infrastructures

MDPI

Article

AI-Based Automated Visual Condition Assessment of Municipal Road Infrastructure Using High-Resolution 3D Street-Level Imagery

Elia Ferrari *, Jonas Meyer and Stephan Nebiker

Institute of Geomatics, FHNW University of Applied Sciences and Arts Northwestern Switzerland, 4132 Muttenz, Switzerland; jonas.meyer@fhnw.ch (J.M.); stephan.nebiker@fhnw.ch (S.N.)
* Correspondence: elia.ferrari@fhnw.ch

Abstract

The effective management of municipal road infrastructure requires up-to-date, standardized and reliable condition information to support sustainable maintenance. While visual road-condition assessment methods based on established standards are widely applied to municipal roads, they remain largely manual, time-consuming, costly and subjective. This study presents an end-to-end workflow for the automated visual inspection and condition assessment of municipal road infrastructure using high-resolution, 3D street-level imagery acquired by professional mobile mapping systems. The proposed approach integrates an efficient preprocessing pipeline for precise road-surface extraction with deep learning models trained for the specific task and an advanced postprocessing method for robust results aggregation. For this purpose, a large dataset covering approximately 352 km of municipal roads across eight municipalities was created by combining street-level imagery with expert-annotated road-condition index (RCI) values. Two neural network variants were implemented: a regression model predicting standardized RCI values and a binary classifier distinguishing between roads requiring maintenance and those in good condition. To ensure decision-oriented outputs at the infrastructure-asset level, frame-based predictions are aggregated into homogeneous road segments using outlier detection and change-point analysis along the road axis. The regression model achieved a mean absolute error of 0.48 RCI values at frame level and 0.40 RCI values at road-segment level, outperforming conventional inter-expert variability, while the binary classification model reached an F1-score of 0.85. These findings demonstrate that AI-based visual road-condition assessment using professional mobile mapping data can provide accurate, standardized and scalable condition information for municipal road infrastructure. The proposed workflow supports maintenance prioritization and infrastructure management decisions without requiring explicit detection of individual pavement defects, offering a practical pathway toward automated, cost-effective road-condition monitoring.

Keywords: pavement condition assessment; convolutional neural network; deep learning; road infrastructure management; mobile mapping; street-level imagery; RGB-D; 3D image spaces

Academic Editor: Linh Truong-Hong

Received: 2 February 2026
Revised: 21 February 2026
Accepted: 5 March 2026
Published: 10 March 2026

1. Introduction

As urbanization accelerates and infrastructure networks age, ensuring the safety, reliability and longevity of critical assets such as bridges, tunnels and utility networks has

become increasingly important. Efficient infrastructure management is essential not only for public safety and economic stability, but also for achieving long-term sustainability goals. Among the various infrastructure domains, road networks play a particularly vital role. However, while well-maintained roads are vital for public safety and the economy, they impose a significant financial burden that demands innovative solutions. Consequently, resource-efficient maintenance relies on up-to-date, comparable, reproducible and cost-effective condition data [1].

Numerous organizations are responsible for the maintenance and development of road infrastructure, basing their management strategies on established standards. These standards typically classify pavement-condition indicators into two categories: visual assessments for surface damage and measurement-based evaluations of geometric characteristics. The latter requires sophisticated and costly equipment, such as skid resistance testing machines or laser-based longitudinal profile measurements [2] and is standardized primarily for high-speed roads, such as highways and interurban routes. Applying these measurement methods to municipal road networks, which have vastly different structural and operational characteristics, is often impractical and not economically feasible. In contrast, standardized visual road-condition assessment methods can be more easily applied to municipal roads, which represent the majority of road networks (between 65% and 80% in most countries) and play a critical role in supporting public transport, private motorized transport, cycling and slow traffic [3–6].

Traditionally, these visual inspections were conducted on-site by experts. Recent advancements in image-based mobile reality capture techniques and cloud technologies have enabled the virtual digitization of such infrastructure using street-level image-based infrastructure management services [7]. However, despite these technological advances, the process remains heavily dependent on manual interpretation, which is time-consuming, resource-intensive and prone to subjectivity. Given the scale and importance of municipal road networks, there is a clear need for automated approaches that can leverage widely adopted professional mobile mapping data. Such methods, in particular deep learning (DL) approaches, could provide a scalable, cost-effective and standardized alternative to traditional interactive image or video analysis by human operators, while still meeting all the requirements for reliable road-condition information.

Our paper investigates the use of 3D street-level imagery from high-end mobile mapping systems (MMSs) for a robust and accurate AI-based assessment of visual road conditions for municipal road networks and features the following main contributions:

- A framework for mapping 3D imagery to object space, featuring an automated process that isolates the road surface by using depth filtering and a 5D clustering algorithm to remove interfering objects.
- A neural network architecture adapted to the specific task of visual road-condition assessment providing a choice of regression RCI or binary classification (road good or poor) output.
- A dataset based on 352 km of municipal roads with professional 3D street-level images and RCIs labelled by experienced road maintenance experts.
- A process for aggregating ‘per frame’ RCI into ‘per segment’ RCI with a process for automatically detecting irregularities in road conditions and for building road segments with homogenous health conditions.
- An output which is fully conformant with established European road-condition standards together with a set of quality indicators to allow road-condition experts to evaluate the quality of the binary classification and of the RCI regression.

2. Related Work

Many existing approaches to visual road-condition assessment focus exclusively on detecting a limited set of asphalt defects, often restricted to cracks, while neglecting the broader context of pavement condition [8–11]. Although valuable, these individual localized defects do not represent overall infrastructure health. As a result, the assessments are often fragmented, producing extensive defect-level outputs that are difficult to interpret and are only partially aligned with standardized evaluation frameworks [12,13]. To support effective infrastructure management and informed decision-making, a more holistic and integrative approach is essential.

A major challenge in this context is the collection and creation of datasets with sufficient scope and quality, as well as the availability of reference data based on established standards. The potential of DL approaches in this field has led to the development of various publicly available datasets. One of the most well-known datasets for damage detection is the German Asphalt Pavement Distress Dataset (GAPs), created by [14] and later extended by [15]. This dataset contains data collected on German national roads using a specialized measurement vehicle with a top-down-perspective view and labelled according to German standards [16]. While datasets like the GAPs adhere to standardized data collection and labelling practices, other datasets with perspective images do not follow any standards. These datasets are often created using low-cost data collection methods and lack the advantages of standardized, accurately georeferenced mobile mapping data [12,13,17,18]. Despite its strengths, the GAPs also has its limitations. The cost-intensive acquisition setup makes it economically feasible only for highways and interurban roads, where data can be gathered efficiently at higher speeds. Moreover, the top-down imaging configuration is optimized for patch-based surface distress detection but cannot be easily extended to further tasks. Table 1 provides a comparison of the main publicly available road-condition datasets. The table highlights their focus on tasks such as individual damage detection, pixel-level segmentation, or patch-based classification, while underscoring the current lack of standardized datasets designed for comprehensive, network-level road-condition evaluation.

Table 1. Comparison of publicly available datasets for pavement-condition evaluation. Task: damage detection (D), damage segmentation (S), patch classification (C). Road types: Highway (H), interurban road (IR), municipal road (MR).

Dataset	Size	Viewing Direction	Data Collection System	Road Types	Task	Damage Categories
GAPs v2 [15]	2468	Top-down	Specialized vehicle	H, IR	D, C	5
Crack 500 [19]	500	Top-down	Low-cost camera	MR	S	Cracks
RDD2022 [12]	47,420	Forward	Low-cost on-board camera	IR, MR	C, D	4
NHA12D [8]	80	Forward, top-down	Specialized vehicle	H	D, S	Cracks
CNRDD [20]	4319	Forward	On-board high-resolution camera	H	C	8
SVRDD [13]	844,432	Forward, top-down	360° roof-mounted camera	H, IR, MR	D	6

Most approaches based on the datasets in Table 1 do not follow established standards for road-condition assessment, which have been developed, validated and refined over

decades. Aligning new assessment methods with such standards is crucial for ensuring broader acceptance in the industry. Most standards distinctly separate visual condition assessment from geometric condition evaluation, primarily because geometric indicators, such as skid resistance or longitudinal evenness, are less applicable or more difficult to interpret in local or municipal road networks. For instance, evaluating longitudinal evenness on municipal roads requires accounting for traffic-calming measures and applying reduction factors due to the lower speed limits (typically below 50 km/h), which not only reduces the reliability of the results but also increases the time and complexity of the evaluation process. Consequently, the assessment of municipal road networks can focus primarily on an evaluation of the road surface based on a visual assessment.

European standards were originally designed for highways and interurban roads, where RCI is typically calculated on a 0.0 to 5.0 scale with a precision of 0.1. The results are subdivided into five main condition classes and incorporated with geometric indicators for the final evaluation [16,21–23]. Outside of Europe, the pavement-condition index (PCI), standardized under the American Society for Testing and Materials standard [24], is widely used. It is based solely on visual assessment and scaled from 0 (failed) to 100 (new road), categorized into seven condition classes. Similarly, the Australian standard for visual road-condition assessment focuses primarily on assessing various pavement distresses based on their severity and extent and aggregating them into an RCI [25]. However, unlike the American PCI, the Australian RCI follows a scale similar to European standards, subdividing the score (0.0–5.0) into five condition classes (Table 2). To capture geometric characteristics, both the American PCI and the Australian RCI must be complemented by additional indicators, such as the International Roughness Index (IRI), which is defined by a separate standard [26].

Table 2. Comparison between American PCI and European/Australian RCI rating scales for road-condition assessment.

Road Condition	Excellent	Good	Fair	Poor	Very Poor	Serious	Failed
US PCI rating scale	100–86	85–71	70–56	55–41	40–26	25–11	10–0
EU/AU RCI rating scale	>0 & <1.0	≥1.0 & <2.0	≥2.0 <3.0		≥3.0 & <4.0	≥4.0 & <5.0	

Despite regional variations, all standards emphasize weighting distress types according to their structural relevance and require the evaluation of both severity and extent. Although the numerical RCI value ranges remain consistent within each standard, the labelling and numbering of condition classes may differ depending on national reporting conventions. Today, however, such in-person inspections are increasingly being replaced by virtual inspections based on mobile mapping data. This shift has improved scalability, accessibility and the potential of an automated pavement-condition evaluation.

3. Materials and Methods

The use of perspective 3D street-level imagery from high-end MMSs with depth information from stereo matching or LiDAR fusion promises several advantages. However, since no datasets with the required characteristics existed, a suitable dataset first had to be created. The overall workflow with the main materials and methods of our investigations with the respective sections of the paper are illustrated in Figure 1. The main aspects include:

- The data acquisition (Section 3.1) with a description of the main characteristics of the high-end mobile mapping data and cadastral road delineation data together with the expert RCI labelling process.
- The data preparation (Section 3.2), including the isolation of the road surface in the 3D imagery and the removal of interfering objects using a 5D clustering process.
- The description of the training and evaluation dataset together with the neural network architecture for RCI regression and binary classification in Section 3.3.
- The postprocessing steps in Section 3.4 aggregating the 'per frame' RCI predictions of the neural network (NN) into RCI values for contiguous and homogeneous road segments.
- Finally, the evaluation and refinement of the results in Section 3.5.

