inventions

MDPI

Article

Road Anomaly Detection with Unknown Scenes Using DifferNet-Based Automatic Labeling Segmentation

Phuc Thanh-Thien Nguyen [1], Toan-Khoa Nguyen [1], Dai-Dong Nguyen [1], Shun-Feng Su [1] and Chung-Hsien Kuo [2,*]

[1] Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan; d10907813@mail.ntust.edu.tw (P.T.-T.N.); m10907803@mail.ntust.edu.tw (T.-K.N.); d10907809@mail.ntust.edu.tw (D.-D.N.); sfsu@mail.ntust.edu.tw (S.-F.S.)
[2] Department of Mechanical Engineering, National Taiwan University, Taipei 106, Taiwan
* Correspondence: chunghsien@ntu.edu.tw

Abstract: Obstacle avoidance is essential for the effective operation of autonomous mobile robots, enabling them to detect and navigate around obstacles in their environment. While deep learning provides significant benefits for autonomous navigation, it typically requires large, accurately labeled datasets, making the data's preparation and processing time-consuming and labor-intensive. To address this challenge, this study introduces a transfer learning (TL)-based automatic labeling segmentation (ALS) framework. This framework utilizes a pretrained attention-based network, DifferNet, to efficiently perform semantic segmentation tasks on new, unlabeled datasets. DifferNet leverages prior knowledge from the Cityscapes dataset to identify high-entropy areas as road obstacles by analyzing differences between the input and resynthesized images. The resulting road anomaly map was refined using depth information to produce a robust drivable area and map of road anomalies. Several off-the-shelf RGB-D semantic segmentation neural networks were trained using pseudo-labels generated by the ALS framework, with validation conducted on the GMRPD dataset. Experimental results demonstrated that the proposed ALS framework achieved mean precision, mean recall, and mean intersection over union (IoU) rates of 80.31%, 84.42%, and 71.99%, respectively. The ALS framework, through the use of transfer learning and the DifferNet network, offers an efficient solution for semantic segmentation of new, unlabeled datasets, underscoring its potential for improving obstacle avoidance in autonomous mobile robots.

Keywords: automated labeling; drivable area and road obstacle detection; mobile robots; semantic segmentation; transfer learning

Citation: Nguyen, P.T.-T.; Nguyen, T.-K.; Nguyen, D.-D.; Su, S.-F.; Kuo, C.-H. Road Anomaly Detection with Unknown Scenes Using DifferNet-Based Automatic Labeling Segmentation. *Inventions* **2024**, *9*, 69. https://doi.org/10.3390/inventions9040069

Academic Editor: Anastasios Doulamis

Received: 14 May 2024
Revised: 25 June 2024
Accepted: 26 June 2024
Published: 28 June 2024

1. Introduction

Autonomous mobile robots, such as wheelchairs, forklifts, and port automation devices, have been developed for independent navigation in various environments. They offer benefits for transportation and driving. Detecting small obstacles and debris is crucial for these robots to avoid damage and navigate safely. Equipped with different camera systems, such as monochrome, stereoscopic, or fisheye cameras, these robots gather detailed environmental information. Research in this field is influenced by advancements in computer vision methods [1–6] including object detection, tracking, 3D mapping, self-localization, lidar–camera fusion, and semantic segmentation.

Deep learning has introduced efficient semantic segmentation networks that detect free-driving space and road obstacles using RGB information. Recent works have shown that incorporating both RGB and depth information improves the segmentation network's performance. However, deep neural networks often require large-scale labeled datasets, which can be time-consuming and labor-intensive to acquire and label.

To address these challenges, the development of a practical automatic labeling approach for accurately generating labels for drivable areas and road obstacles in road segmentation tasks is essential and urgent. Two common techniques that can help overcome this problem are transfer learning (TL) and self-supervised learning (SSL) [7]. TL involves training on a source dataset and transferring the learned knowledge to the target task, while SSL constructs output labels from input images without human-provided labels as in TL. Both TL and SSL have their advantages and disadvantages. SSL is unbiased towards the source task's labels but may have less discriminative learned representations compared with TL. Additionally, SSL outperforms TL when there is a large domain gap between the source and target tasks, while TL is more effective when the domain gap is small. Therefore, an appropriate TL or SSL-based approach for automatic labeling and real-time segmentation could be an ideal system for deployment on current mobile robots, revolutionizing intelligent autonomous robots.

Additionally, the uncertainty in deep learning [8] has become a popular methodology for various purposes. Incorporating the target input as out-of-distribution (OOD) data can enhance the performance of a trained model across various domains. A desirable behavior of a model is to provide a prediction for unknown data that extrapolates far from the observed data. Therefore, we aimed for our trained model to exhibit a high level of uncertainty (low precision or overconfidence in the predictions) for the corresponding inputs. In practical applications, the amount of collected unlabeled road data is huge, with various kinds of road obstacles, leading to huge uncertainty in the model.

To provide a robust detection system for road anomalies, this article proposes an automatic labeling system (ALS) for robust detection of road anomalies, using transfer learning (TL) to generate pseudo-labels for drivable areas and road obstacles. The ALS learns out-of-distribution (OOD) knowledge from the Cityscapes [9] dataset and transfers it to generate labels for new, unseen road scenarios. It uses pretrained semantic segmentation networks and a squeeze-and-excitation (SE) attention-based segmentation network (DifferNet) to estimate maps of road anomalies. The DifferNet is trained using Cityscape [9] with OOD road anomaly classes. The generated pseudo-labels are then used to train off-the-shelf RGB-D semantic segmentation networks for real-time semantic segmentation.

In summary, the contributions of our work are as follows.

1. The framework of TL-based automatic labeling system (ALS) is proposed as an automatic labeling system for unknown road scenes. It takes advantage of the uncertainty in deep learning for training the OOD-based model (DifferNet) to detect road anomalies. The proposed ALS framework could produce a high-quality pseudo-label map within approximately 2 s and can be used for off-the-shelf RGB-D semantic segmentation networks.
2. We propose a new DifferNet module which contains squeeze-and-excitation (SE) blocks based on some common concepts, cooperates with RGB-D input data and uncertainty maps calculated from SoftMax probabilities of semantic maps and the perceptual different between the RGB input image and the resynthesized image. By redefining the OOD labels in the Cityscape dataset, our pretrained model could improve the accuracy of detecting road anomalies without external OOD datasets.
3. The performance of the proposed ALS framework and DifferNet approaches achieved highly accurate and reliable labels for unseen driving scenes compared with the existing methods in terms of comprehensive experiments. We also evaluated our automatic labeling system (ALS) in real-time for the robustness of its capability under semantic segmentation networks for road anomalies.

In general, our proposed automatic labeling approach could use the advantage of the uncertainty in the learning model to extract a reliable pseudo-label map for detecting road anomalies in unknown scenes. The proposed approach is also suitable for several industrial applications, e.g., detection of defects, active learning, etc. Moreover, our ALS framework could be built under the MLOPs framework for a data-centric approach to solve the problem of being time-consuming and labor-intensive.

The remainder of this article is organized as follows. The related works are introduced in Section 2. Section 3 presents the proposed methodology. The experimental setup and the results of our proposed methodology are presented in Section 4. Finally, the conclusions are presented in Section 5.

2. Related Works

2.1. Detection of Drivable Areas

Drivable area segmentation is a crucial component of advanced driving systems (ADAS) and autonomous driving, enhancing the certainty and safety of intelligent vehicle systems. Early approaches [10–12] treated the detection of drivable areas as a diagonal straight-line detection problem. Subsequent methods improved the robustness and accuracy of V-disparity map-based extraction for specific applications, relying on clean depth images for high accuracy. Recent efforts have explored alternative approaches for the detection of drivable areas. Mayr et al. [13] introduced an automatic labeling method for ground plane recognition using deterministic stereo-based techniques, eliminating the need for manual annotation. Ma et al. [14] proposed a novel approach that leveraged LiDAR-based depth information to create large datasets suitable for training DNNs. Other works have focused on improving learning approaches and network architectures. Han et al. [15] developed a semi-supervised learning technique using generative adversarial networks (GAN) for the recognition of roads. Ma et al. [16] used an unsupervised deep learning-based monocular depth estimation method to obtain the stereo disparity map. Then a non-parametric, refined U-V disparity mapping method to extract the road's region of interest. Ali et al. [17] considered roads with deteriorating conditions using 3D LiDAR, presenting a method for segmentation of the drivable area that split smaller point cloud objects based on the number of laser scans and their projection angle, followed by multiple filtration steps to detect the road's boundaries and classify the road's irregularities. Jiang et al. [18] incorporated both salient areas and attention mechanisms into the detection process, inspired by human vision science. The authors formed a triangular area of the road surface using two boundary nodes on the road's edge by computing an attention point that merged salient areas and the region of interest. Asgarian et al. [19] introduced a row-selection approach that achieved faster performance and reduced the computational costs compared with previous methods. By leveraging advanced deep learning techniques, the authors significantly improved the accuracy and efficiency of this critical task, making autonomous driving more reliable and effective.

2.2. Detection of Road Obstacles

Recent advancements in appearance-based obstacle avoidance schemes have led to the development of several novel methods. Rabiee et al. [20] proposed an approach to compare plans generated by vision and the supervisory sensor for detecting failures in stereo-vision-based perception. In contrast, Gosh et al. [21] proposed an obstacle detection method using stereovision-based approaches. However, the disparity maps generated from stereo images struggle in complex environments, and obtaining high-quality stereo images under varying light conditions is challenging.

