ROADVIEW

Robust Automated Driving in Extreme Weather

Project No. 101069576

Deliverable 5.8

XAI design of perception solutions

WP5 – Robust perception system

Authors	Thanh Bui (RISE), Heikki Hyyti (FGI), Eren Aksoy (HH), Idriss Gouigah (HH), Jyri Maanpää (FGI), Abu Mohammed Raisuddin (HH), Hamed Ouattara (CE), Topi Miekkala (VTT)
Lead participant	RISE
Delivery date	29 August 2025
Dissemination level	Public
Type	Document, report

Version 1.0

Funded by the European Union

Innovate UK

Funded by the European Union (grant No. 101069576). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them. UK and Swiss participants in this project are supported by Innovate UK (contract No. 10045139) and the Swiss State Secretariat for Education, Research and Innovation (contract No. 22.00123) respectively.

Revision history

Author(s)	Description	Date
Thanh Bui (RISE)	Draft deliverable	16/07/2025
	Revision 1 (for internal review)	18/08/2025
Heikki Hyyti (FGI)	Review comments	23/08/2025
Tuomas Herranen (LUA)	Review comments 1st round	23/08/2025
Thanh Bui (RISE)	Revision	25/08/2025
Eren Aksoy (HH)	Review comments 2nd round	26/08/2025
Thanh Bui (RISE)	Final version (v1.0)	28/08/2025
Johannes Ripperger (accelCH)	Final checks and formatting	29/08/2025

Contents

List of Figures

List of Tables

Partner short names

HH	Halmstad University
LUA	Lapin Ammattikorkeakoulu
CE	CEREMA
RISE	RISE Research Institutes of Sweden
FGI	Maanmittauslaitos – Finnish Geospatial Research Institute
CRF	Canon Research Centre France S.A.S.
accelCH	accelopment Schweiz AG
VTT	Technical Research Centre of Finland

Abbreviations

ADAS	Advanced Driver-Assistance Systems
AI	Artificial Intelligence
AV	Autonomous Vehicle
CNN	Convolutional Neural Network
D	Deliverable
DL	Deep Leaning
EC	European Commission
EDA	Exploratory Data Analysis
EU	European Union
Grad-CAM	Gradient-weighted Class Activation Mapping
LiDAR	Light Detection and Ranging
LRP	Layer-wise Relevance Propagation
ML	Machine Learning
MSE	Mean Squared Error
ODD	Operational Design Domain
PCA	Principal Component Analysis
RTM	ResNet50-Truncated-MultiTasks
VAE	Variational Autoencoders
WP	Work Package
XAI	eXplainable Artificial Intelligence

Executive summary

Objectives

This deliverable presents the results of Task 5.5 “Trustworthiness and explainability of the perception modules”, completed within WP5 (Robust perception system). It defines the explainability requirement specifications and reports the implementation of XAI design for the AI models developed in Task 5.2 (Collaborative Perception Solutions Integrating Infrastructure-based Perception Systems) and Task 5.3 (Environmental and weather condition state estimator).

Explainable AI (XAI) techniques are crucial for ensuring the safety of Advanced Driver-Assistance Systems (ADAS) and Autonomous Vehicles (AVs). While deep learning (DL) powers core perception tasks, its opacity hinders understanding of why/how a module makes specific decisions. XAI methods increase transparency by revealing which features the DL component focuses on and the reasoning behind its predictions. This allows developers and other stakeholders to identify potential issues such as dataset or model limitations, and to provide the necessary information for safety assessments. By verifying that the perception module bases its decisions on truly relevant and safe features, we can significantly improve the trustworthiness and safety of AI-powered systems, ultimately building safer autonomous vehicles.

Methodology and implementation

This deliverable complements other reports on perception and weather estimation modules within WP5, specifically addressing trustworthiness and explainability. This work was conducted in close collaboration with other tasks within WP5 (especially T5.2 Collaborative perception and T5.3 Environmental condition estimator) to ensure the adoption of appropriate XAI techniques aligned with overall project requirements.

The work begins with a review of the latest standards, guidelines, and advancements in Explainable AI (XAI) techniques, with special focus on supporting the safety assurance process within road vehicle systems. Recognizing the breadth of XAI research and its limited focus on safety-critical applications, we categorized available techniques to facilitate their application in supporting trustworthiness principles throughout both the development lifecycle and operational context. Applicable methods will then be selected and implemented for various perception modules developed within the ROADVIEW project.

Outcomes

This work delivers a comprehensive understanding of Explainable AI (XA)I, establishing a foundation in principles, categories, and relevant literature/standards, defining and allocating safety requirements and detailed specifications for both data and model explainability, together with requirements for operational monitoring. This understanding is then practically applied to collaborative perception models and weather/road condition estimation models within WP5, differentiating between non-safety critical systems like weather noise removals, object segmentation, and safety-critical functionalities including slipperiness detection, visibility estimation, and weather assessment. Detailed implementation examples demonstrate how XAI principles are applied and validated, resulting in a fully documented report detailing the methodology, findings, and implemented solutions.

1 Introduction

This work builds upon state-of-the-art safety assurance processes to systematically derive safety-critical requirements for AI-based perception components. From the high-levels requirement specification specified by T2.4 (ROADVIEW requirement specification), we further defined explainability requirements linked to AI-intended functionality with focus on perception models within WP5, followed by allocated data requirements and model requirements. A key focus is understanding and mitigating risks associated with harsh weather conditions, building upon collaboration with other work packages (T4.3 for dataset and data fusion formats, T4.4 for data filtering model, T6.1 for the integration and trustworthiness requirements). This will be achieved through the application of data/model explanation techniques, including data statistical analysis, feature correlation, feature importance, and other advanced Explainable AI (XAI)methods.

To facilitate the safety assurance process, we evaluate and implement selected state-of-the-art XAI techniques to extract meaningful explanations from the AI modules developed within WP5 (perception and weather estimation), with a particular emphasis on understanding the impact of weather. Generated explanations can serve as evidence within the safety assurance framework. The proposed XAI approach encompasses both post-hoc explanations for “black-box” modules and intrinsic explanation for the “grey-box” modules developed within the ROADVIEW system architecture.

The work has been performed as collaborative work together with other tasks within WP5, and with related tasks in WP2/4/6, prioritizing “trustworthy by design” for AI perception models. It also incorporates noise filtered from raw sensor data to further refine explanation fidelity and collaborates with WP6 on how to support informed control decisions.

2 Relevant standards and best practices

Within the scope of ROADVIEW perception components, AI explainability (hereafter referred to as XAI) is used as a facilitator to build trustworthy systems, informed by best practices, standards, and guidelines. This section summarizes the latest relevant standards and best practices inspiring our work.

The EU AI Act [1] establishes a legal framework for the safe and ethical development of AI systems, categorizing risks and imposing strict requirements on high-risk applications – which cover the AI-based components in WP5. According to the AI Act, the related components are fall within the category «High risk AI systems», as safety components in road traffic. Required measures must be in place including risk management, data governance and record-keeping. Within the scope of ROADVIEW and this task, this reflects on the inspiration of using AI safety assurance frameworks that are built upon conventional functional safety standards and practices.

Several standards are also available on how to achieve explainability. ISO/IEC DTS 6254 [2] describes approaches for achieving explainability throughout the AI lifecycle as defined in ISO/IEC 22989 [3]. At the time of this writing, the standard is under final step of development (publication). The standard IEEE 2894-2024 [4] specifies an architectural framework that facilitates the adoption of XAI. IEEE SA P2976 [5] defines mandatory and optional requirements and constraints that need to be satisfied for an AI method, algorithm, application or system to be recognized as explainable, including partially/fully or strongly explainable methods, algorithms and systems. The standard ISO/IEC 22989:2022 [3] provides some standardization of key AI terms and concepts.

Standards on data and lifecycle management: (i) ISO/IEC 8183:2023 [6] provides a framework for data lifecycle management, (ii) ISO/IEC 42001:2023 [7] specifies requirements for establishing an AI Management System, (iii) ISO/IEC TR 24030:2024 [8], and (iv) ISO/IEC 5339:2024 [9] offer guidance on AI applications and stakeholder engagement, respectively.

Most relevant standards and guidelines for ROADVIEW are the safety-critical applications related standards including: (i) ISO/IEC TR 5469:2024 [10] guides AI integration into safety critical applications. This is complemented by established functional safety ISO 26262 [11] and ISO 21448 [12] (SOTIF – safety of the intended functionalities). ISO/PAS 8800:2024 [13] advances ISO26262 and ISO21448 with a more comprehensive AI-specific safety lifecycle,

and UL 4600 [14] reinforces the need for a robust safety case. The EASA ML Assurance Guidelines [15] and ISO/IEC TR 5469 are the two most relevant guidelines that inspire our work in this task.

While fully compliance with these standards is not the main intention, Task T5.5 leverages the key approaches in these developments and guidelines to inform our application of explainability to the perception AI components developed in WP5.

3 Literature review on explainable AI

Explainable AI (XAI) provides tools for building trust in AI systems, which is particularly important for safety-critical applications. However, these technologies weren't originally designed for this purpose, resulting in a need for a holistic approach to categorize XAI algorithms. This categorization will help select appropriate methods for specific use cases at each phase of the AI safety assurance lifecycle.

3.1 Categorization

XAI is an active research area offering numerous techniques to enhance the transparency of black box AI models. For the objectives of Task 5.5, we categorize XAI methods based on selected dimensions to systematically facilitate the identification of suitable candidates for each AI module within WP5.

XAI techniques (encompassing both intrinsic methods integrated within model architectures and post-hoc analyses applied to existing models) aim to generate interpretable explanations for dataset, data point or model predictions. To ensure the extracted explanations are useful for various purposes within the AI lifecycle, we propose the following identified dimensions and relevant subcategories:

- Target item to be explained: Dataset or AI model.
- Target Audience of the extracted explanation: User of the explanation (AI developers, data analysts, domain/safety experts, other technical component or end users).
- Explanation representation: Text/tabular data, graphs, or images.
- Scope of explanation: Local (single data point or prediction) or global (overall model behaviour).
- Timing Requirement: Time-critical or time-non-critical.
- Interpretability type: Intrinsic (inherently interpretable model, integrated XAI tools in the architecture) or post-hoc (model equipped with additional XAI tools).

These categories facilitate mapping potential XAI methods to phases of the AI lifecycle as in Table 1.

Table 1: Proposed AI candidate mappings.

Phase/activity	Target Item	Target audience	Explanation representation	Scope of explanation	Timing requirement	Interpretability type
Data Management	Dataset	Data analysts, Domain expert	Tabular, graph, images	Local/Global	Non-critical	Posthoc
Model learning management	AI model	AI developers, data analysts	All	Local/Global	Non-critical	Intrinsic
Model inference management	AI model	Safety expert	Tabular, graph	Local	Non-critical	Posthoc
Operation & monitoring	AI model, dataset	Technical component	Tabular, numerical	Local	Critical	Both

3.2 State of the art XAI methods

3.2.1 Data explainers

Explainable AI methods applied to data, referred to as "data explainers", can be used to analyse datasets or specific datapoints within a dataset to support data quality assessments and improvements.

Exploratory Data Analysis (EDA) is a standard approach to support data quality assessment. EDA can range from identifying class imbalances or missing/corrupt data to analysing inherent feature characteristics to understand their potential relevance – both from a domain expert's perspective and through model-driven insights. Numerous tools facilitate EDA, providing statistical summaries and visualizations. Popular options include yData Profiling [16] (formerly pandas profiling) [17], SweetViz [17], and Google Facets [18]. Effective visualizations like Parallel Coordinate Plots (PCP), Principal Component Analysis (PCA), t-distributed Stochastic Neighbour Embedding (t-SNE) [19], and Uniform Manifold Approximation and Projection (UMAP) [20] empower data analysts to uncover patterns and anomalies.

There are also frameworks to systematically evaluate data readiness. Data Readiness Levels (DRL) [21], inspired by Technology Readiness Levels, assess data usability, reliability, and completeness throughout its lifecycle, from collection to application, enabling proactive identification of gaps and improved data management practices.

Feature engineering and extraction are key aspects of data explainability. Domain-specific approaches leverage expert knowledge and EDA insights to identify meaningful features – for example, interpretable representations of camera image data can be created using features such as DAISY [22], HOG [23], Haar [24], LBP [25], CenSurE [26], ORB [27], Gabor [28], SIFT [29], Shape index [30]. Model-based techniques, such as clustering and dictionary learning [31], utilize mathematical models to uncover the inherent structure of the data.

Improving communication between data creators and users is vital. Initiatives promoting standardized dataset documentation, like Datasheets for Datasets [32], Dataset Nutrition Labels [33], and Data Declarations for Natural Language Processing [34], provide frameworks for documenting essential information, including development history, content, collection methods, and ethical considerations.

Since data dimensionality, especially that used within ROADVIEW, is very high, data summarization techniques are important. Prototype selection identifies a small, representative subset of data points (prototypes) that capture the essence of the larger dataset. Tools like ProtoDash [35] optimize for both relevance and diversity, providing a concise and insightful overview. Contrastive Explanation Method [36] (CEM) leverages unsupervised networks to learn from data and provide explanations using these prototype patterns.

Variational Autoencoders [37] (VAEs) offer another approach to data summarization and explainability. By learning a lower-dimensional latent space representation, VAEs can reconstruct input features and provide an interpretable summary of the underlying data structure. Furthermore, techniques like the Disentangled Inferred Prior Variational Autoencoder (DIPVAE) [38] aim to learn high-level, independent features with potential semantic meaning, further enhancing interpretability.

3.2.2 Model explainers

AI models often operate as "black boxes", hindering trust and adoption in safety critical applications. A substantial body of research has emerged focusing on techniques to explain and interpret these models, broadly categorized by their approach and goals. These methods aim to understand «how» models derive the predictions, identify influential inputs, and assess the reliability of predictions.

Feature importance and attribution methods are central to understanding the contribution of individual inputs to model outputs. These techniques highlight the most influential aspects of the decision-making process, bridging the gap between raw inputs and predictions. Methods in this category include gradient-based approaches like Saliency map[39], Integrated Gradient [40], Class Activation Map (CAM) [41], Grad-CAM [42], gradient based feature importance [43], DeepLIFT [44], SmoothGrad [45], Guided backpropagation [46], EigenCAM [47], and perturbation-based techniques (LIME [48], SHAP [49], Anchors [50], Feature occlusion [51], Randomized input sampling RISE

[52], Counterfactual [53]). Complementary approaches analyse feature space (interim feature extracted from specific key layer of a neural network) with data explainers as described in the previous section, with tools supporting the relationship between feature and output (Partial Dependence Plot (PDP) [54], Accumulated Local Effects (ALE) [55]), or dimension reduction tools to analyse the latent space structure.

More advanced semantic understanding methods strive to connect model representations to higher-level, human-interpretable concepts. Example-based methods, such as prototypes and criticisms (ProtoDash [35], ProtoPNet [56], MMD-criticism [57]) and counterfactual explanations[53], enhance the simulatability of explanations by illustrating how the model derives its predictions. Concept-based methods, including Concept Relevance Propagation [58], And-Or Graph (AOG) [59], further bridge the gap by mapping important features to specific concepts.

A proactive approach to explainability involves designing models that are inherently easier to understand. This can be achieved through simulatability, employing simplified models like Generalized Linear Models (GLM[60]), Generalized Additive Models (GAM [61]) Decision Trees (DT [62]), Decision sets [63], rule sets [64], [65], and Random Forests (RF [66]) that allow for mental simulation of the reasoning process. Algorithmic transparency can be incorporated by utilizing structured mathematical functions with clearly interpretable parameters, as seen in Tree regularization [67], ProfWeight [68], Teaching Explanations for Decisions (TED) [69], and Self-Explaining Neural Networks [70]. Decomposability, breaking down complex models into understandable submodules (ModelSpeX [71], ViT-NeT [72]), represents another effective strategy.

Rule extraction techniques transform the complex behaviour of neural networks into simpler, human-readable rule sets. Approaches include reverse engineering trained networks (Anchors [50], RxTEN [73], DEXiRE [74] (for binary classifiers), and ECLAIRE [75]) and training surrogate interpretable models (DeepRED [76], TrePAN [77], RuleFit [78], Two-step CNN rule extractor [79]). Neuro-symbolic approaches (Neural-symbolic learning [80], DL2 [81], DeepProbLog [82], Learning with logical constraints [83]) integrate neural networks with symbolic reasoning, while concept bottleneck models (Concept bottleneck [84], Concept whitening[85]) and attention mechanisms further refine the process.

Understanding if and when a model is likely to be incorrect is crucial for trustworthy AI systems when the results are used to inform safety-critical decisions. Uncertainty quantification techniques assess the confidence in predictions and identify potential edge cases. Bayesian neural networks [86] offer a probabilistic approach, while techniques like Dropout approximate Bayesian inference and provide estimates of prediction uncertainty [87]. Studies have shown a correlation between uncertainty estimates and prediction accuracy [88], demonstrating the potential of these methods to improve model reliability and enhance non-maximum suppression techniques. Furthermore, analysing aleatoric and epistemic uncertainties provides insights into inherent ambiguities in the data and potential model limitations.

Beyond post-hoc explanation and interpretable design, enhanced representation techniques modify model architectures to improve interpretability without complete redesign. Attention mechanisms, like Global-and-local attention (GALA) [89], DomainNet [90], Residual Attention [91], Loss-based attention [92], D-Attn [93], allow models to focus on relevant input parts. Autoencoder-based methods learn compressed, meaningful representations that can be analysed to understand feature importance. Other methods include Adaptive Deconvolutional Networks [94] and Deep Fuzzy Classifiers [95], all aimed at improving model transparency.

It's important to note that research in explainable AI is continuously evolving. Many techniques often overlap or are combined to achieve more comprehensive explanations. The choice of method depends on the specific application, the complexity of the model, and the desired level of interpretability.

3.2.3 XAI evaluation criteria

Besides providing an explanation, one should also assess its quality. Table 2 lists selected evaluation criteria and associated evaluation methods for assessing XAI approaches applicable in the context of ROADVIEW perception modules.

Table 2: XAI evaluation criteria.

Criteria	Description	Evaluation method
Faithfulness	Assesses how accurately the explanation reflects the 'real' reasoning of the model. A faithful explanation should not mislead users into believing the model used features it did not, or that those features had a different impact than they truly did.	Perturbation-based approaches (e.g., removing/adding important features and observing prediction changes), fidelity scores (measuring how well the explanation predicts model behaviour on held-out data), and concept ablation studies.
Robustness	A robust explanation should be less sensitive to small changes in the input or model.	Assessing explanation consistency across similar inputs (e.g., slightly perturbed sensor data), evaluating explanation stability with small model parameter changes, and measuring the variance of explanations across multiple runs with the same input.
Localisation	Granularity and specificity of the explanation. Ideal explanations pinpoint the precise input features and/or data points driving the model's prediction.	Qualitative inspection by domain experts, visualization of feature attributions (e.g., saliency maps), and other quantitative metrics measuring the spatial or temporal compactness of explanations.
Complexity	Explanations should be easily understood by the intended audience.	User studies, complexity of the surrogate model (e.g. number of branches in a Decision Tree or a Random Forest)
Axiomatic	Alignment of explanations with known domain knowledge or desired properties.	Formal verification techniques, logical consistency checks, and domain expert checks
Randomisation	Whether an explanation is based on the model's reasoning or is just random guess. A robust XAI method should produce meaningful explanations.	Comparing explanations generated from random inputs (where the model should produce random outputs) to explanations from real inputs. Statistical tests can determine if the explanations are significantly different.
Saliency map evaluations	Saliency maps visualize different aspects of input data with regards to the final model's predictions (such as neuron activation values, gradients, combination of those and backpropagation rules). An effective saliency map shall provide useful information to human observers (or use as a base for automatically detecting misbehaviours of the model).	Cropping or overlaying important regions based on the saliency map and measuring changes in the model prediction. Comparing with domain knowledge by human experts.
Uncertainty estimation fidelity	For XAI methods that provide uncertainty estimates, evaluating the calibration of these estimates is crucial.	Using calibration curve to visualize the relationship between predicted uncertainty and actual error rates. A well-calibrated estimator will accurately reflect the true level of uncertainty.
Expert feedback	The effectiveness of an XAI approach should be judged by domain experts	Qualitative assessments using expert feedback on the clarity, usefulness, and trustworthiness of explanations. Asking experts to perform tasks (e.g., diagnosing a specific model's behaviour) with and without the aid of explanations, and measuring performance improvements.

