Scenario Classification via Imagery and Video Data for ADAS Trigger Analysis

Degree project report in Information and Communication Technology

WEIRUI ZHAO

DEPARTMENT OF ELECTRICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Degree project report 2026

Scenario Classification via Imagery and Video Data for ADAS Trigger Analysis

WEIRUI ZHAO

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden 2026

Scenario Classification via Imagery and Video Data for ADAS Trigger Analysis
Weirui Zhao

Supervisor: Robin Rahm, Volvo Cars
Examiner: Jennifer Alvén, Department of Electrical Engineering

Degree project report 2026
Department of Electrical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Scenario Classification via Imagery and Video Data for ADAS Trigger Analysis
WEIRUI ZHAO
Department of Electrical Engineering
Chalmers University of Technology

Abstract

Advanced Driver Assistance Systems (ADAS) generate large volumes of trigger events during development and validation. For low-speed reversing functions such as Rear Auto Brake (RAB), these events must be reviewed to distinguish recurring traffic situations and valid interventions from ambiguous or undesired activations. Manual event-by-event inspection is difficult to scale, while extensive ground-truth labels are generally unavailable. This thesis therefore investigates whether pre-trained visual representations and unsupervised clustering can structure industrial RAB trigger data for exploratory analysis.

An offline, configuration-driven pipeline was developed to represent each event by one key frame, extract frozen embeddings using DINOv2 ViT-S/14 or ResNet50, optionally reduce their dimensionality to 128 principal components, and cluster them using HDBSCAN or K-Means. The evaluation comprises 18 controlled experiments on the same 8,017 RAB trigger images. Cluster count, noise ratio, Silhouette score, and Davies–Bouldin index are considered together with PCA and UMAP visualizations.

DINOv2 combined with HDBSCAN produced the strongest internal cluster structure among the evaluated configurations, and PCA improved every paired DINOv2–HDBSCAN experiment. The best internal result reached a Silhouette score of 0.2709 and a Davies–Bouldin index of 1.4676, while assigning 46.8% of the samples to noise. DINOv2 also retained more events and separated them more strongly than ResNet50 in the two matched comparisons, whereas K-Means covered every event but produced substantially weaker separation. The high HDBSCAN noise ratios indicate considerable visual diversity and show that many events did not form sufficiently dense recurring groups.

A manual annotation of 820 events drawn from one reference run examined whether the discovered groups carry consistent semantics. Weighted purity reached 0.931 for trigger validity, 0.947 for scenario cause, and 0.943 for fine object type, while individual clusters still combined opposing validity judgments and the sampled noise group consisted mostly of clear true-positive observations of ordinary static obstacles. Self-supervised embeddings and density-based clustering therefore expose measurable and partly interpretable structure in unlabeled RAB trigger data and can support grouped inspection. The approach remains an exploratory structuring tool rather than an automated scenario classifier, because the semantic evidence came from one run and one reviewer, single key frames omit motion and sensor context, and the effect on review effort was not measured.

Keywords: ADAS, rear auto brake, representation learning, DINOv2, unsupervised clustering, HDBSCAN, scenario discovery, trigger analysis.

Acknowledgements

I would like to express my sincere gratitude to my supervisor at Volvo Cars, Robin Rahm, System Engineer, and to my manager, Shreyas Horantur, for giving me the opportunity to carry out this thesis within the Safe Vehicle Automation (SVA) department, and for their guidance and trust throughout the project.
I would also like to thank Amanda Siklund for her technical support throughout this work, including her help in tracking down and analysing errors, as well as the many discussions that turned dead ends into new directions. My thanks go equally to the other colleagues across the SVA department, who provided the data this thesis is built on, together with the tools, resources, and technical advice that made the work possible.
I would like to extend my thanks to Lakshmanan Swaminathan, Albin Maliqi, Flavio Fagundes, and the many other colleagues who invited me along to fika and group lunches. Thank you for making me feel part of the team from the very first weeks. Your openness made settling into a new workplace and a new country far easier than I had expected.
I am grateful to my examiner, Jennifer Alvén, Assistant Professor, for following this work closely from beginning to end, and for organising the Master Thesis Academy, whose sessions on reporting and academic writing have shaped how this report is structured and written.
I would also like to thank the friends and fellow students I met during these two years, for the explanations before exams, the encouragement when deadlines piled up, and the easy conversations in between, which kept a demanding workload from ever feeling heavy.
My thanks also go to Chalmers University of Technology and to everyone responsible for the Master's Programme in Information and Communication Technology, whose advice on course selection and study planning made these two years in Gothenburg run smoothly and turn out far more memorable than I had expected, and to all the lecturers and teaching assistants I met along the way, who built the foundation this thesis now rests on and who never seemed to tire of my questions after class.
Finally, I would like to thank my parents and my old friends at home. Studying abroad meant two years away from them, and their patience, encouragement, and unwavering support across that distance carried me through every stage of this work.

Weirui Zhao, Gothenburg, August 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

ADAS	Advanced Driver Assistance Systems
CNN	Convolutional Neural Network
CTA/b	Cross Traffic Alert/Brake
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
PCA	Principal Component Analysis
RAB	Rear Auto Brake
SRCB	Short Range Camera Back
UMAP	Uniform Manifold Approximation and Projection
UUID	Universally Unique Identifier
ViT	Vision Transformer

Nomenclature

Below is the nomenclature of indices, sets, parameters, and variables used throughout this thesis.

Indices

i	Index of a trigger event or image sample
k, j	Indices of a cluster
ℓ	Index of a manual semantic class
l	Index of a network layer or residual block

Sets

X	Set of selected RAB key frames used in the analysis
L_k	Set of labeled samples assigned to non-noise cluster k

Parameters

N	Total number of event samples
K	Number of non-noise clusters
d	Original embedding dimension
q	Retained PCA dimension
H, W	Height and width of an input image
m	HDBSCAN neighborhood size, the `min_samples` parameter
N_{-1}	Number of samples assigned the HDBSCAN noise label
$n_{k,\ell}$	Number of labeled samples of manual class ℓ in cluster k

Variables

e_i	Trigger event with index i
b	Visual backbone selected for feature extraction
a	Clustering algorithm selected for an experiment
$\boldsymbol{\lambda}$	Parameter set of the selected clustering configuration
t_i^{s}, t_i^{e}	Start and end of the activation interval selected for event e_i
t_i^*	Time of the selected key frame
F_i	Camera sequence belonging to event e_i
x_i	Selected camera key frame for event e_i
$f_{\boldsymbol{\theta}}$	Frozen feature extractor with parameters $\boldsymbol{\theta}$
$\boldsymbol{z}_i$	Visual embedding of key frame x_i
$\widehat{\boldsymbol{z}}_i$	L2-normalized visual embedding
$\overline{\boldsymbol{z}}^{(b)}$	Mean normalized embedding for backbone b
α	Angle between two normalized embeddings
$\boldsymbol{h}_l$	Input of residual block l
$\mathcal{F}$	Learned residual transformation
$\boldsymbol{Z}$	Centered matrix of normalized embeddings
$\boldsymbol{U}_q$	Matrix of the q leading principal directions
$\boldsymbol{Y}$	PCA-reduced representation
$\boldsymbol{y}_i$	Representation supplied to a clustering algorithm
c_i	Cluster assignment of sample i
$\boldsymbol{\mu}_k$	Centroid of cluster k
$\delta(a, b)$	Distance between two samples
$\delta_{\mathrm{core},m}$	Core distance for neighborhood size m
$\delta_{\mathrm{mreach},m}$	Mutual-reachability distance
r_{noise}	HDBSCAN noise ratio
$a(i)$, $b(i)$	Mean intra-cluster and nearest other-cluster distance of sample i
$s(i)$	Silhouette coefficient of sample i
S_k	Mean distance from samples in cluster k to its centroid
M_{kj}	Distance between the centroids of clusters k and j
p_k	Manual-label purity of cluster k
p_{weighted}	Purity weighted by labeled sample count

Contents

List of Figures

List of Tables

1
Introduction

1.1 Background and Motivation

Advanced Driver Assistance Systems (ADAS) support the driver in situations where limited visibility and nearby road users or obstacles can make maneuvering difficult. Two functions relevant to low-speed reversing are Cross Traffic Alert with auto brake (CTA/b) and Rear Auto Brake (RAB). CTA is intended to detect traffic crossing behind a reversing vehicle, while its auto-brake subfunction can intervene when a collision risk is identified. RAB instead addresses obstacles in or immediately behind the reversing path and can apply the brakes automatically [1]. These functions are complementary. One primarily concerns approaching cross traffic, whereas the other concerns obstacles behind the vehicle.

This thesis is conducted in collaboration with Volvo Cars and addresses the analysis of RAB development data. A trigger event denotes a short recording produced because the vehicle's onboard RAB related algorithm itself decided to brake during reversing; the trigger by itself does not indicate whether the intervention was functionally justified or whether a physical obstacle was actually present. Long-term observation of such trigger events has revealed that a substantial proportion are false positives, i.e., cases in which the vehicle braked even though no safety-critical obstacle was present. Identifying these false positives currently relies on manual review, in which engineers inspect the vehicle log data associated with each event, including the camera recording, vehicle speed, and an occupancy heatmap that indicates the obstacle presence and can reveal the outline of a physical object when the camera image itself is unclear, among other signals. This manual process becomes difficult to scale as the volume of recorded events grows, particularly because the data are predominantly unlabeled and some cases cannot be resolved from a single image alone. Public Volvo documentation corroborates that such false positives occur in practice, noting that conditions such as tall grass can cause unwanted braking interventions [1].

A predefined label set is difficult to construct before the recurring contents of the data are known. Event-by-event labeling would require substantial effort and could still overlook infrequent patterns. An exploratory organization of visually related events can instead provide an intermediate structure for review without assuming that all relevant categories have already been specified.

These conditions motivate the use of learned visual representations and unsupervised clustering. ResNet50 provides a supervised convolutional baseline [2], while DINOv2 provides self-supervised visual features without requiring labels from the

target dataset [3]. Clustering these representations may expose repeated visual structure and support grouped inspection, although the resulting clusters remain hypotheses about the data rather than validated scenario classes or judgments of system correctness.

1.2 Problem Formulation, Aim, and Scope

Building on this engineering bottleneck, this thesis reformulates the manual review problem as a representation-learning and data-structuring task. Each trigger-associated recording is treated as an observation of an initially unknown event category, and the goal is to obtain a compact representation in which events with related visual characteristics can be organized together.
The academic question is therefore whether the intrinsic structure of such industrial trigger data can be recovered without extensive manual labeling, and whether the resulting structure is coherent, interpretable, and useful for systematic analysis. The resulting groups are intended to support exploration rather than provide verified scenario labels, pass–fail decisions, or judgments about functional correctness.
The thesis aims to develop a data-driven approach for automatically structuring and analyzing RAB trigger scenarios from visual data. It investigates how modern image- and video-based representations can support the discovery of recurring false-positive patterns, corner cases, and meaningful scenario categories in an industrial environment. The study also evaluates whether representation-based organization provides a scalable and interpretable basis for ADAS validation even when full automation of scenario classification is not achieved.
The scope is restricted to camera-based data associated with RAB trigger events in low-speed reversing scenarios. Each event is represented by an extracted camera key frame. An RAB event record also contains ultrasonic sensor signals and other vehicle logs, but this study does not use these additional modalities. The work analyzes trigger outputs from existing RAB functions and does not redesign the underlying ADAS algorithms. It emphasizes offline analysis with pretrained feature extractors and evaluation on industrial data that may contain imbalance, noise, and limited labels. Large-scale end-to-end training, real-time system integration, and vehicle deployment are outside the scope of the thesis.

1.3 Related Work

1.3.1 Scenario-Based Validation and Its Data Requirements

Scenario-based validation organizes the operating space of automated vehicles into descriptions that can be analyzed and tested [4]. Ulbrich et al. defined a scenario as a temporal development of scenes containing actions and events [5]. Menzel et al. distinguished functional, logical, and concrete descriptions for different development stages [6]. A useful scenario catalogue therefore depends on recorded observations being converted into searchable descriptions.
Recorded data do not automatically provide such a catalogue. Relevant recordings

must first be identified, represented, and grouped. In this study the identification step is given by the on-board RAB trigger and by metadata filtering of the event database; the work therefore addresses the representation and grouping steps, using one extracted key frame per event. Its image groups may support later scenario definition, but similar key frames do not constitute a complete scenario because temporal development and test parameters remain absent.

1.3.2 Data-Driven Scenario Extraction from Recorded Driving Data

Data-driven scenario clustering also includes work on simulated traffic. Kruber et al. used a modified unsupervised random forest to derive similarities between simulated highway scenarios and train a classifier from the resulting groups [7]. This approach demonstrated automatic category discovery, but it did not extract scenarios from recorded measurements.

Studies using recorded data have examined several structured inputs. Kerber et al. applied spatiotemporal filtering and hierarchical clustering to naturalistic highway data, using a scenario-specific distance measure to organize the scenario space and discuss test coverage [8]. Ries et al. clustered urban sequences containing trajectories of multiple traffic objects [9]. Both represented traffic development through structured motion information rather than raw camera appearance.