Deep learning has been leveraged in many works to segment road anomalies, but RGB-based semantic segmentation networks struggle to differentiate between actual obstacles and changes in appearance. To address this, some methods have suggested using depth maps as additional information. Wang et al. [22] proposed a framework utilizing RGB-D data to detect road obstacles based on differences in the appearance in RGB cues, extracting obstacles from a V-disparity map. However, this method faced challenges in scenarios with low-quality depth images and low contrast between the road obstacles and the background. Additionally, many works considered obstacle detection as a type of uncertainty estimation and an open-set semantic segmentation problem. The uncertainty, interpreted as a pixel-wise anomaly, helps score the detected obstacles on roads. Some methods used simple statistics to estimate the uncertainty from the predicted SoftMax dis-

tribution [23]. Oberdiek et al. [24] suggested using dispersion measures to detect anomalies based on differences in the visual features. Di Biase et al. [25] utilized image resynthesis to combine the reconstruction errors of two uncertainty maps within the segmentation network. Lis et al. [26] also used image resynthesis, using perceptual differences as the loss of reconstruction between the input and the resynthesized images. Liao et al. [27] introduced a contrastive learning approach to handle the covariate shift between the source and target domains in the detection of road anomalies. Lis et al. [28] proposed a GAN-based inpainting method to erase obstacles while preserving the road's texture, generating more realistic resynthesized images in the background regions. However, this approach was complex and computationally intensive, limiting its deployment in real-time applications. Tian et al. [29] enhanced the segmentation of anomalies in urban driving environments by combining pixel-wise abstention learning (AL) with an energy-based model (EBM). The AL directly learned an anomaly class at the pixel level, allowing it to abstain from labeling pixels that did not resemble any predefined inlier classes. Moreover, the energy-based model (EBM) learned the distribution of the inlier pixels and assigned high-energy values to anomalous pixels detected through exposure to outliers. Lis et al. [30] proposed the generation of a scale map, which encoded the apparent size of a hypothetical object at each image's location. Then the perspective information was incorporated into the training process by adding synthetic objects to images of the road. Nayal et al. [31] introduced a method that included region-level classification and a novel outlier scoring function called RbA. This function defined an outlier as any object that was rejected by all known classes. The RbA method enhanced the ability to detect unknown objects by focusing on regions that did not fit into any of the predefined categories. Unlike traditional methods that treat the detection of anomalies as a per-pixel classification problem, that study presented a shift in focus towards the classification of mask. Rai et al. [32] proposed a new method, called Mask2Anomaly, which incorporated a mechanism of detecting anomalies within a mask-classification architecture. This approach offered several key innovations, including a global masked attention module, mask-contrastive learning, and mask refinement techniques. These innovations collectively enhanced the system's ability to accurately identify anomalies while reducing false positives. Katsamenis et al. [33] combined a U-Net structure with recurrent residual and attention modules for automated segmentation of road cracks. During training, the retraining strategy dynamically fine-tuned the U-Net's weights as new rectified samples that were fed into the classifier. Wan et al. [34] tackled the issue of noise interference in the detection of road defects. The authors introduced a novel anti-noise dual-branch network (ADNet), leveraging two backbone networks equipped with dual-branch interaction modules. Li et al. [35] addressed the challenge of detecting surface defects, common road defects that significantly impact traffic's efficiency and safety. They proposed a lightweight object detection network called LHA-Net, which integrated feature-aware modules to ensure both accuracy and real-time performance in detecting surface defects.

3. Proposed Method

The main idea behind our TL-based auto-labeling approach was inspired by the self-supervised RGB-D approach to the segmentation of road anomalies [22] and utilized the benefits of estimations of uncertainty from uncertainty estimation maps. The key module of ALS was developed on the basis of Synboost [25] but also takes the RGB-D images as inputs to extract from high- to low-level feature maps and leverages the uncertainty of segmentation to identify anomalous regions of the drivable area with large variations in the appearance (as opposed to the drivable region). The overview of the proposed approach is illustrated in Figure 1.

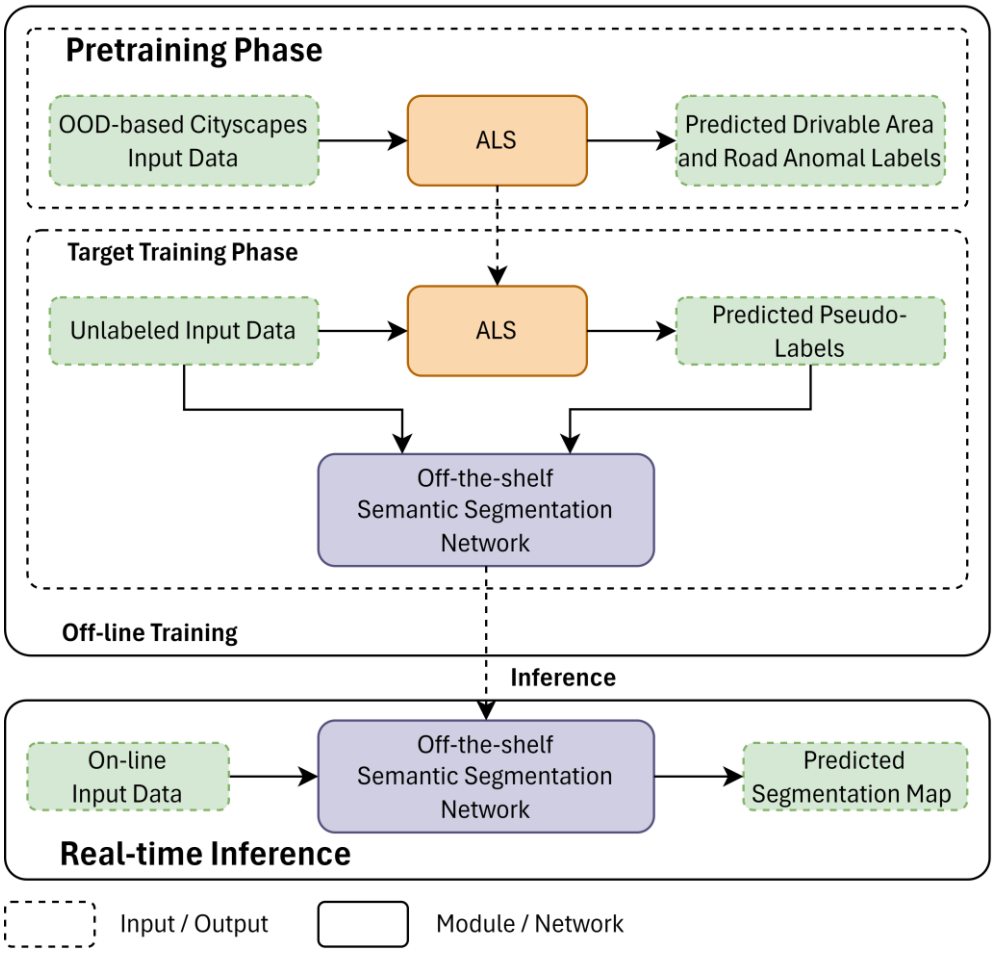

Figure 1. The overview of the proposed transfer learning-based automatic labeling system (ALS) for segmentation of the drivable area and road obstacles.

The proposed transfer learning (TL)-based ALS consists of four sub-modules as follows.

1. **The pretrained semantic segmentation module** was trained on the Cityscapes RGB dataset [9], and served as prior knowledge to compute the maps of the dispersion scores.
2. **The pretrained synthesizing module** was also trained on Cityscapes [9] and was used to compute maps of the dispersion scores.
3. **The discrepancy network** utilized maps of prior knowledge to predict uncertainty in the drivable area.
4. **The postprocessor** was used to estimate the map of pseudo-labels.

The ALS system outputted the final label map, which was then used to train the off-the-shelf semantic segmentation model for unseen road scenes. The data flow of the ALS is illustrated in Figure 2.

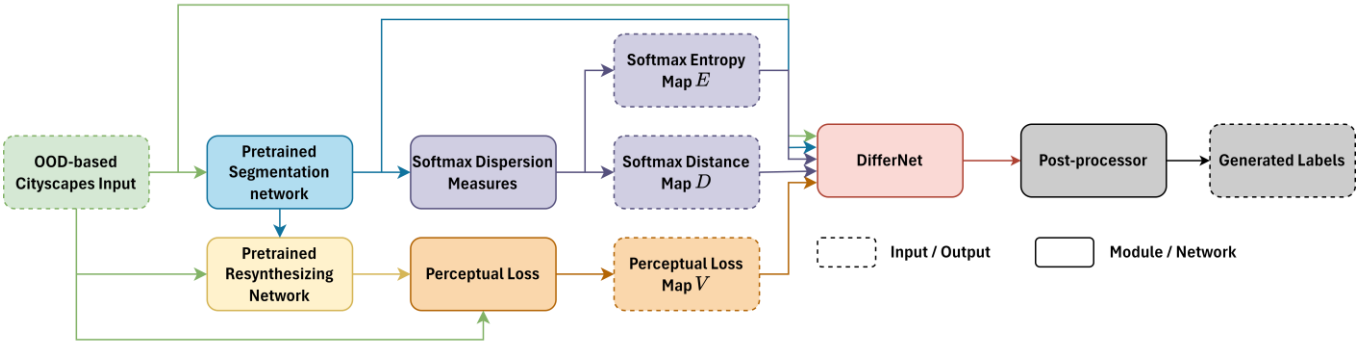

Figure 2. Schematic of the automatic labeling system (ALS) for segmentation of the drivable areas and road anomalies.

3.1. Pretrained Semantic Segmentation Module

This module took an RGB image as input and passed it through the Cityscapes-based pretrained network ICNet [36] to generate a semantic segmentation map. When dealing with road anomalies in the new road dataset, the predicted semantic map had ambiguity around the unknown objects. To have a better understanding of the errors within the segmentation, we computed two pixel-wise dispersion measures, including the SoftMax entropy E [24] and the difference in probability of the difference between the two most significant SoftMax values, known as the SoftMax distance D [37]. A pretrained segmentation network with a SoftMax output layer can be viewed as a statistical model that produces a probability distribution $f_i\ (y|x, w)$ on q class labels y with $y \in C; C = \{y_1, \ldots, y_q\}$, given the weights w and the data x for each pixel i in the input image. The predicted class $\hat{y}_i$ at the pixel i is then given by

$$\hat{y}_i(x,\ w) = \underset{y \in C}{\operatorname{argmax}} f_i(y|x,\ w) \tag{1}$$

The SoftMax dispersion measures of the semantic map for pixel i were determined as follows:

$$E_i(x,\ w) = -\sum_{y \in \mathbb{C}} f_i(y|x,\ w) \log f_i(y|x,\ w) \tag{2}$$

$$D_i(x,\ w) = 1 - f_i(\hat{y}_i(x,\ w)|x, w) + \max_{y \in C \setminus \{\hat{y}_i\}} f_i(y|x,\ w) \tag{3}$$

The SoftMax entropy E is also known as the Shannon information entropy, which computes the probability of unlikely information over all the labels in each pixel, while the SoftMax distance D computes the difference between the two most significant SoftMax values. These SoftMax dispersion measures were then normalized to a range of [0, 1]. The metrics were both used to measure the dispersion or concentration of the randomness of the predicted f in each pixel. The predictive entropy reached its maximum value when all classes were predicted to have an equal uniform probability and its minimum value of 0 when one class had the probability of 1 and all others had a probability of 0, i.e., the prediction was confident. From the observation in Figure 3, the SoftMax distance D emphasized the uncertainty in the edge or corner regions, while the SoftMax entropy E provided the uncertainty in a specific region. Therefore, these feature maps were used to enrich the uncertainty information for training the DifferNet.