3.3 Uncertainty types

For safety-critical applications, proper management of AI model uncertainties is the key. We address this through a two-phase approach (corresponding to the SOTIF guidelines [12]) designed to minimize risk and ensure system safety w.r.t. expected safety goals.

- Development lifecycle phases: Identify and proactively reduce the reducible uncertainties. This encompasses three key phases: data management (ensuring compliance to safety allocated data quality requirements), model learning management (optimizing model architecture, and training processes, V&V activities to report performance variations), and model inference management (ensuring the performance drops by model optimization do not lead to compromising safety goals). This phase should also report safety boundaries where the system is verified as safe.
- Operation and Monitoring stage: At this stage, reported safety boundaries and estimated residual uncertainties will be used to ensure that the overall system will not enter hazardous situations within an acceptable risk limit.

This two-phase approach provides a systematic framework for managing AI model uncertainties throughout the entire lifecycle of a safety-critical application, ensuring both proactive risk reduction and effective mitigation of residual risks during operation.

We adopted the proposal by Brando et al. [96] classifying these uncertainties into three categories (expanding traditional classifications by adding data domain uncertainty):

- Data domain uncertainty: This arises when the training/validation/test datasets fail to adequately represent the full scope of the problem definition. This includes deficiencies in capturing the ODD parameters/boundaries as well as a lack of coverage for critical scenarios and stringent safety requirements. Insufficient or biased data can lead to models that generalize poorly to real-world conditions and exhibit unexpected behaviour in edge cases. Addressing data domain uncertainty requires careful consideration of data collection strategies, data preparation techniques, and V&V activities.
- Model epistemic uncertainty: This results from limitations in the model architecture itself, or suboptimal training procedures. It reflects the model's lack of knowledge regarding the true underlying relationship between inputs and outputs. Epistemic uncertainty can be reduced through techniques like architectural search, hyperparameter optimization, transfer learning, and active learning. Employing more advance models (such as pretrained world models) or incorporating domain-specific knowledge into model architecture design can also help mitigate this type of uncertainty.
- Irreducible aleatoric uncertainty: This represents inherent uncertainties within the problem itself, independent of the model or data. Examples include annotation errors within the training data, inherent noise in sensor readings, and unavoidable environmental impacts (such as harsh weather). Unlike epistemic uncertainty, aleatoric uncertainty cannot be eliminated during the development lifecycle. Instead, it must be monitored/quantified and accounted for in the system's decision-making process in the form of safety envelope supported by safety supervisors. Quantization of the uncertainties w.r.t. input data is thus of importance to support the operation safety architecture.

This classification serves as a foundation for selecting appropriate XAI solutions for the ROADVIEW perception modules to be applied at different phases of the safety assurance lifecycle. This also helps to implement SOTIF guidelines by identifying and reducing uncertainties that are reducible during the development cycle and monitoring residual uncertainties during the operation.

3.4 Data explainers within the ROADVIEW context

To ensure the reliability and safety of ADAS/AV systems, particularly within the context of the ROADVIEW perception modules, properly managing data domain uncertainty is important. This requires proactive activities throughout the data management phase of the development lifecycle, to confirm that defined safety goals, ODD, and critical operational scenarios are consistently supported by the training/ validation and test datasets. The activities include data collection, data annotation, data preparation, and data verification. Data explanations play a crucial role in these activities.

Data explanations enable the following capability:

- Verify the conformity of collected and annotated datasets to the derived data requirements, ensuring adequate coverage of required scenarios, ODD, and safety goals.
- Assess the impact of dataset preparation activities (such as weather denoising, augmentation, or synthetic data generation) on the dataset's compliance to allocated data requirements.
- Identify key features represented within the datasets through data mining techniques, determining their usability for both domain/DL experts and subsequent DL models.

To support these usages, a range of data explainer techniques are employed:

- Data profiling: Calculates descriptive statistics across each input dimension. For example, comprehensively analysing data point distributions across weather parameter dimensions.
- Statistics of annotated data: Data analysis methods to understand the distribution of positive and negative samples across spatial, temporal, and other relevant input dimensions. More advanced data mining techniques are leveraged to discover important features, representative prototypes and criticisms within the dataset, providing valuable insights into its overall structure and content.
- Data descriptors: Utilizing statistical methods or Machine Learning (ML) approaches to summarize the dataset's characteristics. For example, employing Autoencoder families of DL to synthesize representative 'valid' data samples. Such data descriptors can later be used within the Operational stage to verify if runtime input data may be out of distribution resulting in operational risks.

Application of data explainers in ROADVIEW context:

- The required scenario parameter range and expected distribution for harsh weather conditions must be specified, and the collected/prepared datasets mush adequately cover and represent these scenarios.
- Data explainers can help to identify potential risks of bias, unbalanced, or lacking important data features. For example, if sensor data does not provide information that is relevant for accurate predictions.
- Measuring the sensitivity of AI model predictions when changing weather conditions (e.g., snow to heavy snow) or helps understand the model's behaviour under various harsh weather conditions.
- The results of dataset analysis can be used in conjunction with model explainers to support the diagnosis of AI models and identify potential epistemic uncertainty, which can be reduced by analysing if the model uses key features from the dataset and if these features are well-represented or missing.

3.5 Model explainers within the ROADVIEW context

Since WP5 perception modules are working with high dimensional timeseries data, the most applicable XAI methods are mainly falling in the following categories:

- Feature importance and attributions: Understanding the model predictions and its uncertainty (e.g., bounding box of a detected object within a data frame) w.r.t. located key features represented in the input data.
- Intrinsic interpretable: Where applicable a modular design approach should be adopted. Decomposing a monolithic network into semantically meaningful submodules (responsible for a specific subtask) will simplify training, facilitate targeted verification, and improve overall understanding. The modularity allows for independent testing and validation of individual components, reducing the complexity of the overall system. The transparency of these submodules can be achieved with post-hoc XAI methods such as integrated saliency map exporter function, which can highlight the contributions of specific submodules to the final decision. The combination of architectural design and post-hoc analysis provides the required transparency.
- Uncertainty-aware models: For safety-critical modules, the knowledge of model uncertainty w.r.t. specific input is crucial for the decision-making module to make informed safe decisions. It is therefore recommended for the safety-critical modules within WP5 to provide uncertainty estimates together with their predictions as an additional information for the decision-making module (WP6).

3.5.1 Feature importance and attributions

These methods determine the positive or negative influence of each input feature on the model's predictions, measured against the task's performance metrics. This involves carrying out what-if tests that quantify how the prediction changes when a particular feature is present or omitted. The attribution methods are most naturally applied to low-dimensional inputs. Since ROADVIEW perception subsystem works on high dimensional data (sensor data), we first perform an interim feature-extraction step that maps the raw sensor data into a lower dimensional representation suitable for attribution analysis.

Common strategies include:

- Super pixels: employed by techniques such as LIME or SHAP to cluster adjacent pixels via unsupervised image segmentation methods, producing compact homogeneous regions that serve as interpretable input features.
- Object part detections: leverage explicit spatial relationships and logical rules derived from the detected key parts of an object, enabling feature attribution at the level of semantically meaningful parts.

Saliency maps provide a pixel or point-wise explanation of a model's prediction by overlaying a heatmap on the original input. The heatmap value for each location is derived from the model's internal signal values; the exact choice of such signals determines what aspect of the model's reasoning is revealed.

Typical sources of signal per layer include:

- Neuron activations: The magnitude of the feature map at that layer, projected back to the input coordinates
- Gradients: The sensitivity of the output to perturbations of the input, indicating how a value change at that location would affect the prediction
- Layer type: Informs the weighting or propagation rule applied to that layer.

By aggregating these three sources into a single value per pixel or point, different attribution techniques generate distinct heatmaps, each highlighting a different facet of the model's reasoning. Among these, Grad-CAM, which fuses gradients with activations, and Layer-wise Relevance Propagation (LRP), which applies layer-specific propagation rules, are the most widely adopted.

3.5.2 Intrinsic interpretable

Decomposition: Design the model as a composition of conceptually distinct modules. Treat each sub-module as a sub-model and apply data explainers to its input, output for partial verification. This modularity increases the model transparency by design.

Disentanglement: There are different ways to disentangle the extracted feature space. The subspaces can be disentangled by e.g. weather parameters (input) or by model performance metrics (output).

Uncertainty aware models: Core techniques include injecting stochastic perturbations at multiple stages of the processing pipeline (input data or key network layers) to capture model epistemic uncertainty, and modelling aleatoric uncertainty by adding a dedicated estimator head that learns the residual between the prediction and the ground truth. This head accounts for errors arising from noisy sensor measurements or erroneous annotations. In ROADVIEW, such uncertainties must be attached to safety critical output dimensions. For example, providing a probabilistic confidence interval for each detected object's 3D position so that the decision-making module can make informed decisions within acceptable risk level.

4 Explainability requirement specifications

4.1 Objectives and scope

The overarching goal is to ensure transparency, interpretability, traceability, and verifiability of AI systems for diverse stakeholders including AI developers, domain experts, and safety engineers throughout the lifecycle.

The scope is defined by three key dimensions:

- Perception module safety critical levels: Requirements vary based on whether a module is safety-related (directly impacting safety decisions) or non-safety-related (no safety decisions are automatically made).
- Lifecycle Phase: Explainability requirements across 3 phases of the lifecycle
    - Data management: Reducing domain uncertainties by assessing the dataset with data explainers. This may also include the use of model explainers where applicable.
    - Training management: Intrinsic interpretability design, for example domain-based decomposition of model architecture, integrated explanation extractors. Explanations are needed to argue safety w.r.t. model overall behaviours and local behaviours. For the case of ROADVIEW, explanations regarding impacts of weather parameters are relevant. Uncertainty estimates of the trained model w.r.t. individual input shall be made available for later use in Operation stage.
    - Inference management: Understanding input area where optimized models (from the trained models in the previous step) faces significant performance drops, how to predict and if/how the model can be redesigned/retrained to eliminate the potential issues related to safety goals.
    - Operation and Monitoring stage: Monitoring model behaviour and identifying potential anomalies or failures.
- Perception module objectives: Specific explainability targets will be defined based on the objectives of the perception modules within ROADVIEW.

General Objectives:

- Traceability: Establish an auditable chain of evidence linking model detections (outputs) to the model's internal workings and ultimately back to the input data used for training and operation.
- Human-in-the-Loop: Enable human actors to effectively perform their tasks within the AI system lifecycle, including diagnosis of errors, generation of reports, and validation of system behaviour.
- Compliance: Align with the latest developments and standards in functional safety for AI systems, ensuring compliance with relevant regulations and best practices.

4.2 Allocated safety requirements

The primary safety goal within ROADVIEW is to ensure the system operates within acceptable safety limits considering anticipated degradation of perception performance in harsh weather conditions.

The overall system safety goal is then further allocated to safety related requirements for ROADVIEW developed components as follows:

- Perception components:
    - The system shall monitor the performance of all critical perception components (sensors and perception modules).
    - The system shall detect and quantify the level of performance degradation in perception systems due to harsh weather conditions.
    - The system shall provide timely alerts to the vehicle control system when perception performance falls below predefined safety thresholds.
    - The system shall differentiate between temporary and persistent perception degradation.
- Weather condition estimators:
    - The system shall accurately identify weather conditions connected to ODD environmental parameters and ranges, including low visibility (fog, rain, snow) and slippery road surfaces (ice, snow, wet)

As suggested by relevant guidelines (AMLAS, EASA, AI-FSM), the requirements are broken down into requirements for the datasets and requirements for the models

- Allocated to data requirements: The datasets shall maintain quality and diversity that allow for training and testing of perception modules under various weather conditions including harsh weather.
    - Data completeness: The dataset must include sufficient coverage of all object categories relevant to ODD and variations of weather conditions (sunny, rainy, snow, fog) that may affect object recognition performance.
    - Data representativeness: The dataset shall be representative of real-world scenarios, representative distribution of perception scenarios within each weather conditions.
    - Data balance: The dataset shall maintain class balance under each weather parameter range.
- Allocated to model requirements: The trained/optimized model(s) shall maintain performance and robustness that allow decision making module to make informed decisions.
    - Model performance: The model's global behaviours shall correctly model the stated problem
        - Accuracy: The model prediction errors must be within acceptable safe limits.
        - Timeliness: The latency between sensor data capture and output predictions shall be within a predefined time windows.
        - Uncertainty: The model shall provide uncertainty estimate for every prediction it makes.
    - Model robustness: Model shall maintain its required performance/behaviour across all possible data points within the valid data space spanned by ODD/operational scenarios input parameters.
        - Robustness w.r.t. weather parameters: Model performance degradation resulted by weather impacts shall be within a predefined safe limit.
        - Robustness w.r.t. weather and other expected environmental noise: Performance degradation shall be within a safe limit when input noise is within the ODD.
        - Anomaly detector component shall be able to predict whether the model predictions are trustworthy enough (input data, model inner working, output distribution are within the «known» area)

4.3 Requirement specifications

4.3.1 Data explainability requirements

The data management phase includes data requirement specification, data collection, data preparation, and data verification. Data explainability supports verification and, importantly, can guide collection and preparation activities.

Based on these considerations, the following data explainability requirements are specified:

- DAT-EXP-REQ1: Datasets shall be designed to be explainable, enabling verification of completeness, representativeness, and balance with respect to defined data requirements.
- DAT-EXP-REQ2: Individual data points within a dataset shall be explainable when required, to evaluate the presence of key features relevant to AI model prediction accuracy.
- DAT-EXP-REQ3: Data explainers shall facilitate the development of methodologies to determine the membership of new data points within a given dataset.

The following data explainers can be leveraged:

- Dataset Statistics: Provides a summary of dataset characteristics, including key data quality measures such as completeness, accuracy, relevance, and representativeness. These statistics are crucial during data collection and preparation to assess data quality w.r.t. data requirement specification and identify potential improvements.
- Data Point Statistics: Calculates statistical figures for multidimensional data points (e.g., data samples from vehicle sensors). This includes analysing image features or 3D point cloud characteristics, helping to understand the distribution and patterns within individual data points.

- Data Descriptors: Employs statistical or machine learning-based dimension reduction techniques (e.g., PCA, VAE) to represent data points and datasets with lower-dimensional Euclidean vectors. These descriptors enable efficient data comparison and analysis of the dataset representation space.
- Data Anomaly Detection Mechanisms: Leverages data descriptors to distinguish between valid and anomalous data points. By assessing the probability of a new data point belonging to a known dataset based on its descriptor, these mechanisms help ensure data quality during verification and during Operation & Monitoring stage.

4.3.2 Model explainability requirements

Model learning and inference management phases include model selection, design, training, verification, and optimization. AI model explainability is key to understanding model behaviour globally and locally, and for managing model epistemic uncertainty.

AI model explainability requirements are specified as follows:

- MOD-EXP-REQ1: Where applicable, AI models shall be designed with a modular architecture where each submodule performs a distinct, understandable function.
- MOD-EXP-REQ2: The boundaries of the AI model (inputs, outputs, and model limitations) shall be clearly defined and validated.
- MOD-EXP-REQ3: The AI system shall include mechanisms to detect and highlight undesirable bias in the model, including saliency maps composed from neuron activation maps and gradient maps where applicable.
- MOD-EXP-REQ4: The AI system shall support the target audience in identifying errors or potential flaws in the model design, training and optimization.
- MOD-EXP-REQ5: Extracted explanations shall be useful for the target audience to perform their tasks at any stage of the lifecycle, considering assumptions made on the level of qualification and skills.
- MOD-EXP-REQ6: Explanations shall support traceability, allowing diagnosis of a case from prediction all the way back to the input data.

The following XAI techniques can be used:

- Modular model design: decompositional design approach, building the model with distinct, understandable submodules using domain knowledge.
- Disentanglement: utilizing a disentanglement approach that forces the model to predict interpretable concepts (e.g., weather conditions) before deriving the final prediction.
- Feature importance (w.r.t. model performance): quantifying the impact of features by measuring the change in model performance when their values are altered. Specifically analysing the impact of weather conditions on the model performance w.r.t. specific set of operational scenarios.
- Feature Importance: employing techniques like LIME, SHAP, and saliency maps to visualize the influence of input features on model predictions.
- Robustness testing: assessing model stability by introducing random or handcrafted noise to input data and observing the resulting changes in predictions.
- Model internal working: examining neuron activation maps and/or gradient maps to gain insights into the model's internal processing.
- Interpretable surrogate models: approximating the complex AI model with a simpler interpretable model (e.g., a rule-based model) to explain its behaviour.

4.3.3 Operation and Monitoring requirements

During the Operation and Monitoring stage, the system shall operate within verified known conditions, maintaining safety through defined boundaries and proactive monitoring. These boundaries include:

- Known Inputs: The system shall only process inputs within the defined and certified input space.
- Predictable behaviour: The system's behaviour (e.g. model neuron activations) shall remain predictable and within established boundaries.

- Expected outputs: The system shall generate outputs within the defined and certified output space.
- Controlled uncertainty: Residual uncertainty shall be continuously monitored and maintained within safe operating boundaries.

The explainability requirements for Operation and Monitoring (OM) stage are thus specified as:

- OM-EXP-REQ1: The AI system shall support continuous analysis of the model's behaviour in production to detect anomalies and unexpected patterns.
- OM-EXP-REQ2: The AI system shall support safety investigations of hazards where the AI model is involved, providing access to relevant data and explanations.
- OM-EXP-REQ3: The AI system (safety-critical modules) shall provide uncertainty estimates for each prediction to inform inherent limitations of AI models and provide valuable information for decision-making, allowing to assess the reliability of predictions.
- OM-EXP-REQ4: Where applicable, a safe and interpretable surrogate model shall be provided as a safety component within the operational architecture, acting as a fail-safe and provide a fallback mechanism in critical scenarios.

The following XAI techniques can be used:

- Data descriptors (ML-based): use trained Variational Autoencoder (VAE) or other dimensionality reduction techniques to create bi-directional mappings between high-dimensional input data and a lower-dimensional Euclidean feature space. This enables out-of-distribution detection for inputs, predictions, and internal neuron activations.
- Uncertainty estimation: re-design the model to provide epistemic and/or aleatoric uncertainty estimates alongside its predictions. This can be achieved through architectural modifications or by training a separate estimator model.
- Certifiable surrogate models: train a simpler, certifiable surrogate model during development cycle to approximate the main model's behaviour, serving as a fail-safe mechanism.
- Ensemble aggregation with statistical testing: integrate diverse redundant model predictions, uncertainty results, and anomaly scores using ensemble methods and statistical testing to produce a consolidated output (including prediction, uncertainty, and trustworthiness score) for decision-making module.

