NuScenes Detection Score (NDS). These additional metrics provide a more comprehensive assessment of the perception system's performance, capturing aspects such as spatial accuracy, object size estimation, and orientation precision in three-dimensional space. The KPI metrics enable quantitative assessment of OD models, facilitating benchmarking, tracking progress, and identifying strengths and weaknesses. In the thesis, these metrics are utilized to evaluate both single-sensor and fusion solutions, allowing for comprehensive comparisons and guiding further improvements.

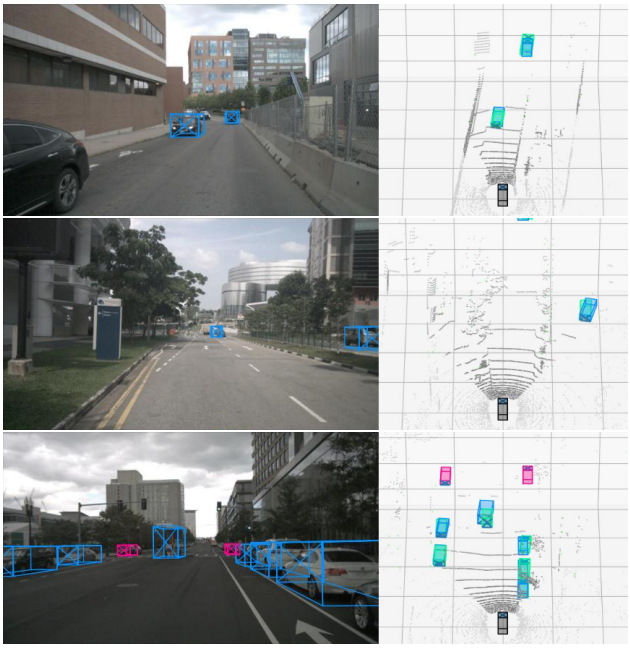

Figure 4.5: The results of the CDSM fusion model for 3D OD using camera and Radar data on the NuScenes test set. Labels are represented in green, positive detections in blue, false detections in magenta, and missed detections in yellow.

The fusion architecture proposed in the thesis integrates camera images and Radar pointcloud data, showing substantial promise in the market due to the utilization of widely available sensors found in production vehicles. While existing literature lacks similar fusion solutions tailored to these specific sensor suites, the model developed here demonstrates potential. The performance of both camera and Radar single-sensor solutions falls short compared to LiDAR-only models in terms of KPIs for 3D OD tasks. These limitations stem from challenges such as inaccurate depth estimation with cameras and the low density of Radar pointcloud data. However, by properly integrating these sensors through fusion, their individual strengths can be harnessed while mitigating weaknesses. Visualization of the fusion model's results (Figure 4.5) confirms its successful synergy between camera and Radar sensors, surpassing the performance of individual single-sensor solutions. The fusion model exhibits high accuracy and precision, particularly in densely populated scenes, showcasing its ability to effectively leverage the strengths of both sensors for reliable object detection.

After training and evaluating all proposed single-sensor architectures and fusion models separately, a comprehensive comparison was conducted to assess their performance and determine the fusion gain achieved, shown in Table 4.1. In addition to single sensor models, the fusion setups were explored, including camera with LiDAR and camera with Radar configurations. The evaluation metrics on the NuScenes dataset showed favourable results for both single-sensor and fusion models. LiDAR-only architecture achieved the best performance among single-sensor models, while the fusion of camera and Radar data exhibited significant improvement over both individual single-sensor methods. Further optimization of the fusion model led to even better performance, nearing that of LiDAR-only models. Visual results confirmed the effectiveness

Table 4.1: The comparison of the KPI metrics for all single-sensor and fusion models trained on the NuScenes dataset.

Method	Sensor	Domain	Association	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓
Vision model	C	2D	IOU20	0.741	-	-	-	-
Vision model	C			0.445	0.439	0.827	0.557	0.315
Pointcloud model	L	3D	DIST2	0.733	0.608	0.524	0.492	0.556
Pointcloud model	R			0.324	0.358	0.811	0.613	0.395
CDSM Fusion	C+L			0.743	0.620	0.487	0.488	0.530
CDSM Fusion	C+R	3D	DIST2	0.523	0.486	0.703	0.551	0.393
CDSM Fusion (FT)	C+R			0.681	0.584	0.623	0.521	0.390

of fusion, with the fusion architecture outperforming camera-only and Radar-only predictions in terms of object detection accuracy and depth estimation, as shown in Figure 4.6. This analysis confirms the success of the fusion approach in significantly enhancing 3D object detection performance, demonstrating the potential of the proposed CDSM fusion method for low-level sensor data fusion in AV perception systems.