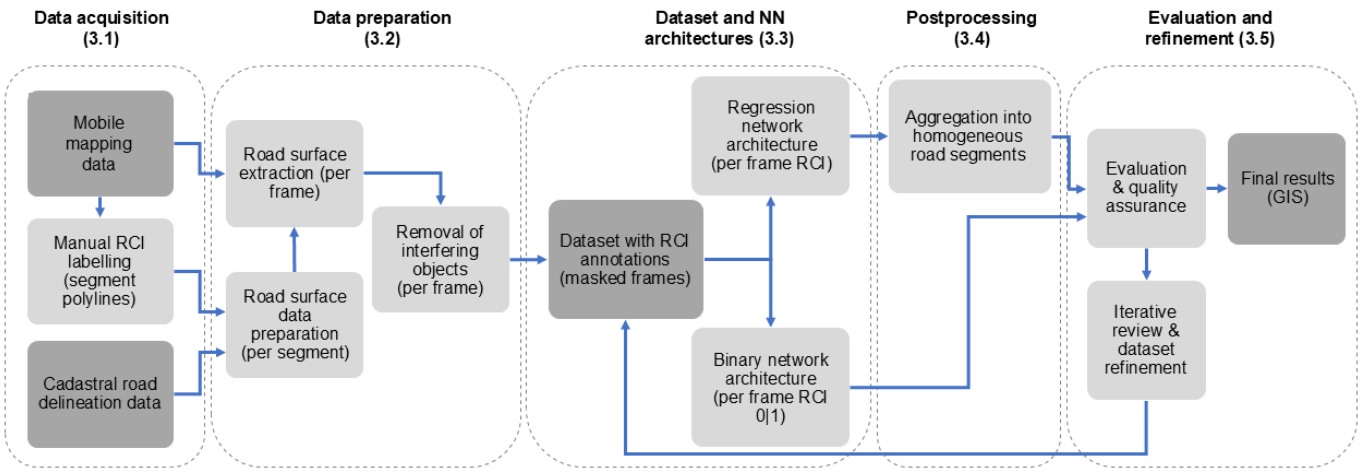


Figure 1. Overview of the proposed workflow for AI-based visual road-condition assessment with the main materials and methods and their respective sections in this paper (data set in dark grey and processes in light grey).

3.1. Data Acquisition

Precise georeferenced 3D street-level imagery forms the foundation of our proposed solution. For the development and training of the RCI-regressor network, labelled RCI values for street segments are also of great importance. To this end, the available data include approximately 352 km of asphalted roads across eight municipalities, annotated with expert-assessed RCI values. In total, around 65,000 images along with the corresponding RCI labels were preprocessed as described in Section 3.2.

3.1.1. Mobile Mapping System and Data

Vehicle-based MMSs are commonly used for efficient infrastructure management along road corridors spanning entire cities or states. For our investigations, we used street-level imagery captured by iNovitas (iNovitas AG, Baden-Dättwil, Switzerland) provided as a cloud-based infrastructure platform. At the time of the study, their fleet of mapping vehicles (Figure 2) featured a multitude of mapping sensors: three high-resolution (16 MP) stereo camera systems and a multi-head panorama camera, a Ladybug5 (Teledyne DALSA, Waterloo, ON, Canada) with a resolution of 5 MP per camera head, all mounted on the roof of the system (Figure 2). Additionally, a high-end profile LiDAR scanner was attached to the rear of the car. The three stereo systems were arranged in three different viewing directions: the main system facing forward and the other two facing the back-right and back-left respectively (see Figure 2). During data acquisition, all camera systems

were triggered simultaneously based on a distance criterion, ensuring a spacing between consecutive images of 4 m.

Figure 2. MMS and sensor configuration employed for the data collection campaign.

For georeferencing, the MMS combines a geodetic GNSS receiver and a tactical-grade IMU. Direct georeferencing using INS and GNSS sensor data fusion [27] is the standard georeferencing procedure in outdoor, vehicle-based mobile mapping applications. However, even direct georeferencing with high-end sensors can be affected by positioning errors of up to several metres due to multipath effects and GNSS signal obstructions [28]. To ensure homogenous georeferencing accuracy within the whole mapping perimeter, an integrated georeferencing approach based on GCPs is employed [29]. This approach allows for precise co-registration of mobile mapping data with cadastral data, even if local distortions in the geodetic reference frame are present [29]. By using optimally distributed, RTK-GNSS-determined GCPs, the final image poses are guaranteed to have a positional error of less than 10 cm.

This process yields precisely georeferenced 3D image spaces (RGB-D images) [7] along the entire mapping trajectory at an average interval of 4 m. Depth maps are computed for all camera systems. By employing a proprietary stereo-matching and depth-completion process, the depth maps for stereo images are computed and completed in areas with low image texture. The depth map completion process involves filling and filtering holes by reprojecting the LiDAR point cloud into the images via known camera poses and camera intrinsics.

After evaluating the available camera systems, the panoramic camera was excluded: its central roof mounting, large vertical field of view and lower geometric resolution lead to insufficient road-surface coverage. The rear-facing stereo systems, while ideal for asset management, were also excluded, since they capture only small portions of the road. In contrast, the front-facing stereo system provides high geometric resolution and, with a horizontal field of view (FOV) of 96°, the left camera covers the road surface optimally. Due to the sensor mounting, the road surface is visible from a depth of more than 3 m. The sensor setup yields ground sampling distances (GSD) between 1.5 mm and 4.1 mm on the road surface for image depths between 3 and 10 m.

3.1.2. Manual RCI Labelling

Assessing the visual condition of municipal road networks is currently a manual process. Due to the inherent complexity caused by various damage patterns and the necessary understanding of the structural interrelationships of roads, road-condition assessments must be carried out by domain experts. However, domain experts do not directly assign RCI to road segments. Instead, they gather information about the presence and frequency of relevant damage patterns, from which the RCI of a road segment is calculated according to established standards. In this study the European RCI convention is applied; RCI values increase monotonically with pavement deterioration. Each segment represents an area with a homogenous RCI, although different damage patterns with varying distributions can occur in a segment. However, segments are not only formed by homogeneous RCIs, but also by additional information. In addition to administrative divisions such as street names, changes in traffic load categories, in particular, lead to new segments, regardless of the RCI, because varying traffic types and intensities lead to different levels of wear over time and therefore to different maintenance measures.

Traditionally, road conditions were assessed on-site by inspecting the entire road network. Recently, professional street-view services have been employed to enable a virtual assessment process. Such services have the advantage of being weather- and time-independent. Viewpoints and zoom can be changed instantly, enabling a detailed inspection of areas of interest with little effort. Additionally, contrast and brightness can be adjusted and tools such as cross-section analysis can be used, allowing for a detailed analysis of visual road conditions.

To create an effective dataset for a DL-based road-condition assessment, senior domain experts selected representative municipalities across Switzerland. They manually labelled four entire municipalities in a standardized fashion using the street-level service infra3D, with identical imagery used for the AI-based assessment. To address the rare occurrence of 'very poor' to 'failed' RCI classes in Switzerland, a targeted assessment of selected roads with poor to very poor conditions in four additional municipalities was added. Since municipal road networks are often not assessed by the same expert, the inter-expert variability was evaluated in collaboration with the experts. This variability was empirically determined to correspond to half an RCI class (± 0.5).

To facilitate the interpretation of the road-condition assessments, the road network is subdivided into segments (see Figure 3). A segment typically represents a road between two network nodes, such as intersections, crossroads, or roundabouts. Segment lengths can vary, and in many cases, the RCI values differ significantly within a single segment due to different maintenance schedules. In such cases, domain experts further subdivide the segment into smaller portions that exhibit relatively uniform RCI values. This refinement provides a clearer overview of which and when specific road segments require maintenance actions. The assessment results are stored as 2D polylines in a GIS-Layer, as shown in Figure 3, together with additional attributes about the occurrence and frequency of relevant damage patterns.

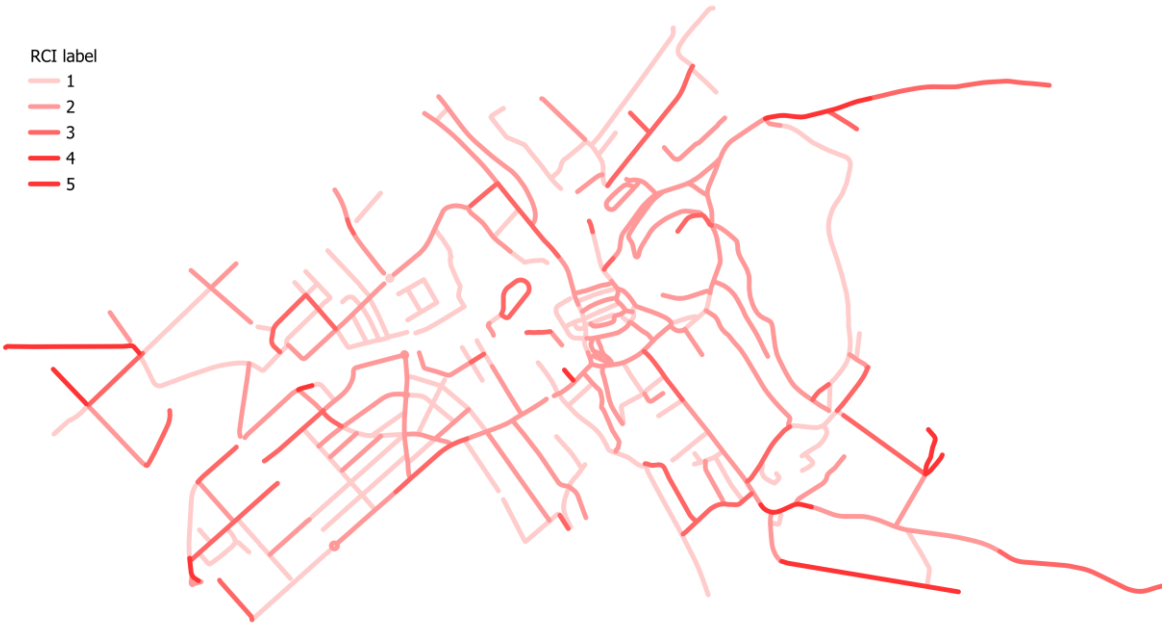


Figure 3. Typical GIS dataset with RCI classes 1–5 assigned to individual road polyline segments with homogeneous health characteristics. The class intervals for the RCI are left inclusive: [0.0, 1.0); [1.0, 2.0); [2.0, 3.0); [3.0, 4.0); [4.0, 5.0].

3.1.3. Cadastral Road Delineation Data

To support the automatic extraction of road surfaces from the mobile mapping imagery, an additional geospatial dataset was integrated and preprocessed. In this context, the official Swiss cadastral land-cover data, publicly available geodata describing both artificial and natural surface types, were used. These data are regularly updated by a nationwide network of licenced professionals and provide positional accuracies ranging from 0.2 m in urban areas to 1.5 m in mountainous regions. From the available land-cover classes, the category representing roads and paths was selected. This class includes all drivable and walkable paved public surfaces but excludes sidewalks. The selected data were then preprocessed as described in the following sections to extract the road surfaces from the 3D street-level imagery.

3.2. Data Preprocessing

The street-level imagery used in our study is of general purpose and covers not just the road surface but the entire streetscape. This also includes vehicles, pedestrians, temporary objects, safety infrastructure, etc., which obstruct the view of the road surface and thus pose potential problems for the training and inference of NNs. For accurate pavement evaluation, it is essential to extract only the relevant and unobstructed road surface portions. This subsection covers the data preprocessing steps to remove all non-road-surface information from the imagery and for assigning the correct RCI value to each image. For privacy and data protection reasons, vehicles and pedestrians are already anonymized (blurred) in the street-level imagery. The masks resulting from the anonymization process were used to completely remove the elements from the imagery.

3.2.1. Road-Surface Data Preparation

Since the RCI annotations were provided as 2D polylines per road segment, assigning the correct RCI value to each image in the dataset requires a precise definition of the road-surface area. To achieve this, we utilize the official Swiss cadastral land-cover data as a reference for delineating road surfaces in the imagery. The subdivision process is

supported by a series of spatial analysis operations depicted in Figure 4: first, the annotated polylines (blue lines in (a)) are buffered to generate polygonal road segments, which ensures sufficient coverage around the centerlines (cyan areas in (b)). These buffered geometries are then clipped with the cadastral land-cover data (orange areas in (a)) to accurately match the official road extents and to exclude unwanted or non-relevant areas, such as private driveways or off-road areas (c). Finally, the resulting road surfaces are subdivided according to the annotated segments, with the corresponding RCI values assigned as attributes. This approach allowed the subsequent extraction of the relevant road portions in the mobile mapping images, while enabling seamless integration with the RCI annotations.