Other work started from vehicle signals or accident records. Montanari et al. segmented multivariate vehicle data, clustered recurring patterns, and interpreted four groups as potential driving scenarios while also observing corner cases [10]. Weber et al. proposed an unsupervised pipeline for urban intersections and compared its clusters with a rule-based reference. Their results exposed a trade-off between catalogue detail and testing effort [11]. Sander and Lübbe clustered reconstructed intersection accidents to derive AEB test cases, but cluster medoids did not reliably reproduce the intervention outcome within each cluster [12]. Together, these validation outcomes show that no single clustering score establishes semantic or functional validity. This motivates the complementary use of internal metrics, sample coverage, projections, and image inspection in this thesis.

The reviewed approaches provide important foundations for data-driven catalogues, yet their inputs mainly consist of trajectories, object states, vehicle signals, map context, or accident variables. The present dataset differs because each sample is an RAB trigger whose functional correctness remains uncertain and the analysis input is an extracted camera key frame. Structured trajectories and object lists are not used and therefore cannot serve as the primary representation.

The application domains also differ. The representative studies above address highway traffic, urban intersections, general driving patterns, or reconstructed forward-driving collisions. They provide limited evidence for low-speed reversing brake-trigger analysis. This difference motivates an image-level investigation that organizes events by visible content without assuming a predefined scenario taxonomy or requiring structured traffic trajectories.

1.3.3 Corner Cases and Anomalies in ADAS Perception

Corner-case research clarifies the limits of image-based event discovery. Breitenstein et al. organized visual corner cases into pixel, domain, object, scene, and scenario levels with increasing detection complexity [13]. Their framework distinguishes effects visible in one image from cases that depend on development over time. Heidecker et al. extended the camera-focused systematization to radar and lidar and introduced method-layer corner cases associated with model or data uncertainty [14]. Bogdoll et al. surveyed anomaly detection across sensor modalities and abstract object-level data [15]. These studies show that detection depends on the sensing modality and the abstraction level at which an event becomes unusual.
These distinctions are relevant to RAB key frames. A single frame can expose repeated visual conditions and object or scene composition. It cannot reconstruct relative motion, the sequence preceding the intervention, or the full sensor evidence used by the vehicle. An infrequent group is not automatically a corner case, and a frequent group is not automatically a false-positive pattern. Clustering therefore supports candidate discovery rather than anomaly certification or functional judgment.

1.3.4 Visual Representation Learning for Unlabeled Collections

Research on unlabeled collections offers different ways to combine representation learning and clustering. DeepCluster jointly optimized convolutional features and cluster assignments on general-purpose images [16]. Zhao et al. learned spatial and temporal features from traffic-agent trajectories and map information before clustering driving sequences [17]. Their results supported self-supervised feature extraction as an alternative to handcrafted scenario variables, although the representation still relied on structured temporal inputs.
The present study takes a constrained transfer-learning route. It keeps the pretrained ResNet50 and DINOv2 backbones fixed and compares their frame-level structures. This design isolates the effects of representation, dimensionality reduction, and clustering configuration. It does not evaluate joint representation learning, fine-tuning, or temporal encoders, but it permits controlled comparison without a large labeled target-domain dataset.
Taken together, prior work establishes scenario-based validation, data-driven scenario extraction, anomaly taxonomies, and unsupervised visual representation learning as relevant foundations. However, the reviewed scenario-extraction studies largely depend on structured temporal data, while the visual clustering studies address general image collections or driving representations constructed from trajectories and maps. The specific problem of organizing proprietary low-speed RAB trigger key frames remains insufficiently studied. This thesis addresses that gap through a controlled comparison of frozen visual representations and clustering configurations, with separate attention to internal cluster structure, sample coverage, and limited semantic interpretation. The resulting evidence concerns the organization of recorded events and does not establish complete scenarios or the functional

correctness of individual interventions.

1.4 Research Questions and Contributions

1.4.1 Research Questions

Based on the engineering problem and the gaps identified in related work, the study examines how learned visual representations and unsupervised clustering can support the analysis of predominantly unlabeled ADAS trigger data. The research questions connect representation quality, the coherence of the discovered groups, the effects of alternative representation strategies, and the practical interpretability of the resulting structure.

1. How do pretrained DINOv2 and ResNet50 representations affect the organization of frame-level RAB trigger events?
2. To what extent can HDBSCAN and K-Means reveal recurring visual structure in predominantly unlabeled RAB trigger data?
3. How do PCA and the main clustering configurations affect cluster compactness, separation, granularity, and sample coverage?
4. To what extent are the discovered groups visually and semantically interpretable, and what limitations arise from single-frame representations and limited annotation?

1.4.2 Contributions

The thesis formulates industrial RAB trigger review as a frame-level unsupervised data-structuring problem and implements a controlled, artifact-preserving pipeline connecting key-frame preparation, frozen feature extraction, optional PCA128, clustering, visualization, and metric aggregation. A controlled 18-run matrix enables paired comparisons on a consistent proprietary event set.

The evaluation provides evidence about the trade-off between cluster geometry and coverage. PCA128 improves the paired DINOv2 HDBSCAN experiments and the K-Means Silhouette scores. HDBSCAN identifies density-based groups but leaves a substantial fraction of samples as noise, whereas K-Means assigns every event and exhibits relatively low internal separation. The thesis further combines internal metrics, projections, organized image groups, and a limited annotation protocol for interpretation. Its contribution lies in the controlled industrial evaluation of established methods, not in a new clustering algorithm or a demonstrated improvement in review time or vehicle safety.

The resulting pipeline is intended as an analysis aid rather than an autonomous decision mechanism. It produces comparable groups, assignments, and visual summaries that can guide subsequent inspection, while responsibility for interpreting an event remains with the engineering analysis and the available contextual evidence.

1.5 Thesis Outline

Chapter 2 presents the preliminaries of the study. It defines the RAB trigger event and its frame-level representation, and then introduces visual representation learning, dimensionality reduction, unsupervised clustering, visualization, and the principles used to evaluate cluster structure.
Chapter 3 describes the industrial trigger data and the analysis methodology. It explains data preparation, feature extraction, event grouping, visualization, and the procedure used to interpret and validate the resulting structure.
Chapter 4 presents the experimental design and the quantitative, qualitative, and semantic results. It relates the findings to the research questions and discusses their industrial implications, limitations, and threats to validity.
Chapter 5 summarizes the main findings and contributions of the thesis. It identifies directions for stronger validation, temporal and multimodal representations, and possible integration into industrial analysis workflows.

2
Preliminaries

This chapter presents the concepts required to understand the proposed analysis of RAB trigger data. It first defines the trigger event as the unit of analysis and clarifies what is retained and lost when an event is represented by one camera frame. It then introduces frozen visual feature extraction with ResNet50 and DINOv2, principal component analysis, K-Means, and HDBSCAN. The final section explains the PCA and UMAP projections and the evaluation principles used to judge whether an unsupervised partition is geometrically coherent and semantically useful. Dataset preparation, parameter choices, and the implemented processing pipeline are described in Chapter 3.

2.1 RAB Trigger-Event Representation

2.1.1 RAB Trigger Events

RAB is a low-speed reversing support function that can brake the vehicle when an obstacle is detected behind it [1]. In this thesis, a trigger event is one recorded log segment produced in association with an RAB braking decision. Each event carries a unique identifier and contains one or more bounded activation intervals during which the RAB brake request is active. The identifier links the rear-facing camera recording, the vehicle signals, and any later annotation to the same record, while the activation intervals locate the system output within that recording. When an event contains several activation intervals, one interval is selected for frame extraction, so that every event contributes exactly one visual sample.

The activation is an observed system output rather than a semantic label. It establishes that the onboard function decided to intervene, but it does not by itself establish why the intervention occurred or whether it was functionally justified. A safety-relevant object may be visible, an environmental pattern may appear more likely to explain an undesired activation, or the available evidence may remain ambiguous. Consequently, terms such as valid activation, false positive, and unclear event describe interpretations supported by later inspection rather than properties encoded in the trigger identifier.

An event record may contain several camera frames and additional vehicle or sensor signals. The camera view provides direct visual evidence about objects, surfaces, illumination, and scene layout, but it does not reproduce all information available to the vehicle function. A single image also lacks explicit motion and depth information. The absence of a clearly visible obstacle therefore does not prove that no relevant

object or sensor response existed, while the presence of an object does not prove that it caused the intervention. The trigger event is thus the engineering observation to be organized, whereas its cause and functional validity remain latent attributes to be examined.

2.1.2 Frame-Level Representation

Let $[t_i^{\mathrm{s}}, t_i^{\mathrm{e}}]$ denote the activation interval selected for event e_i. A frame-level representation selects one camera image at a time $t_i^* \in [t_i^{\mathrm{s}}, t_i^{\mathrm{e}}]$ and defines

$$x_i = F_i(t_i^*), \tag{2.1}$$

where F_i denotes the camera sequence belonging to event e_i, and x_i is its selected key frame. The time t_i^* is taken at the onset of the activation, so that the image shows the scene at the moment of the intervention rather than an arbitrary position in the recording. This construction creates exactly one visual sample for each event. The representation is computationally efficient and gives every event equal weight during feature extraction and clustering. Because the number of retained images does not depend on how long or how often the function intervened, events with long or repeated activations are not over-represented. It also creates a consistent unit for manual inspection because an event can be linked to one image rather than to a variable-length recording. These properties make large collections easier to process and compare.
The simplification changes the meaning of the analysis. A key frame is a static observation of visible scene content rather than a complete driving scenario, since a scenario includes temporal development [5]. The frame does not directly encode relative motion, changes before and after braking, or evidence from unused sensing modalities. It may also miss an object that is occluded or outside the selected view. Frame-level clusters can therefore reveal recurring visual configurations, but they cannot independently reconstruct the full trigger cause or certify functional correctness.

2.2 Pretrained Visual Representations

2.2.1 Image Embeddings and Frozen Feature Extraction

An RGB image with height H and width W contains $3HW$ pixel values. Direct Euclidean comparison in this space is strongly affected by illumination, small translations, and other low-level changes that need not alter scene meaning. A visual feature extractor instead maps each key frame to a lower-dimensional vector

$$\boldsymbol{z}_i = f_{\boldsymbol{\theta}}(x_i), \qquad \boldsymbol{z}_i \in \mathbb{R}^d, \tag{2.2}$$

where $f_{\boldsymbol{\theta}}$ is a neural network with parameters $\boldsymbol{\theta}$, d is the embedding dimension, and $\boldsymbol{z}_i$ is the image embedding. The representation is useful when distances between embeddings reflect visual regularities more effectively than distances between raw pixels.

The extractor is frozen when $\boldsymbol{\theta}$ remains unchanged on the target data. Feature extraction then requires no RAB training labels and avoids fitting a high-capacity model to a small annotated subset. It also permits a controlled comparison of representations because later processing operates on fixed vectors. The corresponding limitation is that the representation cannot adapt to domain-specific cues that were absent from pretraining.
Embeddings with different magnitudes are normalized before distance-based analysis. For a nonzero vector, L2 normalization gives

$$\hat{\boldsymbol{z}}_i = \frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|_2}. \tag{2.3}$$

In this thesis every embedding is normalized in this way before any further processing, so normalization precedes both dimensionality reduction and clustering. For two normalized vectors, squared Euclidean distance equals $2(1 - \cos\alpha)$, where α is the angle between them. Normalization therefore removes magnitude as a source of variation and makes Euclidean comparisons equivalent to ranking by cosine similarity. This equivalence holds in the normalized space itself; a subsequent linear projection does not preserve unit norm, so in the configurations that apply PCA the Euclidean distances used for clustering are no longer equivalent to a cosine ranking. Normalization also does not guarantee that the remaining geometry corresponds to the desired engineering categories.

2.2.2 ResNet50

Convolutional neural networks build image representations by applying learned filters across spatial locations and progressively aggregating local patterns. ResNet introduced residual blocks to support substantially deeper networks [2]. A residual block represents its output as

$$\boldsymbol{h}_{l+1} = \mathcal{F}(\boldsymbol{h}_l) + \boldsymbol{h}_l, \tag{2.4}$$

where $\boldsymbol{h}_l$ is the block input and $\mathcal{F}$ is a learned residual transformation. This form applies when the block input and output have the same shape; where the shape changes, a linear projection is applied on the shortcut. The identity path allows information and gradients to pass directly through the block when the residual transformation is small.
ResNet50 contains 50 learned layers and is commonly pretrained under supervision on ImageNet. Removing its final classification layer converts the network from a fixed-category classifier into a general feature extractor. Global average pooling then produces one 2048-dimensional vector per image. These features reflect object, texture, and spatial patterns learned during supervised pretraining. In this thesis, ResNet50 provides a conventional convolutional baseline against which a self-supervised transformer representation can be compared. Its inclusion does not assume that ImageNet categories coincide with RAB event categories.