Figure 4.6: The comparison of the results for the same scene in the NuScenes test dataset, showcasing the outputs of the camera-only, radar-only, and CDSM fusion models from top to bottom. The presented image highlights the performance gain achieved through the fusion approach when compared to the single-sensor models.

4.3 Grad-CAM adaptation

State-Of-The-Art Neural Network architectures in camera image processing present a challenge due to their "blackbox" nature, lacking transparency. This issue is particularly critical in AD applications due to its regulations. To that end, XAI methods are being developed to provide insights into NN decision-making processes. While XAI methods for camera image processing and OD networks are well-established, applying them to sensor data from LiDAR and Radar poses challenges. Gradient-based methods, such as Class Activation Map (CAM) and Grad-CAM (Figure 4.7), offer visualization techniques to understand NN internal representations, even for convolutional models. These methods provide flexibility by calculating activation weights based on gradient values, making them adaptable to various network architectures and layers. This adaptability holds promise for future applications in understanding pointcloud data and fusion solutions.

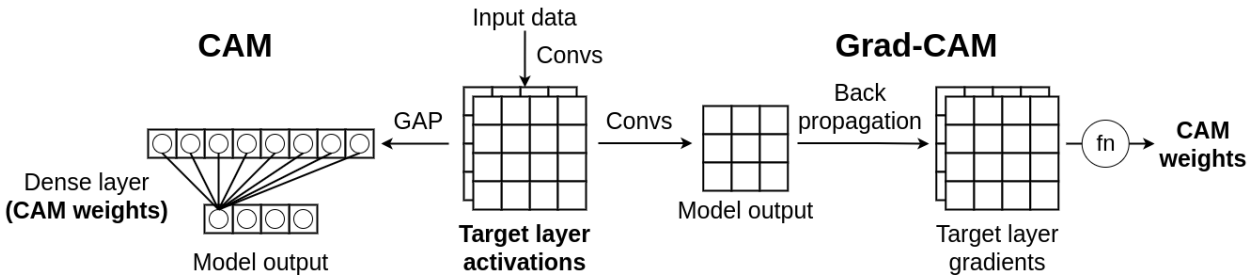

Figure 4.7: Comparison between CAM and Grad-CAM XAI methods.

The adaptation of the Grad-CAM method for visualizing models processing pointcloud data is a pivotal step towards enhancing the analysis of Lidar-only, Radar-only, and fusion solutions. Understanding the internal workings of these models is crucial, especially in domains like autonomous driving, where safety and regulatory compliance are paramount. However, due to the unique architecture and processing requirements of pointcloud data, traditional visualization methods encounter challenges in accurately representing model decisions.

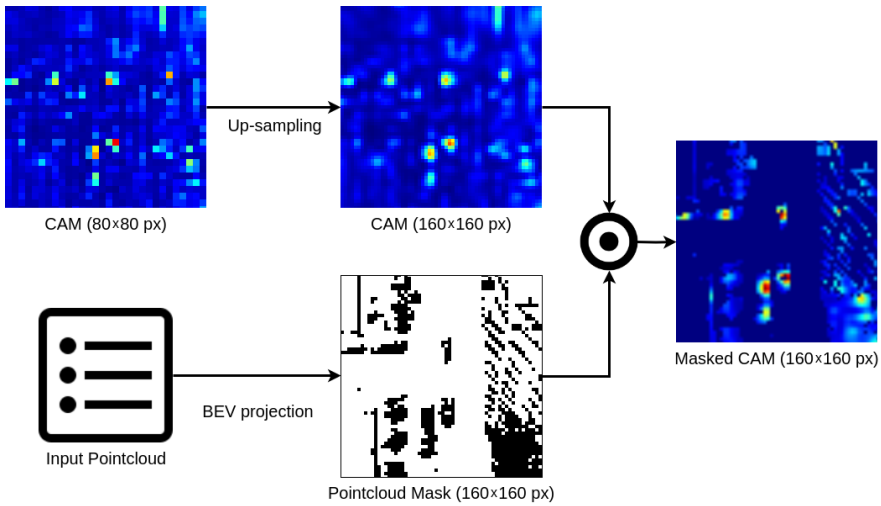

Figure 4.8: An example of generated CAMs and input pointcloud data combination.

The proposed approach addresses these challenges by generating CAM overlays within the BEV perspective. This shift in perspective necessitates careful consideration of how 3D pointcloud input can be integrated with BEV CAMs coherently. To reconcile disparities in resolution between CAMs and pointcloud data, a novel fused visualization method is introduced as shown in Figure 4.8. This method leverages pointcloud masks and up-sampling CAMs to enhance visualization detail. Through experimentation with various mask resolutions, a resolution of 160x160 pixels emerges as a compromise, striking a balance between readability and detail.