(**a**) (**b**) (**c**)

Figure 4. Road segment preparation steps using spatial analysis: (**a**) initial data, (**b**) buffered RCI annotation polylines, (**c**) clipped road surfaces with RCI annotation using land-cover data. Orthophoto: Federal Office of Topography swisstopo.

3.2.2. Road-Surface Extraction in Image Frames

For each road area obtained from the previous step, the images within a buffer of 15 m are selected. Using the available depth maps, the known intrinsics, and the precise camera poses, we construct a point cloud in the superordinate coordinate system for each image. As mentioned in Section 3.1.1, we only consider road surface up to 10 m distance. Hence, depth values larger than 10 m are neglected. The point cloud is then filtered by the 2D road area from the previous step and reprojected into the image space, resulting in a mask representing the relevant road surface. This ensures the only the relevant pixels are presented to the neural network.

3.2.3. Removal of Interfering Objects

Objects such as safety infrastructure, e.g., traffic islands and bollards, and temporary objects, e.g., construction-site barriers, children's toys, garbage, or simply vegetation reaching into the road, are not filtered out by the previous step. Since they are not very frequent, it cannot be expected that the regressor can adequately ignore them. To remove these remaining interfering objects, we construct a point cloud via known intrinsics and the depth map filtered by the mask computed in the previous step. Subsequently, for each point, a normal vector is computed. We expect the camera viewing direction to be approximately parallel to the road surface. Thus, we remove all points with normal vectors that differ more than 45° from the up vector. Additionally, all points above the camera are removed as well. This allows us to filter trees or foliage overarching streets. Finally, we perform DBSCAN clustering [30] to remove the last interfering objects present. Due to the highly variable point density caused by the combination of the perspective sampling of the camera and its mounting (see Section 3.1.1), DBSCAN clustering fails when applied to the point cloud directly. This means that the global density criterion of the DBSCAN algorithm

is easily met by points close to the camera, while it is more difficult or impossible to meet for distant points. To mitigate this problem, to each point we add its image coordinates (u,v) scaled by the maximum GSD of 4.1 mm. The maximum GSD is computed by the known camera intrinsics, the relative orientation of the camera to the street surface and the maximum depth of 10 m. The radius ε, that is used to count the number of neighbouring points, is computed in this five-dimensional space, enforcing vicinity both in image space as well as in the point cloud.

After filtering all interfering objects in the point cloud, we reproject the points into image space to obtain a mask. Small holes measuring less than 1000 pixels are not considered disruptive objects, but rather noise-affected depth values that have been incorrectly removed or entirely missing depth values. Thus, those holes are filled via morphological filtering to obtain a more consistent mask. The threshold of 1000 pixels was determined empirically, based on the observation that holes in the mask usually comprise of up to 500 pixels.

3.3. Dataset and Neural Network Architectures

To automatically assess road condition from mobile mapping imagery, we designed a tailored DL architecture. The following subsections outline the neural network architectures, the chosen loss functions and the evaluation metrics. The architecture builds on a convolutional backbone combined with a regression head to predict continuous RCI values and a variant adapted for binary classification. The second part explains in detail the chosen loss functions for robust optimization and the adopted evaluation metrics for performance assessment.

3.3.1. Neural Network Dataset

The prepared dataset includes approximately 65,000 mobile mapping frames, featuring diverse characteristics such as road widths, markings and lighting conditions. To ensure consistency with established standards, all data were collected under favourable weather conditions, with dry pavement surfaces free of foliage or debris. The integrated annotations include all five European RCI classes from 0.0 (excellent pavement condition) to 5.0 (very poor pavement condition) with a resolution of 0.1 RCI. Figure 5 (top) shows the distribution of annotations at this 0.1 RCI-class precision. The RCI labels are defined at road-segment level, with segment lengths varying across the dataset. The segment length distribution exhibits a first quartile of 91 m, a median of 137 m, and a third quartile of 180 m, with lengths ranging from 50 m to 450 m. When these frame labels are assigned to individual frames and grouped by RCI class (Figure 5, bottom), it is evident that the first four condition classes are well represented, whereas the fifth class (RCI > 4) exhibits a significant imbalance. To mitigate this imbalance, data augmentation techniques, specifically horizontal image flipping, were applied to increase the representation of the fifth class. To ensure the independence of model evaluation, the road segments in the dataset were partitioned into three balanced subsets: 70% for training, 15% for validation and 15% for evaluating final model performance.

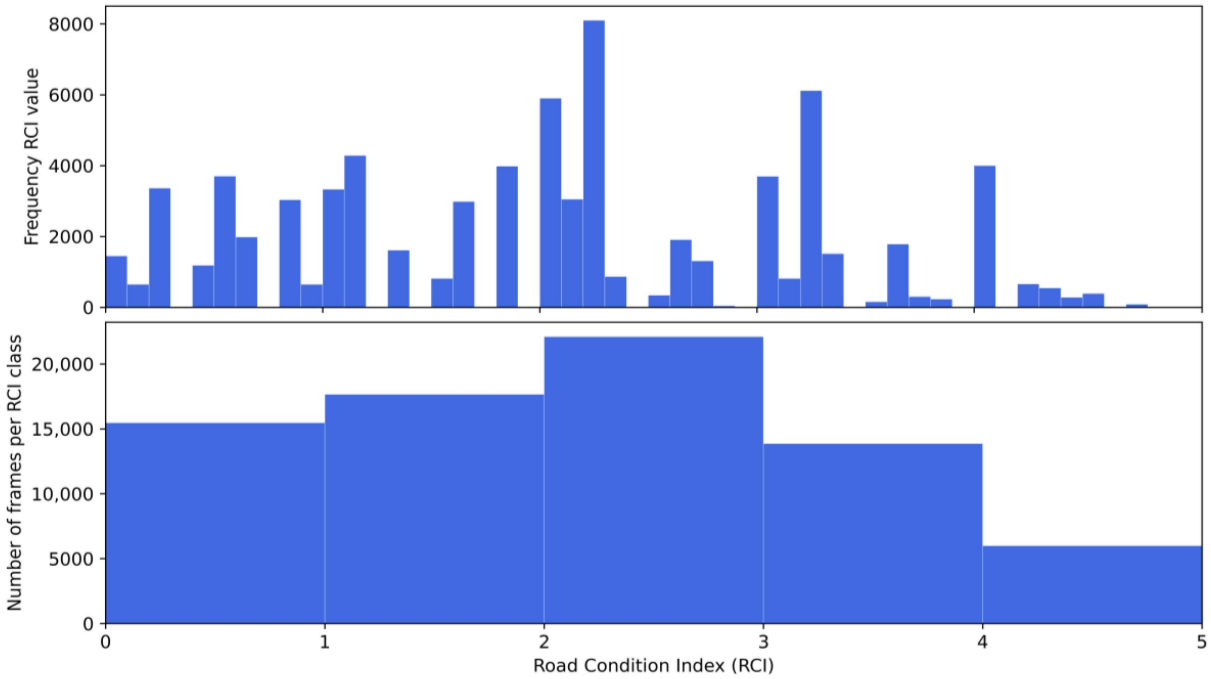


Figure 5. Detailed distribution in increments of 0.1, corresponding to the maximal resolution of the label data (**top**) and dataset distribution grouped by RCI class (**bottom**).

3.3.2. Neural Network Architecture

The proposed neural network architecture is based on convolutional neural networks (CNNs), with an enhanced regression tail composed of two fully connected layers (Figure 6) using a rectified linear unit (ReLU) activation function and dropout regularization. CNNs are well-suited for image-based tasks due to their ability to process variable-sized inputs, unlike traditional fully connected networks that require fixed input dimensions. This flexibility was particularly beneficial for our task and dataset, as the road-surface extraction process during preprocessing generates images with varying dimensions. In addition, the regression tail enables the network to directly output the RCI.

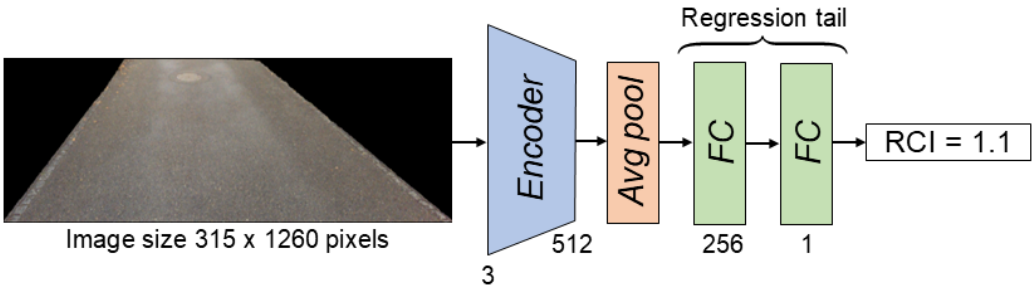


Figure 6. Simplified architecture of the CNN backbone combined with the enhanced regression tail for direct RCI prediction inclusive of the number of channels resulting from each step (Fully Connected layer in green; Average Pooling layer in orange; Encoder backbone in blue).

To identify the most suitable baseline, we evaluated several CNN backbone families and depths: VGG11 [31], DenseNet121 [32], MobileNet V3 [33], Inception V3 [34], ResNet18, ResNet50, and ResNet101 [35], and ResNeXt50 [36]. Given the proven generalization capabilities of CNNs across different tasks, we adopted a transfer learning approach, utilizing ImageNet-pretrained weights [37] for all backbones to accelerate convergence and improve performance. A key consideration after this step was the mismatch in aspect ratio and resolution between ImageNet's standard square input size (240 × 240 pixels)

and the preprocessed road images, which are rectangular, with an average resolution of 840 × 3360 pixels. To address this, we systematically trained the model using a range of image scales, from 630 × 2520 pixels down to 146 × 584 pixels. This approach allowed us to determine the optimal input size for the chosen architecture while preserving critical visual features relevant to road-condition assessment and maintaining computational efficiency.

To further evaluate the network architecture's ability to generalize across the unknown label distribution at the frame level, we implemented an additional variant of the same neural network tailored for a simplified binary classification task. In practical road maintenance planning, it is generally expected that roads with an RCI above 2.5 require maintenance interventions in the near future to preserve serviceability [38]. Based on this principle, the binary classification model was trained to distinguish between roads that do or do not require repair measures, using an RCI of 2.5 as threshold. To adapt the network for this task, a sigmoid activation function was added to the final fully connected layer, ensuring a binary network output as shown in Figure 7.

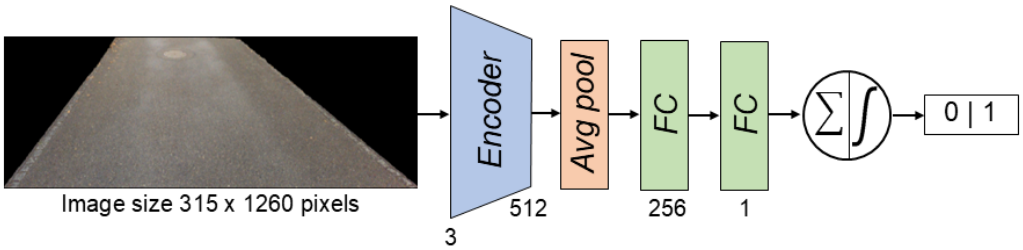


Figure 7. Simplified architecture for binary classification of the CNN backbone combined with two fully connected layers and the sigmoid activation function enabling binary output (Fully Connected layer in green; Average Pooling layer in orange; Encoder backbone in blue).