4.4 Proposed compliance approach to Explainability requirements

4.4.1 Dataset explainability

DAT-EXP-REQ1: Datasets shall be designed to be explainable, enabling verification of completeness, representativeness, and balance with respect to defined data requirements.

This requirement imposes the need to:

- Annotations: Every data point must be annotated with all mandatory ODD attributes and scenario parameters defined in the data specification.
- Representation space: Each annotated point is stored as a vector where each dimension corresponds to a distinct ODD or a scenario parameter.

Applied explainability techniques: Data explainer exported statistical summaries that assess distributional properties:

- Minimum, maximum, mean, and standard deviation per dimension.
- Correlation/interaction matrices.
- Histogram/PDF plots.
- Outlier/anomaly status.

4.4.2 Data point/sample explainability

DAT-EXP-REQ2: Individual data points within a dataset shall be explainable when required, to evaluate the presence of key features relevant to AI model prediction accuracy.

This requirement imposes the need to:

- Unsupervised feature extraction: derive a set of representative low-level features from every data point (a sensor reading at a timestamp).
- Domain expert validation: Extracted features shall be presented to domain experts to assess whether they are relevant to the stated problem.

Applied explainability techniques: Data explainers

- Prototypes/criticisms.
- Traditional signal processing algorithms to extract basic features.

4.4.3 Dataset membership

DAT-EXP-REQ3: Data explainers shall facilitate the development of methodologies to determine the membership of new data points within a given dataset.

This requirement imposes the need to:

- Out-of-Distribution (OOD) Detector: ML-based function to determine whether the data point belongs to a dataset.
- Membership criteria: The OOD detector must evaluate the data point based on
    - Distance based: Distance or similarity metrics to a known data subcluster.
    - Descriptor based: Use dimension reduction model (e.g. Variational Auto Encoder VAE) to map a high dimensional data point to a low dimensional latent vector. The reconstruction fidelity is used to determine whether the data point belongs to the dataset.

Applied explainability techniques:

- Clustering: Group the dataset into clusters (e.g. k-Means [97], DBSCAN [98]…) and compute distance metric (e. g Mahalanobis, cosine) for each new data point to the nearest cluster. The distance value and thresholding will inform the membership assessment.
- Data descriptors: The reconstruction error (e.g. perception based or MSE) can be validated against the accepted tolerance to inform membership assessment.

4.4.4 AI model modularity design

MOD-EXP-REQ1: Where applicable, AI models shall be designed with a modular architecture where each submodule performs a distinct, understandable function.

The following XAI by design principles shall be applied to increase the transparency of AI models:

- Modality based feature extraction: Backbone layers that perform feature extraction per input modality.
- Task based prediction heads: Separate and task specific branches taken input from the shared extracted features and derive task specific predictions.
- Latent/feature disentanglement: Where applicable, the model shall apply disentanglement techniques to partition the latent space into subspaces aligned with ODD parameters or estimated model's performance.

4.4.5 Model safe boundaries

MOD-EXP-REQ2: The boundaries of the AI model (inputs, outputs, and model limitations) shall be clearly defined and validated.

Definition of model safe boundaries:

- Inputs: The space spanned by admissible sensor responses.
- Outputs: The distribution of valid prediction values/sets of values.
- Model: Known activation patterns of neurons.

During the development lifecycle, data descriptors (e.g., VAE based descriptor) shall be derived to validate:

- If the input data is known (data point likely belongs to the admissible space spanned by the training/validation datasets)
- If the output prediction is known (output likely belongs to the expected output distribution).
- If the neuron activation patterns at key network layers belong to the known space of patterns (extracted from the trained model when training/validation datasets are used as input).

These descriptors can be used to describe the safe boundaries (known area) similar to ODD detector described earlier, following recommendation from SOTIF guidelines.

4.4.6 Saliency map

MOD-EXP-REQ3: The AI system shall include mechanisms to detect and highlight undesirable bias in the model, including saliency maps composed from neuron activation maps and gradient maps where applicable.

Perception AI models are typically CNN based with spatial operators. The saliency, activation, and gradient maps must be generated on-the-fly as part of the model method design. They serve as explainability functions to extract required visual information with insights on spatial relationships. Recommended methods include Grad-CAM, EigenCAM and integrated gradients.

4.4.7 Model diagnosability

MOD-EXP-REQ4: The AI system shall support the target audience in identifying errors or potential flaws in the model design, training and optimization.

To detect errors and flaws, explanations can be extracted from the model including:

- Confidence and uncertainty of a prediction.
- Model diagnostic: Activations/Gradients at a specific layer (can use L1 norm for aggregation). Assessment of clipping or vanishing/exploding.
- Training/validation curves per loss component during training runs.

4.4.8 Explanation usability

MOD-EXP-REQ5: Extracted explanations shall be useful for the target audience to perform their tasks at any stage of the lifecycle, considering assumptions made on the level of qualification and skills.

Focus audience within the scope of this document includes:

- Data analyst/Domain expert: Data explainer reports including prototypes/criticisms.
- AI developers: Data explainers, saliency maps, data summary of interim feature space.
- Machine (decision-making module): Uncertainty estimates, anomaly detection scores, ensembles of diverse redundant AI models.

4.4.9 Traceability

MOD-EXP-REQ6: Explanations shall support traceability, allowing diagnosis of a case from prediction all the way back to the input data.

The use of feature importance XAI methods in connection with data explainers, data confidence level logging will facilitate end to end traceability of a specific prediction.

4.4.10 Runtime behaviour anomaly detection

OM-EXP-REQ1: The AI system shall support continuous analysis of the model's behaviour in production to detect anomalies and unexpected patterns.

Out-of-distribution detector models (such as descriptor based or distance based as described in Section 4.4.3) can be used as runtime anomaly detectors to validate model behaviour via its neuron activation patterns.

4.4.11 Diagnosis of hazardous cases

OM-EXP-REQ2: The AI system shall support safety investigations of hazards where the AI model is involved, providing access to relevant data and explanations.

This requirement can be supported using various XAI methods on logged hazardous case data.

4.4.12 Runtime uncertainty quantification

OM-EXP-REQ3: The AI system (safety-critical modules) shall provide uncertainty estimates for each prediction to inform inherent limitations of AI models and provide valuable information for decision-making, allowing to assess the reliability of predictions.

Safety critical models should be designed to provide estimates of uncertainty for every prediction. Practical methods include:

- Ensemble of MC Dropout models: Run several forward passes with random dropout enabled and aggregate the resulting outputs. The standard deviation can be used as uncertainty estimate.
- Dedicated uncertainty head: Adding an extra uncertainty estimation head to predict the irreducible prediction error as a function of ODD or scenario parameters.

4.4.13 Runtime safety backup component

OM-EXP-REQ4: Where applicable, a safe and interpretable surrogate model shall be provided as a safety component within the operational architecture, acting as a fail-safe and provide a fallback mechanism in critical scenarios.

The surrogate model must proceed the same sensor data format used by the primary perception system and generate predictions that match the primary model's output space. It can be chosen from a family of interpretable models, such as regression, decision trees, or random forests.

Since primary perception models operate on high dimensional sensor data, a safe (expert certifiable rule-based) and interpretable surrogate model must be built with the help of a dimension reduction model to transform raw sensor input into a compact vector space. The surrogate model consumes this reduced representation to produce the same prediction while maintaining its transparency and interpretability.

Traditional feature-extraction techniques (e.g. PCA, SIFT, HOG) can be employed to reduce the dimensionality of the sensor data.

5 XAI method candidates for ROADVIEW perception

ISO/IEC TR5469 describes different AI usage levels:

- AI usage level A1: AI technology is used, and it is possible to make an automated decision of the system function using AI technology.
- AI usage level A2: AI technology is used, but it is not possible to make an automated decision of the system function using AI technology (e.g., AI technology is present in the system for diagnostics).
- AI usage level C: AI technology is not part of a safety function but can have an impact on it.
- AI usage level D: AI technology is not part of a safety function and does not have an impact on it due to sufficient segregation and behaviour control.

Similar classifications have also been defined by the EASA AI roadmap [99]:

- EASA Level 1: AI used for assistance to human (human augmentation or cognitive assistance in decision and action selection).
- EASA Level 2: Human-machine teaming (cooperation or collaboration).
- EASA Level 3: Autonomous AI-based decisions and actions.

We leverage these classifications and define the two classifications of AI-based perception modules within ROADVIEW as non-safety-critical modules (AI usage level C-D/EASA level 1-2) and Safety critical modules (AI usage level A1/EASA level 3)

Explainability requirements for perception models within WP5 are categorized into two groups with different audience and usage specifications:

- **Non-safety-critical modules**: Modules where the output predictions do not directly cause impact on decision making that may lead to hazards (harmful to human). For these models, the strict requirements related to risk estimates are not mandatory. The main usage is for improving the performance and robustness of the model w.r.t intended functionalities.
- **Safety critical modules**: Modules that provide outputs as important factor for decision-making module (WP6) to plan or decide vehicle manoeuvre's sequences. The explanations shall be used by the supervisor module (machine audience) to justify the trustworthiness. Requirements include real-time requirements and shall be driven by selected metrics allocated by the safety requirements of the system.

5.1 Common XAI method candidates

For both types, data explainers are needed to ensure that the training and verification of modules are performed with a controlled level of domain uncertainty (data is complete, relevant, and representative). For ROADVIEW specific requirements, it means the following:

- Dataset shall have data samples over defined weather parameter ranges as set forth in ODD:
  - Snow levels
  - Fog levels
  - Rain levels
- The dataset shall include positive and negative (data points falling out of the ODD ranges) samples.
- Data (augmentation, denoising, synthetic) shall be of the types that are relevant as captured by the sensors (mounting position, projection, data format, uncertainty).
- Target object appearance in the dataset shall be representative of the operational scenarios, that include the representative distribution of position, size, aspect ratio, and number of objects in the scene.

Applicable data explainer techniques:

- Data prototypes: The prototypes are extracted from the dataset, with the assumption that they are well presented, of proper quality, related to the problem, and are used by the trained models
  - Model-agnostic data prototypes by extracting representative smaller patches from datasets that are most represented in the dataset.

- Model-specific data prototypes by projection of e.g., CNN conv operators back into the input data space and find the matched patches.
- Data descriptors: Feature extraction techniques that perform data dimension reduction, use to assumably describe the dataset for validity and thus for OOD.

5.2 Specific XAI method candidates for non-safety critical modules

Explainability for the non-safety-critical perception modules developed in WP5 is intended mainly for AI developers. The primary use cases are (i) validating model behaviour against expert knowledge and (ii) verifying the quality of training, validation, and test datasets so that they satisfy the functional and safety requirements of the system. Table 3 lists the XAI method categories selected to support modules such as object detection, free-space detection, and collaborative perception. The categories highlighted in green indicate the recommendations.

Table 3: XAI method mappings for non-safety critical modules

XAI dimension	XAI method categories			
Target item	**Dataset**	**Trained model**		
Target audience	**AI developers**	**Data analysts**	**Domain/safety experts**	End user
Explanation representation	Text	**Tabular/graph**	**Image**	
Scope of explanation	**Local**	**Global**		
Timing requirement	Time critical	**Time noncritical**		
XAI approach	**Intrinsic**	**Post hoc**		

Proposed XAI Techniques:

- Saliency Maps (e.g., CAM, LRP): Visualizing the input data regions most influential to model predictions.
- Decomposition/Disentanglement: Identifying feature subspaces relevant to specific input dimensions (e.g., Operational Design Domain parameters) or performance metrics.
- Ablation/Sensitivity Analysis: Assessing model dependency on input features by systematically removing simulated weather noise and observing its impact on object detection accuracy.

Contrastive techniques to scale the heatmap by 2 classes can be of interest to investigate. An example of YOLOv8, Contrastive LRP is provided in Figure 1.

The contrastive representation will highlight key features of the target class and compare the representation of the same features in the contrastive class: relevance score of target class - relevance score of contrastive class, masked by thresholding relevant scores of target class. This can help to investigate bias between the two classes when shifting focus.

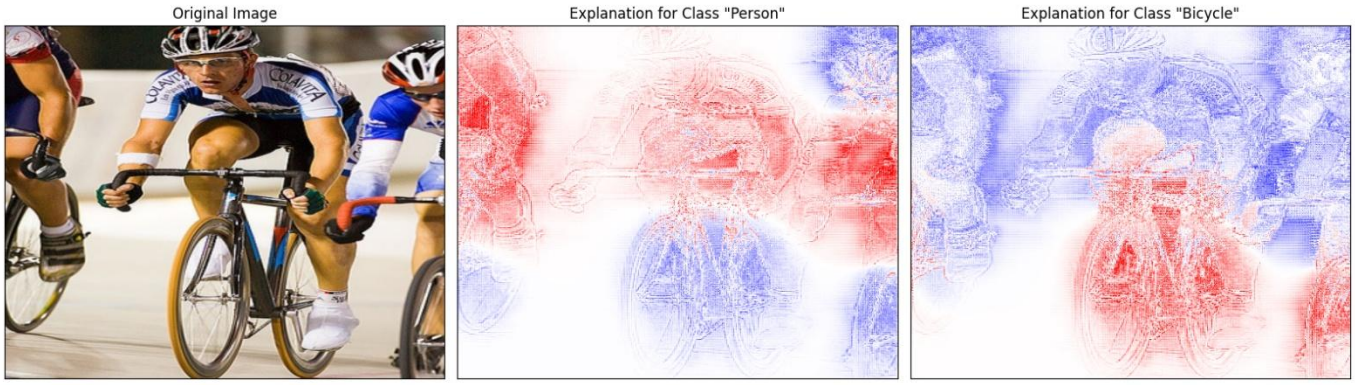

Figure 1: Example of contrastive LRP.

5.3 Specific XAI method candidates for safety-critical modules

Within ROADVIEW, safety-critical modules are those whose outputs directly inform safety-critical decisions, such as degrading the ODD based on weather awareness. These modules include weather parameter estimators and estimators for slipperiness and visibility.

To ensure trustworthiness, the following requirements are established: Modules shall provide explanations as an additional information dimension (separate from the prediction itself) for subsequent modules to verify the prediction's reliability.

Table 4 lists the XAI method categories selected to support safety-critical modules. The categories highlighted in green indicate the recommendations.

The following XAI techniques are therefore suggested:

- Uncertainty estimates: Providing epistemic and aleatoric uncertainty estimates with regard to key performance indicators (KPIs).
- Activation/gradient extraction: Enabling a supervision module to monitor behavioural anomalies.

Table 4: XAI method mappings for safety-critical modules.

XAI dimension	XAI method categories			
Target item	Dataset	**Trained model**		
Target audience	AI developers, data analysts	Domain/safety experts	End user	**Decision making module**
Explanation representation	Text	**Tabular, numeric**	Image	Graph
Scope of explanation	**Local**	Global		
Timing requirement	**Time critical**	Time noncritical		
XAI approach	**Intrinsic**	**Post hoc**		

6 Explainability in ROADVIEW perception modules

This section represents how we apply the proposed approach, as discussed in the previous section, into ROADVIEW perception AI models as a proof of concept.

6.1 Reference architecture

Figure 2 depicts how XAI is integrated into the reference architecture of the ROADVIEW system described in WP2. The core idea is that every stage that transforms raw sensor data or interim processed data into another type of data for the next processing step (including model predictions) should expose required explanations.

Data explainers are recommended to be applied for all components in the data-processing stack to enable end-to-end traceability of data:

- Sensor Calibration and Synchronization: Calibration routines (e.g., extrinsic alignment, time-stamp correction) are often tuned manually or by heuristics. Data explainers can analyse the correlation between each calibration parameter and the final prediction errors.
- Data Filtering: Data explainers can report the impact of the filtering process on dataset quality.
- Low-level Sensor Fusion: Multimodal sensor fusion (camera + LiDAR + radar) involves weighting schemes and feature-level alignments that depend on the selected projections, such as cylindrical view or bird-eye view. To enable data traceability, the projection errors/confidences should also be logged.

We distinguish perception modules by whether their predictions directly affect safety-critical decisions (e.g., graceful degradation, or limit velocity) or only influence higher-level planning (trajectory plan).

- XAI Group 1 – Safety critical: Modules that feed directly into vehicle control are required to expose their confidence and trustworthiness in real time. XAI outputs (e.g., saliency maps, uncertainty estimates, decision-boundary explanations) are presented to the decision-making stack so that safety controllers can make informed, risk-aware choices.
- XAI Group 2 – Non-safety-critical: Modules that operate at the planning level use XAI primarily to reduce the reducible uncertainty during development phase. The target audience of XAI extracted explanations includes data analysts, AI developers, domain experts, and safety engineers. Model explainer XAI tools help these stakeholders design/refine models, validate data pipelines, and ensure trustworthy outputs (given «known» input space).

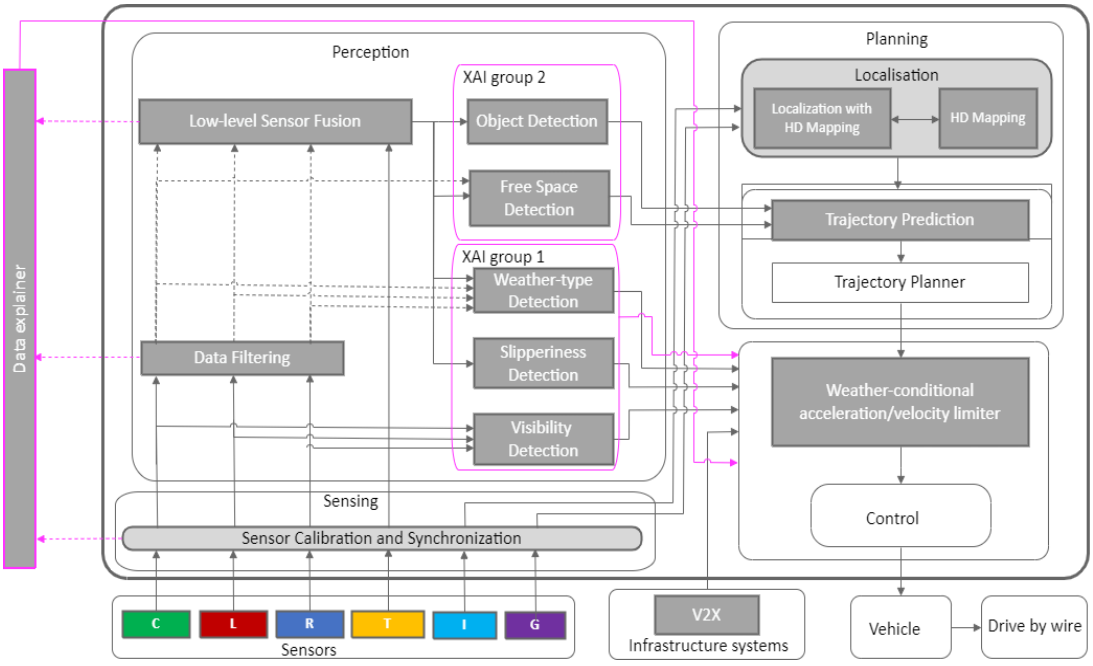

Figure 2: XAI in ROADVIEW reference architecture.