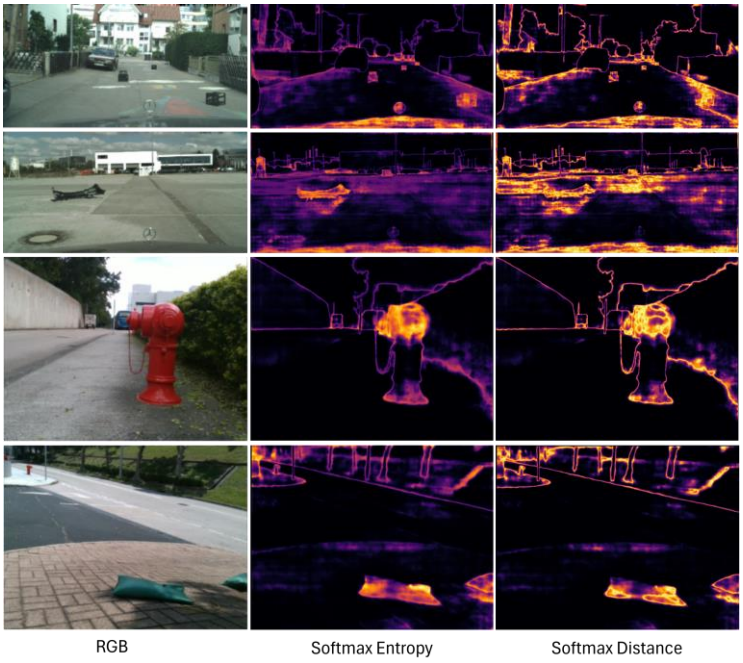

Figure 3. Visualization of the SoftMax dispersion maps, showing the RGB input images and the corresponding SoftMax entropy and SoftMax distance maps (a higher uncertainty score leads to brighter colors).

3.2. Pretrained Synthesizing Module

Similar to Synboost [25], this module used a pretrained Cityscapes-based conditional generative adversarial network [38] to generate a realistic resynthesized image with pixel-to-pixel correspondence from the semantic segmentation image. The pixel-to-pixel correspondence could be recognized by comparing the input to the reconstructed output to maintain the appearance in normal regions while excluding road anomalies.

Since road anomalies were not represented in the semantic classes used for training, their essential information, such as color or appearance, was altered by the resynthesized process. Inspired by the perceptual loss [39], we considered this novel feature map as the feature's score of the difference measure. First, we used the pretrained VGG16 to extract feature maps for identifying the distinct appearance. The distinction between these representations was beneficial for comparing the image's content and spatial organization rather than low-level color and texture-based attributes. The resynthesized image was generated with an inaccurate feature representation if the anomalous object was not identified or categorized correctly, so the perceptual loss calculator was capable of detecting the inconsistencies between the RGB input image I and the resynthesized image R

$$V(\mathrm{I},\ \mathrm{R}) = \sum_{j=1}^{N} \frac{1}{M_j} \left\| F^j(\mathrm{I}) - F^j(\mathrm{R}) \right\|_1 \tag{4}$$

where N is the number of VGG16 layers used to extract the feature maps and F^j represents the jth layer with M_j elements of the VGG network.

Furthermore, we computed the perceptual loss V_j between the output of the jth hidden layer and the resynthesized image using the L1-norm. The perceptual loss map was normalized to the interval of [0, 1] for consistency. Regarding the meaning of the perceptual loss in our system, we tried to leverage the information on the differences in the features between the RGB input image and its reconstructed image, aiming to enhance the capacity of DifferNet for identifying distinct appearances and for detecting out-of-distribution objects (i.e., road anomalies).

3.3. DifferNet Module

The DifferNet module has three main components: the encoder module, the attention fusion module (AFM), and the decoder module, as shown in Figure 4. The purpose of DifferNet is to leverage the feature maps extracted from the previous steps to estimate the uncertainty in the drivable surface.

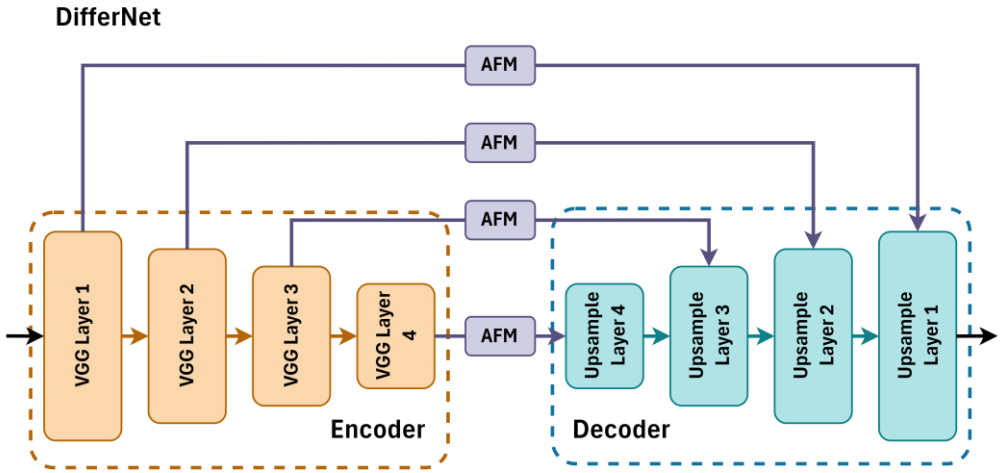

Figure 4. Schematic of the DifferNet Module.

3.3.1. Encoder Module

We use a pretrained VGG network [40] to extract the feature maps separately from the RGB, depth, and resynthesized images. Furthermore, a four-layer CNN network was utilized for encoding the semantic map and a stack of uncertainty maps (SoftMax entropy, SoftMax distance, and perceptual loss). Note that the output in each layer of the designed CNN and VGG network was ensured to be the same size.

3.3.2. Attention Fusion Module (AFM)

This module aimed to guide the network to pay more attention to the high-uncertainty regions in the feature maps. As described in Figure 5, at each level of the feature pyramid, we separated the features into two fusion branches, as described below.

In the first branch, we aimed to extract the differences between the input feature maps. We aimed to extract the differences between the input feature maps in the first branch. First, we focused on the RGB, resynthesized, and semantic feature maps, then applied a 1×1 convolution to compute the difference between the original and the resynthesized features. We also concatenated the dispersion score maps, the semantic feature maps, and the input through a 1×1 convolution to emphasize the uncertainty in the drivable area. The 1×1 convolution layers were able to excavate the correlations between the channels of the input feature maps. Lastly, the feature maps obtained from the two previous steps were further fused at each scale using element-wise multiplication. The output map of the first branch was denoted as S1.

In the second branch, we utilized a squeeze-and-excitation (SE) block [41] as the channel attention method to extract the important salient regions in the input feature maps. The SE block used global information to emphasize the informative channels and suppress the less beneficial ones. We assumed that the input RGB feature map I_{RGB}, the depth feature map I_D, and the resynthesized feature map I_R, with I_{RGB}, I_D, $I_R \in R^{C \times H \times W}$. Finally, we took the summation results of the attention scores of the RGB, depth and resynthesized feature maps; the output feature map S2 of this branch was then expressed as

$$S2 = I_{RGB} \bigotimes \sigma[\phi(I_{RGB})] + I_D \bigotimes \sigma[\phi(I_D)] + \sum_C \left(I_{RGB} \bigotimes I_R \right) \quad (5)$$

where $\sigma[\phi(\cdot)]$ denotes the SE block, $\sum_C(\cdot)$ denotes the summation of the channel, and $\otimes$ denotes the element-wise multiplication.

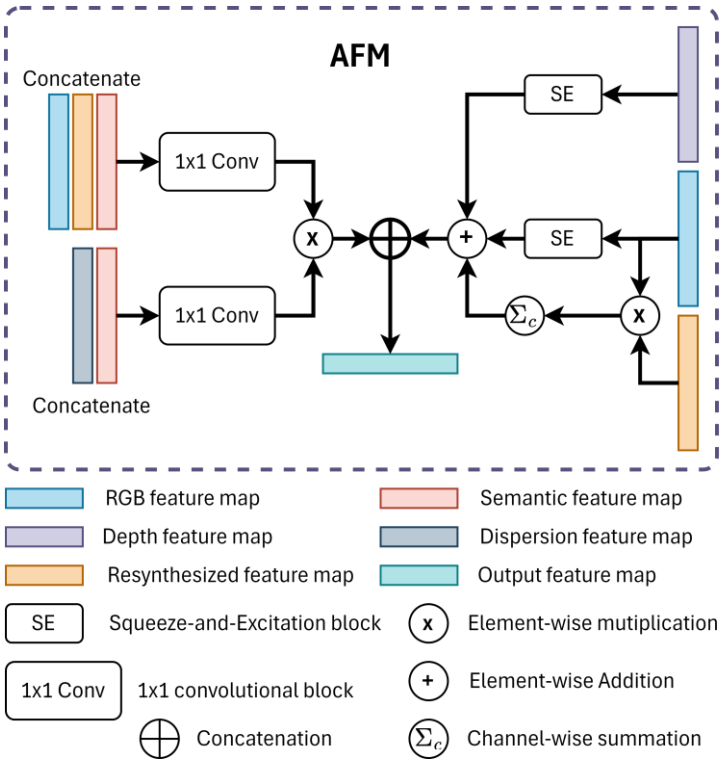

Figure 5. The architecture of the attention fusion module (AFM).

Finally, we concatenated the S1 and S2 feature maps to achieve the final feature maps of the AFM. This type of attention mechanism (an SE block) in RGB-D fusion is capable of elevating the informative features to have higher values of weight, allowing for more effective exploitation of the complementary information from depth images.

3.3.3. Decoder Module

After going through AFM, the attention semantic map was fed into the corresponding decoder block at the same feature level. This module had a decoder function to up-sample each feature map. Then we concatenated it with the matching higher level in the pyramid until the final prediction map for the segmentation of anomalies was obtained. It also included the spatially adaptive normalization (SPADE) [42] procedure, which regulates activations using an input feature map via a spatially adaptive method and can potentially transmit semantic information across the network effectively. Furthermore, this normalization was applied to preserve the semantic information from the common normalization layers because the semantic information tended to be removed during the decoding process when applied to uniform or flat segmentation tasks.