2.2.3 DINOv2

A Vision Transformer divides an image into non-overlapping patches, maps the patches to tokens, adds positional information, and processes the resulting sequence

with self-attention [18]. A learnable class token can aggregate information from all patch tokens into one global image representation. In contrast to the local receptive fields imposed by convolution, self-attention can directly model relations between distant image regions.

DINOv2 is a family of transformer-based models trained to produce transferable visual features without manual class labels [3]. Its training combines self-distillation objectives with additional regularization and data curation so that the pretrained backbone can support multiple downstream tasks. The target RAB frames are not used to update this backbone in the present study. The global class token is instead treated as a fixed embedding for each key frame. This thesis uses the ViT-S/14 variant, whose class token gives a 384-dimensional embedding per key frame.

Self-supervised pretraining is relevant because the target collection is predominantly unlabeled and its recurring categories are not known in advance. Such pretraining avoids tying the representation directly to a predefined target label set. Nevertheless, a general-purpose embedding is not guaranteed to preserve every low-speed reversing cue. Comparison with ResNet50 and subsequent semantic inspection remain necessary to determine whether its geometry is useful for this dataset.

2.3 Dimensionality Reduction and Clustering

2.3.1 Principal Component Analysis

Principal Component Analysis (PCA) replaces correlated variables with orthogonal directions that successively maximize sample variance [19]. Let the centered matrix of normalized embeddings be $\boldsymbol{Z} \in \mathbb{R}^{N \times d}$, with one event per row. If $\boldsymbol{U}_q \in \mathbb{R}^{d \times q}$ contains the eigenvectors associated with the q largest eigenvalues of the sample covariance matrix, the reduced representation is

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{U}_q, \qquad q < d. \tag{2.5}$$

The columns of $\boldsymbol{Y}$ are uncorrelated within the sample, and the retained eigenvalues quantify the variance explained by the projection.

PCA can reduce storage and computation while suppressing directions that contribute little total variance. For Euclidean reconstruction, the leading components also give the best rank-q linear approximation. High variance is not equivalent to semantic relevance, however. A direction that separates meaningful RAB conditions may have low global variance and may be weakened or removed. PCA must therefore be evaluated as an experimental factor rather than assumed to improve clustering.

This thesis distinguishes two uses of PCA. PCA128 denotes a 128-dimensional representation supplied to a clustering algorithm. PCA 2D denotes a separate projection used only to display a result. A visually clear two-dimensional plot does not imply that the 128-dimensional or original feature space has the same separation.

2.3.2 K-Means

K-Means partitions N vectors into a predetermined number K of clusters by minimizing within-cluster squared Euclidean distance [20]. For data vectors $\boldsymbol{y}_i$, assignments c_i, and centroids $\boldsymbol{\mu}_k$, its objective is

$$\min_{\{c_i\},\{\boldsymbol{\mu}_k\}} \sum_{i=1}^{N} \left\| \boldsymbol{y}_i - \boldsymbol{\mu}_{c_i} \right\|_2^2, \qquad c_i \in \{1, \ldots, K\}. \tag{2.6}$$

The usual iterative procedure [21] alternates between assigning each vector to its nearest centroid and recomputing centroids as cluster means. The objective decreases at each step, but the solution can depend on initialization and need not be globally optimal.
Every sample receives a cluster assignment, and K must be specified before optimization. Under Euclidean distance, the resulting decision regions are Voronoi cells around the centroids. The method therefore tends to favor compact, convex groups with comparable dispersion and can divide an elongated group or absorb isolated observations into the nearest cluster. These properties make K-Means unsuitable as an outlier detector but useful as a transparent complete-partition baseline. It provides a contrast with HDBSCAN when the number, density, and shape of recurring event groups are unknown.

2.3.3 HDBSCAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) identifies groups that persist across a hierarchy of density levels [22]. Let $\delta(a, b)$ denote the distance between two samples. For a chosen neighborhood parameter m, the core distance $\delta_{\text{core},m}(x)$ is the distance from x to its m-th nearest neighbor. The mutual-reachability distance between two samples a and b is

$$\delta_{\text{mreach},m}(a, b) = \max\{\delta_{\text{core},m}(a), \delta_{\text{core},m}(b), \delta(a, b)\}\,. \tag{2.7}$$

This transformation increases distances involving sparse regions while leaving sufficiently dense relations largely unchanged. A hierarchy constructed from these distances is condensed according to minimum cluster size, after which stable branches are selected as clusters.
Two parameters have distinct roles. The parameter `min_samples` is the neighborhood size m used in the core distance above, so it controls how local density is estimated. Increasing it generally requires stronger local support and can make the result more conservative. The parameter `min_cluster_size` specifies the smallest group that may persist as a cluster in the condensed hierarchy. HDBSCAN derives the number of selected clusters from the data and these settings rather than requiring it directly.
Samples that do not belong to a selected stable group receive the noise label -1. This mechanism is relevant to heterogeneous and long-tailed event collections because isolated samples are not forced into a centroid-based partition. The noise label has only a density-based meaning. It may contain rare events, transition cases, visually ambiguous observations, or ordinary samples that lack sufficient local support. It

must not be interpreted automatically as a corner case, a false positive, or an invalid recording.

2.4 Visualization and Evaluation

2.4.1 PCA and UMAP Visualization

Two-dimensional projections provide an overview of high-dimensional embeddings and cluster assignments. PCA 2D uses the first two principal components and provides a deterministic linear view of the directions with greatest sample variance. It is useful for identifying broad gradients and overlap, but two components may explain only a limited part of the total variation. Both projections in this thesis are computed from the same feature space that is given to the clustering algorithm, so they display the space in which the clusters were formed rather than the original embedding space.

Uniform Manifold Approximation and Projection (UMAP) builds a neighborhood graph in the feature space supplied to it and optimizes a low-dimensional representation intended to retain local topological structure [23]. It can display nonlinear neighborhoods that a linear projection does not expose. The apparent distance between separated groups, their area, and their orientation in a UMAP plot do not have a direct physical interpretation. The projection also depends on neighborhood, distance, and optimization settings.

PCA and UMAP therefore offer complementary visual evidence rather than proofs of cluster validity. A compact island can be created or exaggerated by projection, while a high-dimensional group can overlap after compression to two dimensions. Cluster labels are overlaid only to support inspection of representative images and geometric patterns. Conclusions about usefulness require evaluation in the clustering space and comparison with manual semantics.

2.4.2 Internal Clustering Metrics

Internal metrics evaluate a partition using only feature-space distances and assignments. The number of non-noise clusters describes granularity but not quality. For HDBSCAN, the noise ratio reports the fraction of samples rejected from selected clusters and is defined as

$$r_{\text{noise}} = \frac{N_{-1}}{N}, \tag{2.8}$$

where N_{-1} is the number of samples labeled -1, and N is the total number of events. A larger cluster count may reflect useful detail or fragmentation, while a lower noise ratio may reflect broader coverage or the inclusion of weakly related samples.

The Silhouette coefficient compares cohesion with separation [24]. For sample i, let $a(i)$ be its mean distance to other samples in the same cluster and let $b(i)$ be the smallest mean distance from i to any other cluster. Its coefficient is

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \qquad -1 \leq s(i) \leq 1. \tag{2.9}$$

Values near one indicate that the sample is closer to its own cluster than to the nearest alternative. Values near zero indicate overlap, and negative values indicate that another cluster is closer on average. The reported Silhouette score is the mean across evaluated samples.
The Davies–Bouldin index compares within-cluster scatter with separation between cluster centroids [25]. Let S_k denote the mean distance from samples in cluster k to centroid $\boldsymbol{\mu}_k$, and let M_{kj} denote the distance between centroids k and j. The index is

$$\mathrm{DB} = \frac{1}{K} \sum_{k=1}^{K} \max_{j \neq k} \frac{S_k + S_j}{M_{kj}}. \tag{2.10}$$

Lower values indicate that each cluster is compact relative to its most similar competing cluster. Both indices are defined through centroids or mean pairwise distances and therefore reward compact, approximately convex groups. Density-based clusters of irregular shape can be penalized by this assumption, so a comparison between a density-based and a centroid-based method is not neutral with respect to cluster geometry. Neither index identifies whether a group has engineering meaning.
Metric coverage must accompany any comparison. K-Means assigns every event, so its internal scores evaluate the complete collection. HDBSCAN scores in this thesis exclude noise because label -1 is not a cluster. Its scores therefore evaluate only the retained subset. A higher Silhouette score or lower Davies–Bouldin index for HDBSCAN may partly result from rejecting difficult samples. Internal scores cannot be compared fairly without also reporting evaluated sample count and noise ratio.

2.4.3 Semantic Evaluation

Semantic evaluation tests whether feature-space groups correspond to interpretable properties of the images. A limited subset is manually annotated with a fixed label guide. The labels may describe visible scene content, an apparent trigger cause, or whether the available frame is sufficient for judgment. Manual labels remain interpretations of the available evidence and should preserve an unclear category when the key frame is insufficient.
Let L_k be the labeled samples assigned to non-noise cluster k, and let $n_{k,\ell}$ be the number carrying manual class ℓ. Per-cluster purity is

$$p_k = \frac{\max_\ell n_{k,\ell}}{|L_k|}. \tag{2.11}$$

Purity equals one when all labeled samples in the cluster share the same manual label. It decreases as the labeled composition becomes more mixed. Weighted purity summarizes the evaluated clusters according to their labeled sample counts and is defined as

$$p_{\text{weighted}} = \frac{\sum_{k=1}^{K} |L_k| p_k}{\sum_{k=1}^{K} |L_k|}. \tag{2.12}$$

Per-cluster label distributions and labeled counts must accompany this aggregate because a high value can be dominated by large or easily interpreted groups. HDBSCAN noise is analyzed separately rather than treated as an ordinary semantic cluster.

Purity depends on the chosen taxonomy and normally increases when an algorithm creates more, smaller clusters. It also reflects only the manually labeled subset. Internal metrics and semantic evaluation therefore answer different questions. The former describe cohesion, separation, and coverage in the selected feature space, while the latter examines whether the discovered groups support consistent human interpretation. Visual inspection, coverage, internal scores, and manual labels must be considered together before a cluster is described as meaningful.

3
Method

This chapter describes the method used to convert proprietary RAB trigger records into a structured collection of visual groups. The study treats each trigger event as one observation and preserves an explicit link from the selected key frame to its event identifier. Frozen visual backbones then transform the frames into embeddings, after which optional PCA and unsupervised clustering produce the candidate groups. The method concludes with internal evaluation, organized image inspection, and a limited semantic evaluation based on manual annotations. The design supports controlled comparison of representation and clustering choices without treating the resulting groups as verified scenario classes.

3.1 Method Overview and Problem Formulation

3.1.1 Pipeline Overview

Figure 3.1 summarizes the processing chain. Event acquisition first identifies RAB records in which the brake request became active. Timestamp matching then selects one rear-camera frame from an activation interval and retains the event identifier in the image name. Two frozen backbones independently encode the same frame collection. The resulting embeddings follow either a direct path or a PCA128 path before HDBSCAN or K-Means assigns cluster labels. The pipeline saves assignments, run configurations, and two-dimensional projections, a later evaluation stage scores the stored runs, and a separate organization step arranges the images for manual inspection. A stratified annotation of one reference run provides additional evidence about semantic consistency and the composition of HDBSCAN noise.
The stages serve different purposes. Event and frame preparation establish a consistent unit of analysis. Frozen feature extraction provides transferable visual descriptions without fitting a model to the target labels. PCA tests whether a compact linear representation improves the geometry available to clustering. HDBSCAN examines recurring density structure and can leave unsupported samples as noise, whereas K-Means supplies a complete-partition baseline. The final evaluation separates geometric evidence from human interpretation so that a visually compact group is not automatically described as a valid engineering category.