3.3.3. Loss Function and Evaluation Metrics

In optimizing the training process of CNNs, both root mean squared error (RMSE) and mean absolute error (MAE) are commonly employed as loss functions. RMSE is sensitive to larger deviations and therefore penalizes significant errors more heavily, while MAE treats all errors linearly, offering greater robustness to outliers. To benefit from the strengths of both metrics, we adopted the Huber loss (1) function during training, which offers smooth and robust optimization for noisy, real-world data. The dataset involves a significant degree of implicit label noise. Within a given segment, locally confined areas with severe damage may be visually diluted by adjacent pavement along the road. Furthermore, empirical analysis of inter-operator variability revealed discrepancies of ±0.5 RCI values (σ = 0.5). The selection of the threshold δ for the Huber loss was guided by the observed inter-operator variability. A sensitivity analysis was conducted for δ values ranging from σ to 2σ. The optimal validation performance was obtained for δ = 0.9, which ensures a model that is less sensitive to larger errors while maintaining accuracy for typical deviations. On the contrary, we used the Binary Cross-Entropy (BCE) loss function for the binary classification variant of the network, which is well-suited for binary outputs.

$$L_\delta(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & for\ |y - \hat{y}| \le \delta \\ \\ \delta\left(|y - \hat{y}| - \frac{1}{2}\delta\right) & otherwise \end{cases} \quad (1)$$

The performance evaluation of the trained models has been conducted on the test dataset. For the RCI regression output, we reported the MAE between the predicted and ground-truth values to provide an interpretable, unit-consistent measure of model

accuracy. On the contrary, we used the F1-score to evaluate the performance of the binary model, which is tailored for classification tasks. To assess the reliability of these results, 95% confidence intervals were calculated by bootstrapping the test set with 1000 iterations.

3.4. Postprocessing: Aggregation of Frames and Formation of Road Segments

The implemented neural network architecture, described in Section 3.3.2, processes individual image frames and predicts an RCI value for each. As a result, multiple predictions are generated along the length of each road segment. To aggregate these frame-level predictions into coherent subsegments while minimizing the influence of outliers, a two-step method was adopted.

First, the image frames are spatially sorted along the road axis. The sequence of predicted RCI values is then analyzed using the Isolation Forest (IF) algorithm [39] to detect and exclude outliers. This step ensures that the subsequent analysis, which is sensitive to anomalies, remains robust. In the second step, the cleaned sequence of predictions is processed using the Pruned Exact Linear Time (PELT) algorithm [40], combined with a Radial Basis Function (RBF) kernel as the cost function. This enables the automatic identification of discontinuities, where significant shifts in road condition occur. As illustrated in Figure 8, these change points are used to partition each road segment. The centroids of the reprojected road surfaces corresponding to the detected change points serve as the division boundaries. To ensure robustness against outliers, the RCI for each newly formed subsegment is determined using the median of all predictions within the latter.

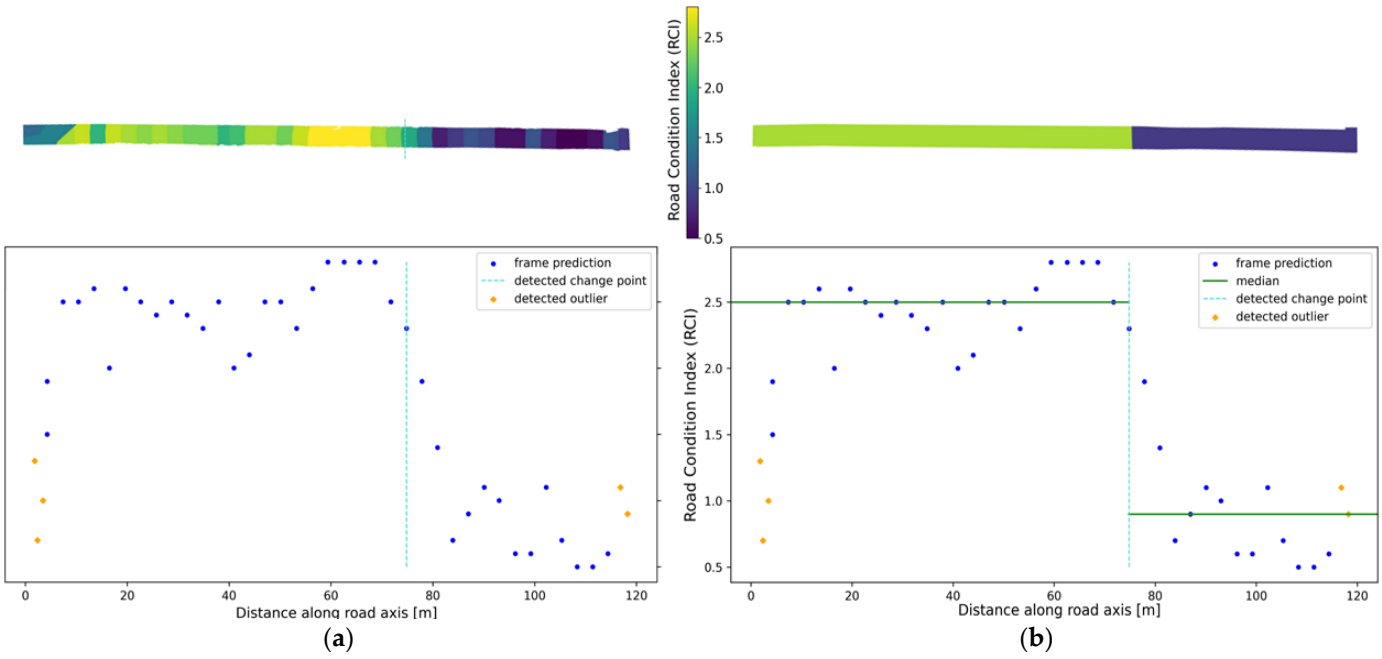


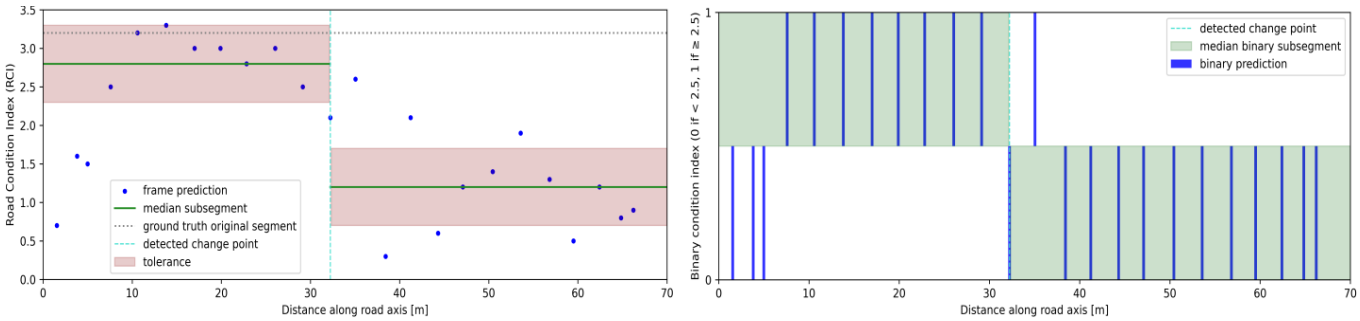
Figure 8. Example of road-segment subdivision based on predicted RCI values and detected outliers by IF. (**a**) Detected change points along the frame-level RCI predictions using the PELT algorithm with an RBF kernel. (**b**) Resulting subsegments, each assigned a median RCI value.

3.5. Iterative Quality Improvement

The training data used in this study was manually annotated by expert engineers. To minimize the impact of human error or variability among operators, an iterative quality assessment process was employed. After training and tuning several models, the final model that achieved the best performance on the test dataset was selected for full-scale inference. This model was applied to the entire dataset to generate predictions and enable the derivation of RCI road segments. Subsequently, the predicted results and RCI segments

were reviewed again by expert engineers to evaluate whether corrections were needed in the original ground-truth annotations. Based on this assessment, any necessary adjustments were made to the dataset, which was then used to retrain the model, thereby enhancing both the quality and consistency of the training data.

To aid expert interpretation, we supplemented the RCI absolute error with additional quality indicators that reflect the prediction variability within each segment and serve as a measure of model uncertainty. Since the newly generated RCI segments do not have ground-truth labels, these indicators provide guidance for expert reviewers in identifying segments where corrections may be necessary. For the binary classification task, uncertainty is represented using the portion of correctly classified frames compared to the subsegment median (Figure 9b). For the regression task, we measured the portion of predictions that fell within ± 0.5 of the calculated subsegment median RCI value, an empirically chosen threshold that aligns with observed inter-expert variability (Figure 9a).



(**a**) (**b**)

Figure 9. Calculated quality indicators to guide expert review. For the regression task (**a**), uncertainty is quantified as the portion of predictions within ± 0.5 of the calculated subsegment median (quality score 58.3%). For the binary classification task (**b**), uncertainty is represented using the portion of correctly classified frames compared to the subsegment median (quality score 83.3%).

4. Experiments and Results

This section presents the experimental evaluation of the proposed methodology. All experiments were conducted on a workstation equipped with an NVIDIA RTX A5000 GPU with 24 GB VRAM, an Intel Xeon i7 processor and 256 GB RAM. We begin by testing the preprocessing workflow (Section 4.1). Next, we assess different neural network architectures and training configurations, including backbone selection, batch size and input image resolution (Sections 4.2.1 and 4.2.2). The resulting models are then evaluated in two variants: binary classification and regression (Sections 4.2.3 and 4.2.4). Finally, we demonstrate the effectiveness of the frame aggregation algorithm and iterative dataset refinement process in Section 4.3.

4.1. Filtering of Interfering Objects

To assess the proposed workflow for the removal of interfering objects, we conducted tests on a set of representative images. Fortunately, such objects are rare in our dataset, so the evaluation was limited to qualitative analysis. The proposed method significantly improves road-surface extraction (Figure 10c) compared to the standard approach (Figure 10b), which relies solely on road-surface delimitation by depth filtering. Vehicles and pedestrians were not included in this evaluation, as they were already filtered out during preprocessing of the street-level imagery (see Section 3.2). The method proved particularly effective for handling residual interfering objects such as vegetation, roadwork barriers and traffic islands. By removing these elements, the workflow ensures that only pixels corresponding

to the actual road surface are passed to the neural network, thereby improving the quality and reliability of the input data.

(**a**) (**b**) (**c**)

Figure 10. Results and intermediate results obtained with the proposed preprocessing workflow for extracting road surface from imagery on representative images. Column (**a**) shows the original street-level imagery; (**b**) depicts the results after filtering the depth maps by the 2D road area (Section 3.2.2); (**c**) illustrates the final road surface obtained by removing points based on normal vectors and subsequent DBSCAN clustering (Section 3.2.3).