3.4. DifferNet's Training Procedure

Road anomalies are an ambiguous phenomenon, since there are a variety of objects with different sizes and shapes that could obstruct the way of a mobile robot during the process of navigation. Therefore, the detection of road anomalies can be solved as an out-of-distribution (OOD) problem. To overcome this problem, [25] used the void classes in the Cityscapes dataset as the anomaly class because the objects belonging to it were not in the segmentation classes used for training and provided high uncertainty pixels to guide the discrepancy network. Moreover, some OOD datasets have also been used to generalize unseen objects.

Our configuration aimed to produce semantic maps of road anomalies from DifferNet for autonomous mobile robots working in various scenarios. The Cityscapes dataset is an excellent prior knowledge learning source because it was captured from 50 cities with urban street scenes with finely annotated labels in 19 classes. However, our desired label map only included three main classes: road anomalies, the drivable area, and the background. The reason is that mobile robots need to detect the driving surface and avoid possible obstacles during navigation. Therefore, we carefully selected foreground classes in the Cityscape dataset as the road anomaly class, including people, riders, cars, trucks, buses, motorcycles, bicycles, poles, traffic lights, and traffic signs, both dynamic and static. Note that the map with labels of the drivable area can be obtained from the segmentation map produced by the previous segmentation module pretrained on Cityscapes (i.e., the road and sidewalk classes). Some examples of maps of road anomaly labels for DifferNet are shown in Figure 6.

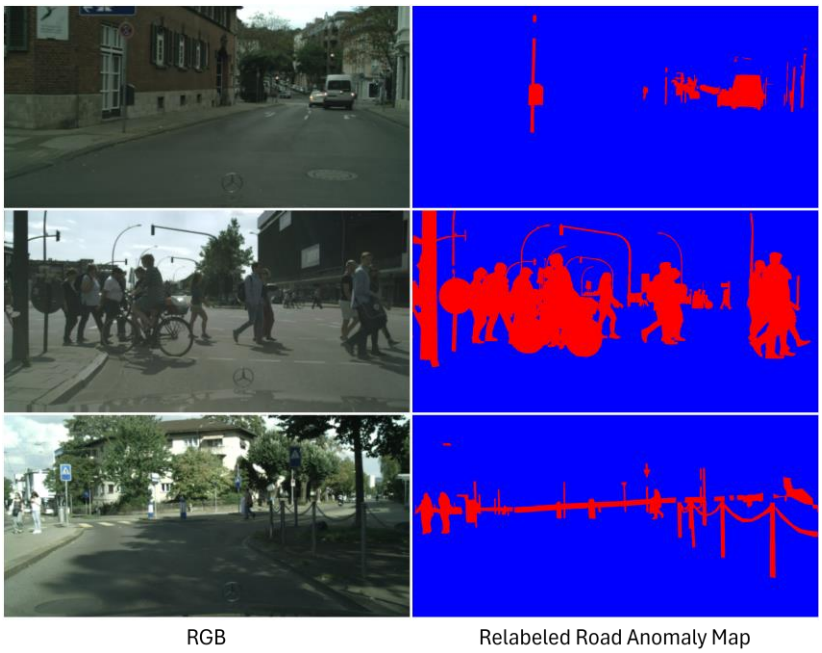

Figure 6. Examples of a map of road anomaly labels for DifferNet. The label map includes two classes: 0 (blue), non-road anomalies; 1 (red), road anomalies.

This configuration assisted our DifferNet in better generalizing the task of segmenting anomalies and enabled it to learn the context well from any extension of the OOD datasets while providing enough prior knowledge for automatic labeling in unseen driving scenarios. Furthermore, since our DifferNet is associated with depth information, we processed the disparity in the images from Cityscapes using the semi-global matching algorithm [43]. We also replicated the training samples and marked items within the empty regions that usually have the same appearance as high-uncertainty pixels. This mechanism guided the DifferNet module to effectively use the uncertainty information, especially the information from the uncertainty maps, including the SoftMax entropy, SoftMax distance, and perceptual loss. To optimize the model during the training procedure, we used binary cross-entropy loss

$$\mathcal{L}_{CE} = \frac{1}{N}\sum_{x}(-y\log(\hat{y}) - (1-y)\log(1-\hat{y})) \tag{6}$$

where N is the total pixels of the input image, and y and $\hat{y}$ are the ground truth and the predicted class in pixel x, respectively.

3.5. Postprocessor

After obtaining the anomaly map from the previous module, we then ran the morphological operators in the postprocessor to eliminate the possible wrong predictions of anomalies in the drivable area of the final pseudo-label map. We normalized the anomaly map generated from DifferNet to the range of [0, 1]. We let h and w denote the input image's height and width, respectively. We assumed that the morphological structuring elements were square. Then the closing operation (dilation followed by an erosion operation) with a structuring element size of $a_1 \times a_1$ was performed on the output anomaly map. Next, the postprocessor module performed another closing operation with a $a_2 \times a_2$ structuring element to the road segmentation map. The structuring element size a_i, with $i \in \{1, 2\}$, was generated by the following formula

$$a_i = f(k_i \times \min(h, w)) \tag{7}$$

where:

$$f(x) = 2 \times \left\lceil \frac{x}{2} + 1 \right\rceil - 1 \tag{8}$$

We found the closest odd integer to x with Equation (8). It was easier to define the origin as the center of the structuring element by assigning an odd value to a_i. A smaller k_i resulted in the module being capable of detecting tiny obstacles, but it possibly increased the rate of missing detection. On the other hand, a larger k_i may reject more tiny obstacles; sometimes, the not-so-small obstacles might be ignored. In our cases, the combination of k_1 and k_2 were chosen empirically to be $1/60$ and $1/48$, respectively, to achieve good results. Finally, the output anomaly map was filtered by a predefined threshold to obtain the road anomalies and combined with the output map of the drivable area. The areas excluding road anomalies and drivable areas were determined to be unknown regions. In the final pseudo-label map, the unknown areas, drivable areas, and road obstacles were indexed as 0, 1 and 2, respectively, as shown in Figure 7.

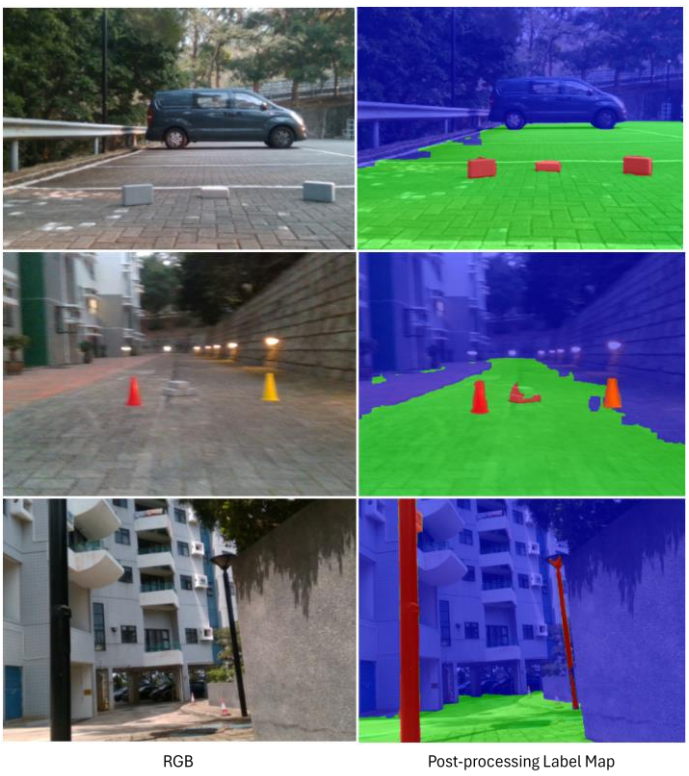

Figure 7. Examples of final label map in the postprocessing stage. The label map includes three classes: 0 (blue), unknown areas 1 (red), road anomalies; 2 (green), drivable area.

4. Experimental Results and Discussion

This section summarizes the experimental results of our TL-based automatic labeling method for segmentation of drivable areas and road obstacle. Practically, we used the DifferNet module to output the segmentation map of anomalies and utilized one of the state-of-the-art semantic segmentation networks to find the pixels of the drivable area. On the basis of this idea, we first explain the training procedure and implementation details of DifferNet. Then we describe DifferNet's experimental setup combined with evaluation in terms of two public datasets [22,44]. We evaluated our proposed method on the task of segmenting the drivable area and road obstacles.

4.1. Datasets

In this section, we evaluated the DifferNet module for the task of detecting road anomalies with two RGB-D semantic segmentation datasets containing road obstacles: Lost and Found [44] and GMRPD [22].

The Lost and Found dataset contains annotated frames (2014) from stereo video sequences (112), and coarse annotations of free-space areas and fine-grained annotations of small road hazards are also included. The test set included 1203 photos with a resolution of 2048 $\times$ 1024. The dataset consisted of various small obstacles existing at a long distance with non-uniform road textures or appearances, and roads with several items in the non-obstacle class functioning as distractors. The depth images from Lost and Found were generated using the semi-global matching algorithm [43].

The GMRPD dataset included 3896 RGB-D images with a resolution of 1280 $\times$ 720 which covered 34 common scenes where robotic wheelchairs usually work and 18 distinct kinds of road obstacles that robots may encounter in their working environments. The depth images from GMRPD were collected using the Intel Realsense D415 RGB-D camera, and the pixels with distances larger than 10 m were removed and labeled with zero values. The testing set was constructed from 4 of 34 common scenes, containing 571 images from different scenes and anomalies from the training and validation sets.

4.2. Training Details

We chose the Adam optimizer for the process of training the pretrained DifferNet. We set 50 epochs with a weight decay of 10^{-4}, a batch size of 4 during training, an initial learning rate of 10^{-4}, and a duration of 10 epochs. In the process of data augmentation, flipping of the vertical axis and normalization using the mean and standard deviation from ImageNet [45] were applied. The training process was trained on a single NVIDIA Tesla V100 GPU. It is noteworthy that the only the Cityscape [9] dataset with defined OOD classes, as mentioned in Section 3.4, was used as training dataset in this phase. The Lost and Found [44] data were used for validation.