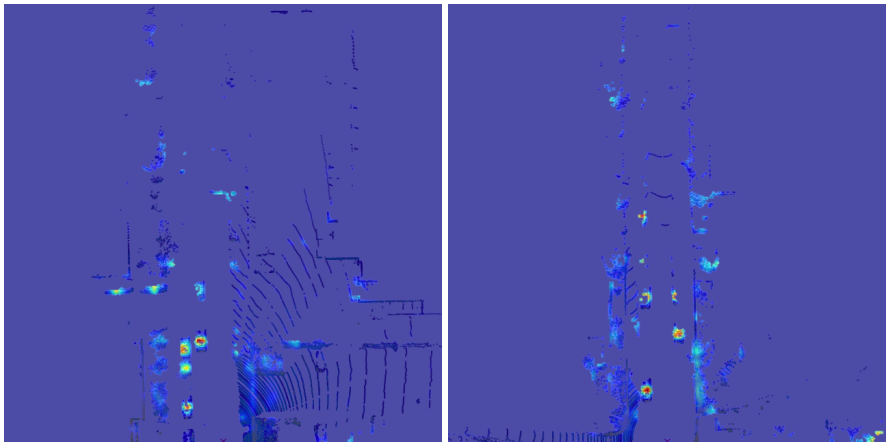

Figure 4.9: The ultimate Grad-CAM adaptation results; a high-resolution, clear, and noise-free CAM heatmaps, which effectively highlight essential areas in the input pointcloud overlay.

Furthermore, this approach not only improves visualization clarity but also widens the RGB spectrum range for relevant CAM values. By excluding certain parts of the activation before normalization, the resulting CAMs exhibit a richer spectrum, enabling more nuanced analysis. The final adaptation for LiDAR pointcloud networks combines voxel-wise processing with 2D Sparsity Invariant Convolutions. This advanced technique produces high-resolution, detailed heatmaps for object detection, free from disruptive noise, and suitable for comprehensive human visual analysis as presented in Figure **??**. In essence, this methodology represents a significant step forward in understanding and interpreting the decisions made by NN models processing pointcloud data, thereby enhancing their transparency and interpretability in critical applications like AD.

5. Summary and contributions

The thesis extensively explores various aspects of AV perception systems, focusing on the fusion of data from automotive sensors to enhance the accuracy and robustness of DL perception models. It introduces AD automation levels, defines perception systems, and outlines the roles of each sensor. It discusses sensor data fusion, encompassing different fusion levels and stages. A survey of OD NN, including single-sensor models and fusion architectures, is provided. Evaluation methods, such as perception KPI metrics and XAI techniques like Grad-CAM, are introduced to assess model performance. The core element of the thesis, the CDSM fusion method, is detailed, utilizing DL techniques for feature-stage LLF. CDSM integrates a novel domain alignment method and distinct fusion strategies. Experiments on two automotive open-source datasets validate the efficacy of the CDSM fusion method through visual and numerical evaluations.

The conclusions affirm the thesis's objective, demonstrating the potential of DL-based LLF solutions to advance AV perception systems. The CDSM fusion consistently outperforms single-sensor model architectures, with the fusion of camera and Radar showcasing significant enhancements. Notably, the fusion compensates for each sensor's weaknesses, providing synergistic improvements in perception outcomes.

The author highlights the following notable accomplishments throughout the thesis:

- The development and implementation of the CDSM fusion architecture is a significant contribution to the AV perception research domain. This LLF method offers a unique approach to aligning sensor data features from different domains, potentially applicable beyond fusion, such as in the presented 3D monocular camera architecture.

- The CDSM architecture introduces novel fusion techniques enabled by the domain alignment component. Among these techniques, the range-based approach, utilizing FOV-based features aggregation and refinement, demonstrates superior performance in terms of KPI metrics.

- The complete CDSM fusion architecture yields improved perception outcomes and could potentially serve as an alternative to current state-of-the-art approaches.

- The successful adaptation of the Grad-CAM analysis technique to pointcloud models addresses visualization challenges in complex and Radar models. This adaptation enhances interpretability, augmenting their applicability and development process.

- Research efforts resulting in scientific publications and patent applications underscore the practical relevance and industrial implications of ML and AV perception research.

Looking ahead, promising avenues for further exploration include data augmentation to boost model performance, utilization of raw Radar data for fusion, and extension of fusion XAI visualization techniques. By addressing these challenges, future research can innovate the domain of AV perception systems, building upon the foundations laid in this work.