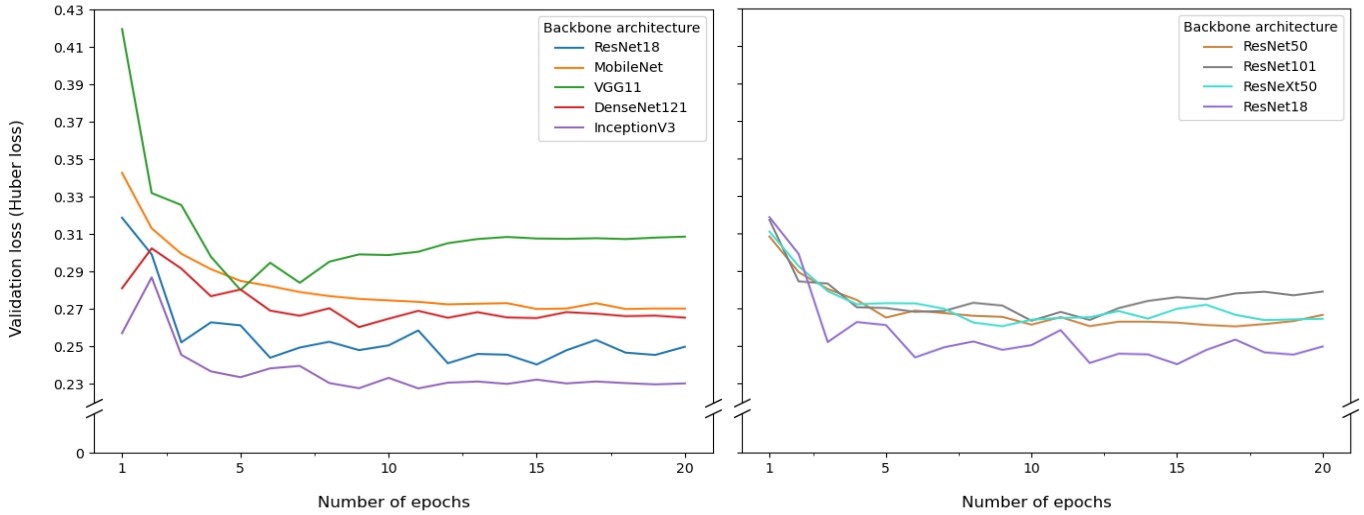
4.2. Neural Network Architecture Evaluation

We first evaluated the optimal backbone architecture for the neural network in Section 4.2.1 and subsequently investigated the impact of different batch size and input image dimensions in Section 4.2.2. After tuning the key hyperparameters on the validation set, including the learning rate, learning rate decay strategy and dropout rate for regularization layers, we evaluated the final performance on the test dataset for both the binary classification model in Section 4.2.3 and the RCI regression model in Section 4.2.4. All experiments were conducted on the reported hardware. Training required approximately 80 h with a batch size of 32.

4.2.1. Backbone Architecture

The evaluation of different backbone architectures was conducted using the same dataset split (70/15/15) for training, validation and test as well as identical hyperparameter configurations across all models. Each model was trained until overfitting was observed by

a divergence between the training and validation loss curves. The best-performing model checkpoint (based on validation loss) was then used for inference on the test set. Figure 11 presents a comparison of validation loss across different architecture families (a) and within various depths of Residual Networks (b).



(**a**) (**b**)

Figure 11. Comparison of validation loss (Huber loss) across different backbone architectures. (**a**) Performance of various architecture families and (**b**) performance comparison within the Residual Network family at different depth levels.

As shown in Table 3, the ResNet18 backbone achieved the best performance, with an MAE at frame level of 0.51 RCI values. Although the InceptionV3 backbone performed well during validation, it exhibited inferior generalization on the test dataset, yielding a higher MAE at frame level than ResNet18. Based on these findings, ResNet18 was selected as the backbone architecture for subsequent experiments.

Table 3. Performance comparison of different backbone architectures on the validation and test datasets using the validation loss (Huber) and MAE for RCI prediction are reported. The best result achieved for each measure is highlighted in bold.

Backbone Architecture	Best Validation Loss ↓	MAE Test
DenseNet121	0.261	0.59
Inception V3	**0.227**	0.53
MobileNet V3	0.272	0.61
ResNet18	0.241	**0.51**
ResNet50	0.262	0.59
ResNet101	0.263	0.60
ResNeXt50	0.261	0.58
VGG11	0.281	0.62

4.2.2. Batch and Input Image Size

For the selected backbone architecture, we systematically evaluated the effect of varying batch and input image resolution. To identify the optimal batch size, experiments were conducted with different values while keeping the input image size fixed at 210 × 840 pixels. As shown in Table 4, both the validation loss and the test MAE at frame level decreased as the batch size increased up to 32, after which performance began to degrade. Based on this trend, a batch size of 32 was selected for all subsequent experiments.

Table 4. Effect of training batch size on model performance during validation and testing process (MAE at frame level). In bold, the lowest value reached. The best result achieved for each measure is highlighted in bold.

Batch Size	Best Validation Loss	MAE Test
8	0.192	0.599
16	0.185	0.573
32	**0.181**	**0.546**
64	0.192	0.636
128	0.194	0.659

Given that the extracted road-surface images have an average aspect ratio of approximately 4:1, the next step was to determine the minimum resolution at which critical features could still be preserved while maintaining this aspect ratio. The goal was to balance input size, feature retention and computational efficiency. As summarized in Table 5, higher resolutions generally improved validation performance; however, the best test performance was achieved at a resolution of 315 × 1260 pixels. Consequently, this resolution was adopted for all further experiments.

Table 5. Effect of input image size on model performance. It reports the best validation loss and the corresponding MAE at frame level on the test set for different input resolutions. The best result achieved for each measure is highlighted in bold.

Image Input Size [px]	Best Validation Loss	MAE Test
146 × 584	0.249	0.576
210 × 840	0.255	0.541
315 × 1260	0.231	**0.507**
420 × 1680	0.230	0.524
525 × 2100	0.233	0.532
630 × 2520	**0.226**	0.512

4.2.3. Binary Classification

The binary classification model was designed to distinguish between road segments that require planned maintenance and those that do not. Based on VSS 40 730b (2019), segments with an RCI value below 2.5 were labelled as 'good' (no immediate maintenance required), while segments with RCI $\geq$ 2.5 were labelled as 'bad' (requiring repair). The model outputs a binary prediction reflecting this classification, considering samples labelled as 'bad' positive and the samples labelled as 'good' negative. On the test dataset, the classifier achieved an accuracy of 0.88 (95% CI: 0.87–0.89) and an F1-score of 0.85 (95% CI: 0.84–0.86), demonstrating strong overall performance. However, the performance differed across classes. As shown in Figure 12a, the model achieved a higher score of 0.90 for segments in bad condition, while the detection of segments not requiring repair was comparatively more challenging, with a score of 0.87. Furthermore, the influence of segments with RCI values near the threshold was examined by removing segments with RCI values between 2.25 and 2.75 from the test dataset. The new confusion matrix (Figure 12b) shows that the evaluation without these segments improved the performance by 3.5%.

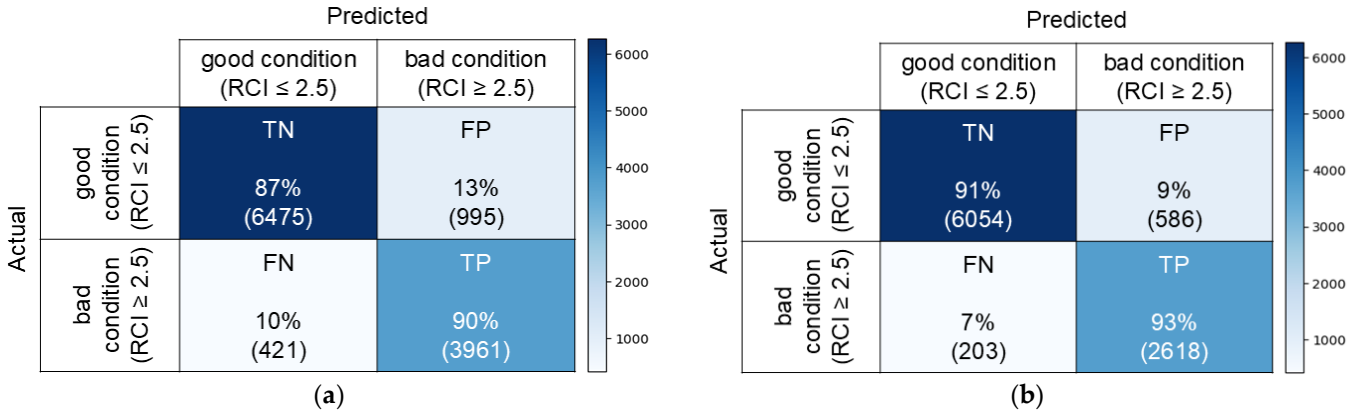


Figure 12. Confusion matrices of the binary classifier. (**a**) Confusion matrix on the full test dataset. (**b**) Confusion matrix after removing segments with RCI values near the threshold (2.25 > RCI < 2.75). Classes: true positive (TP), false positive (FP), true negative (TN), false negative (FN).

4.2.4. Regression Model

The regression model was trained using the same architecture and training strategy as the binary classifier, but with a continuous output predicting the RCI directly. The model achieved an MAE of 0.48 RCI values at frame level (95% CI: 0.473–0.491) on the test dataset, indicating accurate prediction performance across the test dataset, even better than the observed inter-expert variability. In terms of computational efficiency, the model achieved an average inference time of one minute per kilometre (250 images), demonstrating its suitability for large-scale infrastructure monitoring. Figure 13 shows the distribution of prediction errors across the five RCI classes. The model performed best in the first three classes (RCI $\leq$ 3.0), where the road condition is typically acceptable. Performance decreased for higher RCI values (RCI > 3.0), corresponding to more severe road damage, with MAEs at frame level of 0.62 RCI values and 0.94 RCI values for the fourth and fifth classes, respectively. Furthermore, a correlation between the number of training samples and the achieved MAE per RCI class is observable (Figure 13a). Overall, approximately 87% of the test samples were predicted within one RCI class of the ground truth and nearly 61% fell within the empirically determined half-class scattering range. Achieving the reported performance was only possible after refining the dataset through the iterative quality assurance process described in the next Section 4.3.

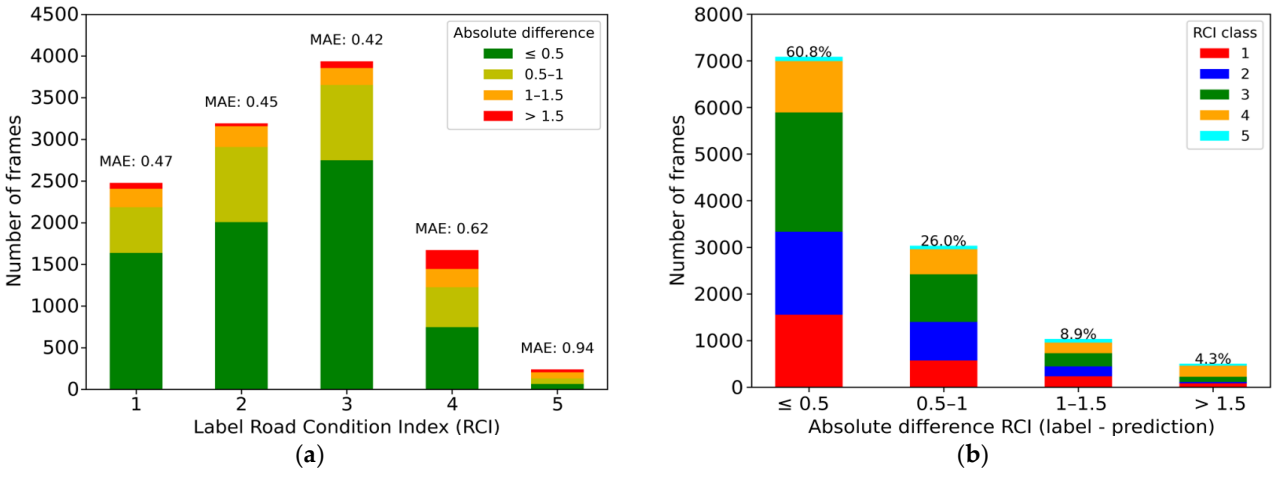


Figure 13. Resulting model prediction errors on the test dataset. (**a**) Errors per RCI class with MAE and (**b**) number of frames within the MAE range subdivided by RCI class. The MAE intervals represented are right-inclusive ($\leq$0.5; (0.5, 1]; (1, 1.5]; >1.5).

4.3. Frames Aggregation Algorithm and Dataset Refinement

The implemented outlier detection method ensured robust filtering of anomalous predictions, which is critical for the change-point detection algorithm due to its sensitivity to outliers. This process enabled accurate subdivision of road segments into homogeneous RCI subsegments. As illustrated in Figure 14, the two displayed road portions clearly belong to different RCI classes and must therefore be aggregated separately. These qualitative examples highlight that, without outlier removal with IF, reliable detection of RCI change points along the road axis with PELT and RBF only would not be possible.