For the process of transfer learning, this study utilized a stochastic gradient descent (SGD) optimizer to train the model with a learning rate of 10^{-3} and 400 epochs. It should be noted that the resolution of the inputs of the GMRPD [22] was downscaled to 640 $\times$ 480 pixels. Practically, we used four different scenes in the GMRPD, including Scene 12, Scene 13, Scene 14, and Scene 29, with a total of 571 images for testing. In addition, we used three off-the-shelf RGB-D data-based neural networks for semantic segmentation, namely FuseNet [46], RTFNet [47] with ResNet-18 as the backbone, and Depth-aware CNN [48], to evaluate our ALS labels, SSLG labels, and manual labels to make a comparison. After generating pseudo-labels, we processed the GMRPD training set on the ALS.

For a fair benchmark comparison, validation of the detection of road anomalies by different methods was achieved on a desktop integrated with a single NVIDIA Tesla V100 GPU. Meanwhile, for the real-time detection of road anomalies, the trained model was set up on a laptop (ASUS ROG ZEPHYRUS, Taipei, Taiwan) equipped with a GeForce GTX 1070 Max-Q, 8 GB of RAM, and an OAK-D camera (Luxonis, Denver, CO, USA) to control a self-built mobile robot.

4.3. Comparison of the Results of Detecting Road Anomalies with Public Datasets

A visualization of the results of DifferNet's outputs on the Lost and Found and GMRPD datasets are illustrated in Figures 8 and 9. To evaluate the performance of the proposed DifferNet module against the existing methods, this study applied the same metrics for detecting anomalies, including the mean precision (AP) and the false positive rate at a 95% true positive rate (FPR95). We did not use the receiver operating curve (ROC) as the metric for the segmentation of anomalies because ROC is improper for highly imbalanced problems such as detecting anomalies; [49] explained the details.

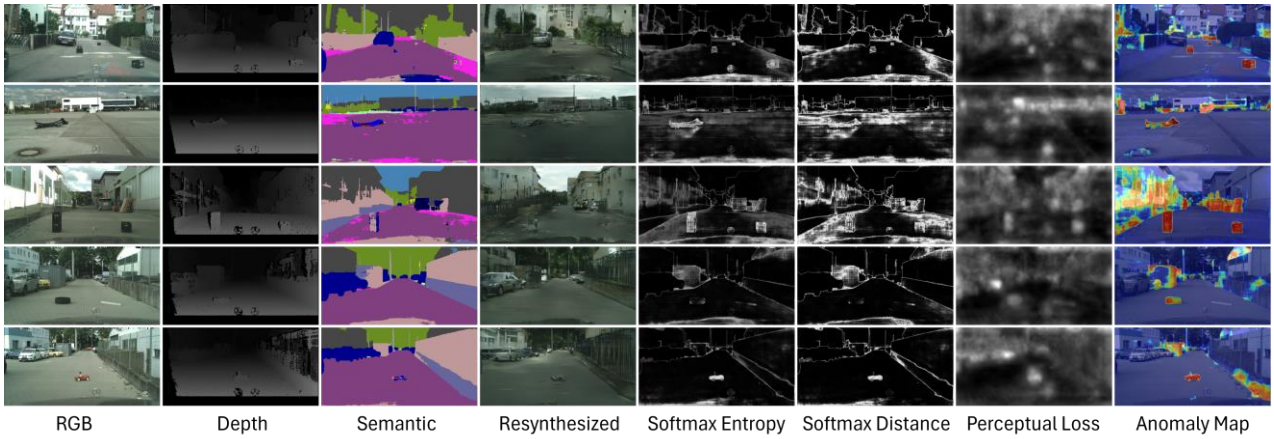

Figure 8. Examples of DifferNet's predictions of anomalies with the Lost and Found dataset [44], with RGB-D images, resynthesized images SoftMax entropy, SoftMax distance, and perceptual loss as the inputs and the anomaly map as the output.

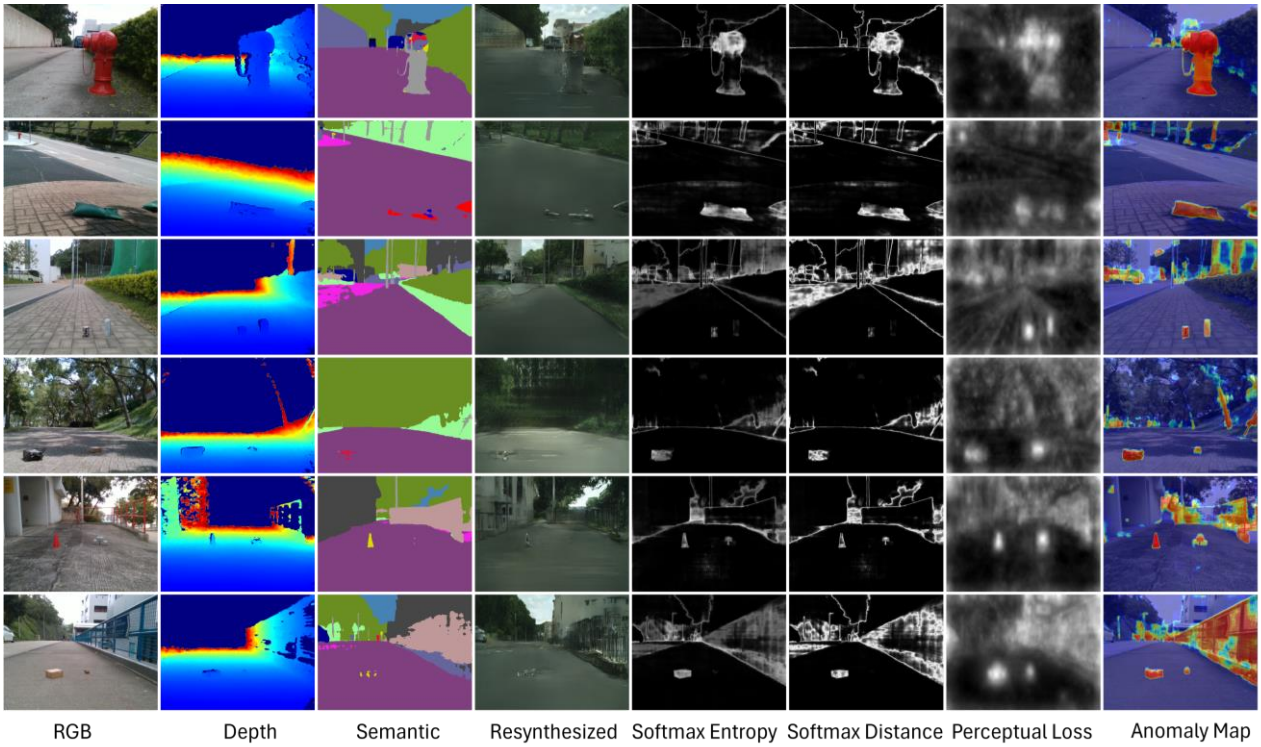

Figure 9. Examples of DifferNet's predictions of anomalies with the GMRPD dataset [22], with RGB and depth images, resynthesized values, SoftMax entropy, SoftMax distance, and perceptual loss as the inputs and the anomaly map as the output.

All evaluations were computed using the region of the road only, similar to prior works [26,28,44]. We also compared our method with Synboost because our framework was based on Synboost. We used the same pretrained segmentation and resynthesized networks to ensure a fair comparison. Specifically, we applied the semantic segmentation network in [50] and the resynthesized network trained by [38] with the Cityscapes dataset.

Table 1 shows the quantitative comparison of our DifferNet module with some state-of-the-art methods of detecting anomalies on RGB/RGB-D data. Our approach was compatible with the state of the art in terms of its performance in detecting anomalies, particularly in terms of minimizing the false positive rates across a wide range of operating points. Among the methods evaluated, Synboost [25] stood out for its high AP on both datasets, indicating its effectiveness in identifying anomalies. However, it showed a relatively high FPR95 on the GMRPD dataset, suggesting room for improvement in reducing false positives. On the other hand, PEBAL [29] achieved the lowest FPR95 on the Lost and Found dataset, demonstrating its ability to minimize false positives despite having a slightly lower AP. Mask2Anomaly [31] achieved the highest AP on the Lost and Found dataset, indicating its strong performance in identifying anomalies by taking advantage of the SOTA backbone's Vision Transformer. However, its FPR95 was high compared with our method. Our previous work, AGSL [45], utilizing RGBD data with Wider-ResNet, performed well across both datasets, balancing a high AP with a low FPR95.

Table 1. Comparison of DifferNet and state-of-the-art approaches to detecting anomalies with two standard datasets. The best results are highlighted in **bold** font.

Method	Architecture	Data	Lost and Found [44]		GMRPD [22]	
			AP ↑	FPR95 ↓	AP ↑	FPR95 ↓
Resynthesis [26]	PSPNet	RGB	61.90	46.60	-	-
JSR-Net [51]	ResNet-101	RGB	79.40	3.60	-	-
Synboost [25]	Wider-ResNet	RGB	83.51	1.39	73.37	12.50
Erasing road obstacles [28]	ResNeXt-101	RGB	82.30	68.5	-	-
PEBAL [29]	WireResnet38	RGB	78.29	**0.81**	-	-
Mask2Anomaly [32]	Vision Transformer	RGB	**86.59**	5.75	-	-
AGSL [49]	Wider-ResNet	RGBD	82.85	2.92	78.03	6.67
DifferNet w/o depth (our)	Wider-ResNet	RGB	82.38	1.10	77.02	**6.37**
DifferNet w/o SE block (our)	Wider-ResNet	RGB-D	83.00	1.21	77.72	6.49
DifferNet (our)	Wider-ResNet	RGB-D	84.19	1.10	**79.82**	6.97

Our proposed DifferNet method, especially the full model with Wider-ResNet and RGB-D data, achieved competitive results. It demonstrated high APs of 84.19% and 79.82% and low FPR95 values of 1.10 and 6.97, respectively, on both datasets. Furthermore, DifferNet improved the mean accuracy by 3.57% and reduced the mean FPR95 by 2.92% compared with Synboost [25]. In particular, the comparison revealed that the accuracy and false-positive rates obtained by using the DifferNet module without using depth images as input were 1.26% and 3.22% higher than those of Synboost. We can attribute this better performance to the fact that our method explicitly optimizes the approach for detecting anomalies based on an efficient fusion module in investigating the correlation of the input feature maps with the attention mechanism as an SE block.