3.1.2 Mathematical Formulation

Let the prepared dataset be

$$X = \{x_i\}_{i=1}^{N}, \tag{3.1}$$

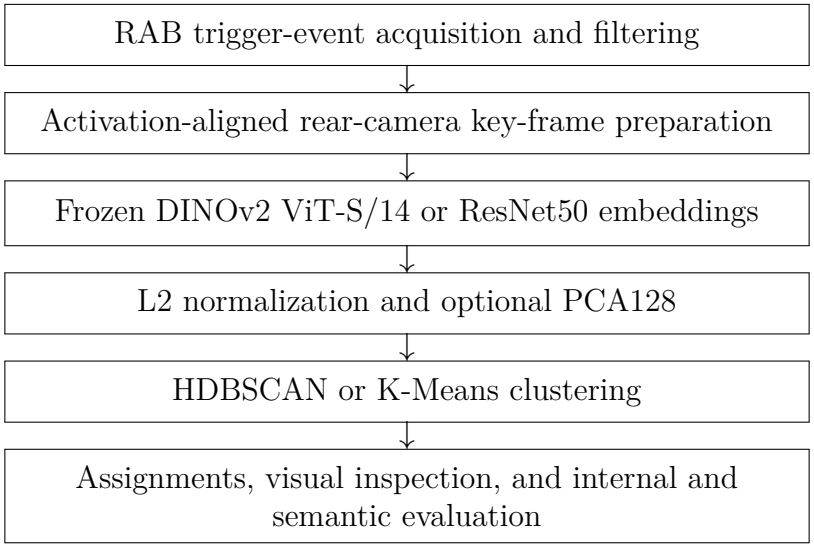


Figure 3.1: Processing pipeline from RAB trigger records to evaluated visual groups.

where x_i is the key frame selected for event e_i and $N = 8{,}017$. For a selected backbone b, frozen feature extraction produces

$$\boldsymbol{z}_i^{(b)} = f_{\boldsymbol{\theta}_b}(x_i), \qquad \widehat{\boldsymbol{z}}_i^{(b)} = \frac{\boldsymbol{z}_i^{(b)}}{\|\boldsymbol{z}_i^{(b)}\|_2}, \tag{3.2}$$

where b denotes DINOv2 ViT-S/14 or ResNet50 and the parameters $\boldsymbol{\theta}_b$ remain fixed. The image name stored at row i preserves the correspondence between x_i, its embedding, and the original event identifier.
Each experiment supplies one of two representations to clustering. The no-PCA path uses

$$\boldsymbol{y}_i = \widehat{\boldsymbol{z}}_i^{(b)}, \tag{3.3}$$

whereas the PCA128 path centers the normalized embeddings and applies the 128 leading principal directions,

$$\boldsymbol{y}_i = \left(\widehat{\boldsymbol{z}}_i^{(b)} - \overline{\boldsymbol{z}}^{(b)}\right) \boldsymbol{U}_{128}. \tag{3.4}$$

The clustering algorithm then maps each representation to a label

$$c_i = g_{a,\boldsymbol{\lambda}}(\boldsymbol{y}_i), \tag{3.5}$$

where a identifies HDBSCAN or K-Means and $\boldsymbol{\lambda}$ contains the parameters of the selected configuration. For K-Means, $c_i \in \{0, \ldots, K-1\}$ for every event. For HDBSCAN, $c_i = -1$ denotes a sample outside the selected stable clusters. This formulation makes the visual backbone, PCA choice, clustering method, and clustering parameters explicit experimental factors while keeping the event collection fixed.

3.2 Data Preparation

3.2.1 Event Acquisition and Filtering

The source data consist of RAB trigger records stored in Volvo Cars' internal data lake. Each record is indexed by an event UUID and contains a rear-camera log together with a vehicle-signal payload. The camera log uses ZVLF, a proprietary Volvo Cars container that stores an ordered sequence of compressed camera images together with a system timestamp for each image. The format is documented internally and is read here through the vendor decoder, so the thesis describes only the properties the method depends on, namely frame ordering and per-frame timestamps. The payload contains a brake-request source signal and its timestamps. This categorical signal identifies the vehicle function associated with the active brake request. Access to the event database, decoded-log storage, and the extraction container requires the internal computing environment.

To evaluate the method, the acquisition query selected Volvo RAB events whose signal summary indicated that RAB was identified as the source of the active brake request at least once. The records were additionally filtered according to the legal ground required for use of the internal data. The query accepts one legal ground per execution, so the extraction ran separately for each applicable legal ground and the resulting images were merged into one dataset. The UUID identifies the event at every stage. Because it also forms the prefix of the key-frame file name, the extraction records, image file, embeddings, cluster assignment, and later annotation all refer to the same engineering observation.

The filtering step established that an event contained an active brake request attributed to RAB, but it did not assign a semantic class. Events remained in the collection regardless of whether later inspection suggested a valid intervention, a false positive, or insufficient evidence. This decision prevented the unlabeled clustering input from being pre-filtered according to the outcome that the later analysis aimed to examine.

3.2.2 Key-Frame Selection and Image Preparation

The SRCB rear-camera data do not provide a frame at every vehicle-signal sample, and the camera frames are not indexed directly by the samples of the brake-request source signal. Inspecting the signal waveform alone therefore cannot determine which camera frame corresponds to the onset of braking. The extraction process addressed this difference through timestamp-based temporal alignment. It obtained timestamps for the RAB-attributed brake-request intervals and the decoded rear-camera frames independently, converted them to a common time unit, and matched the frames to the intervals. The selected key frame consequently provides an approximate visual description of the scene when RAB-triggered braking began, with its temporal precision limited by the camera sampling period.

For each acquired event, the extraction process used the brake-request source signal to identify every continuous interval during which the active brake request was attributed to RAB. Payload timestamps were converted from nanoseconds to mi-

croseconds to match the rear-camera clock. The rear-camera ZVLF log was decoded into an ordered frame sequence with one camera timestamp per frame.
The intervals were examined in chronological order. The selected key frame was the first camera frame satisfying

$$t_i^{\mathrm{s}} \leq t_i^{*} < t_i^{\mathrm{e}}, \tag{3.6}$$

for one of the event's RAB-attributed brake-request intervals. If an early interval contained no camera timestamp, the search continued to the next interval. This situation can occur because the camera period is longer than some brief brake-request intervals. The rule therefore selects the first available camera observation inside the earliest interval that contains a frame, rather than assuming that a frame exists exactly at the interval onset.
Each selected image was saved as an RGB PNG. Its filename encoded the UUID, the zero-based interval index, and the decoded-frame index. The interval distribution contained 7,365 frames from interval 0, 534 from interval 1, 101 from interval 2, 12 from interval 3, and 5 from interval 4. These counts sum to 8,017 and show that 652 events required a later activation interval because no camera frame fell inside an earlier one.
The final dataset contains one PNG for each of 8,017 unique events. No temporal stacking, occupancy heatmap, ultrasonic signal, or other sensing modality enters the visual clustering input. Model-specific resizing and normalization occur during feature extraction rather than during dataset construction, which keeps the stored image collection common to both backbones.

3.3 Representation and Clustering Pipeline

3.3.1 Visual Feature Extraction

The same ordered list of 8,017 images was processed independently by DINOv2 ViT-S/14 [3] and ResNet50 [2]. Both networks operated in evaluation mode and gradient computation was disabled, so no parameter update used the target events or their later manual labels. The DINOv2 path used the final global class token as a 384-dimensional embedding. The ResNet50 path removed the final classifier and used the 2048-dimensional output after global average pooling.
Each extractor saved an embedding matrix and a row-aligned array of image names. Sorting the input paths before inference made the ordering deterministic within each extraction run. Downstream stages loaded the two arrays together and used the image name, which carries the event UUID as its prefix, as the frame-level join key. This design allowed the evaluation to verify that embeddings, labels, assignments, and annotations referred to the same frames rather than relying only on array position.
The extractors stored their raw pretrained features without target-domain normalization. The clustering stage applied L2 normalization to every row before either PCA or direct clustering. Keeping this operation in the shared downstream stage ensured that the PCA and no-PCA branches started from the same normalized representation for a given backbone.

3.3.2 PCA Transformation

The experimental design compared a full-dimensional path with a PCA128 path. The former supplied the normalized embedding directly to clustering. The latter fitted PCA [19] to the normalized embeddings of the complete event collection and retained 128 components. DINOv2 was therefore reduced from 384 to 128 dimensions, while ResNet50 was reduced from 2,048 to 128 dimensions.
The paired design changed only the PCA choice within a comparison. The image set, backbone output, L2 normalization, clustering method, and remaining clustering parameters stayed fixed. This control allowed differences in cluster count, coverage, and internal geometry to be attributed to the PCA transformation within the evaluated configuration rather than to a simultaneous change of data or algorithm.
PCA was fitted separately for each backbone because their feature dimensions and geometries differ. The PCA coordinates were not normalized a second time before clustering. Consequently, the PCA branch used Euclidean distances in the projected space, while the no-PCA branch used Euclidean distances between unit-normalized embeddings.

3.3.3 HDBSCAN and K-Means Clustering

HDBSCAN [22] served as the main discovery method because it does not require a predetermined number of clusters and can leave samples outside the selected density structure. The method used Euclidean distance and returned one label per image. Non-negative labels identified selected clusters and label -1 identified noise. The experiment sweep varied minimum cluster size while holding the neighborhood setting fixed, which examined how the required recurring-group size affected granularity and coverage.
K-Means [21] supplied a fixed-K baseline on the DINOv2 representations. It assigned every event to one of the requested clusters and repeated centroid initialization before retaining the solution with the lowest objective among those starts. The sweep varied K, allowing comparison of increasingly fine complete partitions. Because K-Means cannot withhold an assignment, it provided a useful contrast to the density-based rejection performed by HDBSCAN.
Both clustering paths saved the label array, a text mapping from image name to label, and a snapshot of the run configuration. The internal evaluation measures reported in this thesis were not produced during clustering but by a separate evaluation stage applied afterwards to the stored artifacts. That stage reloads the frozen embeddings, rebuilds each run's feature space from its configuration snapshot, and verifies that embeddings, cluster labels, and the image-to-label mapping agree row by row before any score is computed. It then records the number of non-noise clusters, the number and ratio of samples labeled -1, the Silhouette score [24], and the Davies–Bouldin index [25], all defined in Section 2.4.2. Both indices are computed in the same feature space that was supplied to the clustering algorithm, and for HDBSCAN they exclude the noise samples, so the evaluated sample count accompanies every reported score. Applying one evaluator to all runs keeps the measures mutually comparable and makes the alignment between representation and labels

an explicit precondition of the reported numbers. Section 4.1.3 lists the parameter values swept for each algorithm, and the experiment identifiers recorded in Chapter 4 link the stored artifacts to the tested backbone, PCA choice, and clustering parameters. The comparison treats cluster granularity and retained coverage as joint outcomes rather than assuming that either a larger cluster count or a lower noise ratio is independently preferable.

3.4 Result Interpretation

3.4.1 Cluster Organization

The image-to-label assignment file is the authoritative output used for inspection. A separate organization step reads this mapping and creates one directory for each non-noise label and a distinct directory for label -1. The local inspection view uses symbolic links by default to avoid duplicating the image collection, although copied files can be produced when a portable view is needed.
Grouping the frames in this way allows a reviewer to scan repeated backgrounds, visible objects, surface conditions, and failure patterns within one candidate group. The folders are an inspection aid rather than a new dataset or a source of labels. Cluster identifiers have no ordinal meaning, and their numerical values are not compared across independently fitted runs.

3.4.2 Visual Inspection

Every experiment produced a PCA 2D plot and a UMAP 2D plot [23] from the same feature space supplied to clustering. For a PCA128 experiment, both plots therefore originate from the 128-dimensional representation. For a no-PCA experiment, they originate from the normalized full-dimensional embedding. Cluster labels were overlaid after projection and did not influence either projection.
The plots were used to locate broad overlap, isolated regions, density variation, and relationships between the selected labels and the underlying feature geometry. Inspection then returned to the original images in each organized cluster. This second step was necessary because a two-dimensional layout cannot establish what visual property produced a separation. Representative and contradictory examples were used to determine whether a cluster appeared visually coherent and whether similar appearance corresponded to a consistent trigger interpretation.
The interpretation did not use apparent UMAP distance, orientation, or island area as a quantitative measure. Internal scores were computed in the clustering space, and semantic claims were based on image inspection and manual labels. This separation reduced the risk of treating projection artifacts as evidence of cluster validity.

3.4.3 Noise Inspection

HDBSCAN noise was organized and inspected separately from the selected clusters. The inspection considered whether samples showed uncommon visual content,

transitions between common appearances, incomplete views, low image quality, or ordinary scenes without sufficient local density support. The same manual annotation fields and decision rules were applied to noise and clustered samples.

No event received a corner-case or false-positive interpretation solely because HDBSCAN assigned label -1. A rare event may fall into noise, but a common semantic category can also appear there when viewpoint, background, or other visual variation disperses its embeddings. The method therefore treats noise as a density outcome whose engineering meaning requires separate evidence.

3.5 Manual Semantic Evaluation

The visual groups produced by clustering do not by themselves establish whether their events support a consistent engineering interpretation. The study therefore used manual annotation as a limited external reference for the semantic evaluation introduced in Section 2.4.3. The annotation examined whether events grouped by H6 received consistent judgments about apparent trigger validity, the associated mechanism, and the relevant visible entity. It also provided a structured basis for examining the composition of HDBSCAN noise. This stage evaluated the interpretability of the discovered structure rather than assigning verified scenario classes to the complete dataset.

H6 served as the reference run because its 31 non-noise clusters provided a practical basis for cluster-level review. The annotated sample contained 20 events from each non-noise cluster together with 200 events from the noise group. These selections produced 820 annotated events. The sample was selected manually rather than by a seeded random procedure. Sampling across groups supported comparison of semantic consistency within the H6 partition, but the resulting label distribution does not estimate prevalence in the full collection because small and large clusters received similar review effort.

Each annotation record represented one event and retained the event UUID, keyframe name, source run, and H6 cluster label for alignment with the stored assignments. The reviewer used the activation-aligned key frame from Section 3.2.2 as the primary visual evidence and inspected the events cluster by cluster. The five manual fields in Table 3.1 separated the validity judgment, the attributed mechanism, the corresponding visual entity, the usability of the image evidence, and supplementary information that the structured labels did not capture.

Table 3.1: Manual fields recorded for each annotated event.