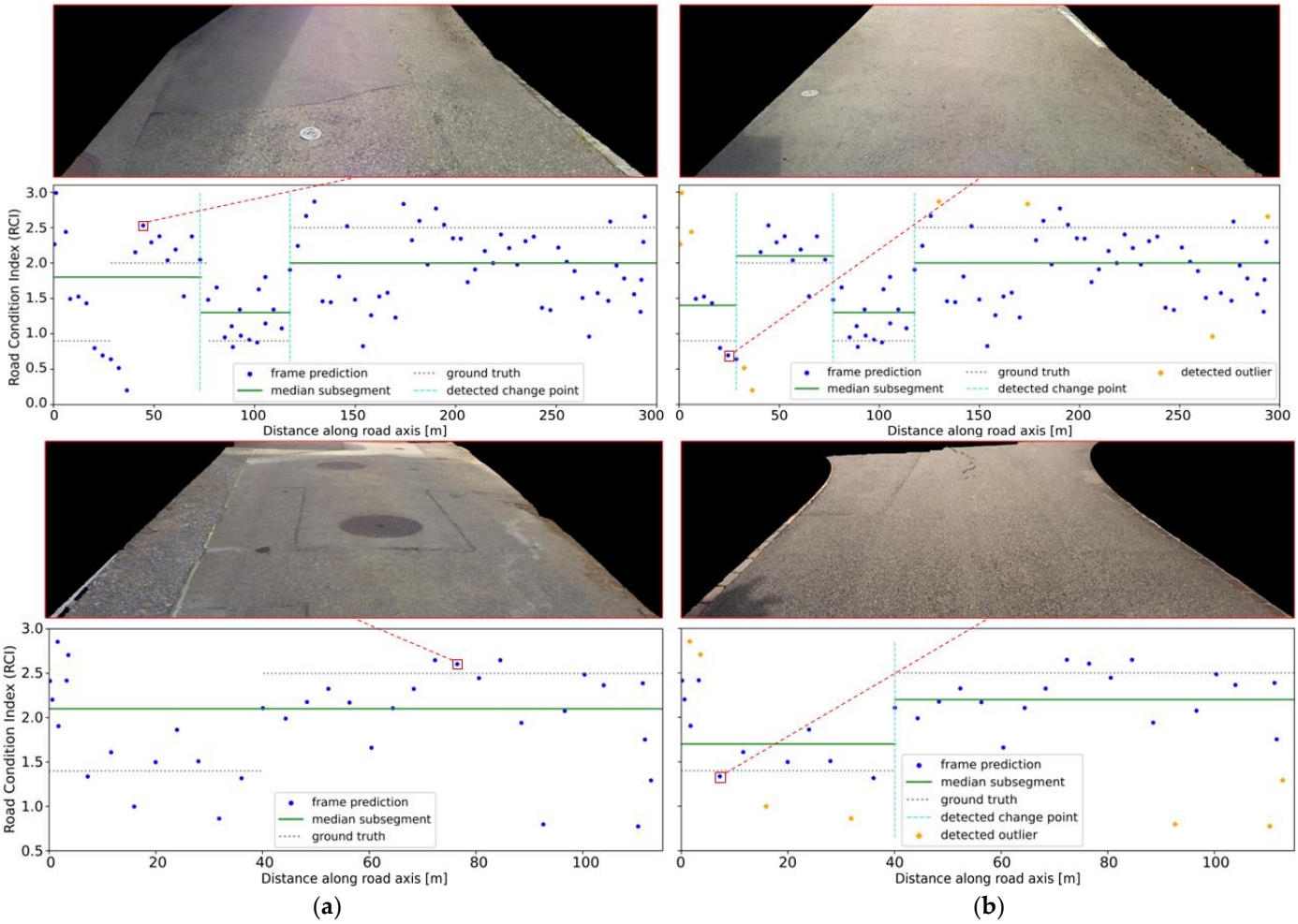


Figure 14. Qualitative result examples of the detected change points for road-segment subdivision with (**b**) and without (**a**) outlier detection using IF. The sampled road-surface images on the top show how different the pavement conditions are in the subsegments.

Supported by outlier removal, the subsequent change point detection identified 76 new road-segment boundaries, which were validated by experts and incorporated into the dataset. This corresponds to an increase of approximately 5% in the total number of segments compared to the original dataset. The subdivision into subsegments led to a reduction in intra-segment variability, from an average variance of RCI predictions of 0.22 to 0.15, showing dispersion improvement. In addition, the average MAE of RCI predictions within these segments decreased from 0.53 to 0.47 RCI values. For comparison, the segments identified as homogeneous by the algorithm, and thus not subdivided, exhibited an average variance of RCI predictions of 0.17 and an average MAE of 0.48 RCI values, demonstrating that the new subsegments now align closely with the characteristics of the naturally homogeneous segments. Qualitative examples from the test dataset are shown in Figure 15.

It illustrates the impact of introducing subsegments derived from frame-level predictions with outlier removal with a comparison between (a) the median of the predictions along the road axis in the original segment and (b) the medians of subsegments generated with the proposed approach. These examples clearly demonstrate that different portions of the same road segment can exhibit substantially different RCI values, highlighting the importance of finer-grained subdivision for accurate assessment. The generation of subsegments changed the original segment length distribution, presented in Section 3.3.1, which now exhibits a first quartile of 72 m, a median of 118 m, and a third quartile of 163 m, with lengths ranging from 30 m to 401 m. The subsegments shown in Figure 15b were evaluated by field experts, who provided new ground-truth values to integrate the refined segments into the dataset.

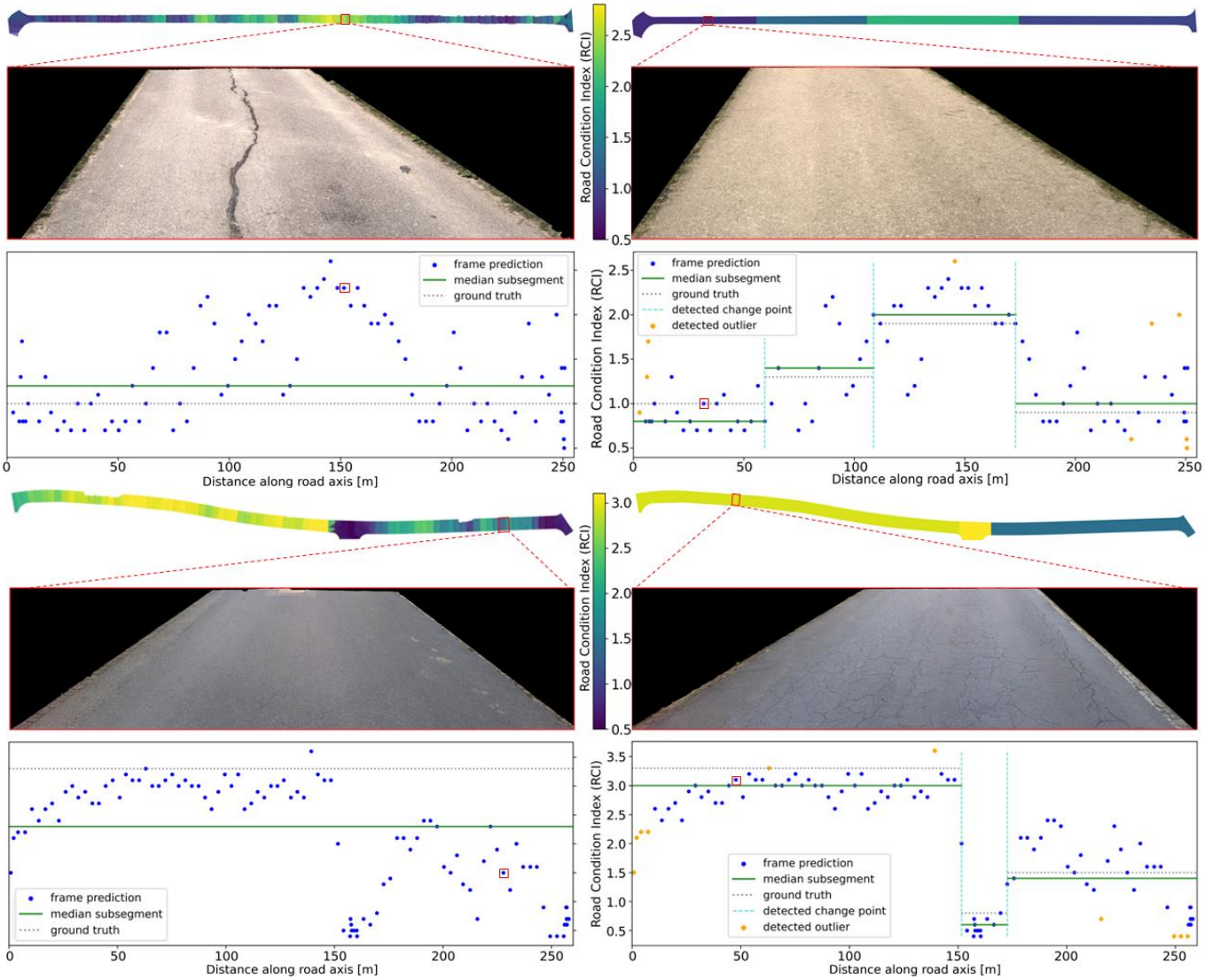


Figure 15. *Cont.*

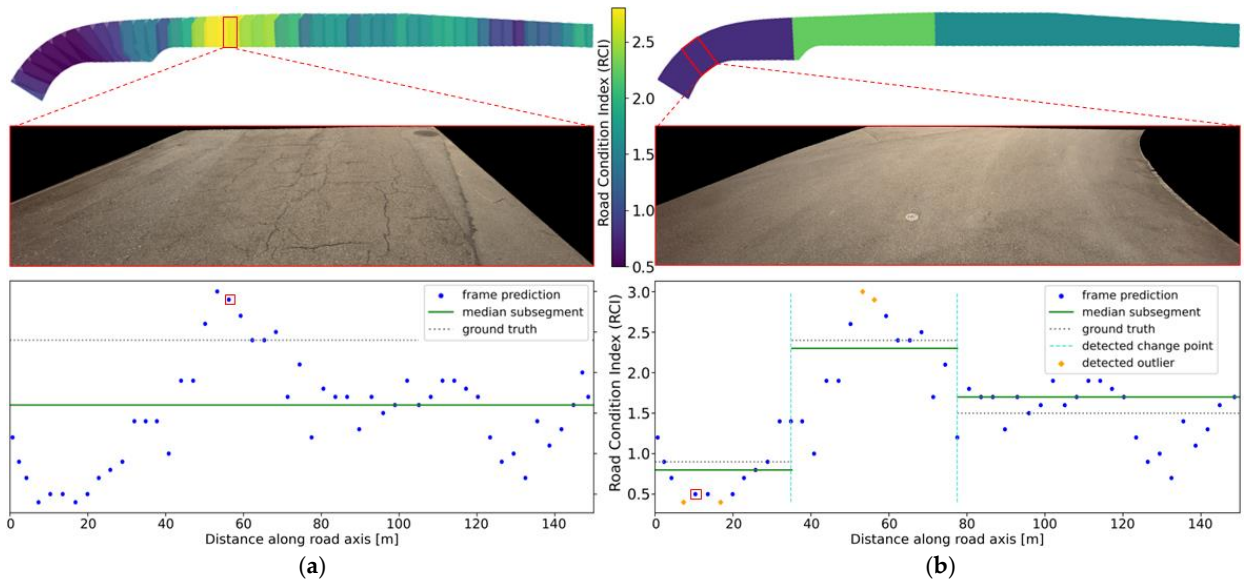


Figure 15. Qualitative examples of the generated road subsegment using the implemented aggregation algorithm as a combination of PELT and RBF. The original segments' predictions at frame level (**a**) are aggregated into subsegments (**b**) improving the accuracy of the evaluation. The graphs show the distribution of the predictions and detected outliers along the road axis as well as the calculated RCI median value of the segment (**a**) and subsegments (**b**). The sampled road-surface images in between show how different the pavement conditions are in specific frames (highlighted in red).

The dataset was refined in collaboration with experts through two iterative cycles, during which corrected RCI labels and newly generated segment subdivisions were integrated. This refinement led to a substantial improvement in model performance, reducing the MAE at frame level on the test dataset from 0.57 to 0.48 RCI values. The quantitative results presented in Figure 16 highlight a marked improvement in prediction accuracy for the first two RCI classes, while maintaining similar MAE values at frame level for the third and fourth classes. A slight decrease in performance was observed for the highest RCI class, likely due to its limited representation in the dataset. Overall, the proportion of frames predicted within the empirically defined half-class scattering range increased by more than 4%, while the proportion of frames with errors exceeding 0.5 RCI values decreased accordingly.

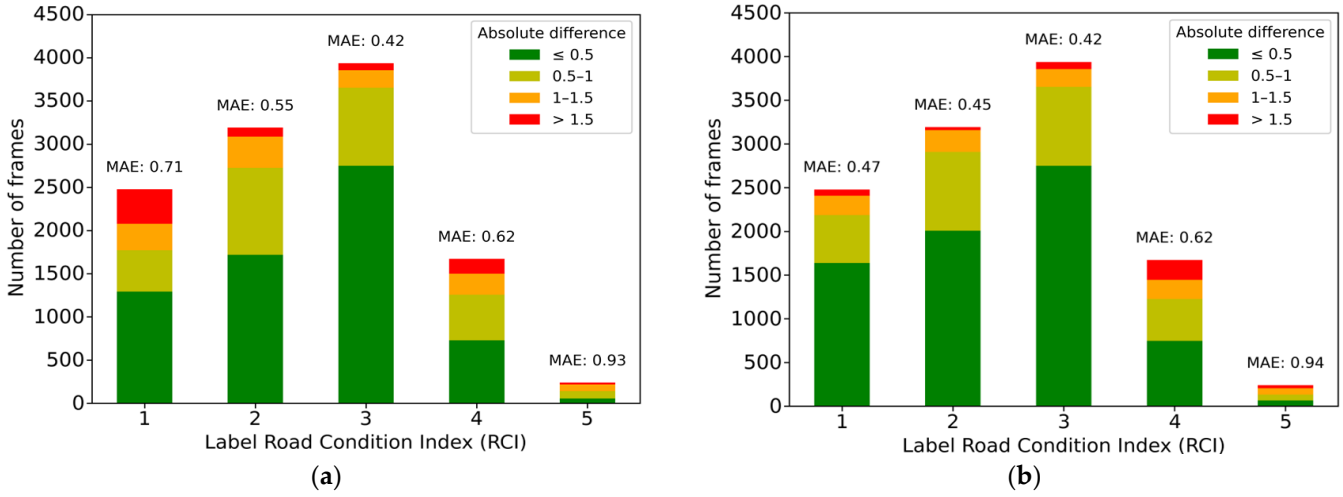


Figure 16. *Cont.*

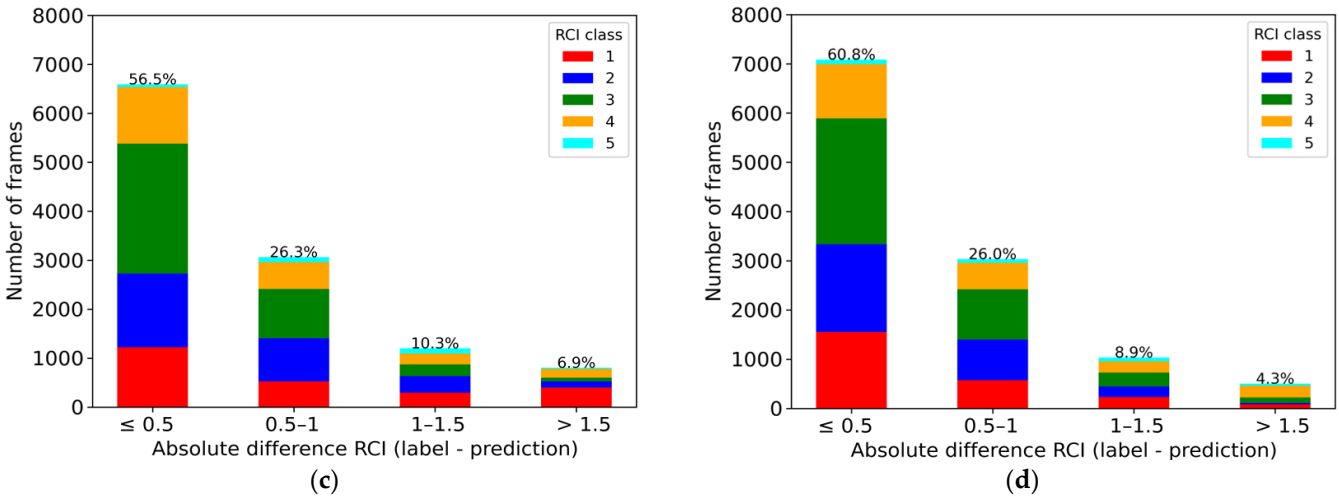


Figure 16. Comparison of MAE on the test dataset at frame level before (**a**,**c**) and after (**b**,**d**) dataset refinement. The subdivisions by RCI class (**a**,**b**) show the MAE for each class subdivided by error range while the subdivisions by error range (**c**,**d**) show the percentage of frames as well as the subdivision by RCI class. The MAE intervals represented are right-inclusive (≤0.5; (0.5, 1]; (1, 1.5]; >1.5).

After aggregating the final frame-based predictions into the segments' RCI values, further improvements were achieved through the application of the median as a robust estimator for outlier filtering. As a result, an overall MAE of 0.40 RCI values at segment level (95% CI: 0.366–0.444) was reached on the test dataset. In addition, 73.2% of the segments fell within the empirically defined half-class scattering range, while over 94% of the segments were within one full RCI class of error (Figure 17).

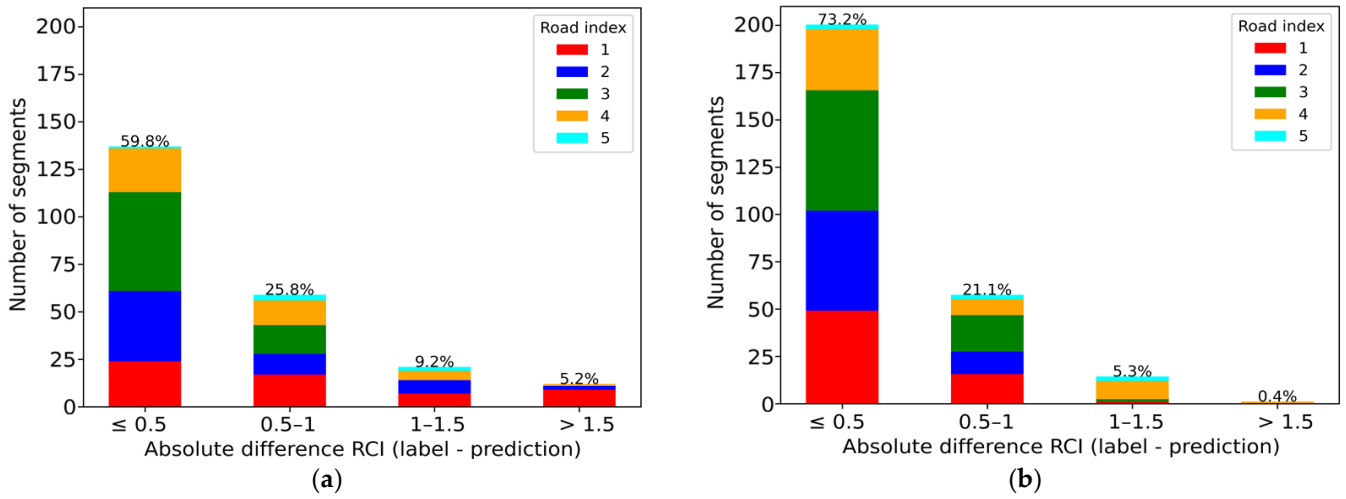


Figure 17. Comparison of MAE on the test dataset at segment level, showing results after prediction aggregation into segments (**a**) before and (**b**) after dataset refinement. The MAE intervals represented are right-inclusive (≤0.5; (0.5, 1]; (1, 1.5]; >1.5).

5. Discussion

The results of this study demonstrate the potential of neural networks combined with professional, precise georeferenced mobile mapping data for automated road-condition assessment. A key strength of our workflow lies in the sophisticated preprocessing pipeline, which ensures that only relevant pixels are presented to the neural network. The approach integrates cadastral land-cover data for automated road-surface extraction and

consequently filters the pavement information by removing interfering objects. However, this dependency on external cadastral data introduces limitations. These data may lack the accuracy required for reliable road-surface extraction or may be not always available or may be outdated, affecting transferability to other countries. Reducing this dependency is critical, and alternative methods, such as direct road-surface extraction through semantic segmentation of the imagery, should be investigated.

The availability of high-quality mobile mapping imagery and expert-annotated RCI values for each road segment is one of the strengths of our work. This enabled the creation of a dataset annotated according to established standards and represents a strong foundation for model training and evaluation. However, the preprocessing pipeline, as well as the neural network architectures, operate at the frame level, whereas the RCI labels are available at segment level. This structural difference introduces an inherent source of implicit label noise, as individual frames within a segment are assigned a uniform RCI value, although the distribution of pavement conditions within each segment is not explicitly known. To mitigate this effect, the iterative expert refinement process helped identify and correct segments exhibiting high internal variability. Despite this measure, the absence of explicit frame-level annotations remains a limitation. Future work will therefore focus on deriving labels at frame level. In addition, although the dataset includes diverse road and environmental conditions across multiple municipalities, all data were acquired using the same mobile mapping system. Consequently, the robustness of the proposed workflow with respect to imagery obtained from different acquisition units or sensor configurations could not be evaluated within the scope of this study. Cross-system validation using independently acquired and annotated datasets represents an important direction for future research to further assess generalization performance and practical transferability.

Moreover, the dataset is affected by a strong class imbalance, particularly for the fifth RCI class (RCI > 4). This class is severely underrepresented due to the practical reality that roads in critical condition are typically repaired quickly and are thus significantly less frequently available for data collection. Despite their scarcity, these classes have a disproportionate impact on model performance. The regression task showed a higher MAE in these ranges, highlighting the challenge of generalizing across visually diverse and structurally severe damages with limited training examples. Although the binary classification model was less affected by this imbalance and data augmentation partially mitigated this issue, further solutions must be investigated. Promising strategies include the use of class-aware loss functions, targeted data collection and the integration of complementary indicators like transverse unevenness derived from MMS LiDAR data, which is often associated with severely distressed pavements but not visible in RGB imagery.

The choice of convolutional architectures and transfer learning from ImageNet accelerated convergence and yielded lower MAE values for the implemented tasks. The regression model achieved an MAE of 0.48 RCI values, which is below the inter-expert variability of 0.5 RCI values. Similarly, the binary classification variant achieved an F1-score of 0.85, effectively distinguishing between roads requiring maintenance and those in good condition and thereby offering a practical tool for rapid prioritization in maintenance planning. However, each architecture implementation presents trade-offs. On the one hand, the regression approach provides continuous RCI output but, as mentioned before, is more sensitive to label noise at frame level. On the other hand, the binary classification model exhibited sensitivity to segments with RCI values near the RCI threshold of 2.5. This sensitivity results in unstable predictions for RCI values close to the threshold. Despite having little practical impact on maintenance decisions in real-world scenarios, this suggests that binary outputs are less informative in borderline cases.

The frame-to-segment aggregation algorithm significantly improved prediction results and their interpretability. The combined use of IF for outlier removal and PELT-based change-point detection enabled automatic aggregation into homogeneous subsegments that align well with expert assessments. The implemented approach further improved the obtained MAE at frame level, reducing it from 0.48 to 0.40 at segment level. In addition, the implemented method mitigates the effects of the aforementioned distribution uncertainty within a segment, using outlier filtering and a median-based estimator. However, intra-segment variability remains an open challenge, as use of the median could lead to an underestimation of localized deteriorations. This suggests that advanced aggregation strategies, such as sequence learning models that can explicitly learn variability across frames, could further enhance segment-level consistency. Moreover, the surfaces for the aggregation still depend on the initial road-surface geometry derived from cadastral data, making it indirectly subject to the same external dependencies discussed earlier.

The iterative quality-improvement process was essential for aligning predictions with expert assessments. The inclusion of quality indicators enhanced transparency and interpretability for experts, allowing the incorporation of corrections into subsequent training cycles. This improved dataset consistency, reduced inter-expert variability and resulted in a decrease in MAE from 0.57 RCI values to 0.48 RCI values at frame level. This iterative process highlighted the importance of combining automated predictions with human refinement to maintain accuracy and robustness.

6. Conclusions and Outlook

This study presented a DL-based workflow for automated visual road-condition assessment using professional 3D street-level imagery. The proposed framework integrates road-surface extraction and the filtering of interfering objects, robust neural network training, automated segment aggregation and iterative dataset refinement supported by expert feedback. The resulting models achieved high accuracy, with regression predictions outperforming inter-expert variability with an MAE of 0.48 RCI values at frame level and an MAE of 0.40 RCI values at segment level, as well as binary classification reaching an F1-score of 0.85. Importantly, segment-level aggregation and iterative quality improvement significantly enhanced the reliability of the outputs, making them well-suited for practical infrastructure management.

Despite these promising results, several challenges remain. The strong class imbalance continues to limit generalization in the most relevant ranges for maintenance prioritization. Targeted data collection campaigns, the use of class-aware loss functions, as well as the integration of a transverse unevenness information cloud help mitigate this limitation. Second, future work should explore direct road-surface extraction via semantic segmentation, reducing reliance on cadastral data. Third, integrating individual damage detection could enable comparisons with patch-based methods and enhance interpretability for maintenance planning. Finally, the reliability of frame-to-segment aggregation could be improved through advanced sequence learning models that explicitly capture intra-segment variability.

In conclusion, the proposed approach represents a significant step toward scalable, standardized and cost-effective road-condition assessment. By bridging automated DL methods with advanced mobile mapping data preprocessing, it provides a pathway for robust, reproducible and decision-oriented infrastructure monitoring that can support municipalities and road authorities in maintaining safer and more sustainable transportation networks.

Author Contributions: Conceptualization, S.N. and E.F.; methodology, E.F. and J.M.; software, J.M. and E.F.; validation, E.F. and J.M.; formal analysis, E.F. and J.M.; investigation, E.F. and J.M.; data curation, E.F. and J.M.; writing—original draft preparation, E.F.; writing—review and editing, J.M. and S.N.; visualization, E.F. and J.M.; supervision, S.N.; project administration, S.N.; funding acquisition, S.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Innosuisse, grant number 104.661 IP-ENG.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from iNovitas AG and are available from Elia Ferrari with the permission of iNovitas AG.

Acknowledgments: We would like to thank our project partners Hannes Eugster and Simon Haumann, iNovitas AG, as well as Anja Herlyn, Dorothea Zuleger and Gregor Zographos, WIF Partner AG, for making this project possible and for their numerous important inputs. We would further like to acknowledge the hints and support of Denis Jordan in finalizing the neural network architecture.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Doll, C.; Van Essen, H. *Road Infrastructure Cost and Revenue in Europe*; CE Delft: Delft, The Netherlands, 2008.
2. Maurer, P. *Straßenzustandserfassung Mit Dem RoadSTAR: Messsystem und Genauigkeit/Verf*; Research Report/Arsenal Research; Österreichisches Forschungs- und Prüfzentrum Arsenal: Wien, Austria, 2002.
3. Bundesamt für Statistik. Infrastruktur und Streckenlänge. Available online: https://www.bfs.admin.ch/bfs/de/home/statistiken/mobilitaet-verkehr/verkehrsinfrastruktur-fahrzeuge/streckenlaenge.html (accessed on 27 September 2022).
4. Department for Transport Road Length Statistics (RDL). Available online: https://www.gov.uk/government/statistical-data-sets/road-length-statistics-rdl#road-length-in-miles-rdl01 (accessed on 3 February 2025).
5. Statistik Austria. *Verkehrsstatistik 2023*; Statistik Austria: Wien, Austria, 2024.
6. Statistisches Bundesamt Straßen des Überörtlichen Verkehrs. Available online: https://www.statistikportal.de/de/transport-und-verkehr/ueberoertlicher-verkehr (accessed on 3 February 2025).
7. Nebiker, S.; Cavegn, S.; Loesch, B. Cloud-Based Geospatial 3D Image Spaces—A Powerful Urban Model for the Smart City. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 2267–2291. [CrossRef]
8. Huang, Z.; Chen, W.; Al-Tabbaa, A.; Brilakis, I. NHA12D: A New Pavement Crack Dataset and a Comparison Study of Crack Detection Algorithms. *arXiv* **2022**, arXiv:2205.01198v1. [CrossRef]
9. Liu, Z.; Gu, X.; Chen, J.; Wang, D.; Chen, Y.; Wang, L. Automatic Recognition of Pavement Cracks from Combined GPR B-Scan and C-Scan Images Using Multiscale Feature Fusion Deep Neural Networks. *Autom. Constr.* **2023**, *146*, 104698. [CrossRef]
10. Yang, N.; Li, Y.; Ma, R. An Efficient Method for Detecting Asphalt Pavement Cracks and Sealed Cracks Based on a Deep Data-Driven Model. *Appl. Sci.* **2022**, *12*, 10089. [CrossRef]
11. Zhu, J.; Zhong, J.; Ma, T.; Huang, X.; Zhang, W.; Zhou, Y. Pavement Distress Detection Using Convolutional Neural Networks with Images Captured via UAV. *Autom. Constr.* **2022**, *133*, 103991. [CrossRef]
12. Arya, D.; Maeda, H.; Ghosh, S.K.; Toshniwal, D.; Sekimoto, Y. RDD2022: A Multi-National Image Dataset for Automatic Road Damage Detection. *arXiv* **2022**, arXiv:2209.08538. [CrossRef]
13. Ren, M.; Zhang, X.; Zhi, X.; Wei, Y.; Feng, Z. An Annotated Street View Image Dataset for Automated Road Damage Detection. *Sci. Data* **2024**, *11*, 407. [CrossRef]
14. Eisenbach, M.; Stricker, R.; Seichter, D.; Amende, K.; Debes, K.; Sesselmann, M.; Ebersbach, D.; Stoeckert, U.; Gross, H.-M. How to Get Pavement Distress Detection Ready for Deep Learning? A Systematic Approach. In *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*; IEEE: Piscataway, NJ, USA, 2017; pp. 2039–2047.
15. Stricker, R.; Eisenbach, M.; Sesselmann, M.; Debes, K.; Gross, H.-M. Improving Visual Road Condition Assessment by Extensive Experiments on the Extended GAPs Dataset. In *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–8.
16. *ZTV ZEB-StB*; Zusätzliche Technische Vertragsbedingungen und Richtlinien zur Zustandserfassung und -Bewertung von Straßen. Forschungsgesellschaft für Straßen- und Verkehrswesen FGSV: Cologne, Germany, 2006.
17. Maeda, H.; Kashiyama, T.; Sekimoto, Y.; Seto, T.; Omata, H. Generative Adversarial Network for Road Damage Detection. *Comput.-Aided Civ. Infrastruct. Eng.* **2021**, *36*, 47–60. [CrossRef]
18. Maeda, H.; Sekimoto, Y.; Seto, T.; Kashiyama, T.; Omata, H. Road Damage Detection Using Deep Neural Networks with Images Captured Through a Smartphone. *Comput. Aided Civ. Eng* **2018**, *33*, 1127–1141. [CrossRef]

19. Yang, F.; Zhang, L.; Yu, S.; Prokhorov, D.; Mei, X.; Ling, H. Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 1525–1535. [CrossRef]
20. Zhang, C.; Nateghinia, E.; Miranda-Moreno, L.F.; Sun, L. Pavement Distress Detection Using Convolutional Neural Network (CNN): A Case Study in Montreal, Canada. *Int. J. Transp. Sci. Technol.* **2022**, *11*, 298–309. [CrossRef]
21. *RSV 13.01.11*; Zustandsbeschreibung und Mögliche Schadensursachen von Asphalt- und Betonstraßen. Forschungsgesellschaft Straße-Schiene-Verkehr: Wien, Austria, 2009.
22. The British Standards Institution (BSI). *Road Condition Monitoring (RCM) Data. Specification; Definitive*; BSI: Singapore, 2024.
23. *VSS 40 925b*; Erhaltungsmanagement der Fahrbahnen (EMF)—Zustandserhebung und Indexbewertung. Schweizerischer Verband Strassen- & Verkehrsfachleute Schweizerischer Verband Der Strassen- und Verkehrsfachleute (VSS): Zurich, Switzerland, 2019.
24. *ASTM D6433-20*; Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys. ASTM International: West Conshohocken, PA, USA, 2024. [CrossRef]
25. Western Australian Local Government Association WALGA. *Road Visual Condition Assessment Manual*; Western Australian Local Government Association WALGA: West Leederville, Australia, 2016.
26. *ASTM E1926-08*; Standard Practice for Computing International Roughness Index of Roads from Longitudinal Profile Measurements. ASTM International: West Conshohocken, PA, USA, 2021. [CrossRef]
27. Schwarz, K.P.; Martell, H.E.; El-Sheimy, N.; Li, R.; Chapman, M.A.; Cosandier, D. VIASAT—A Mobile Highway Survey System of High Accuracy. In *Proceedings of the Vehicle Navigation and Information Systems Conference*; IEEE: Piscataway, NJ, USA, 1993; pp. 476–481.
28. Cavegn, S.; Nebiker, S.; Haala, N. A Systematic Comparison of Direct and Image-Based Georeferencing in Challenging Urban Areas. *ISPRS—Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *XLI-B1*, 529–536. [CrossRef]
29. Eugster, H.; Huber, F.; Nebiker, S.; Gisi, A. Integrated Georeferencing of Stereo Image Sequences Captured with a Stereovision Mobile Mapping System—Approaches and Practical Results. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **2012**, *XXXIX-B1*, 309–314. [CrossRef]
30. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*; AAAI Press: Portland, OR, USA, 1996; Volume 96, pp. 226–231.
31. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556. [CrossRef]
32. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2018**, arXiv:1608.06993.
33. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
34. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *arXiv* **2015**, arXiv:1512.00567. [CrossRef]
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
36. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. *arXiv* **2017**, arXiv:1611.05431. [CrossRef]
37. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
38. *VSS 40 730b*; Erhaltung von Fahrbahnen; Kopfnorm; Massnahmenkonzept. Schweizerischer Verband Strassen- & Verkehrsfachleute Schweizerischer Verband der Strassen- und Verkehrsfachleute (VSS): Zurich, Switzerland, 2019.
39. Liu, F.T.; Ting, K.M.; Zhou, Z.-H. Isolation Forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*; IEEE: Piscataway, NJ, USA, 2008; pp. 413–422.
40. Killick, R.; Fearnhead, P.; Eckley, I.A. Optimal Detection of Changepoints with a Linear Computational Cost. *J. Am. Stat. Assoc.* **2012**, *107*, 1590–1598. [CrossRef]