6.2 Explainability for dataset membership

Variational Auto-Encoder (VAE) models have been trained with ROADVIEW datasets to be used as a data descriptor. Its encoder stacks three convolutional layers, followed by two fully connected layers that output the mean and log variation of the latent vector. The decoder mirrors the encoder and up-samples the resolution to the original input data size. The training optimizes the evidence lower bound (ELBO) by combining pixel/voxel-wise L2 errors and a KL divergence term (to capture variation).

Once trained, the VAE model can be used as a validation data descriptor to validate a data point or describe safe known boundaries.

An example of using the VAE data descriptor for the FGI ROADVIEW dataset release to validate dataset membership is provided in Figure 3. Each row shows input image data, its reconstructed version using the data descriptor, and the related anomaly score computed as L2 reconstruction error. The first row shows an outlier where the data point is taken from another ROADVIEW dataset (AVL dataset).

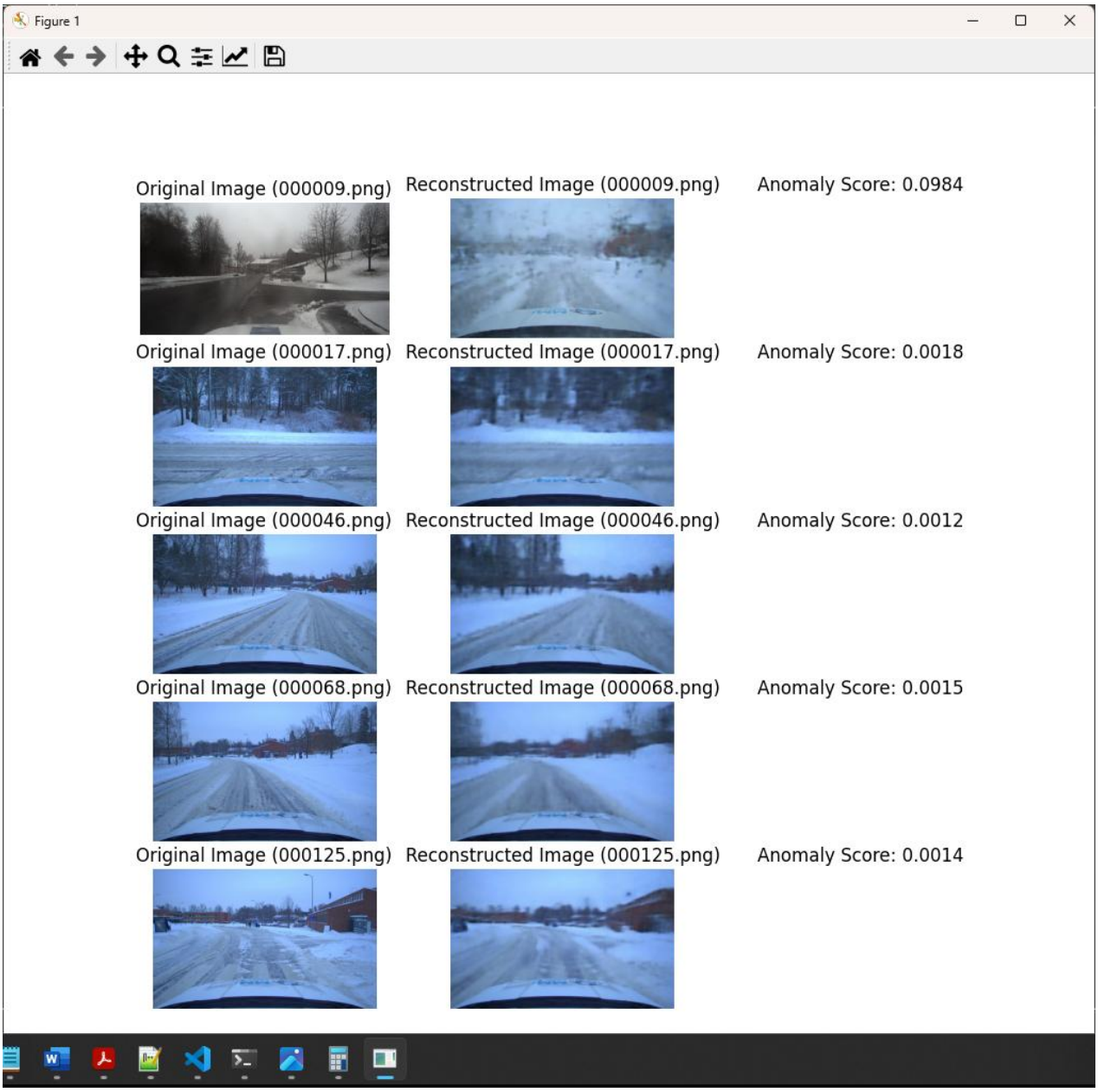

Figure 3: VAE-based data descriptor (trained for FGI ROADVIEW dataset release).

6.3 Explainability for non-safety critical modules

6.3.1 Weather noise filtering

3D-OutDet [100] is a lightweight 3D outlier detector designed (Task T4.4, see deliverable D4.8) to label precipitation anomalies such as snow particles. It leverages a nearest-neighbour convolutional architecture that delivers competitive accuracy while reducing memory usage and inference latency. Integrated into ROADVIEW's object-detection stack, 3D-OutDet boosts weather-robustness and overall prediction performance as comprehensively described in D4.8 in WP4. Because the model operates as part of a data-processing stack, data explainers are essential for understanding its behaviour, diagnosing filtering effects, and ensuring that the downstream perception modules receive clean and trustworthy inputs.

This section illustrates how data explainers can be leveraged to diagnose model performance. The experiments use a representative point-cloud subset extracted from a LIDAR point cloud within the WADS dataset [101], whose contents are summarized in Table 5.

Table 5: Experiment point cloud subset description.

Point set	Description	Variable name
Falling snow (groundtruth)	Points labeled as snow by the WADS dataset	point_cloud_snow
All points	The full LiDAR response	point_cloud_all
Grad-CAM focus	Points that the model's Grad-CAM map assigns the highest importance	point_cloud_grad
Prediction	Points classified as snow by 3D-OutDet	point_cloud_pred

Intersection over Union (IoU) distance has been computed between the subset pairs (Table 6). The observation is as follows:

- Strong alignment between model attention provided by Grad-CAM and model prediction (96%): This indicates that the Grad-CAM heatmap is a reliable explainer for the detector's reasoning.
- Reasonable Precision/Recall trade-off: The achieved number 74% is comparable with many state-of-the-art LiDAR-based segmentation methods for non-rigid, low-density classes such as snow or rain.
- Sparse nature of snow class: The 13 % IoU between all points and snow labelled points demonstrates that snow is a minority class in the LiDAR response. The low 10 % IoU of predictions against the full point cloud indicates that the 3D-OutDet likely captures this sparsity.

Table 6: Distribution distances (IoU metric).

IoU	point_cloud_pred	point_cloud_grad	point_cloud_all	point_cloud_snow
point_cloud_pred	1	0.969649067	0.10036898	0.740729219
point_cloud_grad	0.969649067	1	0.103510624	0.726802241
point_cloud_all	0.10036898	0.103510624	1	0.128521055
point_cloud_snow	0.740729219	0.726802241	0.128521055	1

Spherical harmonics experiments

Spherical harmonics [102] are a family of orthogonal functions that form an orthonormal basis on the unit sphere. By projecting 3D geometric structures, such as LiDAR point clouds, into this basis, one obtains a compact and rotation-invariant representation. This property makes spherical harmonics particularly attractive for point cloud analysis, similar to the use of traditional Fourier or wavelet transforms in 2D image processing.

Table 7 demonstrates how spherical harmonics can be used to analyse the LIDAR point cloud subsets. By extracting and comparing the harmonic signatures of snow-labelled points, the entire scene, and the model's attention/predictions, one obtains a quantitative measure of the model's fidelity to snow-specific geometric cues.

The following observations are found:

- Low prediction and attention distance: The Grad-CAM focus is almost identical to the model's predictions in the harmonic feature space, confirming that the model's internal attention aligns well with its final output.
- Very low prediction and ground truth distance: The spherical harmonic signature of the predicted snow points closely matches that of the ground truth snow. This indicates that the network is successfully extracting the rotation-invariant geometric cues that define snow in the scene.
- Medium distance between snow and all scene points: Snow forms a distinct sub-space compared to the full LiDAR cloud. It demonstrates that snow is geometrically separable from the background and other objects.
- Medium snow and Grad-CAM distance: Slightly larger than the prediction ground truth distance, meaning that the model focus better than its prediction performance.

Table 7: Distribution distance (PCA on spherical coordinates).

PCA (0)	point_cloud_pred	point_cloud_grad	point_cloud_all
point_cloud_grad	0.047093756		
point_cloud_snow	0.0077358494	0.054762963	0.039700802

Table 8 shows statistical properties of the point cloud subsets. Some of the first insights can be drawn as follows:

- Snow labelled points (ground truth): Snow points are thin, spread along the road, and sparsely distributed. The model captures a subset that is more concentrated and less spread, potentially missing distant or diffuse snow.
- Attention alignment: Grad-CAM and predictions largely overlap in spatial extent and density, validating the use of Grad-CAM as a useful model explainer for this purpose.
- Distribution statistics: High skewness and kurtosis in the full scene highlight the challenge of distinguishing snow from background clutter. While the Grad-CAM has negative kurtosis in X and Y (more spread), the predictions have positive kurtosis (more concentrated around the mean value)

Table 8: Statistical properties of different point cloud distributions.

Metric	point_cloud_snow	point_cloud_pred	point_cloud_grad	point_cloud_all
Number of Points	21122	14317	14752	207149
Centroid	21122	14317	14752	207149
Spread (STD)	[1.04, 3.96, -0.42]	[0.28, 2.26, 0.12]	[0.23, 1.90, 0.16]	[0.78, 5.08, -0.70]
Bbox (W, H, D)	[14.13, 4.65, 0.90]	[5.81, 3.68, 0.49]	[6.13, 4.43, 0.62]	[28.75, 22.45, 1.36]

Metric	point_cloud_snow	point_cloud_pred	point_cloud_grad	point_cloud_all
Density	[138.74, 81.18, 7.73]	[52.40, 38.60, 7.85]	[187.64, 55.64, 7.99]	[367.71, 221.02, 17.10]
3D bbox volume	0.24	0.90	0.18	0.15
Skewness	87010.85	15882.83	83394.35	1389802.13
Kurtosis	[-0.46, -0.19, -0.00]	[0.01, 0.05, 2.26]	[-0.53, -1.04, 3.24]	[0.39, 1.10, 1.69]

Figure 4 shows the Grad-CAM heatmap in a 3D point cloud view. The points that have higher attention values are masked with yellow colour, while other points are masked with green colour.

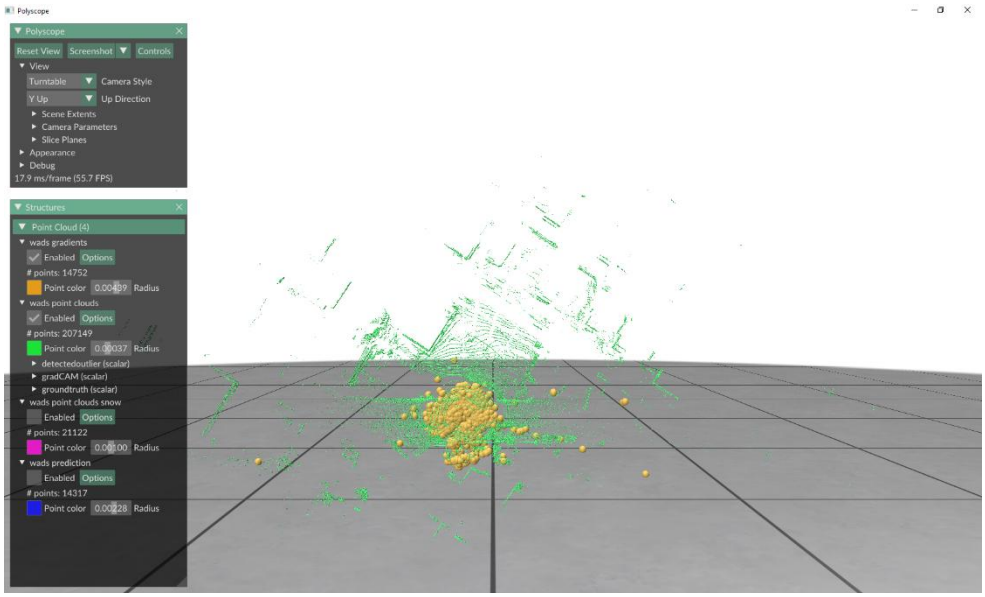

Figure 4: Grad-CAM maps on point cloud.

Data explainers with Grad-CAM

Figure 5 illustrates how a data explainer, together with a local model explainer, uncovers a corner case for the 3D-OutDet model (Task 4.4). From left to right, the panels display the correlation matrices of the point cloud coordinates (X, Y, Z) for three point-sets: the ground-truth snow points, the Grad-CAM-selected points, and the model's predicted snow points.

The data explainer shows that the ground truth snow points exhibit moderate correlations between (X, Y) and (Y, Z). By contrast, both the Grad-CAM and the final predictions display only weak correlations along these axes, indicating that the model's attention and decision-making are not aligned with the true geometric relationships. This misalignment explains why a subset of snow points remains undetected.

These findings motivate two potential improvements:

- Model architecture: modify the convolutional operator to operate on bounded neighbourhoods, incorporate local 3D coordinates, and/or optionally set the convolution kernel's centre weight to zero to avoid over-reliance on the central point.
- Revise the existing data normalization: applies a normalization scheme that preserves the aspect ratio of the 3D coordinates during augmentation (instead of normalizing to the same range for all spatial dimensions), ensuring that the spatial relationships present in the ground truth are retained during training.

These improvements may restore the missing correlations and enhance the model's sensitivity to sparse snow patterns.

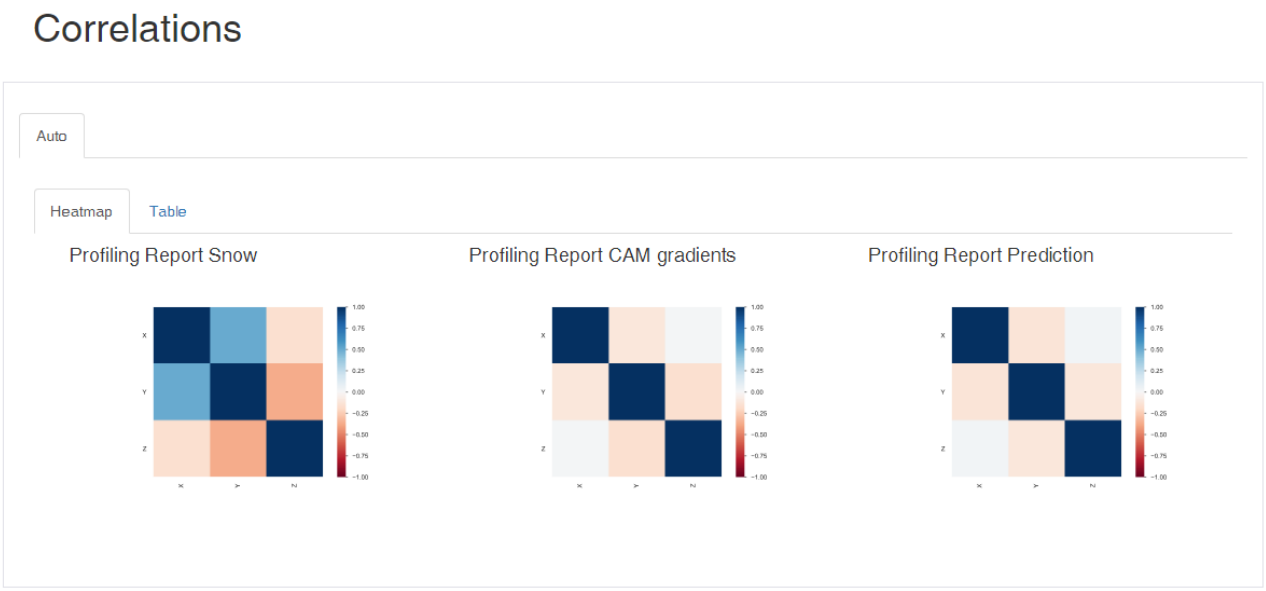

Figure 5: Data explainer (profiling) using yData with sequence 12, WADS LiDAR point cloud dataset (snow ground truth, 3DOutDet detected snow points and Grad-CAM point clouds).

6.3.2 Reference explainability embedded design for 3D point cloud segmentation

As the WP5 perception models (within task T5.1, T5.2) evolve through continual design iterations, a generic set of XAI techniques is required to keep pace with new architectures and improvements. The most common 3D point cloud segmentation models can be broadly categorized into point-based, voxel-based, and polar-based approaches. Voxel- and polar-based models use dense feature representations, which makes the adoption of e.g. saliency map techniques relatively straightforward. However, the sparsity of point cloud data processed by point-based models make the XAI adoption less trivial. In this section, we introduce a practical approach for point-based models, illustrated through an example using a representative architecture.

Figure 6 shows an example technique for generating explanations in point-based models.

The standard PointNet++ architecture [103] has been extended with functions to extract interim features from different abstraction layers (multi-scale abstraction layers, global abstraction layer, and feature propagation layer), highlighted by arrows and a star in the figure. At each selected layer, the L1 norm of the feature multidimensional output tensor is computed, resulting in a saliency map capturing the relative importance of neuron activations per point. The L1 norm provides compact, interpretable relevance scores. These scores are then projected back onto the point cloud (at different point resolution per extracted layer), producing a 3D heatmap representation that reveals how geometric cues propagate through the network, highlighting influential regions for the prediction. An illustrative experiment was performed with a data sample from the ShapeNet[104] point cloud dataset, with generated heatmaps provided in the figure corresponds to each interim feature.

Observations:

- The method is fast and can be used to highlight regions where the model is focusing, informing targeted data collection/preparation strategies.
- The choice of L1-norm is heuristic, based on an assumption that the absolute magnitude of neuron activations reflects importance. Future work could investigate alternative weighting schemes or dimensionality reduction techniques as alternative choices for the attribution signal.