Field	Interpretation
`trigger_validity`	Whether the visible evidence supports an apparent true positive, false positive, or unclear judgment
`scenario_cause`	Mechanism associated with the apparent activation, selected from a closed vocabulary
`object_type`	Visible entity corresponding to the attributed mechanism
`visibility_quality`	Usability of the key frame as visual evidence
`comment`	Supplementary information not covered by the structured fields and useful for judging trigger validity

The fields express related but distinct judgments. Trigger validity records whether the frame provides a visible reason for braking, while scenario cause records how the relevant scene element relates to the apparent activation and object type records what that element is. A wall has object type `wall` whether it occupies the reversing path or lies outside it. Its cause is `static_obstacle` in the first case and `real_object` in the second. Visibility quality describes the strength of the image evidence rather than the event outcome. The comment field supplements the structured labels with context, ambiguity, or other observations useful for the validity judgment. Appendix A gives the complete cause vocabulary and the mapping from fine object types to the coarse families used in evaluation.

The semantic evaluation joined the annotations to the H6 assignments by image name. For non-noise clusters, it calculated separate label distributions and purity values for trigger validity, scenario cause, and object type. Fine object types came from the annotation table, except that a record naming several entities was reduced to its leading entity because purity assumes one class per sample. The coarse object taxonomy was applied during evaluation and was not a choice made by the reviewer. The analysis treated label -1 separately and summarized its annotation composition rather than interpreting noise as an ordinary semantic cluster.

Several design choices limit the resulting evidence. The cluster-by-cluster review exposed the reviewer to the H6 grouping and may have encouraged similar judgments within a folder. A single reviewer produced the labels, and no independent agreement measurement was available. The judgments also relied primarily on one key frame and therefore describe apparent validity in the recorded image rather than the internal decision process of the braking function. The reported purity consequently measures semantic consistency within the reviewed H6 sample and not classification accuracy against independently collected ground truth.

4

Experimental Evaluation

This chapter evaluates the representation and clustering pipeline on the 8,017 RAB trigger events. The evaluation combines internal cluster metrics, sample coverage, two-dimensional projections, organized image inspection, and manual semantic labels. These sources answer different questions. Internal metrics describe geometry in the evaluated feature space, while manual labels and image inspection address whether that geometry corresponds to useful visual or engineering interpretations. No single measure is treated as sufficient evidence of clustering quality.

4.1 Evaluation Design

4.1.1 Dataset and Experimental Environment

Every experiment used the same 8,017 rear-camera key frames described in Chapter 3. Each frame represents one unique event UUID. DINOv2 and ResNet50 processed the same ordered image set, and row-aligned image-name arrays provided an explicit check that the two embedding matrices covered the same events. The DINOv2 matrix contained 384 features per event and the ResNet50 matrix contained 2,048 features per event.

The dataset came from one proprietary Volvo Cars RAB event collection. It was unlabeled when the clustering experiments were conducted and contained an imbalanced, long-tailed mixture of repeated backgrounds, objects, surface conditions, and ambiguous observations. The study therefore did not assume that clusters had equal sizes or that every event belonged to a recurring category. The later H6 annotation covered a manually selected subset drawn from every H6 group and did not alter any embedding or cluster assignment.

Feature extraction and clustering ran on an internal Volvo Cars high-performance computing cluster through configuration-controlled scripts. Event acquisition and log decoding depended on separate internal services and the extraction container described in Section 3.2. The delivered experiment artifacts contain the embeddings, assignments, configurations, projections, and metric files needed to repeat the downstream evaluation without rerunning the confidential acquisition stage.

4.1.2 Evaluation Measures

The evaluation recorded the number of non-noise clusters as a measure of granularity. For HDBSCAN, it also recorded the number and ratio of samples assigned label -1.

The retained count equals the number of samples on which the HDBSCAN internal metrics were calculated. Cluster count alone does not indicate quality because a larger value can represent useful detail or unnecessary fragmentation, while a smaller value can represent useful consolidation or excessive merging.
The Silhouette score and Davies–Bouldin index followed the definitions in Section 2.4.2. A larger Silhouette score indicates stronger cohesion relative to separation, whereas a smaller Davies–Bouldin index indicates a more favorable ratio of within-cluster scatter to between-cluster separation. Both metrics describe the supplied feature space rather than semantic correctness. K-Means metrics cover all 8,017 events. HDBSCAN metrics exclude noise and therefore describe only the retained subset, which makes the noise ratio and evaluated count necessary parts of every comparison.
Semantic evaluation used per-cluster purity and weighted purity as defined in Section 2.4.3. Purity was computed independently for trigger validity, scenario cause, and object type. It can increase when a method creates more clusters, and the cluster-stratified H6 sample gives similar annotation weight to clusters with very different population sizes. The evaluation therefore combines purity with label distributions and specific mixed-cluster cases. Every reported measure consequently falls into one of two groups, namely the internal scores computed over all evaluated samples of a run and the semantic scores computed over the labeled H6 subset, and no measure in this chapter is derived from a two-dimensional projection.

4.1.3 Implementation Details

Both backbones remained frozen and ran under PyTorch [26]. The DINOv2 implementation loaded the Facebook `dinov2-small` checkpoint through the Transformers library [27], corresponding to ViT-S/14, and used its 384-dimensional class-token output. The ResNet50 implementation loaded the pretrained torchvision model [28], removed its classifier, and retained the 2,048-dimensional global-average-pooled feature. Both extractors used a batch size of 32 in the experiment configurations.
The two paths preprocessed the same stored PNG differently, because each backbone requires the transformation used during its pretraining. Every image was first read as RGB. The ResNet50 path then resized the shorter side to 256 pixels, took a central crop of 224 by 224 pixels, and applied the standard ImageNet channel statistics with means 0.485, 0.456, and 0.406 and standard deviations 0.229, 0.224, and 0.225. The DINOv2 path delegated resizing, cropping, and normalization to the image processor shipped with the checkpoint. The central crop matters for interpretation because it removes the outer margins of the rear-camera field of view, so content at the left and right edges of a key frame can be absent from the representation even though it is present in the delivered image.
The clustering scripts L2-normalized the saved features. PCA experiments retained 128 components and did not normalize the projected coordinates a second time. The implementation left `svd_solver` at the scikit-learn [29] default of `auto` and supplied random state 42. The dependency range spans versions that can resolve `auto` differently for the recorded matrix shapes, and the repository did not archive the installed version or fitted PCA objects. The exact solver used in the completed

runs therefore cannot be recovered from the saved artifacts. HDBSCAN [30] used Euclidean distance, `min_samples`=5, and `min_cluster_size` values of 30, 50, 70, or 90, and selected clusters with the default excess-of-mass criterion. K-Means used K values of 8, 12, 20, or 30 together with `n_init`=10, the default k-means++ initialization, and random state 42. UMAP used two components, Euclidean distance, and the library defaults of 15 neighbours and a minimum distance of 0.1. These two settings were not recorded in the saved run configuration and have to be read from the implementation. PCA, K-Means, and UMAP all used random state 42 where the implementation exposed one. The Silhouette computation carried a sampling cap of 10,000 evaluated samples, which never took effect because no run exceeded 8,017 samples, so every reported Silhouette score is a mean over all evaluated samples.

4.1.4 Experiment Matrix

Table 4.1 summarizes the 18 controlled runs. The H-series isolates HDBSCAN granularity and PCA effects on DINOv2. The K-series supplies matched fixed-partition baselines. The two R-series runs provide limited backbone comparisons at `min_cluster_size`=50. These experiment blocks support the research questions in complementary ways. The R-series comparisons address the backbone choice, the HDBSCAN and K-Means results address the form of recurring visual structure, the paired PCA runs address representation and configuration effects, and the later H6 analysis addresses semantic interpretability.

Table 4.1: Experiment matrix for the 18 clustering runs.

Runs	Backbone	Method	PCA	Main values	Comparison purpose
H1–H4	DINOv2	HDBSCAN	No	30, 50, 70, 90	Minimum-cluster-size sensitivity without PCA
H5–H8	DINOv2	HDBSCAN	128	30, 50, 70, 90	Minimum-cluster-size sensitivity with PCA
K1–K3, K7	DINOv2	K-Means	No	8, 12, 20, 30	Fixed-partition sensitivity without PCA
K4–K6, K8	DINOv2	K-Means	128	8, 12, 20, 30	Fixed-partition sensitivity with PCA
R1	ResNet50	HDBSCAN	128	50	Backbone comparison with H6
R2	ResNet50	HDBSCAN	No	50	Backbone comparison with H2

The paired PCA comparisons are H1 against H5, H2 against H6, H3 against H7, and H4 against H8 for HDBSCAN. The corresponding K-Means pairs are K1 against K4, K2 against K5, K3 against K6, and K7 against K8. The K-series identifiers follow execution order. K7 and K8 were added after K1–K6 to extend the comparison to $K = 30$. The manual semantic evaluation used H6 as its reference run, and Section 4.2.1 explains the result-based trade-off behind that selection.

4.2 Internal Clustering Results

4.2.1 HDBSCAN Parameter Sensitivity

Table 4.2 reports the ten HDBSCAN results. Increasing `min_cluster_size` consistently reduced the DINOv2 cluster count. Without PCA, the count fell from 41 in H1 to 17 in H4. With PCA128, it fell from 45 in H5 to 18 in H8. The same change also reduced coverage. Noise rose from 51.6% to 58.4% without PCA and from 46.8% to 57.4% with PCA.

Table 4.2: Internal HDBSCAN results for the eight DINOv2 runs and the two matched ResNet50 runs. PCA128 improved every matched DINOv2 run, while larger minimum cluster sizes generally reduced retained coverage.

Run	Backbone	PCA	Min. size	Clusters	Retained	Noise %	Silhouette / DB index
H1	DINOv2	No	30	41	3,880	51.6	0.254 / 1.544
H2	DINOv2	No	50	28	3,790	52.7	0.223 / 1.705
H3	DINOv2	No	70	22	3,750	53.2	0.212 / 1.768
H4	DINOv2	No	90	17	3,336	58.4	0.211 / 1.744
H5	DINOv2	128	30	45	4,265	46.8	0.271 / 1.468
H6	DINOv2	128	50	31	4,063	49.3	0.263 / 1.537
H7	DINOv2	128	70	24	3,910	51.2	0.257 / 1.600
H8	DINOv2	128	90	18	3,413	57.4	0.259 / 1.605
R1	ResNet50	128	50	28	3,362	58.1	0.227 / 1.681
R2	ResNet50	No	50	24	3,676	54.1	0.155 / 1.869

The metric trend was not monotonic without PCA. H1 achieved the strongest no-PCA geometry and the highest coverage in that group, while the larger minimum sizes reduced cluster count without improving both metrics. In the PCA group, H5 achieved the highest Silhouette score, the lowest Davies–Bouldin index, and the lowest noise ratio of every HDBSCAN run. It is therefore the strongest HDBSCAN configuration according to the recorded internal evidence. After this comparison, H6 was selected for manual analysis because it traded 2.5 percentage points of retained coverage and a small metric reduction against a decrease from 45 to 31 clusters. The selection reduced the cluster-by-cluster annotation burden while retaining 4,063 events, but it was a result-based practical choice rather than evidence that H6 was the globally optimal configuration.

These results show a granularity and coverage trade-off rather than a universally optimal minimum cluster size. A stricter minimum removes small recurring groups and can place additional observations into noise. The metric values describe the surviving groups, so a setting cannot be selected from separation alone when half of the original events are excluded from those scores.

4.2.2 Effect of PCA

PCA128 improved every paired DINOv2 HDBSCAN comparison. Relative to H1–H4, H5–H8 increased the Silhouette score by 0.017, 0.040, 0.045, and 0.048. The

Davies–Bouldin index decreased by 0.077, 0.168, 0.169, and 0.139. PCA also reduced the noise ratio by 4.8, 3.4, 2.0, and 1.0 percentage points and produced between one and four additional clusters. The advantage was largest in internal geometry at the larger minimum sizes, while the coverage gain was largest at `min_cluster_size`=30.

The consistency across the four DINOv2 pairs suggests that some directions in the full embedding increased within-group dispersion without providing equally useful separation under Euclidean distance. PCA [19] retains the directions with the greatest dataset-level variance, so the 128-dimensional representation can concentrate the dominant variation and produce more compact local neighborhoods. This interpretation is consistent with the simultaneous increase in Silhouette score and decrease in Davies–Bouldin index. It does not establish that the discarded components were semantically irrelevant, because PCA preserves variance rather than optimizing cluster quality or engineering meaning.

The paired K-Means results show the same directional effect. PCA increased the Silhouette score at every matched K by values between 0.012 and 0.025. It reduced the Davies–Bouldin index at $K = 8$, 12, 20, and 30 by approximately 0.152, 0.233, 0.394, and 0.371. Because K-Means evaluated the same 8,017 samples in both branches, these paired changes do not involve a coverage difference.

The K-Means pairs therefore provide the clearest evidence that the improvement came from a change in feature geometry rather than from excluding difficult samples. The larger Davies–Bouldin reductions at $K = 20$ and $K = 30$ further suggest that the compact representation helped most when the algorithm had to maintain a larger number of centroid-based partitions. The absolute scores nevertheless remained weak, which means that PCA made the tested partitions more favorable without turning the collection into a set of well-separated spherical groups.