We also provide the results of analyzing the contribution of the individual components in the proposed DifferNet. We first found that adding the depth image and SE block significantly improved the framework. The improvement due to the input data confirmed the potential of the depth information and this useful method for adding depth information to the fusion block of the framework. Further improvements due to the SE block elevated the complementary information from depth images more efficiently by acting as an attention mechanism in the process of RGB-D fusion to emphasize the informative channels. Finally, the performance of removing the depth information and the SE block was clearly shown, reducing the precision by 2.31% and 1.65%, respectively.

4.4. Results of ALS Segmentation

We evaluated the segmentation of drivable areas and road obstacles using our validation set constructed from the GMRPD dataset. We also compared our method with SSLG [22] because this approach first evaluated the GMRPD. Initially, we reimplemented SSLG for further comparison, and the Hough transform dealt with the exception errors from the line detection. Then we used our version of SSLG to evaluate the constructed evaluation set for an evaluation of the segmentation under multiple RGB semantic segmentation networks because the baseline method did not publicize the testing set.

Comparisons of the results of segmentation for indoor and outdoor scenarios are shown in Figures 10 and 11. In these figures, our proposed method outperformed the re-implemented SSLG method with the same datasets. Furthermore, our experimental results on the GMRPD dataset showed that our approach efficiently found the drivable area. Moreover, the proposed method could detect road obstacles beyond the manual labels. The performance of the task of segmenting the drivable area and road obstacles was evaluated quantitatively on the basis of the performance measures of precision (Pre), recall (Rec), and intersection-over-union (IoU) for three classes, including unknown areas, drivable areas, and road obstacles (road anomalies), as shown in Table 2. The "Overall" column represents the mean precision, recall, and IoU of all three segmentation classes. In the first part, we evaluated the performance of the labels generated from the GMRPD datasets. Then our ALS approach was compared with the original SSLG method for evaluation. As a consequence, our reimplementation, SSLG++, was improved compared with the original results of SSLG presented in [22], especially for drivable areas, which led to a significant boost in the mean performance, i.e., a mean precision (Pre) of 73.29% and 66.16%, a mean recall (Rec) of 75.54% and 63.40%, and a mean intersection-over-union (IoU) of 4.82% and 52.33%, respectively.

Table 2. Comparison of the results of segmentation (%) among the ALS labels (ours), the original SSLG labels [34], the reimplemented SSLG++ labels, FSL++ (FuseNet trained on the reimplemented SSLG++ labels), FAL (FuseNet trained on the ALS labels), FML (FuseNet trained on manual labels), DSL++ (Depth-aware CNN trained on reimplemented SSLG++ labels), DAL (Depth-aware CNN trained on ALS labels), DML (Depth-aware CNN trained on manual labels), RSL++ (RTFNet trained on reimplemented SSLG++ labels), RAL (RTFNet trained on ALS labels), RML (RTFNet trained on manual labels). The best results without using manual labels are highlighted in **bold** font.

Method	Unknown Area			Drivable Area			Road Obstacles			Overall		
	Pre	Rec	IoU	Pre	Rec	IoU	Pre	Rec	IoU	Pre	Rec	IoU
SSLG [22]	89.62	80.36	75.09	75.70	86.91	65.87	33.15	22.92	16.03	66.16	63.40	52.33
SSLG++ (ours)	**97.82**	90.25	**87.80**	88.63	**96.46**	**84.42**	33.41	39.91	22.23	73.29	75.54	64.82
ALS (ours)	94.87	**90.91**	86.65	**89.44**	93.62	84.30	**56.62**	**68.72**	**45.02**	**80.31**	**84.42**	**71.99**
FSL++	99.67	87.58	87.32	83.59	**97.99**	82.17	29.26	43.90	21.30	70.84	76.49	63.60
FAL	96.99	89.90	87.47	86.04	95.30	82.54	71.64	82.44	62.15	84.89	89.21	77.39
FML	98.59	99.13	97.75	98.69	97.37	**96.13**	82.80	99.63	**82.55**	**93.36**	98.71	**92.14**
DSL++	**99.83**	85.77	85.64	81.49	97.13	79.58	26.72	50.73	21.21	69.35	77.88	62.14
DAL	98.07	92.11	90.47	88.97	96.39	86.10	58.21	79.50	50.62	81.75	89.33	75.73
DML	98.44	96.96	95.50	95.40	97.03	92.69	78.11	96.23	75.79	90.65	96.74	87.99
RSL++	99.37	98.07	**97.47**	97.64	97.51	95.26	54.04	84.78	49.26	83.68	93.45	80.66
RAL	95.34	**98.97**	94.40	**98.84**	93.64	92.63	64.39	94.30	61.98	86.19	95.64	83.00
RML	99.58	98.00	97.60	97.04	98.69	95.80	**77.52**	**99.64**	77.30	91.38	**98.78**	90.23

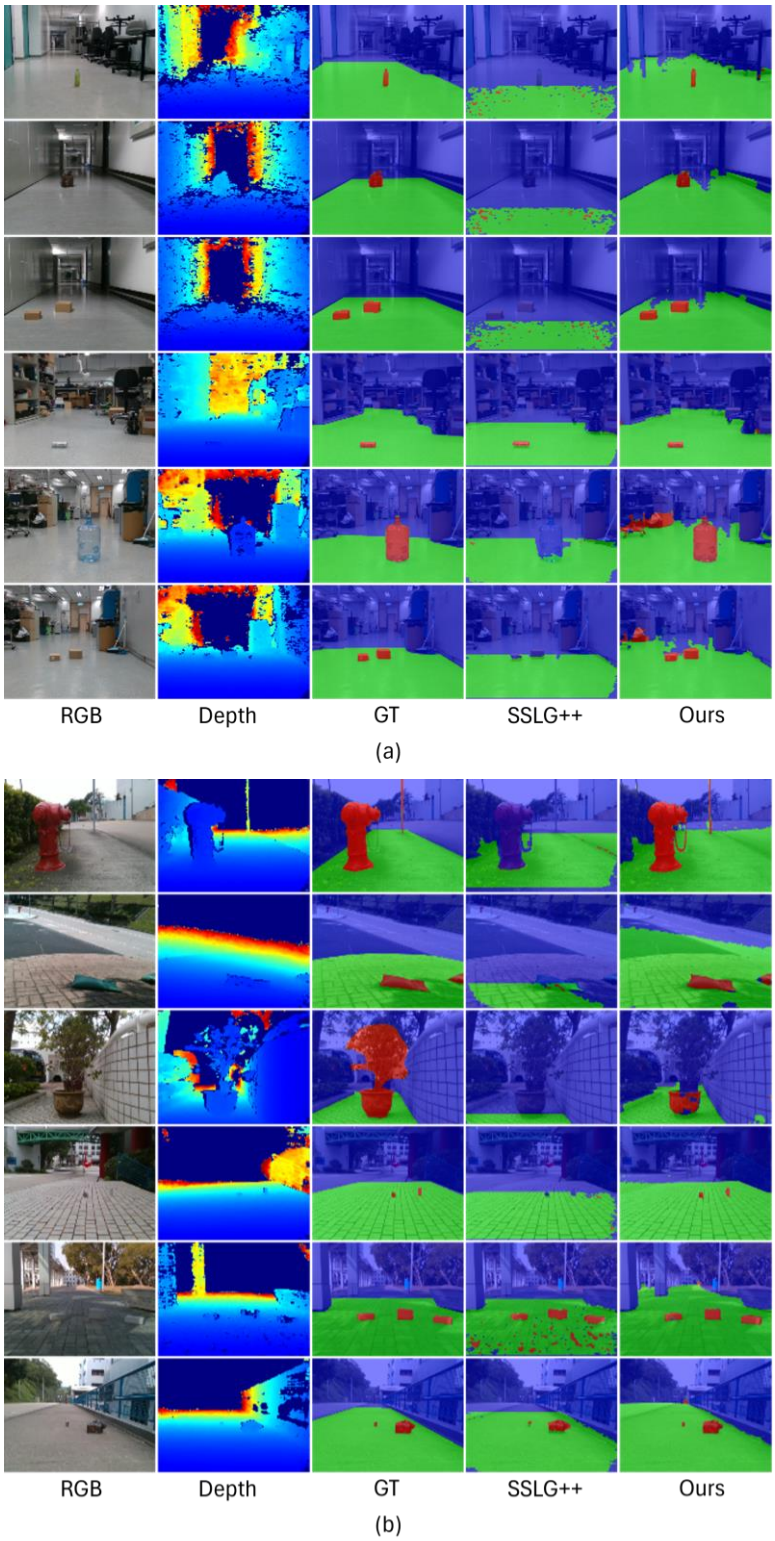

Figure 10. Comparison of the results of segmentation in indoor (**a**) and indoor (**b**) scenarios among GT (manual labels), reimplemented SSLG++ labels, and our proposed method, where the red area refers to road obstacles, the blue area refers to the unknown area, and the green area refers to the drivable areas. The figure is best viewed in color.

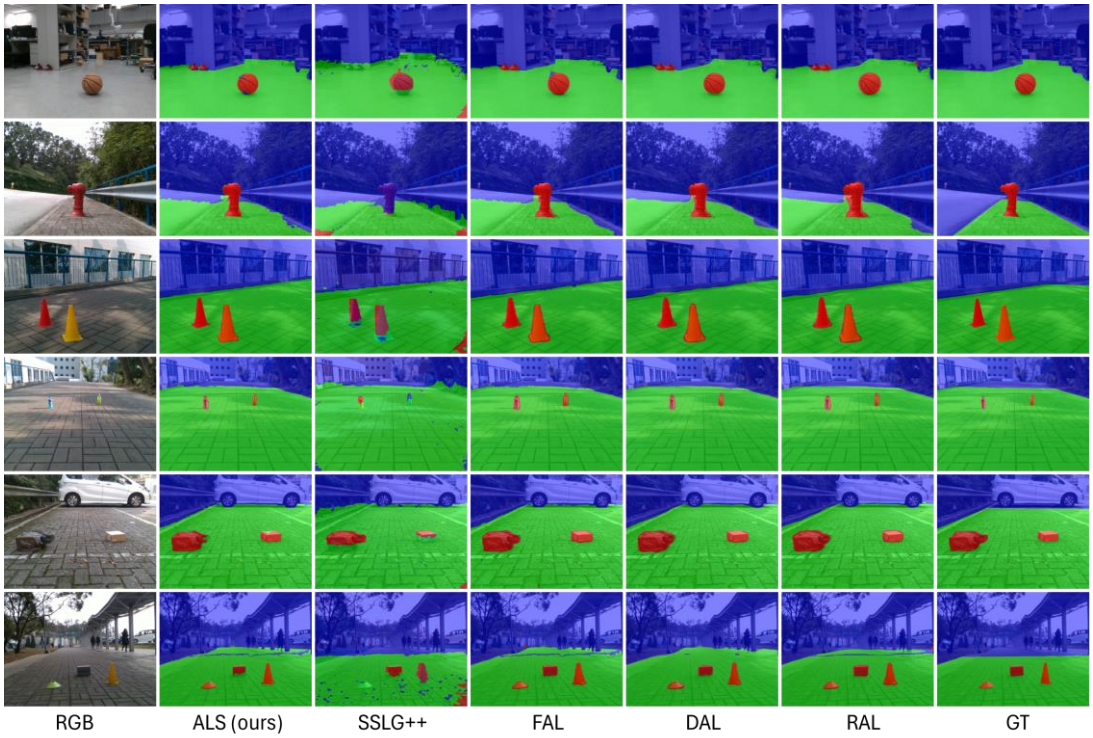

Figure 11. Comparison of the results of segmentation among the ALS labels, SSLG++ labels (the reimplementation of SSLG), FAL (FuseNet trained on the ALS labels), DAL (Depth-aware CNN trained on the ALS labels), RAL (RTFNet trained on the ALS labels), and GT (ground truth labels), where the red area refers to road obstacles, the blue area refers to the unknown area, and the green area refers to the drivable areas. The figure is best viewed in color.

The proposed reimplemented version (SSLG++) outperformed the original SSLG, and the reason is that we appropriately chose parameters for the in-depth processing pipeline of the Hough transform, as well as the RGB processing pipeline of the sigma value in the constructed Gaussian filter. The limitation of the original SSLG method was affected by the abovementioned parameter settings. In addition, the quality of the depth images significantly affected the SSLG framework, which may cause imperfections in the output segmentation maps. It was also noted that the SSLG method usually fails to detect the drivable area whenever it is unsuccessful in tracing the Hough transform's diagonal straight line from the calculated V-disparity map. Therefore, an adaptive approach could be taken to adjust the hyperparameters for specific conditions to obtain the best result.

By taking advantage of the learning-based prior knowledge, our ALS with the proposed DifferNet module determined regions of the image with high uncertainty, giving significantly better results than the baseline (SSLG method) and our reimplementation SSLG++ in terms of the out-of-distribution category in detecting in road obstacles. Consequently, our proposed method achieved 80.31% for the mean precision, 84.42% for the mean recall, and 71.99% for the mean IoU, which were 14.15% (Pre), 21.02% (Rec), and 19.66% (IoU) higher than the original SSLG method. A comparison of the results of segmenting six images with different scenarios is shown in Figure 10. Three off-the-shelf RGB-D networks trained on ALS labels outperformed the ALS labels regarding human visual supervision. This confirmed that our approach of automatic pseudo-labeling could learn the patterns in representations of road anomalies and ignore the effects of noisy depth images. In addition, adding the squeeze-and-excitation block as an attention module in our DifferNet elevated the ability to learn the relationships between individual feature channels to obtain the weight of each channel. It allowed our framework to focus more on channels with valuable information and suppress unimportant ones.

Regarding road anomalies, the evaluation of training on three RGB-D networks using ALS labels demonstrated significant improvements compared with the baseline method (SSLG labels), which led to obvious improvements, with a mean precision of 84.89%, 81.75%, and 86.19% and a mean IoU of 77.39%, 75.73%, and 83% for FuseNet, Depth-aware CNN, and RTFNet, respectively. Such an improvement in performance was obtained when we investigated the high-to-low features of the inputs, regions of uncertainty, and depth information. Moreover, the results of road segmentation using a pretrained state-of-the-art semantic segmentation method on the Cityscapes dataset gave better predictions. Although our proposed method was not entirely comparable with manual labels, it still yielded remarkable auto-labeling results with impressive speed while reducing the need for labor-intensive manual labeling, with notable results.

4.5. Execution Experiment

In terms of auto-labeling for unlabeled datasets, for each RGB-D input, our proposed TL-based approach could produce a pseudo-label map within 2.58 s on average, which was comparable with the SSLG approach, which can be processed within 2–3 s, but our method gave better semantic results. Specifically, the processing times for maps of the dispersion measure, such as entropy maps, SoftMax entropy maps, and perceptual maps were 1.72 s, 0.001 s, and 0.56 s, respectively.

Additionally, we conducted experiments on real-time segmentation using a laptop equipped with a GeForce GTX 1070 Max-Q graphics card, 8 GB of RAM, and an OAK-D camera while controlling a mobile robot, as shown in Figure 12. The autonomous system worked on the robot operating system (ROS). Our results indicated that off-the-shelf models trained on pseudo-labels generated by our ALS could achieve up to 30 and 20 frames per second (FPS) with RTFNet-18 and FuseNet, respectively.

Figure 12. The setup of the proposed method for real-time applications.

4.6. Limitation

While uncertainty in deep learning was beneficial for our ALS framework in identifying road anomalies in unfamiliar road scenes based on prior knowledge of the pretrained model, it struggles to produce accurate label maps during inference under certain conditions. This issue may stem from the quality of the captured images, which can be easily distorted by environmental factors such as noise, blurring, reflections, or occlusions. Additionally, the nature of the unlabeled data can also contribute to this problem, especially when the visual features differ significantly from the prior knowledge, leading to overconfidence in the estimation of uncertainty. Figure 13 illustrates some instances of failure in our ALS. The first row depicts an indoor scene that differed greatly from the pretrained Cityscapes data. The second row shows an example heavily affected by reflection, as well as noise and blurring. Another challenge arises when objects blend in with the background scene due to having similar colors.

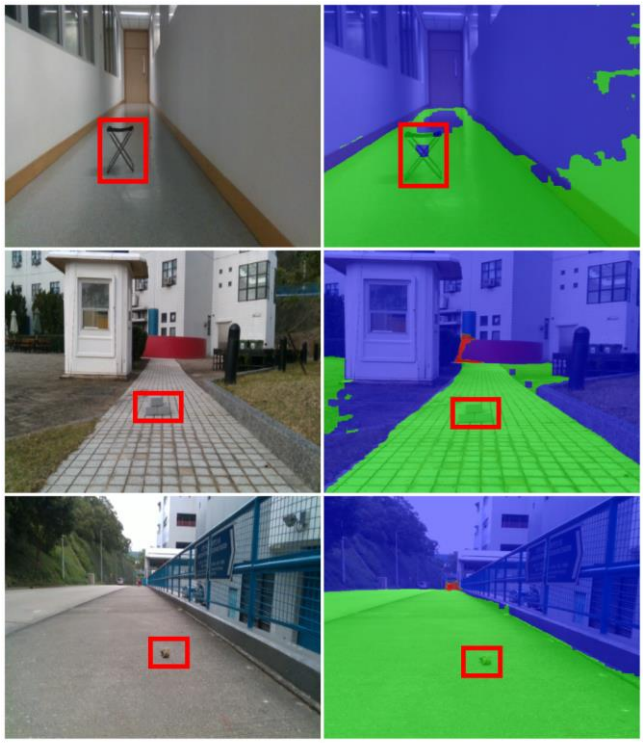

Figure 13. Some cases of failure when applying our ALS framework. The red boxes indicate the location of the road anomalies in the images. The red, green, and blue colors represent the obstacle, drivable area and background, respectively

5. Conclusions

This study presents an efficient automatic labeling framework called ALS for segmentation of the drivable area and road anomalies based on RGB-D data, as typically found in autonomous mobile robots. The key to our method is the use of the defined discrepancy network, DifferNet, which differs from our previous work with the automatic generating segmentation label system [49]. DifferNet contains squeeze-and-excitation blocks to investigate areas of high uncertainty in the input images and dispersion measures, including SoftMax entropy, SoftMax distance, and perceptual loss, which are suited for semantic segmentation of the free-driving space and unknown objects by mobile robots. Extensive experiments and the results of visualization indicated that the effectiveness of our proposed ALS framework overcame the limitations of manual labeling. In the future, we will extend our method by designing an end-to-end and lightweight approach with real-time capability for detecting road anomalies to elevate the performance of autonomous mobile robots in navigating real-world applications.

Author Contributions: Conceptualization, S.-F.S. and C.-H.K.; methodology, P.T.-T.N. and T.-K.N.; software, P.T.-T.N. and T.-K.N.; validation, P.T.-T.N., T.-K.N. and D.-D.N.; writing—original draft preparation, P.T.-T.N. and T.-K.N.; writing—review and editing, P.T.-T.N.; visualization—P.T.-T.N. and T.-K.N.; supervision, C.-H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was finically supported by the Ministry of Science and Technology, Taiwan, R.O.C., under Grant MOST 109-2221-E-002-210-MY3 (corresponding author: Chung-Hsien Kuo).

Data Availability Statement: The original data presented in the study are openly available in [9,22].