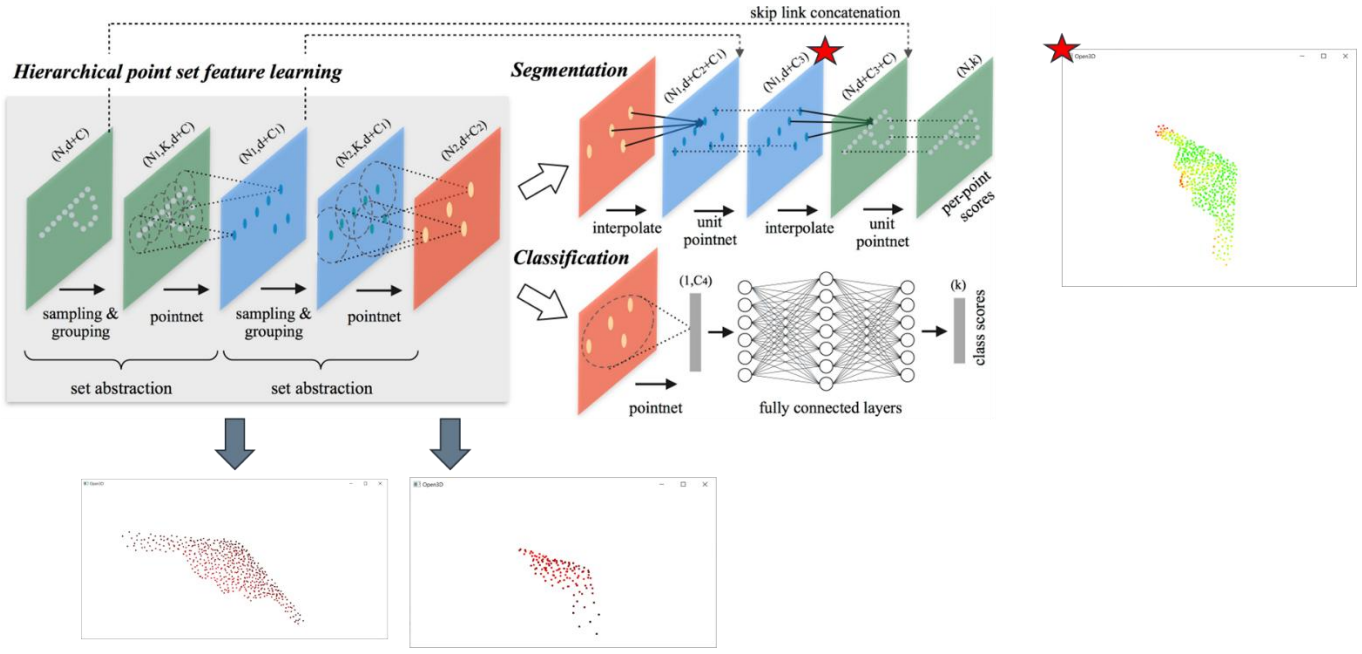

Figure 6: Point cloud segmentation explanations.

By integrating a decompositional design, choosing key layers as checkpoints lets stakeholders inspect the internal logic of each sub-module. In Figure 6, the checkpoints were positioned at the stages where the point-cloud resolution is reduced.

6.3.3 Infrastructure integrated collaborative perception

This section outlines how explainability techniques were applied to the AI models within collaborative perception module developed under T5.2 (see deliverable D5.4).

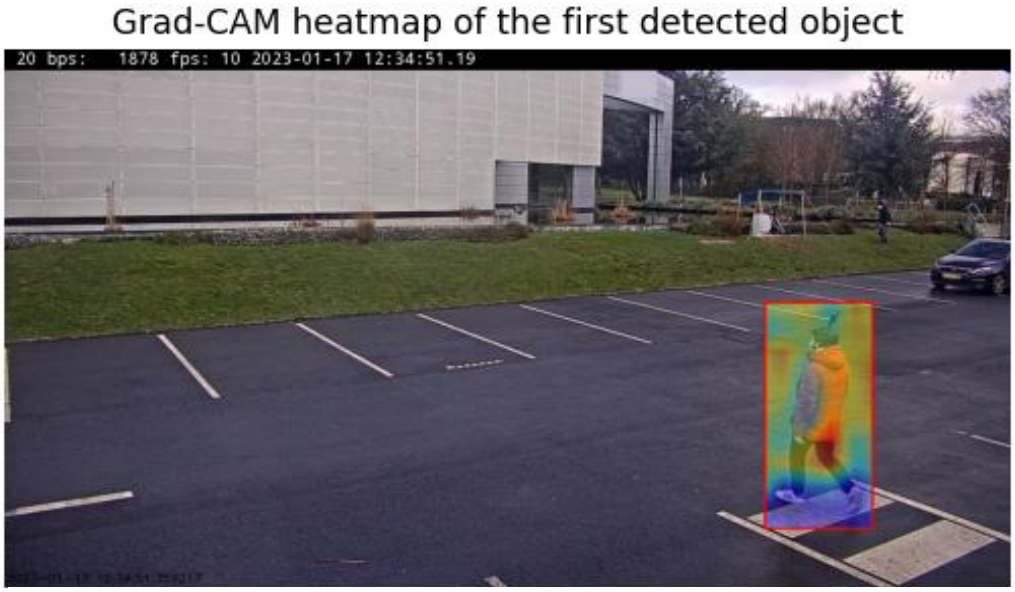

Figure 7: Grad-CAM heatmap of pedestrian detector (V2X module).

Figure 7 displays a Grad-CAM heatmap for CRF's pedestrian detection module, which processes camera images. Because Grad-CAM is originally formulated for classification, we restrict the visualization to the area inside each detected pedestrian bounding box and overlay the resulting heatmap on the original image. The heatmap thus highlights the image regions that the detector emphasizes when making a pedestrian detection decision.

Figure 8 illustrates an alternative approach in which Grad-CAM is applied not to the full input image but to a cropped region that contains only the target object.

6.3.4 BEV-based Sensor fusion for object detection model

This section describes XAI application for Sensor fusion for object detection model developed within T5.1 and also be used for collaborative perception in T5.2 (D5.1, D5.2).

A typical object detection model (such as SSD [105] or Faster R-CNN [106]) first extracts feature maps from a convolutional backbone. On each of these feature maps a set of predefined anchor boxes is tiled, and two lightweight heads are applied in parallel to every anchor:

- Classification head: outputs a probability distribution over all object classes (plus background),
- Regression head: predicts four offsets that refine the anchor to better match a ground truth bounding box.

The anchors are matched to the ground truth bounding boxes during training, via a joint loss combining a cross-entropy for the classification predictions and L1 or logistic loss for the bounding-box offsets. During inference the non-overlapping detections are selected via non-maximum suppression, producing the final set of classified objects and their localised bounding boxes.

Figure 8: Grad-CAM applied for cropped image.

To redesign an object detection model to also provide uncertainty estimates, we add an extra head in parallel with the existing classification and regression heads. This (namely) uncertainty head predicts a log variance for each of the four bounding box coordinate offsets. During training, the residual between the predicted bounding box and its matched ground truth bounding box is treated as a sample from a Gaussian distribution whose variance is given by the head's output. The loss can be chosen as the Gaussian negative log likelihood function (NLL). By minimizing NLL, the model is encouraged to achieve accurate predictions while simultaneously learning a faithful uncertainty estimate. The extra uncertainty head will thus learn to predict the distribution of bounding box localization errors.

Figure 9 shows the modified architecture of the original SSDLite[105] model with an additional bbox uncertainty head (in green). This extra branch produces for every predicted box a log variance (representing uncertainty) for each of the four box coordinates, providing a measure of localization uncertainty estimate.

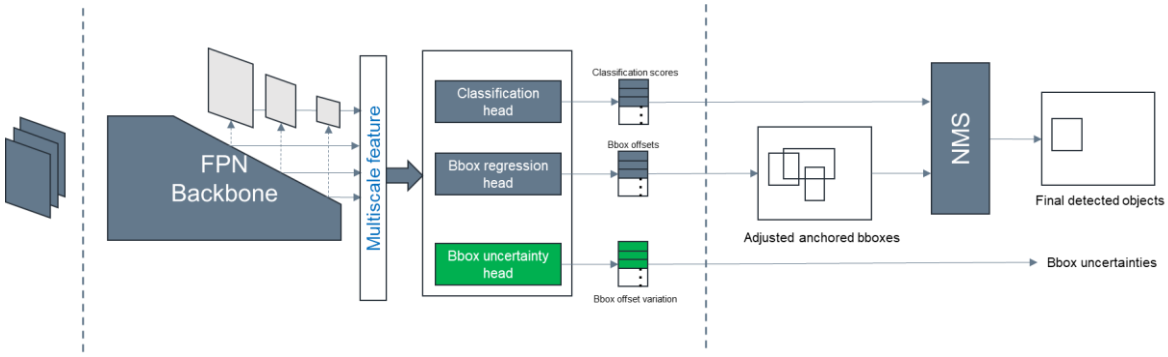

Figure 9: Modification of the SSD Object detection model to include uncertainty estimation head.

The residual uncertainty of a 3D detector (that is used in ROADVIEW sensor fusion for object detection model) results from the irreducible uncertainties (sensor noise, human annotation error) and the remaining epistemic uncertainty.

To provide the subsequent decision-making module this information as a trustworthiness measure, we introduce a lightweight surrogate model that learns to predict a per bounding box uncertainty after the main model has been trained.

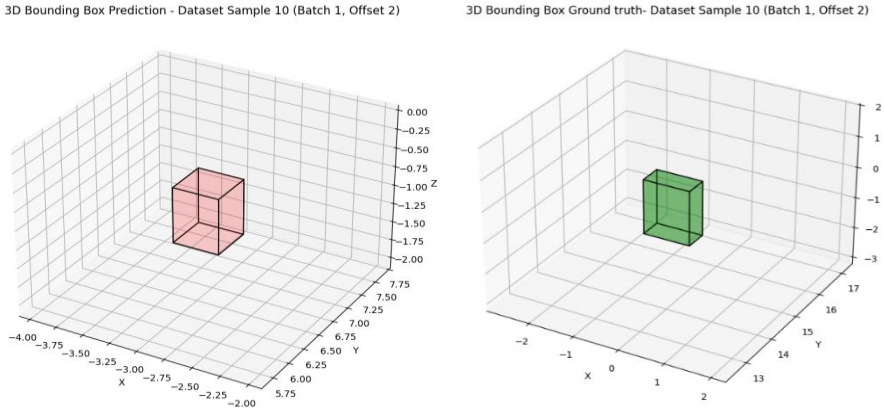

Figure 10: Sensor fusion object detection model prediction (left) and ground truth (right).

The surrogate is trained to map the same extracted features that the main detector uses, i.e. the convolutional backbone activations masked by the predicted bounding box (ROI align feature) and the five-dimensional box coordinates (x, y, z, w, h, d) to a log-variance that characterizes the distribution of localization errors. We assume that the prediction errors follow a univariate Gaussian in each of the spatial coordinates (Delta_x, Delta_y, Delta_z), so that the surrogate outputs a single scalar log(sigma^2) per box, representing the variations from the predicted values. The surrogate model is therefore can be built as a lightweight regressor (e.g., a few fully-connected layers) using NLL loss. Figure 10 shows an example of a predicted bounding box and related ground truth for a data sample. The distance distributions between ground truths and predictions will be estimated as a Gaussian distribution reporting uncertainties by the surrogate model.

Training the surrogate in a separate stage improves the training convergence compared to jointly optimizing both the classification/regression and uncertainty heads. The surrogate therefore inherits the learned representation of the main model while adding only a lightweight computational overhead.

The resulting uncertainty estimates enable safety-critical decision module to reason about the probability that a detected object lies on a collision course. For example, the decision making can construct 95 % confidence intervals for the bounding box coordinates (Figure 11) by sampling two bounds from the predicted Gaussian and propagate these into trajectory optimisation.

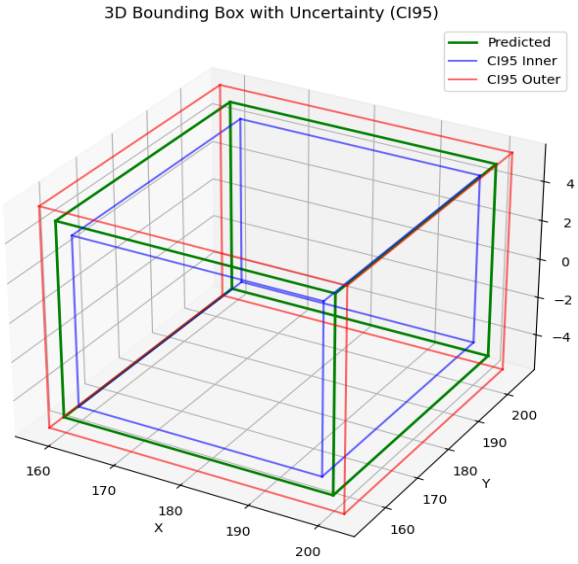

Figure 11: Object bounding boxes with CI95 intervals

6.4 Explainability for safety critical modules

6.4.1 Slipperiness estimation model

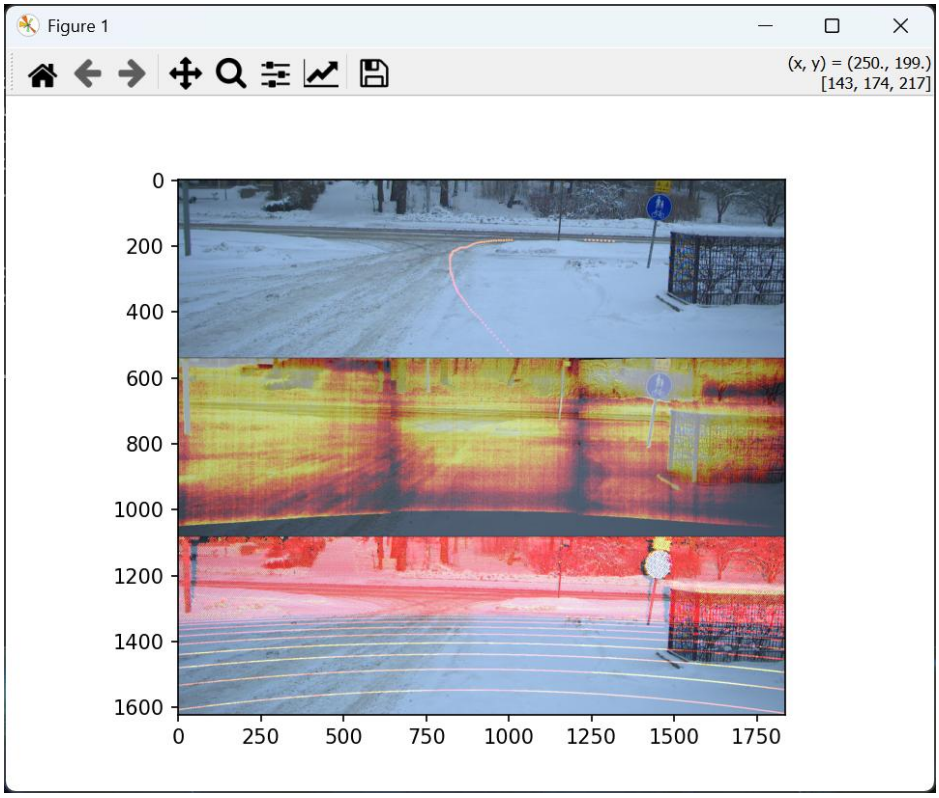

Figure 12: Captured sensor data (camera, thermal, lidar) vs measured ground truths.

The Dense road surface grip map prediction model [107] (Task T5.3, see deliverable D5.5) was built with intrinsic explainability by incorporating an uncertainty estimate. Instead of treating grip as a direct regression target, the problem is reformulated as a surface state segmentation task. The model first predicts per-pixel probabilities for each road surface class (dry asphalt, ice, snow, etc.). Each class is associated with a known grip distribution, so the overall grip prediction is a weighted mixture of these distributions, where the weights come from the class probabilities. The variance of this mixture provides a grip uncertainty estimate. A full technical description of this intrinsic uncertainty-aware model has been published in [108].

In this section, we describe a set of XAI analyses that were performed to better understand the behaviour of the grip prediction model and to quantify the intrinsic uncertainty in the data.

- Exploration of the ground-truth distribution: visualizing where the true grip values are in the spatial domain of each scene.
- Prototype and critical patch analysis: extracting representative and outlier image patches that carry the most information for the model.

Ground-truth distribution data analysis

The spatial distribution of the ground-truth grip values can give insight into the spatial distribution of aleatoric uncertainty. For each recorded scene, we projected the grip annotations onto the 2D XY plane of the camera view. The resulting 2D heatmap indicates the density of non-default grip values. We generated these heatmaps for the two datasets collected at Otaniemi (Figure 14) and Northern Hilantie (Figure 13), both in Finland. The heatmap shows a concentration of grip values in a tight central region. The non-drivable area of the image contains very few annotated grips, indicating that the model will likely have high aleatoric uncertainty in these zones.

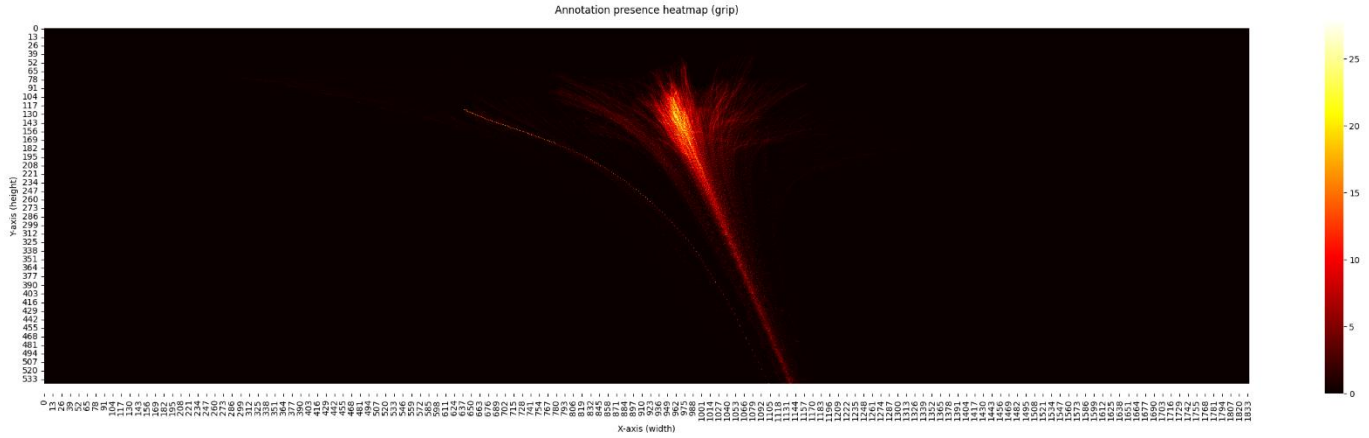

Figure 13: Heatmap of grip ground truth value distribution on XY plane (Northern Hilantie).

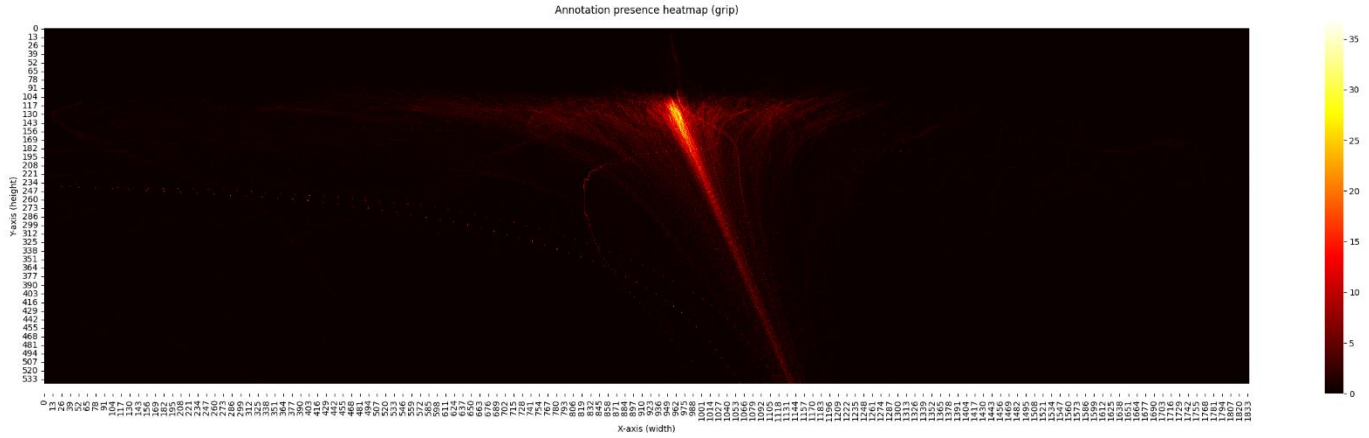

Figure 14: Heatmap of grip ground truth value distribution on XY plane (Otaniemi).

Prototype and critic patch analysis

While the heatmaps provide a coarse view of where grip data exists, we also need to understand if the provided dataset can provide some key information that the model is learning from. By extracting small patches centred on pixels with known grip values, we can discover prototypes (representative samples that the model can rely on) and critics (samples that most deviate from the norm that may confuse the model). For each ground truth pixel, we extracted a 70×70 pixel patch centred on the pixel. The choice of 70×70 is empirical. We then applied k-means clustering to identify clusters of similar patches. The number of clusters (k=10) was chosen empirically. The cluster centroid patches are then selected as the prototype's patches. Patches with a distance to their nearest centroid exceeding a threshold (chosen as the 95th percentile of all distances) were flagged as critics. Figure 14 shows computed data prototype patches, while Figure 15 shows computed data criticism patches, computed for the Northern Hilantie dataset.

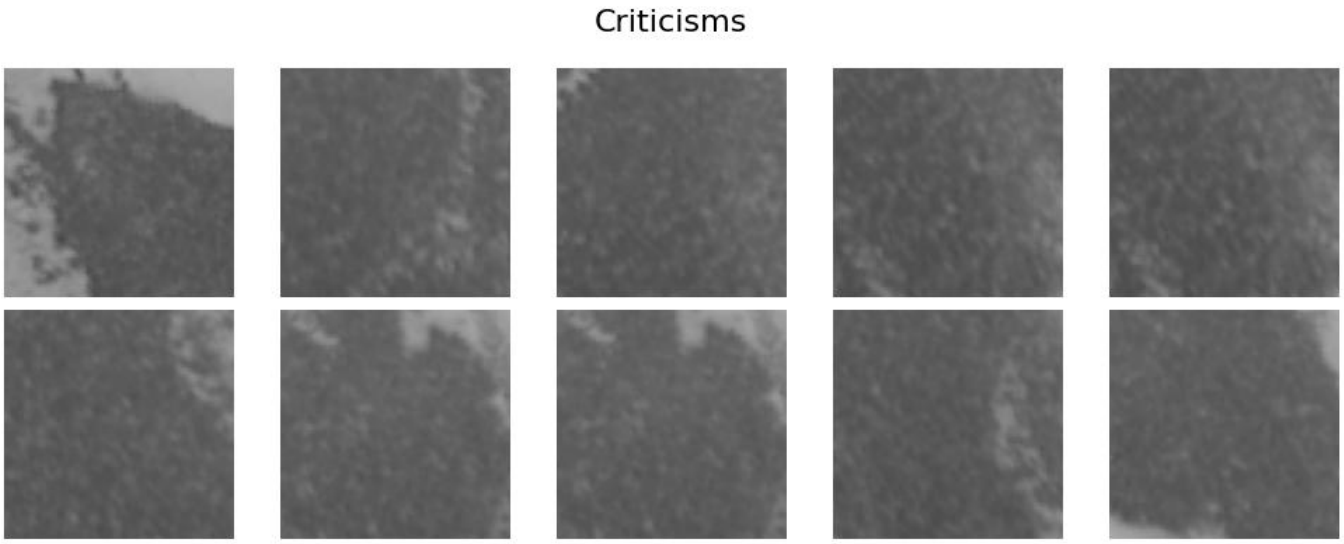

Figure 15: Criticisms patches (Northern Hilantie).

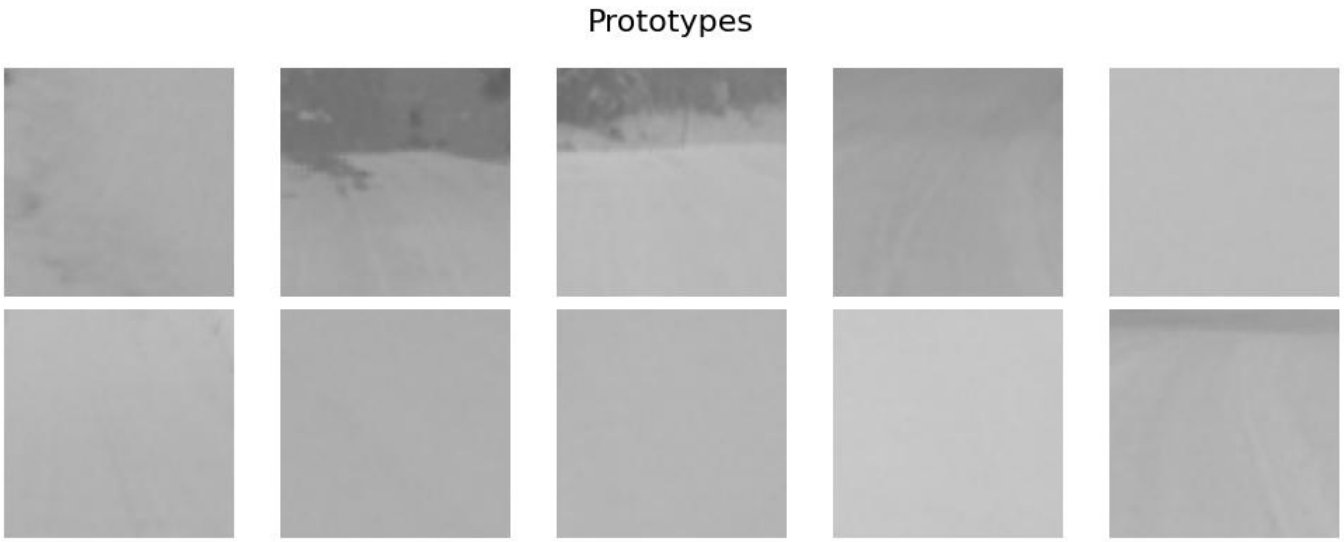

Figure 16: Prototype patches (70x70) using kMean cluster approach (Northern Hilantie).

EDA analysis of the patch dataset

For every 70 × 70 pixel patch in the dataset, we extracted a comprehensive set of low-level image descriptors. Each patch was first converted to grayscale and summarized by mean, standard deviation, minimum, maximum, and median pixel values. A 10-bin intensity histogram was computed, together with a local binary pattern (LBP) histogram (10 bins) to characterize texture. We described the spatial entropy of each patch using Shannon entropy computed for every pixel in the patch and then computed the mean and standard deviation of the resulting entropy map. We also calculated an edge density metric via the mean absolute Sobel response. Finally, each patch was annotated with the ground truth grip statistics (mean, standard deviation, and count of contributing points) and the centre coordinates. These features provide a statistical description of every patch extracted from the dataset.

Figure 17 shows a correlation matrix computed for the extracted low-level descriptors. Analysing this correlation matrix results in key features representing the patch datasets, including: Intensity distribution, entropy distribution, and edge density. The texture feature computed by LBP does not seem to provide enough meaningful information.

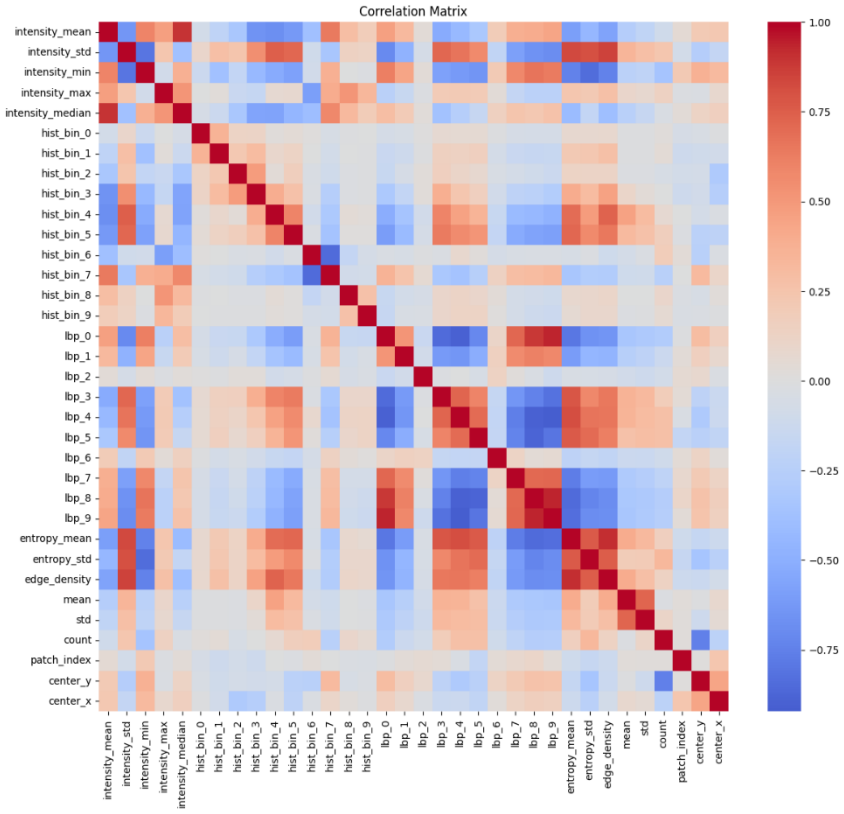

Figure 17: Correlation matrix (patch image statistic features impact on grip values).

Figure 18 provides three histograms (using KDE smoothing). The left histogram shows the distribution of the average grip value reported for each patch. The middle histogram shows the spread (std) of grip ground truth values within a patch, and the right histogram shows the statistics of how many patches contain a specific number of ground truth pixels.

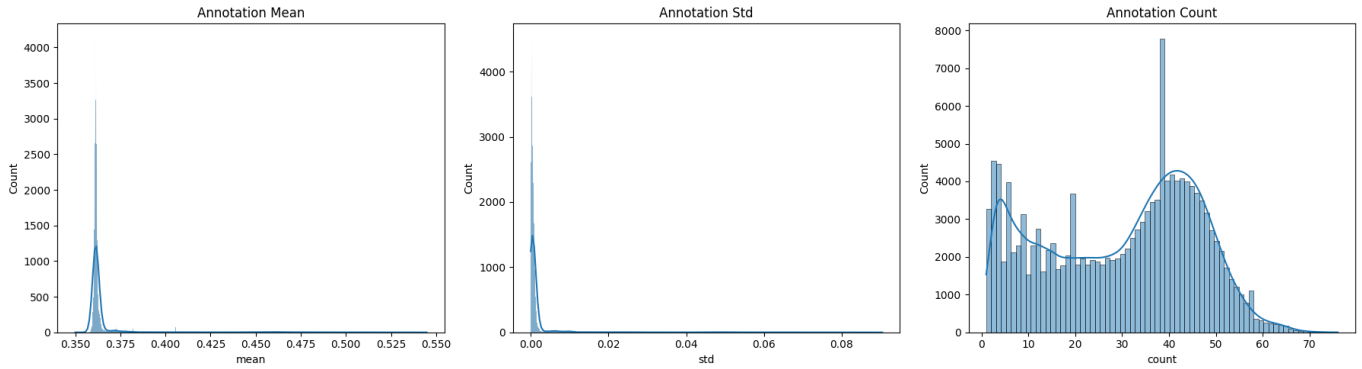

Figure 18: EDA on patch dataset.

Model inner working transparency

Figure 19 displays a 2D heatmap visualizing the mean activation values over all channels of the feature map produced by a bottleneck block of the encoder.

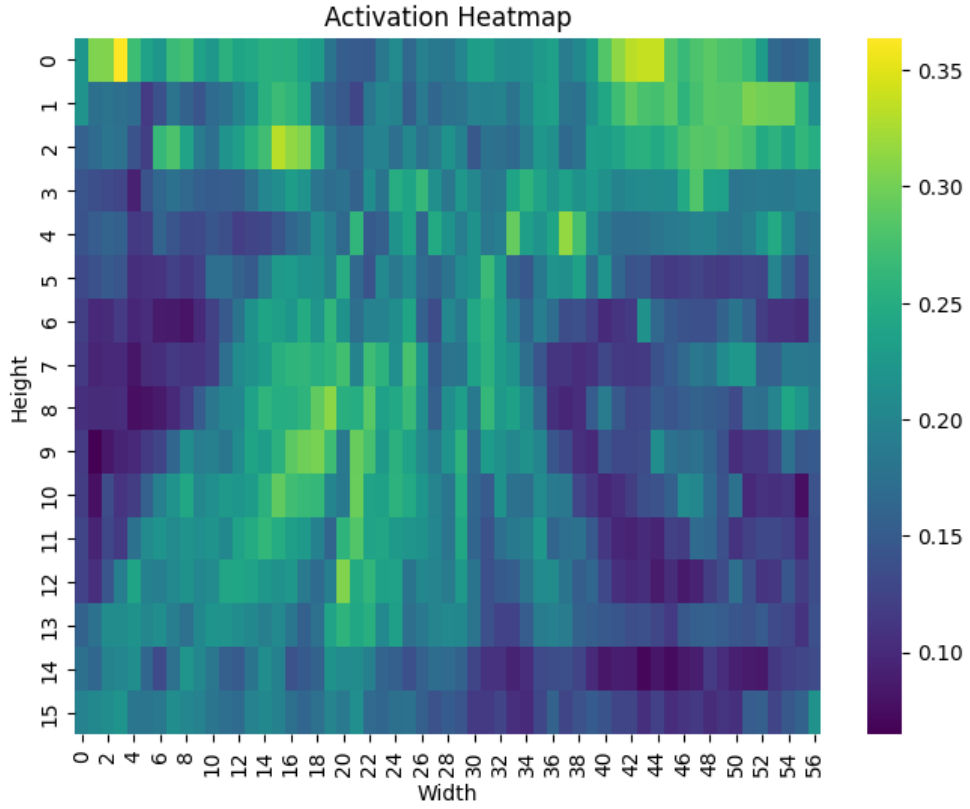

Figure 19: Activation patterns.

Figure 20, Figure 21, Figure 22, and Figure 23 visualize the sensitivity of the model's output (grip, water, ice and snow respectively) with respect to the raw pixel values of the input image. The heatmaps highlight the spatial locations whose perturbation would most strongly affect the model's confidence in predicting the output values. Red regions pinpoint the image areas that the network relies on when estimating that part of the image for grip/water/ice/snow estimates. This is an example of a gradient-based saliency map interpreting which pixels drive the network's decision for the corresponding road condition under the imposed mask constraint.

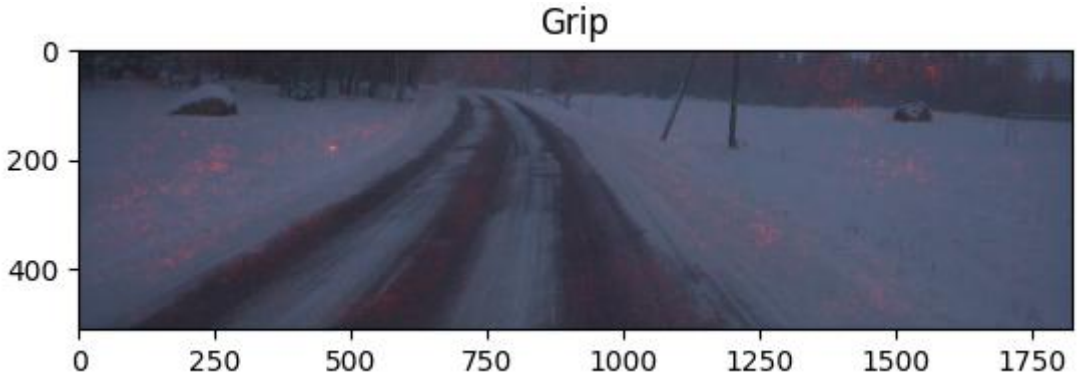

Figure 20: Attribution map (grip prediction).

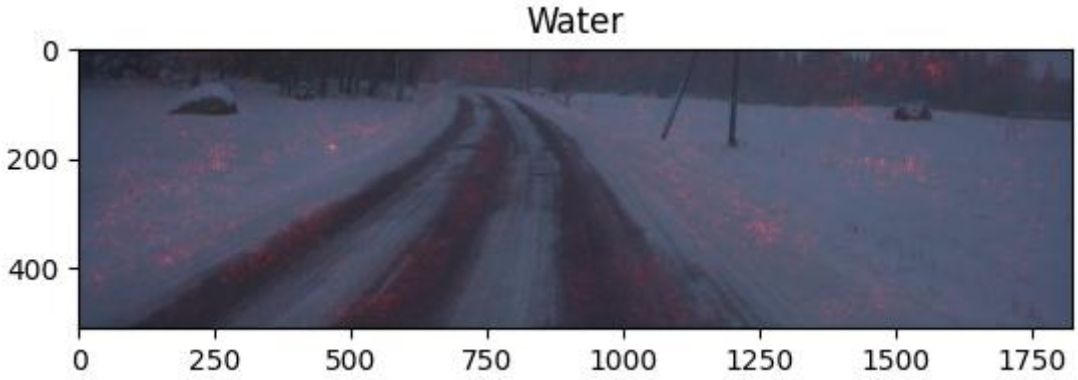

Figure 21: Attribution map (water prediction).

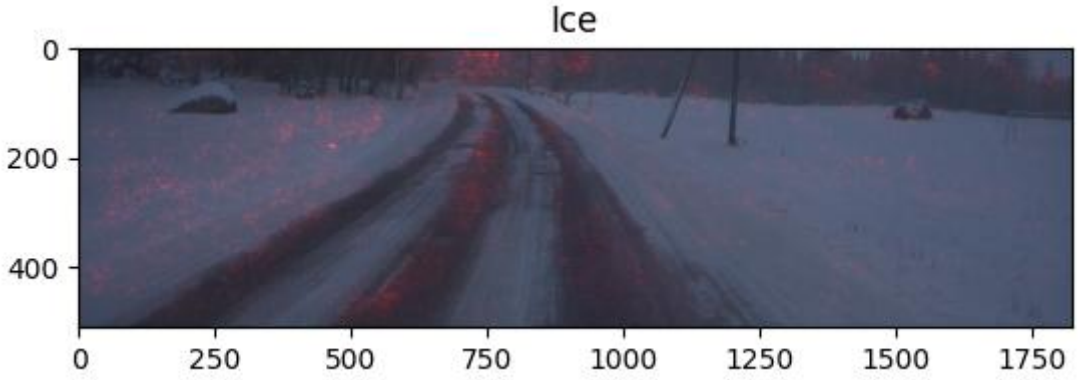

Figure 22: Attribution map (ice prediction).

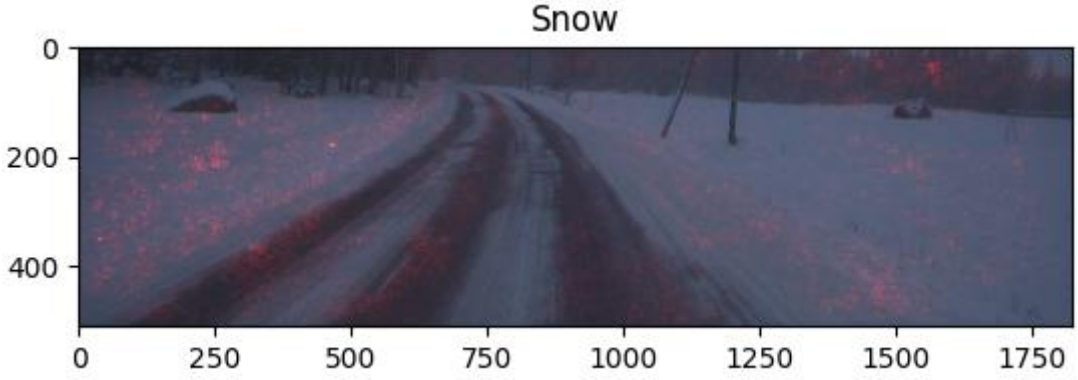

Figure 23: Attribution map (snow prediction).