The ResNet50 comparison requires a different interpretation. R1 improved the Silhouette score from 0.155 to 0.227 and reduced the Davies–Bouldin index from 1.869 to 1.681 relative to R2. It also increased the cluster count from 24 to 28. However, the noise ratio increased from 54.1% to 58.1%, which removed 314 additional events from the metric calculation. PCA therefore improved the geometry of the retained ResNet50 groups while reducing their coverage.

The opposite coverage response for ResNet50 shows that dimensionality reduction interacts with the backbone representation rather than producing a universal density effect. PCA can make the surviving groups more compact while changing local neighborhoods enough to leave additional borderline samples without density support. The stronger internal scores for R1 consequently describe a smaller retained subset and cannot be interpreted as an unconditional improvement over R2.

For the present project, PCA128 is a well-supported default for the DINOv2 clustering path because it improved both internal geometry and coverage in every paired HDBSCAN run and improved both K-Means metrics without changing coverage. It also reduces each DINOv2 vector from 384 to 128 values, which lowers the representation size and the dimensionality of subsequent distance calculations, although this study did not benchmark runtime or storage savings. The result supports using PCA128 when generating groups for engineering review, while the H5–H6 comparison shows that the final operating point must still balance metric quality, coverage, and the number of groups that reviewers need to inspect. Because only 128 com-

ponents were tested, alternative dimensions and stability on newly collected events remain open validation steps.

4.2.3 K-Means Baseline

Table 4.3 reports the eight K-Means baselines. Increasing K generally increased the Silhouette score, but the absolute values remained between 0.078 and 0.130. The Davies–Bouldin values remained between 2.326 and 2.961. K8 obtained the largest K-Means Silhouette score, while K6 obtained the smallest Davies–Bouldin index. The disagreement between those two choices illustrates that the two measures reward different aspects of the partition.
The upward Silhouette trend is expected when additional centroids allow samples to lie closer to their assigned center and reduce some within-cluster variation. This mechanism does not guarantee that the new boundaries correspond to additional scenario types. The Davies–Bouldin index improved up to $K = 20$ with PCA and then increased slightly at $K = 30$, which indicates that the extra subdivision in K8 reduced average within-cluster distances without improving centroid separation to the same degree. The result explains why selecting K from one internal score alone would be unreliable.

Table 4.3: Internal K-Means results. PCA128 improved both metrics at every matched K, while all runs retained the complete dataset.

Run	PCA	K	Evaluated	Silhouette	Davies–Bouldin
K1	No	8	8,017	0.078	2.961
K2	No	12	8,017	0.083	2.746
K3	No	20	8,017	0.092	2.720
K7	No	30	8,017	0.105	2.720
K4	128	8	8,017	0.090	2.809
K5	128	12	8,017	0.099	2.514
K6	128	20	8,017	0.117	2.326
K8	128	30	8,017	0.130	2.349

K-Means assigned every sample, which makes it useful when complete coverage is mandatory. Its substantially lower Silhouette scores and higher Davies–Bouldin values indicate weak spherical partitions in the tested DINOv2 spaces. The RAB collection contains continuous changes in viewpoint, object scale, background, surface appearance, and scene composition, so forcing all events into centroid-centered groups can join observations that share only part of their visual content. This explanation is consistent with the low internal scores, but the present experiment cannot identify which source of visual variation contributed most strongly.
A direct ranking against HDBSCAN would still be misleading because HDBSCAN removed 46.8% to 58.4% of the events before its scores were calculated. The evidence instead supports a method trade-off. HDBSCAN produced stronger internal separation within its density-supported retained subsets, while K-Means provided complete but less separated partitions. No semantic annotation was performed for the K-Means runs, so the internal results do not show whether their mixed geometry corresponds to less useful engineering groups.

In the project workflow, K-Means offers a predictable number of groups and guarantees that every event appears in a review folder. These properties can support complete inventory, broad sampling, and workload planning when leaving events unassigned is unacceptable. The weak separation also means that reviewers should not interpret a K-Means folder as a homogeneous scenario category without inspecting its contents. The results suggest a possible complementary workflow in which HDBSCAN identifies the most density-supported recurring patterns and K-Means supplies a complete secondary organization of the remaining collection. This combined workflow was not evaluated in the thesis and should therefore be treated as an application hypothesis rather than a demonstrated improvement.

4.2.4 DINOv2 versus ResNet50

The two matched comparisons favored DINOv2. With PCA128, H6 retained 4,063 events in 31 clusters, while R1 retained 3,362 events in 28 clusters. H6 had 8.7 percentage points less noise, a 0.035 higher Silhouette score, and a 0.145 lower Davies–Bouldin index. Without PCA, H2 retained 114 more events and four more clusters than R2. Its Silhouette score was 0.068 higher and its Davies–Bouldin index was 0.164 lower.

The agreement between the PCA and no-PCA comparisons makes the observed advantage less likely to result from the dimensionality-reduction branch alone. One plausible explanation is that the self-supervised DINOv2 representation [3] preserves global scene layout and visual relationships that are useful for grouping rear-camera observations, whereas the supervised ResNet50 representation [2] was optimized originally for ImageNet object classification. The DINOv2 class token may therefore organize similarities in background, surface, and object configuration more favorably for this dataset. The experiment does not isolate this mechanism, because architecture, pretraining objective, preprocessing, and embedding dimensionality all changed together between the two backbones.

More clusters and greater retained coverage do not by themselves imply more meaningful scenario categories. They show that DINOv2 provided local neighborhoods from which HDBSCAN selected more density-supported structure under the tested settings. Manual semantic evaluation was completed only for the DINOv2 H6 run, so the study cannot determine whether an equally sampled ResNet50 partition would receive lower purity or be less useful to a reviewer.

These comparisons indicate that DINOv2 provided more favorable density structure for the tested RAB key frames. They do not establish a general superiority of self-supervised ViT features over supervised CNN features. ResNet50 was evaluated only at one minimum cluster size and was not tested with K-Means, alternative distances, or a parameter sweep. The internal metrics also do not measure classification accuracy or functional trigger validity.

For this project, the paired evidence supports DINOv2 as the primary candidate for constructing HDBSCAN review groups from the current key-frame collection. This choice is specific to the evaluated data and should be reassessed when the event distribution, camera preprocessing, or review objective changes. An operational comparison should also include repeated-run stability, inference cost, storage

requirements, and reviewer-assessed group usefulness. The present experiments did not measure those criteria, so they support a representation choice for continued analysis rather than a final production decision.

4.3 Semantic Evaluation and Case Analysis

4.3.1 Annotation Summary

The H6 annotation table contains 820 events, one record per event, drawn from the 31 non-noise clusters and the noise group. Every non-noise cluster contributed 20 records and noise contributed 200. The sample therefore covers all discovered groups without reproducing their relative sizes in the complete dataset, and it gives every cluster the same annotation weight regardless of its population.

Trigger validity is recorded for every event and comprises 440 apparent true positives, 328 apparent false positives, and 52 unclear cases, so unresolved events account for 6.3% of the sample. Scenario cause and object type are recorded for 817 events, and three records carry no entry in either field. Two further records retain a compound cause value that the consolidated vocabulary does not contain, and Section 4.3.2 reports them under that value rather than folding them into a neighboring class.

A further 67 of the 817 object records name more than one visible entity, of which 44 fall inside the non-noise clusters. Purity assumes one class per sample, so the evaluation keeps the leading entity of such a record as its object class. This rule applies only to the object field, because no value of the consolidated cause vocabulary contains a separator. Section 4.3.2 reports the resulting sensitivity of the object purity values, and Appendix A lists the closed cause vocabulary and the object families.

Visibility quality is clear for 797 events, partial for 22, and none for one. Weak keyframe evidence is therefore rare, which matters for the analyses below because it removes image quality as a general explanation for the observed label distributions.

4.3.2 Cluster Semantics and Purity

Table 4.4 summarizes the H6 non-noise evaluation. Trigger validity covered 620 labeled rows and reached weighted purity 0.931. Scenario cause covered 619 rows over ten classes and reached 0.947, and object type covered the same rows and reached 0.943 over 25 fine classes and 0.952 over nine coarse families. The class counts state how many classes occur in these clusters and not how large the annotation vocabulary is. The ten observed causes are static obstacle, real object, physical connection, surface pattern, surface transition, surface photometric effect, drivable structure, early stop with no identified cause, unspecified, and the retained compound value. Dynamic obstacle is the one closed-vocabulary cause that no labeled non-noise frame received, and the noise group in Section 4.3.3 contains further object classes that these clusters do not show.

Table 4.4: Semantic purity of the labeled H6 non-noise clusters. Most clusters were pure for each field, while coarse object grouping changed weighted purity by less than one percentage point. The class counts are the classes observed in these clusters rather than the size of the annotation vocabulary.

Field	Taxonomy	Classes	Labeled	Weighted purity	Pure clusters
Trigger validity	Fine	3	620	0.931	22 of 31
Scenario cause	Fine	10	619	0.947	22 of 31
Object type	Fine	25	619	0.943	23 of 31
Object type	Coarse	9	619	0.952	23 of 31

The high values indicate that many H6 clusters carried consistent manual semantics in the labeled subset. Twenty-two clusters were perfectly pure for trigger validity, 22 for scenario cause, and 23 for object type. The coarse object taxonomy improved weighted purity by less than one percentage point relative to the fine taxonomy, which suggests that most clusters already separated the recorded object labels rather than merely separating broad object families. The three fields are not directly comparable, because purity decreases as a vocabulary becomes finer, and scenario cause here carries ten classes against three for trigger validity.

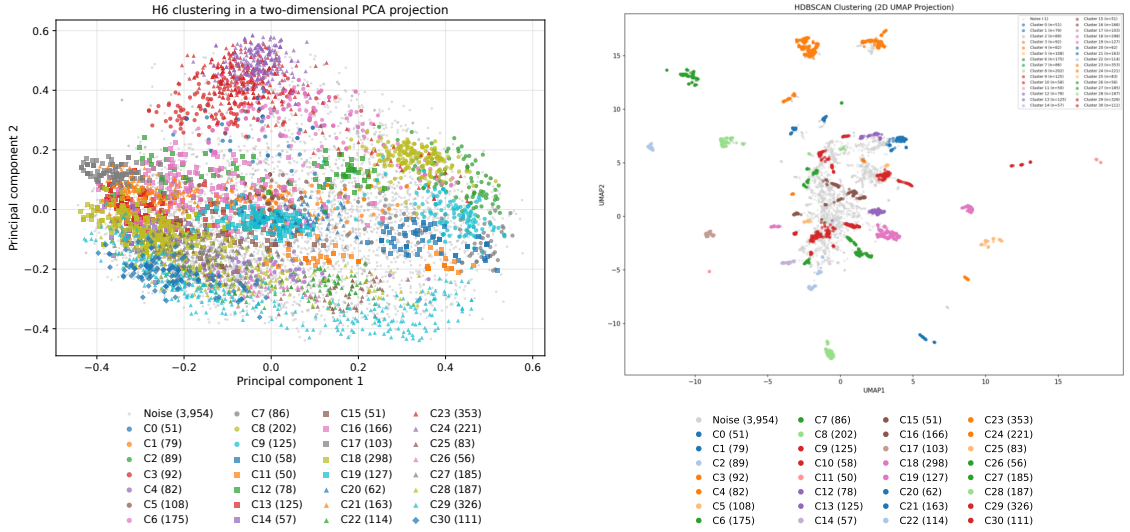
The object values are also sensitive to the multi-entity rule introduced in Section 4.3.1. Treating every multi-entity string as a class of its own instead of keeping its leading entity raises the number of observed fine object classes from 25 to 28, lowers the fine weighted purity from 0.943 to 0.927, and moves one cluster out of the perfectly pure group. The rule therefore accounts for about 1.6 percentage points of the reported fine object purity.

The aggregate also hides important mixed cases. Nine trigger-validity clusters were not perfectly pure and six fell below 0.9. Cluster 27 was the least consistent group in every field, since it contained 11 true positives and nine false positives and therefore reached purity 0.550. Clusters 0 and 29 followed at 0.650, and cluster 4 reached 0.700 with 14 true-positive and six false-positive labels. Cluster 27 combined vegetation with road edges and ground carrying no identifiable entity, while cluster 4 brought together vegetation with reflections and light surface strips on indoor tiled floors. These cases show that membership in one visual cluster does not by itself establish a consistent validity judgment.

Cluster 27 also shows what the consolidated cause vocabulary exposes. Its 20 labeled frames spread over four causes, namely 11 static obstacles, five real objects, three unspecified records, and one photometric surface effect, so its cause purity is 0.550 as well. The group is mixed in every recorded field rather than coherent in one and mixed in another, and one visual cluster therefore carries several distinct activation mechanisms.