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

References

1. Ozkan, Z.; Bayhan, E.; Namdar, M.; Basgumus, A. Object Detection and Recognition of Unmanned Aerial Vehicles Using Raspberry Pi Platform. In Proceedings of the 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Türkiye, 21 October 2021; pp. 467–472.
2. Tao, M.; Zhao, C.; Wang, J.; Tang, M. ImFusion: Boosting Two-Stage 3D Object Detection via Image Candidates. *IEEE Signal Process. Lett.* **2024**, *31*, 241–245. [CrossRef]
3. Wang, X.; Li, K.; Chehri, A. Multi-Sensor Fusion Technology for 3D Object Detection in Autonomous Driving: A Review. *IEEE Trans. Intell. Transp. Syst.* **2023**, *25*, 1–18. [CrossRef]
4. Zhang, C.; Zheng, S.; Wu, H.; Gu, Z.; Sun, W.; Yang, L. AttentionTrack: Multiple Object Tracking in Traffic Scenarios Using Features Attention. *IEEE Trans. Intell. Transport. Syst.* **2024**, *25*, 1661–1674. [CrossRef]
5. Xing, Y.; Wang, J.; Chen, X.; Zeng, G. Coupling Two-Stream RGB-D Semantic Segmentation Network by Idempotent Mappings. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1850–1854.
6. Bakalos, N.; Katsamenis, I.; Protopapadakis, E.; Montoliu, C.M.-P.; Handanos, Y.; Schmidt, F.; Andersson, O.; Oleynikova, H.; Cantero, M.; Gkotsis, I.; et al. Chapter 11. Robotics-Enabled Roadwork Maintenance and Upgrading. In *Robotics and Automation Solutions for Inspection and Maintenance in Critical Infrastructures*; Loupos, K., Ed.; Now Publishers: Delft, The Netherlands, 2024; ISBN 978-1-63828-282-2.
7. Yang, X.; He, X.; Liang, Y.; Yang, Y.; Zhang, S.; Xie, P. Transfer Learning or Self-Supervised Learning? A Tale of Two Pretraining Paradigms. *arXiv* **2020**, arXiv:2007.04234. [CrossRef]
8. Gawlikowski, J.; Tassi, C.R.N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R.; et al. A Survey of Uncertainty in Deep Neural Networks. *Artif. Intell. Rev.* **2023**, *56*, 1513–1589. [CrossRef]
9. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
10. Cong, Y.; Peng, J.-J.; Sun, J.; Zhu, L.-L.; Tang, Y.-D. V-Disparity Based UGV Obstacle Detection in Rough Outdoor Terrain. *Acta Autom. Sin.* **2010**, *36*, 667–673. [CrossRef]
11. Dixit, A.; Kumar Chidambaram, R.; Allam, Z. Safety and Risk Analysis of Autonomous Vehicles Using Computer Vision and Neural Networks. *Vehicles* **2021**, *3*, 595–617. [CrossRef]
12. Park, J.-Y.; Kim, S.-S.; Won, C.S.; Jung, S.-W. Accurate Vertical Road Profile Estimation Using V-Disparity Map and Dynamic Programming. In Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 16–19 October 2017; pp. 1–6.
13. Mayr, J.; Unger, C.; Tombari, F. Self-Supervised Learning of the Drivable Area for Autonomous Vehicles. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 362–369.
14. Ma, F.; Liu, Y.; Wang, S.; Wu, J.; Qi, W.; Liu, M. Self-Supervised Drivable Area Segmentation Using LiDAR's Depth Information for Autonomous Driving. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1 October 2023; pp. 41–48.
15. Han, X.; Lu, J.; Zhao, C.; You, S.; Li, H. Semisupervised and Weakly Supervised Road Detection Based on Generative Adversarial Networks. *IEEE Signal Process. Lett.* **2018**, *25*, 551–555. [CrossRef]
16. Ma, W.; Zhu, S. A Multifeature-Assisted Road and Vehicle Detection Method Based on Monocular Depth Estimation and Refined U-V Disparity Mapping. *IEEE Trans. Intell. Transport. Syst.* **2022**, *23*, 16763–16772. [CrossRef]
17. Ali, A.; Gergis, M.; Abdennadher, S.; El Mougy, A. Drivable Area Segmentation in Deteriorating Road Regions for Autonomous Vehicles Using 3D LiDAR Sensor. In Proceedings of the 2021 IEEE Intelligent Vehicles Symposium (IV), Nagoya, Japan, 11 July 2021; pp. 845–852.
18. Jiang, F.; Wang, Z.; Yue, G. A Novel Cognitively Inspired Deep Learning Approach to Detect Drivable Areas for Self-Driving Cars. *Cogn. Comput.* **2024**, *16*, 517–533. [CrossRef]
19. Asgarian, H.; Amirkhani, A.; Shokouhi, S.B. Fast Drivable Area Detection for Autonomous Driving with Deep Learning. In Proceedings of the 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA), Kashan, Iran, 28 April 2021; pp. 1–6.
20. Rabiee, S.; Biswas, J. IVOA: Introspective Vision for Obstacle Avoidance. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 4–8 November 2019; pp. 1230–1235.
21. Ghosh, S.; Biswas, J. Joint Perception and Planning for Efficient Obstacle Avoidance Using Stereo Vision. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 1026–1031.
22. Wang, H.; Sun, Y.; Liu, M. Self-Supervised Drivable Area and Road Anomaly Segmentation Using RGB-D Data For Robotic Wheelchairs. *IEEE Robot. Autom. Lett.* **2019**, *4*, 4386–4393. [CrossRef]
23. Rahman, Q.M.; Sunderhauf, N.; Corke, P.; Dayoub, F. FSNet: A Failure Detection Framework for Semantic Segmentation. *IEEE Robot. Autom. Lett.* **2022**, *7*, 3030–3037. [CrossRef]

24. Oberdiek, P.; Rottmann, M.; Fink, G.A. Detection and Retrieval of Out-of-Distribution Objects in Semantic Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 1331–1340.
25. Di Biase, G.; Blum, H.; Siegwart, R.; Cadena, C. Pixel-Wise Anomaly Detection in Complex Driving Scenes. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 16913–16922.
26. Lis, K.; Nakka, K.K.; Fua, P.; Salzmann, M. Detecting the Unexpected via Image Resynthesis. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2152–2161.
27. Liao, J.; Xu, X.; Nguyen, M.C.; Goodge, A.; Foo, C.S. COFT-AD: COntrastive Fine-Tuning for Few-Shot Anomaly Detection. *IEEE Trans. Image Process.* **2024**, *33*, 2090–2103. [CrossRef] [PubMed]
28. Lis, K.; Honari, S.; Fua, P.; Salzmann, M. Detecting Road Obstacles by Erasing Them. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 2450–2460. [CrossRef] [PubMed]
29. Tian, Y.; Liu, Y.; Pang, G.; Liu, F.; Chen, Y.; Carneiro, G. Pixel-Wise Energy-Biased Abstention Learning for Anomaly Segmentation on Complex Urban Driving Scenes. In *Computer Vision–ECCV 2022*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Lecture Notes in Computer Science; Springer Nature: Cham, Switzerland, 2022; Volume 13699, pp. 246–263, ISBN 978-3-031-19841-0.
30. Lis, K.; Honari, S.; Fua, P.; Salzmann, M. Perspective Aware Road Obstacle Detection. *IEEE Robot. Autom. Lett.* **2023**, *8*, 2150–2157. [CrossRef]
31. Nayal, N.; Yavuz, M.; Henriques, J.F.; Güney, F. RbA: Segmenting Unknown Regions Rejected by All. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1 October 2023; pp. 711–722.
32. Rai, S.N.; Cermelli, F.; Fontanel, D.; Masone, C.; Caputo, B. Unmasking Anomalies in Road-Scene Segmentation. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1 October 2023; pp. 4014–4023.
33. Katsamenis, I.; Protopapadakis, E.; Bakalos, N.; Varvarigos, A.; Doulamis, A.; Doulamis, N.; Voulodimos, A. A Few-Shot Attention Recurrent Residual U-Net for Crack Segmentation. In *Advances in Visual Computing*; Bebis, G., Ghiasi, G., Fang, Y., Sharf, A., Dong, Y., Weaver, C., Leo, Z., LaViola, J.J., Kohli, L., Eds.; Lecture Notes in Computer Science; Springer Nature: Cham, Switzerland, 2023; Volume 14361, pp. 199–209, ISBN 978-3-031-47968-7.
34. Wan, B.; Zhou, X.; Sun, Y.; Wang, T.; Lv, C.; Wang, S.; Yin, H.; Yan, C. ADNet: Anti-Noise Dual-Branch Network for Road Defect Detection. *Eng. Appl. Artif. Intell.* **2024**, *132*, 107963. [CrossRef]
35. Li, G.; Zhang, C.; Li, M.; Han, D.-L.; Zhou, M.-L. LHA-Net: A Lightweight and High-Accuracy Network for Road Surface Defect Detection. *IEEE Trans. Intell. Veh.* **2024**, 1–15. [CrossRef]
36. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In *Computer Vision–ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11207, pp. 418–434, ISBN 978-3-030-01218-2.
37. Rottmann, M.; Colling, P.; Paul Hack, T.; Chan, R.; Huger, F.; Schlicht, P.; Gottschalk, H. Prediction Error Meta Classification in Semantic Segmentation: Detection via Aggregated Dispersion Measures of Softmax Probabilities. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–9.
38. Liu, X.; Yin, G.; Shao, J.; Wang, X.; Li, H. Learning to Predict Layout-to-Image Conditional Convolutions for Semantic Image Synthesis. *Adv. Neural Inf. Process. Syst.* 2020; 32, 570–580.
39. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Computer Vision– ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2016; Volume 9906, pp. 694–711, ISBN 978-3-319-46474-9.
40. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556. [CrossRef]
41. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
42. Park, T.; Liu, M.-Y.; Wang, T.-C.; Zhu, J.-Y. Semantic Image Synthesis With Spatially-Adaptive Normalization. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–17 June 2019; pp. 2332–2341.
43. Hirschmuller, H. Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 10–26 June 2005; Volume 2, pp. 807–814.
44. Pinggera, P.; Ramos, S.; Gehrig, S.; Franke, U.; Rother, C.; Mester, R. Lost and Found: Detecting Small Road Hazards for Self-Driving Vehicles. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Republic of Korea, 9–14 October 2016; pp. 1099–1106.
45. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
46. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. In *Computer Vision–ACCV 2016*; Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2017; Volume 10111, pp. 213–228, ISBN 978-3-319-54180-8.

47. Sun, Y.; Zuo, W.; Liu, M. RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes. *IEEE Robot. Autom. Lett.* **2019**, *4*, 2576–2583. [CrossRef]
48. Wang, W.; Neumann, U. Depth-Aware CNN for RGB-D Segmentation. In *Computer Vision–ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11215, pp. 144–161, ISBN 978-3-030-01251-9.
49. Nguyen, T.-K.; Nguyen, P.T.-T.; Nguyen, D.-D.; Kuo, C.-H. Effective Free-Driving Region Detection for Mobile Robots by Uncertainty Estimation Using RGB-D Data. *Sensors* **2022**, *22*, 4751. [CrossRef] [PubMed]
50. Zhu, Y.; Sapra, K.; Reda, F.A.; Shih, K.J.; Newsam, S.; Tao, A.; Catanzaro, B. Improving Semantic Segmentation via Video Propagation and Label Relaxation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–17 June 2019; pp. 8848–8857.
51. Vojir, T.; Sipka, T.; Aljundi, R.; Chumerin, N.; Reino, D.O.; Matas, J. Road Anomaly Detection by Partial Image Reconstruction with Segmentation Coupling. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 15631–15640.