Figure 24 provides a channel-by-channel snapshot of how the segmentation head is behaving for a specific class, blending both the learned parameters (weight) and their importance during back propagation (gradient). This explanation is useful for interpreting and diagnosing the inner workings of the model.

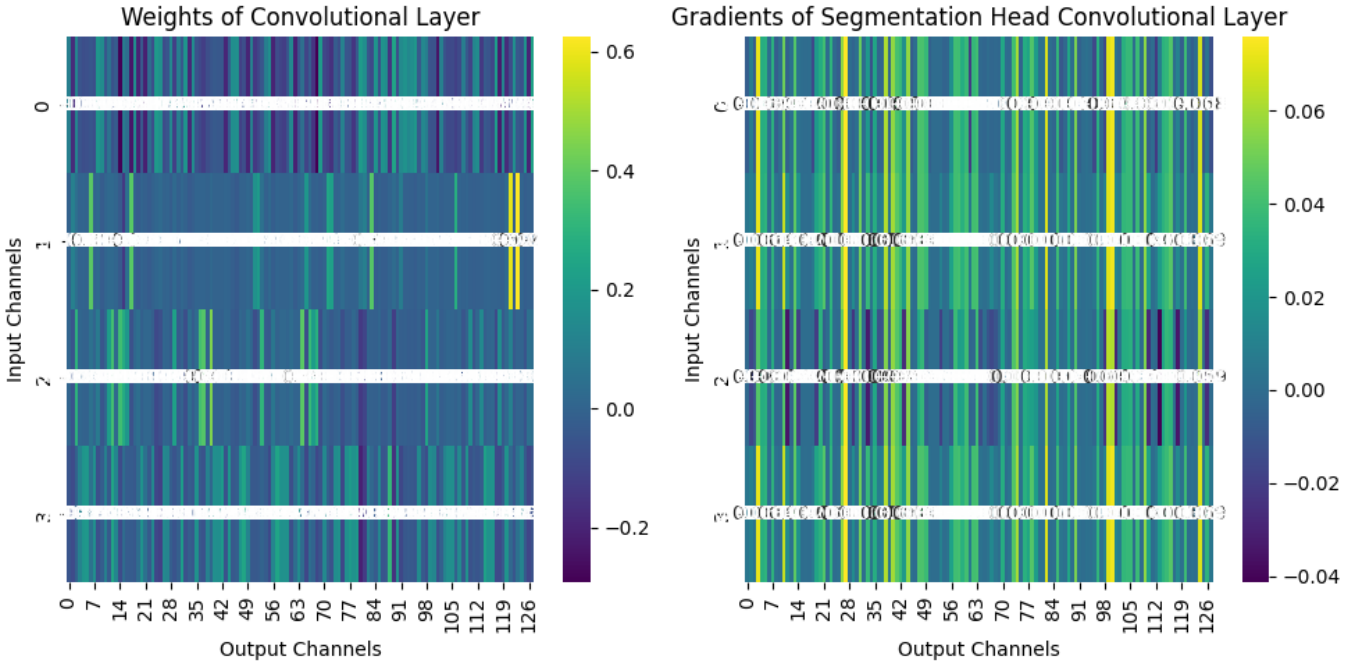

Figure 24: Importance of convolutional filter (weight, left) and sensitivity (gradient, right).

6.4.2 Visibility estimation model

Visibility estimate model (Task T5.3, see deliverable D5.5) takes as input from the LIDAR sensor and returns visibility estimates for a set of discrete angular sectors surrounding the vehicle. Each output value corresponds to one angular sector, quantifying how far the system can see in that direction based on the observed longest object's surface found from the observed point cloud. The model produces 25 values for every 5° step between (30°, 150°), whereas 90° is the straight-ahead direction of the vehicle (Figure 25).

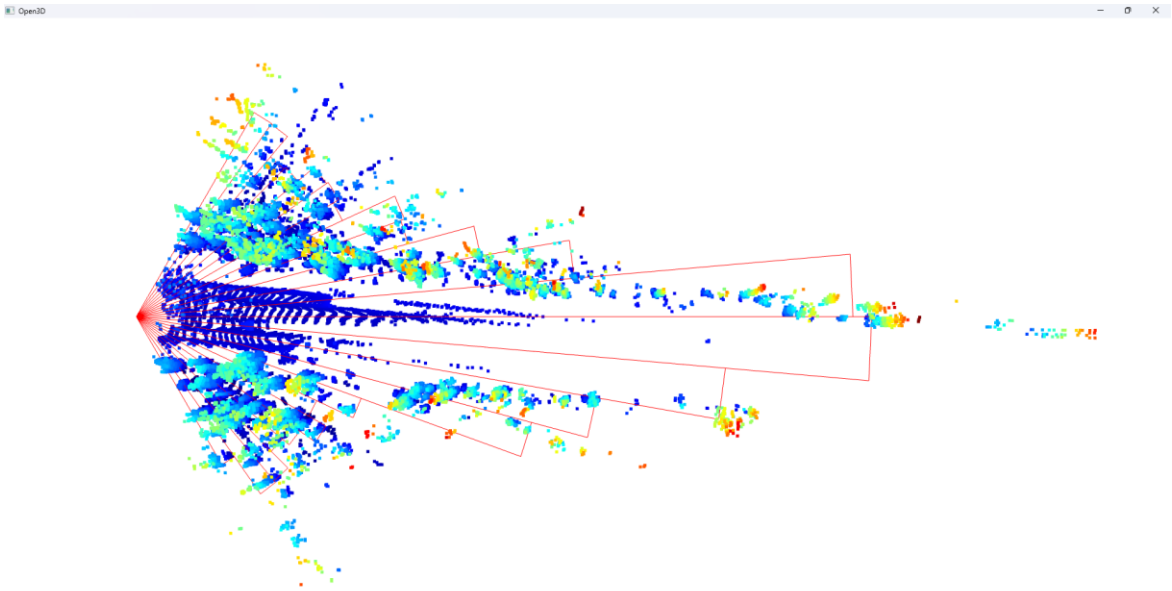

Figure 25: Example LIDAR point cloud and Visibility estimates per angular grid.

Figure 26 shows visibility estimates, smoothed by spline interpolation, the uncertainty measures are taken as angular uncertainty, computed as standard deviation over a window size of 3. The figure reveals how visibility estimates vary with the scan angle and related uncertainty represented in the light blue area.

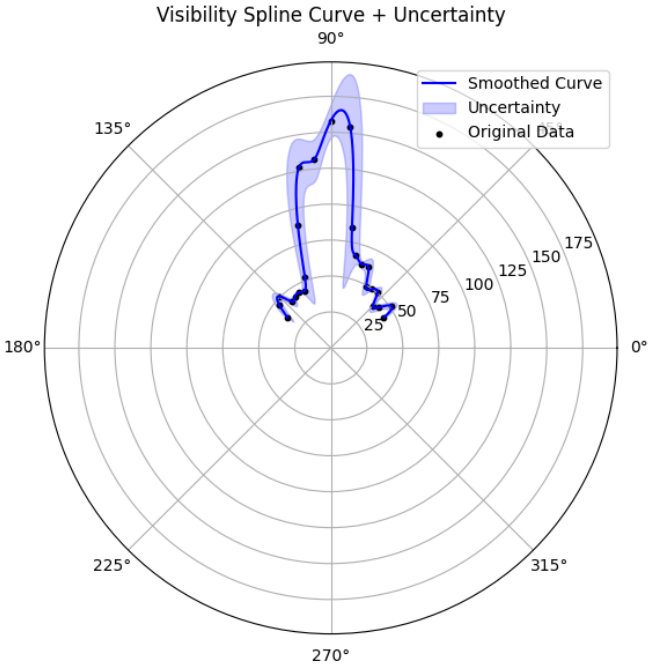

Figure 26: Visibility estimates with uncertainty.

Figure 27 visualizes a 2D colour map of the visibility estimates (the first 50 time steps) for all angular bins (in columns). Warm colours are where visibility is high, and cool colours are where it is low. The warmer hues in the heatmap along the road region reveal that visibility in the driving direction is higher, suggesting that obstacles in the surrounding non-drivable areas may be influencing visibility estimates at other angles. The cool colour regions (darker colours) in the driving direction show the impact of moving objects on the road on the visibility estimates.

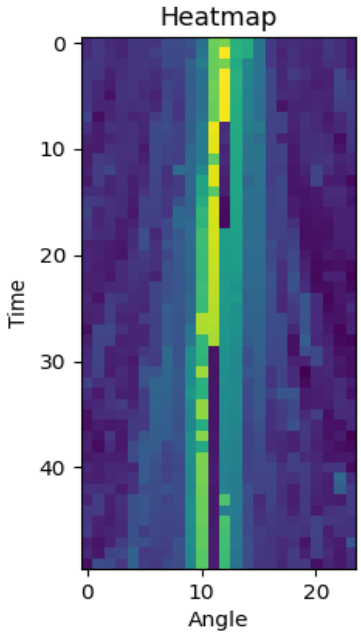

Figure 27: Heatmap of visibility estimates over time window.

6.4.3 Camera-based weather detection model

A camera-based weather detection model has been designed within Task T5.2 and subsequently reused for Collaborative perception in Task T5.3 (see deliverables D5.4, D5.2 respectively). The model is named ResNet50-Truncated-Multitasks (RTM), relying on the first few convolutional layers of ResNet50 as its backbone to extract features from the input image for downstream tasks to estimate weather parameters.

To select a proper CNN layer for truncating, several feature importance methods have been used to identify which layer may provide the most important information for the subsequent multi-tasks. Saliency maps have been computed for every key CNN layer for the comparison.

Figure 28 shows an example of a Grad-CAM heatmap computed for convolutional block 2-9, illustrating the spatial regions that most strongly influence the model's predictions for that block.

Figure 29 shows another example of an integrated gradient saliency map computed for blocks 2-11 (left) and 1-12 (right), providing a complementary view of pixel-wise attribution.

More information on this model's explainability can be found in D5.4.

Figure 28: Grad-CAM heatmap (layer 2-9) for snow class.

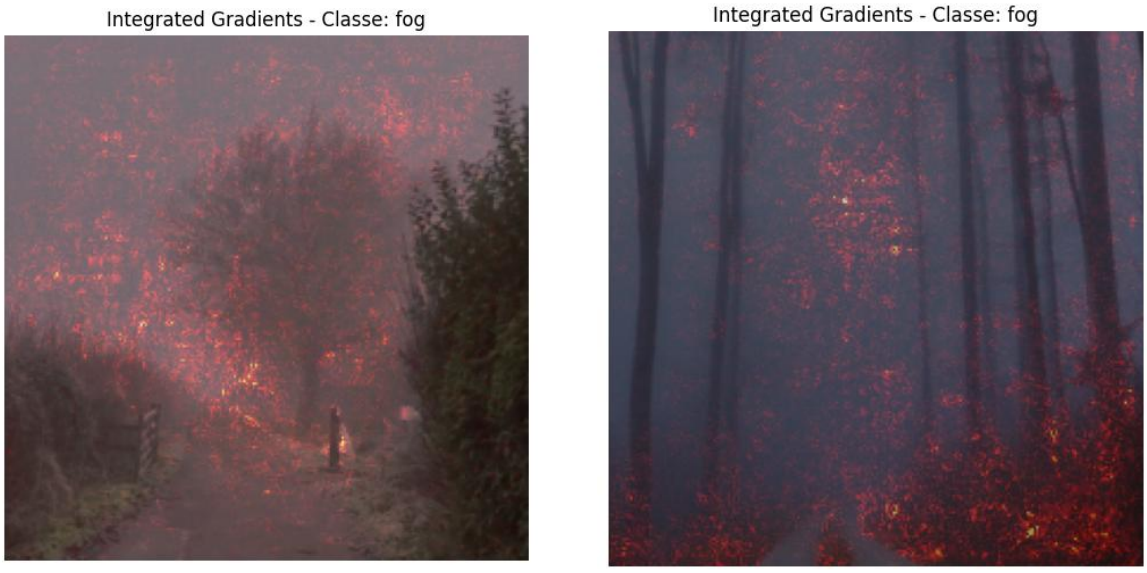

Figure 29: Integrated gradients for fog class (left: layer 2-11, right: layer 1-12).

6.5 Explainability requirement compliance

The use of data explainers together with model explainers facilitates compliance of ROADVIEW perception models with the specified explainability requirements (Section 4). Details of the applied methods are provided in Table 9.

Table 9: Techniques implemented to enable compliance with the specified explainability requirements.

Requirements	Applied methods in ROADVIEW
DAT-EXP-REQ1: Datasets shall be designed to be explainable, enabling verification of completeness, representativeness, and balance with respect to defined data requirements.	Data explainers have been applied to slipperiness and other datasets.
DAT-EXP-REQ2: Individual data points within a dataset shall be explainable when required, to evaluate the presence of key features relevant to AI model prediction accuracy.	Data prototypes and criticisms have been applied to slipperiness datasets.
DAT-EXP-REQ3: Data explainers shall facilitate the development of methodologies to determine the membership of new data points within a given dataset.	The VAE-based data descriptor has been trained and used as a membership validator.
MOD-EXP-REQ1: Where applicable, AI models shall be designed with a modular architecture where each submodule performs a distinct, understandable function.	All AI models within WP5 have been designed with these principles.
MOD-EXP-REQ2: The boundaries of the AI model (inputs, outputs, and model limitations) shall be clearly defined and validated.	Boundaries are defined by the admissible space spanned by the trained data descriptor(s).
MOD-EXP-REQ3: The AI system shall include mechanisms to detect and highlight undesirable bias in the model, including saliency maps composed from neuron activation maps and gradient maps where applicable.	Activation pattern extractor, gradient extractor are provided, and Grad-CAM, integrated gradients have been applied to several models (Data filtering, weather estimator).
MOD-EXP-REQ4: The AI system shall support the target audience in identifying errors or potential flaws in the model design, training, and optimization.	Saliency maps, activation patterns are applied in several models. An example of the diagnosis of Data filtering has been provided in detail.
MOD-EXP-REQ5: Extracted explanations shall be useful for the target audience to perform their tasks at any stage of the lifecycle, considering assumptions made on the level of qualification and skills.	In this report, we describe extracted explanations tailored for data analysts (during data management), AI developers (Saliency maps), and Decision-making (uncertainty-aware models)
MOD-EXP-REQ6: Explanations shall support traceability, allowing diagnosis of a case from prediction all the way back to the input data.	The use of data explainers and model explainers to diagnose a case to improve the 3D-OutDet model has been provided.
OM-EXP-REQ1: The AI system shall support continuous analysis of the model's behaviour in production to detect anomalies and unexpected patterns.	Data descriptor is used for the activation pattern space instead of the input data space.
OM-EXP-REQ2: The AI system shall support safety investigations of hazards where the AI model is involved, providing access to relevant data and explanations.	This requirement is related to the management process of runtime extracted explanation logging. The methods are similar to those used in the development lifecycle.
OM-EXP-REQ3: The AI system (safety-critical modules) shall provide uncertainty estimates for each prediction to inform inherent limitations of AI models and provide valuable information for decision-making, allowing for assessing the reliability of predictions.	Safety-critical models developed within WP5 are designed with uncertainty estimates.

7 Conclusions

One of the toughest issues in adoption of AI for safety-critical functions in ADAS/AVs is the immaturity state of technology and applicable approaches for:

- Explainability: Black box models can achieve high performance but are opaque and thus cannot properly guarantee the required performance for all possible traffic situations.
- Stakeholders such as AI developers, domain experts or regulators must know why a specific decision has been made.
- Current functional safety standards are not ready for AI-based systems (they impose overly strict requirements)
- Existing AI explainability techniques are not developed for this purpose
- Road traffic environments are highly unpredictable with irreducible uncertainty.
- Perception systems in road vehicles must proceed very high dimensional sensor data as inputs

Without the unrealistic ambition of fully solving the issue, Task T5.5 focused on raising awareness of the challenges among consortium partners as representative ecosystem, and across different tasks as representative aspects of building safe AI in automotive domain. We presented and discussed potential solutions through extensive inter-task and inter-work package workshops.

This report bridges the research area in Explainable AI (XAI) with the safety assurance process for AI-based perception systems developed within WP5. Through a systematic review of relevant standards, guidelines, and state of the art explainability techniques, we have developed a taxonomy that maps XAI methods to specific usages within the ROADVIEW perception stack. Explainability requirement specification has been refined from the initial specification in Task 2.4, and potential XAI enabled approaches that support compliance have been identified as the foundation for adopting explainability in practice for the relevant AI models in WP5.

All AI models developed in WP5 (specifically T5.1-5.3) adopted a decompositional architecture as part of XAI by design principles. Data explainability methods have been applied to support ROADVIEW's AI models such as weather noise filtering and slipperiness estimation. Safety-critical modules (including weather estimation, road surface condition estimation, visibility prediction) incorporate uncertainty aware design principles to provide real-time uncertainty estimates. Non-safety-critical modules (including BEV-based sensor fusion, infrastructure perception, weather noise filtering) integrated embedded explainability tools such as activation/gradient extraction and saliency maps for visual feature attribution. Combined with data explainers, these techniques enable end-to-end traceability for diagnoses. Operational monitoring component has also been prototyped as an AI-based Out-of-Distribution detector, serving as a key component within the safety runtime architecture.

The findings demonstrate that XAI can be integrated throughout the safe AI development lifecycle. Explainability supports identification of epistemic uncertainty during development lifecycle and supports mitigation. XAI by design, such as uncertainty estimates in safety-critical AI models, provides evidence for assessing the trustworthiness of AI predictions to the decision-making modules for the informed and safe decisions on road.

The established XAI framework aligned with the safety assurance process, complying with European regulatory expectations for AI-driven road vehicle systems. The proposed implementation strategy also facilitates the integration of new perception modules into the same explainability pipeline.

References

[1] "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)," *Official Journal of the European Union*, vol. L 1689, pp. 1–100, July 2024.

[2] International Organization for Standardization, *Information technology — Artificial intelligence — Objectives and approaches for explainability and interpretability of machine learning (ML) models and artificial intelligence (AI) systems*, 2025. [Online]. Available: https://www.iso.org/standard/82148.html

[3] International Organization for Standardization, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology (ISO/IEC 22989:2022)*, 2022. [Online]. Available: https://www.iso.org/standard/74296.html

[4] IEEE Computer Society, "IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence." 2024. [Online]. Available: https://standards.ieee.org/standard/2894-2024.html

[5] IEEE Computational Intelligence Standards Committee, "Standard for Explainable Artificial Intelligence - for Achieving Clarity and Interoperability of AI Systems Design." 2021. [Online]. Available: https://standards.ieee.org/standard/P2976.html

[6] International Organization for Standardization, "Information technology — Artificial intelligence — Data life cycle framework." 2023. [Online]. Available: https://www.iso.org/standard/83002.html

[7] International Organization for Standardization, "Information technology — Artificial intelligence — Management system." 2023. [Online]. Available: https://www.iso.org/standard/81230.html

[8] International Organization for Standardization, "Information technology — Artificial intelligence (AI) — Use cases." 2024. [Online]. Available: https://www.iso.org/standard/84144.html

[9] International Organization for Standardization, "Information technology — Artificial intelligence — Guidance for AI applications." 2024. [Online]. Available: https://www.iso.org/standard/81120.html

[10] International Organization for Standardization, "Artificial intelligence — Functional safety and AI systems (ISO Standard No. 5469:2024)," ISO, 2024. [Online]. Available: https://www.iso.org/standard/81283.html

[11] International Organization for Standardization, *Road vehicles — Functional safety (ISO Standard No. 26262:2018)*, 2018. [Online]. Available: https://www.iso.org/publication/PUB200262.html

[12] International Organization for Standardization, *Road vehicles — Safety of the intended functionality (ISO Standard No. 21448:2022)*, 2022. [Online]. Available: https://www.iso.org/standard/77490.html

[13] International Organization for Standardization, *Road vehicles — Safety and artificial intelligence (ISO Standard No. 8800:2024)*, 2024. [Online]. Available: https://www.iso.org/standard/83303.html

[14] Underwriters Laboratories, *Standard for Evaluation of Autonomous Products (UL 4600)*, 2023. [Online]. Available: https://ulstandards.ul.com/standard/?id=4600

[15] M. Consortium, "EASA Research – Machine Learning Application Approval (MLEAP) Final Report," EASA, May 2024. [Online]. Available: https://www.easa.europa.eu/sites/default/files/dfu/mleap-d4-public-report-issue01.pdf

[16] F. Clemente, G. M. Ribeiro, A. Quemy, M. S. Santos, R. C. Pereira, and A. Barros, "ydata-profiling: Accelerating data-centric AI with high-quality data," *Neurocomputing*, vol. 554, p. 126585, Oct. 2023, doi: 10.1016/j.neucom.2023.126585.

[17] F. Bertrand, *sweetviz: A pandas-based library to visualize and compare datasets.* Python. Accessed: Feb. 26, 2024. [OS Independent]. Available: https://github.com/fbdesignpro/sweetviz

[18] J. Wexler, "Facets: An open source visualization tool for machine learning training data," *Google Open Source Blog*, 2017.

[19] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[20] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018, doi: 10.21105/joss.00861.

[21] "Data Readiness – ianmarsh.org." Accessed: Feb. 26, 2024. [Online]. Available: https://ianmarsh.org/data-readiness/

[22] E. Tola, V. Lepetit, and P. Fua, "DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 815–830, May 2010, doi: 10.1109/TPAMI.2009.77.

[23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, June 2005, pp. 886–893 vol. 1. doi: 10.1109/CVPR.2005.177.

[24] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Kauai, HI, USA: IEEE Comput. Soc, 2001, p. I-511-I–518. doi: 10.1109/CVPR.2001.990517.

[25] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, Jan. 1996, doi: 10.1016/0031-3203(95)00067-4.

[26] M. Agrawal, K. Konolige, and M. R. Blas, "CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching," in *Computer Vision – ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, Eds., Berlin, Heidelberg: Springer, 2008, pp. 102–115. doi: 10.1007/978-3-540-88693-8_8.

[27] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International Conference on Computer Vision*, Nov. 2011, pp. 2564–2571. doi: 10.1109/ICCV.2011.6126544.

[28] G. H. Granlund, "In search of a general picture processing operator," *Computer Graphics and Image Processing*, vol. 8, no. 2, pp. 155–173, Oct. 1978, doi: 10.1016/0146-664X(78)90047-3.

[29] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Sept. 1999, pp. 1150–1157 vol.2. doi: 10.1109/ICCV.1999.790410.

[30] J. J. Koenderink and A. J. van Doorn, "Surface shape and curvature scales," *Image and Vision Computing*, vol. 10, no. 8, pp. 557–564, Oct. 1992, doi: 10.1016/0262-8856(92)90076-F.

[31] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22071–22080, Oct. 2019, doi: 10.1073/pnas.1900654116.

[32] T. Gebru *et al.*, "Datasheets for Datasets," Dec. 01, 2021, *arXiv*: arXiv:1803.09010. doi: 10.48550/arXiv.1803.09010.

[33] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski, "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards," May 09, 2018, *arXiv*: arXiv:1805.03677. doi: 10.48550/arXiv.1805.03677.

[34] E. M. Bender and B. Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 587–604, 2018, doi: 10.1162/tacl_a_00041.

[35] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, "Efficient Data Representation by Selecting Prototypes with Importance Weights," Aug. 12, 2019, *arXiv*: arXiv:1707.01212. doi: 10.48550/arXiv.1707.01212.

[36] A. Dhurandhar *et al.*, "Explanations based on the missing: towards contrastive explanations with pertinent negatives," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, in NIPS'18. Red Hook, NY, USA: Curran Associates Inc., Dec. 2018, pp. 590–601.

[37] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114 [cs, stat]*, May 2014, Accessed: Dec. 06, 2021. [Online]. Available: http://arxiv.org/abs/1312.6114

[38] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational Inference of Disentangled Latent Concepts from Unlabeled Observations," presented at the International Conference on Learning Representations, Mar. 2023. Accessed: Mar. 23, 2023. [Online]. Available: https://openreview.net/forum?id=H1kG7GZAW

[39] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014. [Online]. Available: http://arxiv.org/abs/1312.6034

[40] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," June 12, 2017, *arXiv*: arXiv:1703.01365. doi: 10.48550/arXiv.1703.01365.

[41] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," Dec. 13, 2015, *arXiv*: arXiv:1512.04150. Accessed: Mar. 23, 2023. [Online]. Available: http://arxiv.org/abs/1512.04150

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.

[43] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences," *arXiv:1605.01713 [cs]*, Apr. 2017, Accessed: July 12, 2021. [Online]. Available: http://arxiv.org/abs/1605.01713

[44] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, in ICML'17. Sydney, NSW, Australia: JMLR.org, Aug. 2017, pp. 3145–3153.

[45] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "SmoothGrad: removing noise by adding noise," *CoRR*, vol. abs/1706.03825, 2017, [Online]. Available: http://arxiv.org/abs/1706.03825

[46] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for Simplicity: The All Convolutional Net," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6806

[47] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class Activation Map using Principal Components," in *2020 International Joint Conference on Neural Networks (IJCNN)*, July 2020, pp. 1–7. doi: 10.1109/IJCNN48605.2020.9206626.

[48] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *arXiv:1602.04938 [cs, stat]*, Aug. 2016, Accessed: July 12, 2021. [Online]. Available: http://arxiv.org/abs/1602.04938

[49] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 4768–4777.

[50] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Art. no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11491.

[51] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 818–833. doi: 10.1007/978-3-319-10590-1_53.

[52] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," in *British Machine Vision Conference (BMVC)*, 2018. [Online]. Available: http://bmvc2018.org/contents/papers/1064.pdf

[53] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *SSRN Journal*, 2017, doi: 10.2139/ssrn.3063289.

[54] J. H. Friedman, "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991, doi: 10.1214/aos/1176347963.

[55] D. W. Apley and J. Zhu, "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 82, no. 4, pp. 1059–1086, June 2020, doi: 10.1111/rssb.12377.

[56] J. Donnelly, A. J. Barnett, and C. Chen, "Deformable ProtoPNet: An Interpretable Image Classifier Using Deformable Prototypes," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, June 2022, pp. 10255–10265. doi: 10.1109/CVPR52688.2022.01002.

[57] R. K. Been Kim and S. Koyejo, "Examples are not Enough, Learn to Criticize! Criticism for Interpretability," in *Advances in Neural Information Processing Systems*, 2016.

[58] R. Achtibat *et al.*, "From attribution maps to human-understandable explanations through Concept Relevance Propagation," *Nat Mach Intell*, vol. 5, no. 9, pp. 1006–1019, Sept. 2023, doi: 10.1038/s42256-023-00711-8.

[59] Q. Zhang, R. Cao, Y. N. Wu, and S.-C. Zhu, "Growing interpretable part graphs on ConvNets via multi-shot learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, in AAAI'17. San Francisco, California, USA: AAAI Press, Feb. 2017, pp. 2898–2906.

[60] P. McCullagh and J. Nelder, *Generalized Linear Models*. London, UK: Chapman & Hall / CRC, 1989.
[61] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, "Accurate intelligible models with pairwise interactions," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, in KDD '13. New York, NY, USA: Association for Computing Machinery, Aug. 2013, pp. 623–631. doi: 10.1145/2487575.2487579.
[62] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, pp. 81–106, 1986.
[63] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable Decision Sets: A Joint Framework for Description and Prediction," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1675–1684. doi: 10.1145/2939672.2939874.
[64] J. Jung, C. Concannon, R. Shroff, S. Goel, and D. G. Goldstein, "Simple Rules for Complex Decisions," Feb. 16, 2017, *Rochester, NY*: 2919024. doi: 10.2139/ssrn.2919024.
[65] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille, "A Bayesian Framework for Learning Rule Sets for Interpretable Classification," *Journal of Machine Learning Research*, vol. 18, no. 70, pp. 1–37, 2017.
[66] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Aug. 1995, pp. 278–282 vol.1. doi: 10.1109/ICDAR.1995.598994.
[67] M. Wu, M. C. Hughes, S. Parbhoo, M. Zazzi, V. Roth, and F. Doshi-Velez, "Beyond sparsity: tree regularization of deep models for interpretability," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, in AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press, Feb. 2018, pp. 1670–1678.
[68] A. Dhurandhar, K. Shanmugam, R. Luss, and P. A. Olsen, "Improving Simple Models with Confidence Profiles," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/972cda1e62b72640cb7ac702714a115f-Paper.pdf
[69] M. Hind *et al.*, "TED: Teaching AI to Explain its Decisions," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, in AIES '19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 123–129. doi: 10.1145/3306618.3314273.
[70] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, in NIPS'18. Red Hook, NY, USA: Curran Associates Inc., Dec. 2018, pp. 7786–7795.
[71] U. Schlegel, E. Cakmak, and D. A. Keim, "ModelSpeX: Model Specification Using Explainable Artificial Intelligence Methods," in *Machine Learning Methods in Visualisation for Big Data*, D. Archambault, I. Nabney, and J. Peltonen, Eds., The Eurographics Association, 2020. doi: 10.2312/mlvis.20201100.
[72] S. Kim, J. Nam, and B. C. Ko, "ViT-NeT: Interpretable Vision Transformers with Neural Tree Decoder," in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., in Proceedings of Machine Learning Research, vol. 162. PMLR, July 2022, pp. 11162–11172. [Online]. Available: https://proceedings.mlr.press/v162/kim22g.html
[73] M. G. Augasta and T. Kathirvalavakumar, "Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems," *Neural Processing Letters*, vol. 35, no. 2, pp. 131–150, Apr. 2012, doi: 10.1007/s11063-011-9207-8.
[74] V. Contreras *et al.*, "A DEXiRE for Extracting Propositional Rules from Neural Networks via Binarization," *Electronics*, vol. 11, no. 24, 2022, doi: 10.3390/electronics11244171.
[75] M. E. Zarlenga, Z. Shams, and M. Jamnik, "Efficient Decompositional Rule Extraction for Deep Neural Networks," Nov. 24, 2021, *arXiv*: arXiv:2111.12628. Accessed: Mar. 03, 2024. [Online]. Available: http://arxiv.org/abs/2111.12628
[76] J. R. Zilke, E. Loza Mencía, and F. Janssen, "DeepRED – Rule Extraction from Deep Neural Networks," in *Discovery Science*, T. Calders, M. Ceci, and D. Malerba, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 457–473. doi: 10.1007/978-3-319-46307-0_29.
[77] M. Craven and J. Shavlik, "Extracting Tree-Structured Representations of Trained Networks," in *Advances in Neural Information Processing Systems*, MIT Press, 1995. Accessed: Mar. 24, 2025. [Online]. Available:

https://proceedings.neurips.cc/paper_files/paper/1995/hash/45f31d16b1058d586fc3be7207b58053-Abstract.html

[78] J. H. Friedman and B. E. Popescu, "Predictive Learning via Rule Ensembles," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 916–954, 2008.

[79] G. Bologna and S. Fossati, "A Two-Step Rule-Extraction Technique for a CNN," *Electronics*, vol. 9, no. 6, 2020, doi: 10.3390/electronics9060990.

[80] "Neural-Symbolic Learning Systems," in *Neural-Symbolic Cognitive Reasoning*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 35–54. doi: 10.1007/978-3-540-73246-4_4.

[81] M. Fischer, M. Balunovic, D. Drachsler-Cohen, T. Gehr, C. Zhang, and M. Vechev, "DL2: Training and Querying Neural Networks with Logic," in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, May 2019, pp. 1931–1941. Accessed: Mar. 24, 2025. [Online]. Available: https://proceedings.mlr.press/v97/fischer19a.html

[82] R. Manhaeve, S. Dumančić, A. Kimmig, T. Demeester, and L. De Raedt, "Neural probabilistic logic programming in DeepProbLog," *Artificial Intelligence*, vol. 298, p. 103504, Sept. 2021, doi: 10.1016/j.artint.2021.103504.

[83] E. Giunchiglia, M. C. Stoian, and T. Lukasiewicz, "Deep Learning with Logical Constraints," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, Vienna, Austria: International Joint Conferences on Artificial Intelligence Organization, July 2022, pp. 5478–5485. doi: 10.24963/ijcai.2022/767.

[84] P. W. Koh *et al.*, "Concept Bottleneck Models," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., in Proceedings of Machine Learning Research, vol. 119. PMLR, July 2020, pp. 5338–5348. [Online]. Available: https://proceedings.mlr.press/v119/koh20a.html

[85] Z. Chen, Y. Bei, and C. Rudin, "Concept whitening for interpretable image recognition," *Nature Machine Intelligence*, vol. 2, no. 12, pp. 772–782, Dec. 2020, doi: 10.1038/s42256-020-00265-z.

[86] E. Goan and C. Fookes, "Bayesian Neural Networks: An Introduction and Survey," vol. 2259, 2020, pp. 45–87. doi: 10.1007/978-3-030-42553-1_3.

[87] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Proceedings of The 33rd International Conference on Machine Learning*, PMLR, June 2016, pp. 1050–1059. Accessed: Apr. 30, 2022. [Online]. Available: https://proceedings.mlr.press/v48/gal16.html

[88] F. Kraus and K. Dietmayer, "Uncertainty Estimation in One-Stage Object Detection," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, Oct. 2019, pp. 53–60. doi: 10.1109/ITSC.2019.8917494.

[89] D. Linsley, D. Shiebler, S. Eberhardt, and T. Serre, "Learning what and where to attend," June 11, 2019, *arXiv*: arXiv:1805.08819. Accessed: Mar. 03, 2024. [Online]. Available: http://arxiv.org/abs/1805.08819

[90] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA: IEEE, June 2015, pp. 842–850. doi: 10.1109/CVPR.2015.7298685.

[91] F. Wang *et al.*, "Residual Attention Network for Image Classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA: IEEE, July 2017, pp. 6450–6458. doi: 10.1109/CVPR.2017.683.

[92] X. Shi *et al.*, "Loss-Based Attention for Interpreting Image-Level Prediction of Convolutional Neural Networks," *IEEE Transactions on Image Processing*, vol. 30, pp. 1662–1675, 2021, doi: 10.1109/TIP.2020.3046875.

[93] S. Seo, J. Huang, H. Yang, and Y. Liu, "Interpretable Convolutional Neural Networks with Dual Local and Global Attention for Review Rating Prediction," in *Proceedings of the Eleventh ACM Conference on Recommender Systems*, Como Italy: ACM, Aug. 2017, pp. 297–305. doi: 10.1145/3109859.3109890.

[94] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proceedings of the 2011 International Conference on Computer Vision*, in ICCV '11. USA: IEEE Computer Society, Nov. 2011, pp. 2018–2025. doi: 10.1109/ICCV.2011.6126474.

[95] M. Yeganejou, S. Dick, and J. Miller, "Interpretable Deep Convolutional Fuzzy Classifier," *Trans. Fuz Sys.*, vol. 28, no. 7, pp. 1407–1419, July 2020, doi: 10.1109/TFUZZ.2019.2946520.

[96] A. Brando, I. Serra, E. Mezzetti, F. J. Cazorla, and J. Abella, "Standardizing the Probabilistic Sources of Uncertainty for the sake of Safety Deep Learning," in *Proceedings of the Workshop on Artificial Intelligence Safety 2023 (SafeAI 2023) co-located with the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI 2023), Washington DC, USA, February 13-14, 2023*, G. Pedroza, X. Huang, X. C. Chen, A. Theodorou, J.

Hernández-Orallo, M. Castillo-Effen, R. Mallah, and J. A. McDermid, Eds., in CEUR Workshop Proceedings, vol. 3381. CEUR-WS.org, 2023. [Online]. Available: https://ceur-ws.org/Vol-3381/11.pdf

[97] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, doi: 10.1109/TIT.1982.1056489.

[98] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, in KDD'96. Portland, Oregon: AAAI Press, Aug. 1996, pp. 226–231.

[99] European Union Aviation Safety Agency, "Artificial Intelligence Roadmap 2.0: Human-centric approach to AI in aviation." European Union Aviation Safety Agency, May 2023. [Online]. Available: https://www.easa.europa.eu/en/ai

[100] A. M. Raisuddin, T. Cortinhal, J. Holmblad, and E. E. Aksoy, "3D-OutDet: A Fast and Memory Efficient Outlier Detector for 3D LiDAR Point Clouds in Adverse Weather," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, June 2024, pp. 2862–2868. doi: 10.1109/IV55156.2024.10588582.

[101] A. M. Kurup and J. P. Bos, "Winter adverse driving dataset for autonomy in inclement winter weather," *Optical Engineering*, vol. 62, no. 3, p. 031207, 2023, doi: 10.1117/1.OE.62.3.031207.

[102] "SPHERICAL HARMONICS," in *Quantum Theory of Angular Momentum*, pp. 130–169. doi: 10.1142/9789814415491_0006.

[103] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," June 07, 2017, *arXiv*: arXiv:1706.02413. doi: 10.48550/arXiv.1706.02413.

[104] A. X. Chang *et al.*, "ShapeNet: An Information-Rich 3D Model Repository," Dec. 09, 2015, *arXiv*: arXiv:1512.03012. doi: 10.48550/arXiv.1512.03012.

[105] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.

[106] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Jan. 06, 2016, *arXiv*: arXiv:1506.01497. doi: 10.48550/arXiv.1506.01497.

[107] J. Maanpää *et al.*, "Dense Road Surface Grip Map Prediction from Multimodal Image Data," in *Pattern Recognition*, A. Antonacopoulos, S. Chaudhuri, R. Chellappa, C.-L. Liu, S. Bhattacharya, and U. Pal, Eds., Cham: Springer Nature Switzerland, 2025, pp. 387–404. doi: 10.1007/978-3-031-78447-7_26.

[108] J. Maanpää, J. Pesonen, I. Melekhov, H. Hyyti, and J. Hyyppä, "Road Grip Uncertainty Estimation Through Surface State Segmentation," in *Image Analysis*, J. Petersen and V. A. Dahl, Eds., Cham: Springer Nature Switzerland, 2025, pp. 231–244. doi: 10.1007/978-3-031-95911-0_17.