Cluster 14 shows the opposite arrangement. All 20 labeled frames received an apparent false positive, yet their causes split into 13 unspecified records, six surface transitions, and one real object. A cluster can therefore be functionally uniform while the reason for the intervention stays unresolved. Cluster 0 inverts that pattern once more, since all 20 frames received `physical_connection` with a rear-mounted carrier rack, while the attachment occluded the area behind it and left 13 of the 20 validity verdicts unclear.

Cluster 5 is the clearest case in which agreement between the fields still leaves the engineering question open. Its 20 labeled frames are pure in all three of them, with an apparent false positive, the cause `early_stop_no_cause`, and an unresolved entity in every record. The clustering recovered a recurring appearance for which the key frame shows no candidate obstacle at all, which is a useful group to inspect as a batch precisely because the single image does not explain the activation.
Figure 4.1 provides a supporting projection-based check for the quantitative and semantic evaluation. The PCA panel retains broad overlap between several clusters and noise, which is consistent with the mixed cases identified above, while the UMAP panel shows that local islands coexist with a substantial central noise region. The projections help locate groups for inspection and corroborate the presence of both local structure and overlap. They do not determine cluster quality or contribute to any reported metric, so the conclusions remain based on coverage, internal scores, and the labeled H6 subset.



(a) PCA projection **(b)** UMAP projection

Figure 4.1: Supporting two-dimensional projections of H6 with cluster assignments overlaid. Gray points denote HDBSCAN noise, and legend counts give the number of events. The panels support qualitative inspection only, and projected distances are not used as quantitative evidence.

4.3.3 Noise Analysis

The labeled noise subset contained 200 events. Trigger validity consisted of 152 true positives, 42 false positives, and six unclear labels, so true positives accounted for 76.0% of the subset. The labeled non-noise clusters were far more balanced, with 288 true positives, 286 false positives, and 46 unclear labels among their 620 records, so their true-positive share was 46.5%. The events that H6 discarded were therefore more likely to be genuine obstacle encounters than the events it retained. Scenario cause was available for 198 events. Static obstacles dominated with 148 labels, corresponding to 74.7%, followed by 22 physical connections and ten real objects.

The three surface causes together accounted for 11 events, drivable structures for five, and the unspecified and dynamic-obstacle classes for one each.

The fine object vocabulary was heterogeneous. Vegetation was the largest individual label with 72 of 198 events, which represented only 36.4%. Under the coarse mapping, vegetation accounted for 72 events, built structures for 56, street furniture for 31, vehicles for 12, surface features for nine, and unidentified entities for eight, while image effects, ego equipment, cluttered scenes, one other object, and one vulnerable road user covered the remaining ten. This subset alone carried 28 fine object classes and 11 coarse families, which exceeds the counts observed in the clusters and shows that noise held classes the clusters never recovered.

Visibility was weaker in noise than in the clusters, but not weak enough to explain the assignments. Of the 200 sampled noise events, 185 were clear, 14 partial, and one recorded as `none`, so 92.5% carried clear key-frame evidence against 98.7% in the labeled non-noise clusters. Poor image evidence is therefore about six times more frequent in noise, and it still applies to fewer than one in ten sampled noise events.

These results contradict an interpretation of noise as a direct corner-case reservoir or a collection dominated by poor images. Most sampled noise events were clear true-positive observations involving ordinary static obstacles. Their label -1 indicates insufficient local density under H6, which can arise from variation in viewpoint, background, scale, or appearance within common semantic categories. Some rare or ambiguous events may still occur in noise, but rarity and engineering value require evidence beyond the density label.

4.3.4 Representative Success and Failure Cases

Table 4.5 presents representative semantic patterns derived from the labeled H6 clusters. The success cases combine pure trigger validity with pure or near-pure scenario and object labels. The failure cases show that one aspect can remain coherent while another is mixed. These examples are based on label distributions rather than publicly reproduced camera frames because the raw images require internal approval.

Table 4.5: Representative H6 semantic successes and limitations.

Cluster	Validity result	Dominant semantics	Interpretation
1	20 false positive	Real object, wall	Retaining walls beside a narrow driveway stayed outside the driven path
2	20 true positive	Static obstacle, wall	A visible built structure aligned with a consistent validity judgment
12	20 true positive	Static obstacle, vehicle	Vehicle appearance produced a pure labeled group
25	20 true positive	Static obstacle, vegetation	Vegetation formed a repeated obstacle pattern in the labeled subset
5	20 false positive	Early stop with no identified cause, unresolved entity	A group that is pure in every field while the key frame shows no candidate obstacle
0	13 unclear, 7 false positive	Physical connection, rear-mounted carrier	A pure cause label left most validity verdicts unresolved
4	14 true positive, 6 false positive	Static obstacle and photometric surface, vegetation and reflection	Similar visual content mixed opposing validity judgments
14	20 false positive	Unspecified and surface transition, unresolved entity	A uniform validity verdict coexisted with an unresolved cause
27	11 true positive, 9 false positive	Four causes, vegetation and road edge	One visual family carried several activation mechanisms
30	15 true positive, 5 unclear	Static obstacle, vegetation	Single-frame evidence left part of an otherwise coherent group unresolved

The successful cases provide evidence that frozen embeddings can organize recurring visual patterns that support batch inspection. The mixed cases identify the boundary of that contribution. Viewpoint, trajectory relation, and temporal development can change the functional meaning of frames with similar content, and the single image does not expose every signal needed for an event-level judgment.

4.4 Discussion

4.4.1 Answers to the Research Questions

This section answers the four research questions stated in Section 1.4.1 in the order given there.

The first research question concerned the effect of DINOv2 and ResNet50 representations. In the two matched HDBSCAN comparisons, DINOv2 achieved higher retained coverage, higher Silhouette scores, lower Davies–Bouldin indices, and more clusters. The evidence supports DINOv2 as the more suitable of the two tested

backbones for this frame-level structuring task. The asymmetric experiment matrix limits the conclusion to the evaluated configurations.

The second research question asked whether HDBSCAN and K-Means reveal recurring visual structure. HDBSCAN identified density-supported groups with Silhouette scores from 0.211 to 0.271 for DINOv2, while K-Means provided complete coverage but substantially weaker internal separation. The H6 analysis further showed that many of its retained groups carried consistent manual semantics, although no corresponding semantic evaluation was completed for the K-Means partitions. The evidence therefore favors HDBSCAN for discovering cleaner retained subsets in the tested spaces and K-Means when every event requires an assignment, but it does not establish a semantic ranking between the algorithms.

The third research question addressed PCA and clustering configuration. PCA128 improved every paired DINOv2 HDBSCAN comparison and every paired K-Means result according to both internal metrics. H5 provided the strongest internal HDBSCAN result, while H6 reduced the number of groups enough to support a manageable annotation study. Increasing the HDBSCAN minimum cluster size reduced granularity and usually reduced coverage. Increasing K improved K-Means Silhouette scores but did not produce strong separation. Configuration choice therefore changes geometry, coverage, and review burden together.

The fourth research question concerned visual and semantic interpretability. H6 reached weighted purities of 0.931 for trigger validity over three observed classes, 0.947 for scenario cause over ten, and 0.943 for fine object type over 25 on the labeled non-noise subset. These values indicate substantial within-sample consistency under the manual labels used in the study. However, cluster 27 remained close to an even true-positive and false-positive split and spread over four scenario causes, cluster 4 mixed vegetation with photometric surface effects, and the noise group contained mostly clear true-positive static-obstacle events at a higher true-positive rate than the clusters themselves. The groups are interpretable as recurring visual patterns, but they do not replace event-level engineering judgment.

4.4.2 Interpretation of High Noise Ratios

Every HDBSCAN run assigned between 46.8% and 58.4% of the dataset to noise. This range indicates that the event collection did not form a small set of uniformly dense groups under the tested single-frame representations. Long-tailed visual content, continuous changes in viewpoint, different backgrounds around the same object, and images between denser modes can all reduce local density.

The method also contributes to the observed ratio. HDBSCAN with Euclidean distance and a fixed neighborhood setting applies one density model across the representation. PCA changed the ratio but did not remove the large noise group. The semantic analysis further shows that noise is not synonymous with poor evidence, false positives, or rare cases. A more useful interpretation is that the retained clusters represent the most density-supported recurring patterns, while the noise group combines common but dispersed categories with less frequent and ambiguous observations.

4.4.3 Industrial Implications

The pipeline can support cluster-assisted review by placing recurring visual appearances into navigable groups and preserving an event-level link back to the source records. Pure groups such as off-path retaining walls, static walls, vehicles, and vegetation provide candidates for batch inspection and for selecting representative events. Mixed clusters can direct attention to patterns for which visible similarity does not resolve functional validity.

The approach can also support future labeling. A reviewer can sample across clusters to cover diverse recurring patterns or inspect noise separately to search for poorly represented variants. Cluster assignments and semantic summaries can help define a more stable label vocabulary before supervised model development. The present study did not measure review time, reviewer agreement, or downstream model performance, so these implications describe supported uses of the artifacts rather than demonstrated efficiency gains.

4.5 Limitations and Threats to Validity

4.5.1 Data and Annotation Limitations

The dataset came from one industrial source and one RAB application. Its distribution may depend on Volvo-specific sensing, event definitions, operating conditions, and collection practices, which limits external validity. Confidentiality prevents distribution of the raw images and internal acquisition infrastructure. The delivered extraction also lacks the complete metadata needed to reproduce every filtering and skip statistic outside the internal environment.

The semantic evidence came from one H6 sample and one annotation table. Sampling across clusters improved group coverage but did not represent the full dataset distribution. The 20 events reviewed in each cluster were also selected manually rather than by a seeded random draw, so visually typical or easily judged frames may be over-represented inside a group, and a random draw of the same size could produce different per-cluster purity. Reviewing the events cluster by cluster may in addition have encouraged consistent judgments within a folder and inflated purity. Without a second annotator or an agreement statistic, the study cannot separate repeatable application of the label guide from consistency specific to one reviewer. Five records also remain outside the consolidated vocabulary, since three carry no cause and no object and two retain a compound cause value, and the object purity depends on the rule that keeps the leading entity of a multi-entity record. Purity therefore describes this annotation table rather than adjudicated ground truth.

Single images also limit construct validity. A clear key frame can still omit motion, trajectory, activation timing, and information from the original log. Trigger validity and scenario cause are therefore apparent judgments based on the available evidence. The analysis does not claim to establish the functional correctness of an intervention from an image alone.

4.5.2 Methodological Limitations

The study compared DINOv2 ViT-S/14 with ResNet50. ResNet50 had two HDBSCAN runs rather than the full sweep used for DINOv2, so backbone conclusions rely on two matched points. The study fixed PCA at 128 components, HDBSCAN `min_samples` at 5, and the clustering distance at Euclidean. It did not test cosine distance, alternative PCA dimensions, other density methods, or learned target-domain representations.

The reported clusterings came from one execution per configuration. K-Means used ten initializations internally, but the study did not measure repeated-run, bootstrap, or subsampling stability. UMAP plots used a fixed seed and support inspection rather than reproducibility claims about global layout. HDBSCAN and K-Means internal metrics cover different sample sets, which prevents a simple algorithm ranking. Semantic purity was calculated only for H6, so the study cannot determine whether the internal advantages of H5 or the PCA pairs translate into higher semantic consistency.

No frozen-embedding linear probe was completed. The evaluation therefore establishes geometric structure and limited semantic alignment but does not measure whether the same labels are linearly predictable from DINOv2 or ResNet50 features. The loose dependency constraints and missing exact hardware manifest further limit bitwise reproduction, although the saved assignments and metrics preserve the reported outcomes.

4.5.3 Scope and Practical Limitations

The experiments did not include CTA/b events, radar, ultrasonic measurements, occupancy heatmaps, or temporal clips. They did not evaluate real-time deployment, incremental updates, or integration with a production review interface. The pipeline remained an offline analysis workflow.

The study also did not conduct a controlled reviewer-timing experiment, measure changes in engineering decisions, or evaluate downstream safety performance. Consequently, it cannot claim that clustering reduced review time or improved the vehicle function. It provides structured evidence and inspection artifacts that can motivate those evaluations in future work.

5
Conclusion and Future Work

5.1 Conclusion

This thesis investigated how frozen visual representations and unsupervised clustering can support the organization of rear-camera key frames from RAB trigger events. The study processed 8,017 proprietary events through a controlled pipeline that combined DINOv2 and ResNet50 feature extraction, optional PCA128, HDBSCAN and K-Means clustering, two-dimensional visualization, internal cluster metrics, and limited manual semantic evaluation. The 18-run experiment matrix supported paired comparisons of backbone, dimensionality reduction, and configuration on a common event set, with the backbone comparison limited to two matched HDBSCAN runs and the K-Means baselines limited to the DINOv2 representation. The next two subsections summarize the answers developed in Section 4.4.1, in the order of the four research questions.

5.1.1 Representation and Clustering Structure

The internal evaluation showed that DINOv2 produced more favorable density structure than ResNet50 in the two matched HDBSCAN comparisons. PCA128 improved both recorded internal metrics in every paired DINOv2 HDBSCAN comparison and every paired K-Means comparison. HDBSCAN retained subsets with stronger internal separation but assigned between 46.8% and 58.4% of the events to noise. K-Means assigned every event, although its partitions showed substantially weaker internal separation in the tested DINOv2 spaces. Because the two methods computed their metrics over different sample sets, these values describe a trade-off between coverage and separation rather than a ranking of the algorithms. H5 achieved the strongest recorded HDBSCAN result according to the internal metrics, whereas H6 reduced the number of retained groups from 45 to 31 with a limited loss of coverage and internal performance, as reported in Section 4.2.1. H6 therefore provided a practical basis for the annotation study rather than a globally optimal configuration.

5.1.2 Semantic Consistency of the Discovered Groups

The semantic evaluation covered 820 manually selected events, with 20 records from each of the 31 H6 clusters and 200 from the noise group, so every cluster received equal annotation weight regardless of its size. Many H6 clusters carried consistent labels within this annotated subset. Weighted purity reached 0.931 for trigger validity, 0.947 for scenario cause, and 0.943 for fine object type in the labeled non-noise

sample, where the object value depends on the multi-entity labeling rule and falls to 0.927 under the alternative treatment. These aggregate values did not imply uniform functional meaning. Some clusters combined opposing validity judgments or several activation mechanisms, and the sampled noise group consisted mostly of clear true-positive events involving ordinary static obstacles, at a higher true-positive rate than the clusters themselves. Visual similarity and local feature density can therefore organize recurring appearances without fully resolving why the system activated or whether an individual intervention was functionally appropriate.

5.1.3 The Annotation Scheme as a Study Outcome

The annotation table produced for that evaluation is a further outcome of the study. It was built to measure cluster purity, but it also serves as a small worked example of how RAB trigger events can be described in a structured way. Each of its 820 records keeps the event UUID, the key-frame name, and the H6 cluster label alongside five manual fields, and its labels follow a closed vocabulary of ten scenario causes together with an object mapping onto eleven coarse families, both listed in Appendix A. That vocabulary was derived from the events actually observed in the clusters and in noise rather than assumed in advance, which is the sense in which the clustering contributed to it. The table therefore offers a first working classification scheme for this event type. It gives a shared vocabulary for stating why a trigger appears to have occurred, it separates the validity judgment from the attributed mechanism and the visible entity, and it records how ambiguous, unresolved, and multi-entity cases were handled. It also leaves behind a labeled seed set and a repeatable protocol that later work can extend to more events or more runs. It remains a demonstrator produced by one reviewer on one reference run, so it indicates a workable direction for a review taxonomy rather than an established standard.

5.1.4 Contributions and Scope

The main contribution of the thesis is a controlled industrial evaluation of an end-to-end workflow for structuring predominantly unlabeled RAB trigger key frames. The workflow connects event preparation, frozen representation extraction, clustering, visual summaries, and semantic inspection while preserving links to individual source events. Its outputs can support repeated-pattern discovery, representative-event selection, and cluster-assisted review, and the annotation scheme described above gives the resulting groups a consistent description. They remain analysis aids rather than autonomous validity decisions. The study did not demonstrate shorter review time, improved engineering decisions, or increased vehicle safety, and its conclusions remain limited to the data, representations, configurations, and annotation sample evaluated in Chapter 4.

5.2 Future Work

5.2.1 Validation of the Discovered Structure

The first priority is stronger validation of the discovered structure. A second annotator and an explicit agreement measure would show whether the label guide produces repeatable judgments rather than consistency specific to one reviewer. Equivalent semantic evaluation of K-Means, H5, and additional HDBSCAN configurations would test whether differences in internal geometry correspond to differences in semantic consistency. Repeated executions, bootstrap analysis, or subsampling experiments could quantify clustering stability. Linear probes on the frozen representations could further test whether the annotated labels are predictably encoded even when unsupervised cluster boundaries do not isolate them.

5.2.2 Multimodal Representations

Future representations should incorporate information that a single key frame omits. Short temporal clips, vehicle trajectory, activation timing, and selected log signals could help distinguish visually similar events with different functional meanings. Vehicle speed is a concrete candidate, since it is recorded alongside every event and the key frame does not express it. Radar, ultrasonic measurements, occupancy heatmaps, or other available signals could extend the analysis beyond image appearance. A multimodal representation that concatenates or jointly encodes visual and log features could then be clustered and evaluated under the protocol used here, which would also show whether the added signals separate the mixed clusters that a single frame could not. Additional visual backbones, alternative PCA dimensions and distance measures, and target-domain adaptation could be compared under the same coverage, stability, and semantic criteria.

5.2.3 Label-Efficient and Continual Learning

The annotated subset also enables label-efficient learning. Its records cover every H6 cluster and could seed semi-supervised methods that propagate labels from the annotated events to the unlabeled remainder, for example through pseudo-labeling, self-training, or graph-based propagation over the frozen embeddings, with a held-out annotated portion used to check whether propagation preserves label quality. A light classification head trained on the same frozen features would provide a directly comparable supervised reference. Because trigger events keep accumulating, such a model should also be examined under continual learning, in which new events, later annotation batches, and possibly new cause or object values are incorporated without retraining on the complete history and without losing performance on earlier categories. The practical questions are how many additional labels are needed before propagation becomes reliable, which events are most informative to annotate next, and how a shift in the recorded event distribution can be detected.

5.2.4 Cluster Structure and Trigger Validity

The relationship between the discovered groups and trigger validity deserves separate study. The present evaluation measured how consistently a cluster carries one validity label, but it did not test whether cluster identity predicts that label. Estimating a false-positive rate per cluster with an interval, testing whether validity is predictable from the embedding or from the cluster assignment alone, and examining which causes dominate the mixed clusters would show whether the structure can prioritize review rather than only summarize it. Cluster 27 and the noise group are the informative cases, since one remained close to an even split between opposing validity judgments and the other carried a higher apparent true-positive rate than the clusters themselves. Such an analysis needs a larger and randomly drawn validity sample, because the present sample weighted every cluster equally and cannot estimate prevalence in the full collection.

5.2.5 Industrial Integration and Deployment

Industrial follow-up should evaluate the workflow in the review process itself. Automatic representative-sample selection, contact sheets, and an interactive cluster browser could preserve traceability while reducing navigation effort. A controlled reviewer study should measure inspection time, decision consistency, and the types of events found with and without cluster assistance. Deployment would raise further engineering questions that this offline study did not address, including incremental processing so that newly recorded events are embedded and assigned without reclustering the whole collection, a persistent store for embeddings and assignments that preserves the link to the source event, monitoring for representation and distribution drift, an agreed treatment of ambiguous clusters and noise inside the review procedure, and execution within the internal environment given the confidentiality of the data. Such integration becomes meaningful only after the reviewer study establishes a reliable benefit.

Bibliography

[1] Volvo Cars, "Interventions and warnings when reversing," *Volvo Support*, Apr. 4, 2025. [Online]. Available: https://www.volvocars.com/uk/support. [Accessed: Jul. 26, 2026].

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[3] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024.

[4] S. Riedmaier, T. Ponn, D. Ludwig, B. Schick, and F. Diermeyer, "Survey on scenario-based safety assessment of automated vehicles," *IEEE Access*, vol. 8, pp. 87456–87477, 2020, doi: 10.1109/ACCESS.2020.2993730.

[5] S. Ulbrich, T. Menzel, A. Reschka, F. Schuldt, and M. Maurer, "Defining and substantiating the terms scene, situation, and scenario for automated driving," in *Proceedings of the IEEE 18th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 982–988, 2015, doi: 10.1109/ITSC.2015.164.

[6] T. Menzel, G. Bagschik, and M. Maurer, "Scenarios for development, test and validation of automated vehicles," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pp. 1821–1827, 2018, doi: 10.1109/IVS.2018.8500406.

[7] F. Kruber, J. Wurst, E. S. Morales, S. Chakraborty, and M. Botsch, "Unsupervised and supervised learning with the random forest algorithm for traffic scenario clustering and classification," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pp. 2463–2470, 2019, doi: 10.1109/IVS.2019.8813994.

[8] J. Kerber, S. Wagner, K. Groh, D. Notz, T. Kühbeck, D. Watzenig, and A. Knoll, "Clustering of the scenario space for the assessment of automated driving," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pp. 578–583, 2020, doi: 10.1109/IV47402.2020.9304646.

[9] L. Ries, P. Rigoll, T. Braun, T. Schulik, J. Daube, and E. Sax, "Trajectory-based clustering of real-world urban driving sequences with multiple traffic objects," in *Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 1251–1258, 2021, doi: 10.1109/ITSC48978.2021.9564636.

[10] F. Montanari, R. German, and A. Djanatliev, "Pattern recognition for driving scenario detection in real driving data," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pp. 590–597, 2020, doi: 10.1109/IV47402.2020.9304560.

[11] N. Weber, C. Thiem, and U. Konigorski, "Toward unsupervised test scenario extraction for automated driving systems from urban naturalistic road traffic

data," *SAE International Journal of Connected and Automated Vehicles*, vol. 6, no. 3, pp. 263–281, 2023, doi: 10.4271/12-06-03-0017.

[12] U. Sander and N. Lübbe, "The potential of clustering methods to define intersection test scenarios: Assessing real-life performance of AEB," *Accident Analysis & Prevention*, vol. 113, pp. 1–11, 2018, doi: 10.1016/j.aap.2018.01.010.

[13] J. Breitenstein, J.-A. Termöhlen, D. Lipinski, and T. Fingscheidt, "Systematization of corner cases for visual perception in automated driving," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pp. 1257–1264, 2020, doi: 10.1109/IV47402.2020.9304789.

[14] F. Heidecker, J. Breitenstein, K. Rösch, J. Löhdefink, M. Bieshaar, C. Stiller, T. Fingscheidt, and B. Sick, "An application-driven conceptualization of corner cases for perception in highly automated driving," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pp. 644–651, 2021, doi: 10.1109/IV48863.2021.9575933.

[15] D. Bogdoll, M. Nitsche, and J. M. Zöllner, "Anomaly detection in autonomous driving: A survey," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4488–4499, 2022.

[16] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149, 2018.

[17] J. Zhao, J. Fang, Z. Ye, and L. Zhang, "Large scale autonomous driving scenarios clustering with self-supervised feature extraction," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pp. 473–480, 2021, doi: 10.1109/IV48863.2021.9575644.

[18] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[19] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, Art. no. 20150202, 2016, doi: 10.1098/rsta.2015.0202.

[20] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.

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

[22] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, "Hierarchical density estimates for data clustering, visualization, and outlier detection," *ACM Transactions on Knowledge Discovery from Data*, vol. 10, no. 1, Art. no. 5, pp. 1–51, 2015, doi: 10.1145/2733381.

[23] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, Art. no. 861, 2018, doi: 10.21105/joss.00861.

[24] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[25] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[26] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pp. 8024–8035, 2019.

[27] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020, doi: 10.18653/v1/2020.emnlp-demos.6.

[28] TorchVision maintainers and contributors, "TorchVision: PyTorch's computer vision library," GitHub repository, 2016. [Online]. Available: https://github.com/pytorch/vision. [Accessed: Aug. 11, 2026].

[29] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, vol. 2, no. 11, Art. no. 205, 2017, doi: 10.21105/joss.00205.

A

Annotation Taxonomies

This appendix provides the closed scenario-cause vocabulary and the object-type mapping referenced in Section 3.5.

A.1 Scenario-Cause Vocabulary

Table A.1: Allowed `scenario_cause` values used in manual annotation.

Value	Definition
`static_obstacle`	A stationary element occupies the reversing path
`dynamic_obstacle`	A moving road user enters the reversing path
`physical_connection`	Equipment attached to the ego vehicle enters the rear view
`real_object`	A real object is visible but lies outside the driven path, remains distant, or is traversable
`surface_pattern`	Colour or texture contrast within a single flat drivable surface
`surface_transition`	Flat boundary between two ground materials, including indoor-to-outdoor thresholds
`surface_photometric`	Shadow, reflection, or glare, that is, an illumination effect rather than a material property
`drivable_structure`	Real three-dimensional structure that the vehicle is intended to traverse
`early_stop_no_cause`	The vehicle stops before reaching any obstacle and the frame shows no candidate cue
`unspecified`	The cause could not be resolved, and the record carries a re-inspection flag

A.2 Object-Type Mapping

Table A.2: Mapping from primary `object_type` values to the coarse families used in evaluation. Compound raw values are represented by their leading entity.

Family	Primary entities
built_structure	wall, building_structure, fence_barrier, metal_platform_structure, opened_door, gate, curb_road_edge, gas_station, door, charging_station
vegetation	vegetation
surface_feature	dark_surface_strip, grass_paver_pattern, indoor_outdoor_threshold, paving_material_edge, grass_edge, light_surface_strip, snow_cover, brick_color_pattern, low_surface_protrusion
street_furniture	bicycle_rack, trash_bin, flexible_bollard
ego_equipment	rear_mounted carrier_rack, mounting_bracket, ego_vehicle_component
vehicle	vehicle, bicycle
unidentified	unknown_object, none
cluttered_scene	messy_room, clutter
image_effect	reflection, shadow
other_object	string_on_the_floor
vulnerable_road_user	person

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY