Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 4684669, 20 pages
https://doi.org/10.1155/2022/4684669

Research Article

Improvement of Multiclass Classification of Pavement Objects Using Intensity and Range Images

Elham Eslami and Hae-Bum Yun

Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL, USA

Correspondence should be addressed to Hae-Bum Yun; haebum@mac.com

Received 25 February 2022; Accepted 23 June 2022; Published 9 August 2022

Academic Editor: SeyedAli Ghahari

Automated recognition of road surface objects is vital for efficient and reliable road condition assessment. Despite recent advances in developing computer vision algorithms, it is still challenging to analyze road images due to the low contrast, background noises, object diversity, and variety of lighting conditions. Motivated by the need for an improved pavement objects classification, we present Dual Attention Convolutional Neural Network (DACNN) to improve the performance of multiclass classification using intensity and range images collected with 3D laser imaging devices. DACNN fuses heterogeneous information in intensity and range images to enhance distinguishing foreground from background, as well as to improve object classification in noisy images under various illumination conditions. DACNN also leverages multiscale input images by capturing contextual information for object classification with different sizes and shapes. DACNN contains an attention mechanism that (i) considers semantic interdependencies in spatial and channel dimensions and (ii) adaptively fuses scale-specific and mode-specific features so that each feature has its own level of contribution to the final decision. As a practical engineering project, dataset are collected from road surfaces using 3D laser imaging. DACNN is compared with four deep classifiers that are widely used in transportation applications. Experiments show that DACNN consistently outperforms the baselines by 22–35% on average in terms of the F-score. A comprehensive discussion is also presented regarding computational costs and how robustly the investigated classifiers perform on each road object.

1. Introduction

Automation in road condition assessment is a crucial yet challenging task in smart transportation management. The goal is to label various road objects in pavement images and to establish appropriate maintenance and repair strategies to ensure road serviceability and safety. Manual road assessment, however, is labor intensive, time-consuming, and inconsistent. Automated road object detection is an alternative way for objective and scalable assessment of road networks. Fast and accurate automated road assessment can be used as quantitative data for optimal maintenance and rehabilitation practices to improve road performance and decrease the overall life-cycle cost.

To automate the road condition assessment, data are usually collected by surveying vehicles equipped with digital cameras that acquire images from pavement surfaces at high speed. There are two main high-resolution imaging techniques frequently used in road survey projects: (i) two-dimensional (2D) imaging technology in which line-scanning cameras are used to generate 2D intensity images; (ii) three-dimensional (3D) imaging technology that provides additional range (depth) images in addition to the intensity images. Recently, the 3D imaging technology has been increasingly adopted by state and local transportation agencies for data collection of road networks [1, 2]. The 3D imaging equipment employs high-resolution laser imaging devices associated with a high-precision inertial measurement unit (IMU) to capture 3D pavement surface profile data at highway speed. One of the main advantages of the 3D technology is that it is less sensitive to light effects and less prone to noises coming from oil or water stains, dirt or sand, skid marks, etc. Furthermore, the combination of intensity and range images provides additional information to model

object boundaries and global layouts and to better recognize pavement defects.

Despite those advantages of new 3D imaging technology, existing kinds of literature [3–6] lack investigations to quantify improved performance in road object detection due to 3D technology using additional range images, compared to traditional 2D technology relying on intensity images only. Existing studies address the recognition of pavement defects, mostly cracks, using intensity images by employing deep convolutional neural networks (CNNs) [7–9]. CNNs have been successfully employed for various visual recognition tasks including image classification [10, 11], object detection [12], and semantic segmentation [13]. Although CNNs have demonstrated good performance on pavement defects recognition using intensity images, the performance tends to be degraded when detecting defects in complex scenes. The complexity comes from intensity inhomogeneity, low contrast, background noises, objects diversity in terms of shape and size, variety of lighting conditions, etc., when using intensity images only. For example, when there exists low contrast between cracks (as the foreground) and asphalt (as the background) or when dealing with thin cracks, it is difficult to distinguish between background and foreground based on only intensity data. In the case of objects with similar color and texture (such as crack seals and patches), it is easy to misclassify those objects into the same categories. Moreover, intensity-based features extracted from pavement 2D images are sensitive to illumination differences among images. The abovementioned limitations motivate the joint use of range and intensity images to enhance the classification of pavement objects. Figure 1 shows a surveying vehicle installed with a 3D laser imaging device developed by Korea Institute of Civil Engineering and Building Technology (KICT) used in this study, and a sample of intensity and range images collected by the system.

We present the novel Dual Attention Convolutional Neural Network (DACNN) to utilize additional range of input images along with intensity images to improve pavement objects classification. In this paper, DACNN classifies pavement tiles into 8 classes, including crack, crack seal, patch, pothole, marker, manhole, curbing, and asphalt. DACNN leverages multiscale input tiles that capture scale-sensitive information for multiclass classification of various road objects with different sizes and shapes. Furthermore, DACNN adopts two attention modules to effectively fuse heterogeneous features in terms of (i) scales (multiscale input tiles) and (ii) modes (range and intensity tiles). The scale and mode attention modules focus on spatial and channel-related informative features and suppress the noninformative ones for performance improvement. The dual attention mechanism is designed to identify semantic image regions relevant to specific pavement objects. Pruning feature maps in both spatial and channel dimensions enhance the quality of feature representation, contributing to more accurate and efficient object classification.

The contribution of this study is not only limited to the architectural design of DACNN. We also evaluate the effectiveness of the additional range of data in 3D technology over 2D technology through quantitative comparison using different CNN models, including VGG16, VGG19, ResNet50, DenseNet121, as well as the DACNN. The goal of the above comparisons is (i) to understand the effects of the additional range data to improve object classification, (ii) to understand how the scale and mode attention modules can effectively fuse heterogeneous information to improve objects classification, and (iii) to understand the effects of CNN model selection to the number of trainable variables, training time, inference time, and classification accuracy. Our main contributions in this paper are summarized as follows:

We present the new DACNN framework to systematically utilize both intensity and range images collected with 3D imaging devices for multiclass classification of pavement images. Considering the variety of pavement objects and surveying field conditions, DACNN extracts scale-specific and mode-specific features from images robustly. The dual attention mechanism used in DACNN is designed to adaptively fuse multiscale multimodal features, helping the network to capture discriminative object-specific features related to their spatial and channel information.

The classification performance comparison is conducted for 8 different pavement objects using CNN models. The results show that our DACNN outperforms other models for all road object classes. We also present quantitative comparisons to understand how the additional range of images in 3D technology can improve object classification performance for compared CNN models.

2. Related Works

2.1. Deep Learning in Pavement Assessment. Conventional image processing and more recent deep learning methods are two main approaches for automated pavement image analysis. The image processing methods can be considered as feature engineering techniques in which images are represented with human-specified feature vectors. They can be sorted into intensity-thresholding [14], edge detection [15], wavelet transforms [16, 17], and texture-analysis [18, 19]. A major problem with the conventional methods is that the prediction performance mainly relies on the validity of human-specified features. Extracting those features can be subjective, domain-specific, and inefficient, which makes the detection process ungeneralizable and tedious. Especially in pavement applications, hand-crafted features are not robust enough to detect distresses in the complex background with high variations. For instance, thresholding approaches for crack detection only achieve acceptable results under certain scenarios. If there exists a complex background or the illumination changes, either the parameters should be adjusted or the method is not applicable to the new scene.

Deep learning methods overcome the drawbacks of conventional image processing methods by automatically capturing complex structures of data with multiple

Figure 1: 3D laser imaging system developed by Korea Institute of Civil Engineering and Building Technology (KICT); sample of high-resolution intensity and range road surface images.

processing layers. CNNs are the most studied deep learning models using vision-based input data in which automated feature learning is done at many different levels of abstraction to catch the topology of input images. Partial connections, sharing weights, and pooling layers in CNNs not only decrease the computations but also demonstrate state-of-the-art results in computer vision tasks [20, 21]. Detection, classification, and segmentation of pavement distress, especially cracks, are the main three branches of deep learning research in automated pavement assessment. Alfarrarjeh et al. [22] employed YOLO [23] as the object detection method to detect distresses, including cracks, potholes, and rutting, in pavement images. Maeda et al. [24] adopted SSD [25] as the training algorithm to detect the same defects on pavement surfaces. Song et al. [26] utilized Faster R–CNN [27] algorithm to detect pavement distresses, including cracks, potholes, and bleeding. Li et al. [28] presented a CNN model to classify pavement tiles into different types of cracks including longitudinal, transverse, alligator, and block cracks. Gopalakrishnan et al. [29] utilized a pretrained VGG16 [30] on ImageNet and then fine-tuned it on a pavement dataset for a binary crack classification. Lau et al. [31] proposed a U-Net [32] based model in which the encoder is a pretrained ResNet34 [33] to segment pavement crack images. Inspired by SegNet [34], Chen et al. [35] proposed a fully convolutional neural network (FCNN) to detect pavement cracks at pixel level.

2.2. Attention in Deep Learning. The performance of deep learning-based approaches has been constantly improving by developing new architectural designs, and the attention mechanism is one of them. The main idea behind an attention mechanism is to give higher weights to relevant features while minimizing the irrelevant ones by giving lower weights. Focusing on the distinctive parts when processing large amounts of information, the attention mechanism enhances the quality of feature representation, contributing to a more accurate and efficient performance of the designed network. Attention was initially proposed by [36] for machine translation. Then, it was employed for various tasks, such as action recognition [37–39], speech recognition [40, 41], image captioning [42, 43], and recommendation [44, 45]. More specifically, the attention mechanism is investigated in computer vision community in three aspects: (i) spatial attention in which the network learns the locations that should be focused on [46, 47]; (ii) channel attention in which the network adaptively recalibrates channel-wise features by modeling interdependencies between channels [48, 49]; and (iii) Self-attention in which long-range dependencies are captured by the network [50, 51]. In pavement applications, attention modules have been also applied for defect detection. Song et al. [52] presented a channel of attention to detect and classify different types of cracks in pavement images. Wan et al. [53] proposed an encoder-decoder network, called CrackResAttentionNet, containing spatial and channel attention modules after each block in the encoder to segment pavement cracks. Similarly, Qiao et al. [54] proposed CrackDFANet in which a channel-spatial attention module is designed to increase the generalization ability of the model in predicting cracks under different conditions of roads. Wang et al. [55] proposed using DenseNet121 as an encoder and a spatial attention module to combine multiscale features. Eslami et al. [56] designed a channel-spatial attention module to adaptively fuse multiscale features for pavement image classification. Zhou et al. [57] presented a VGG16-based network to predict crack maps, and employed spatial and channel attention modules to further refine the model. Qu et al. [58] employed Res2Net [59] along with an attention module to capture global context and long-range dependency for a better pavement segmentation. *Pan* et al. [60] proposed SCHNet with VGG19 as the base net in which a self-attention module is designed to global as well

as semantic interdependencies in the channel and spatial dimensions. Finally, Li et al. [61] proposed a self-attention module along with a scale-attention module to enhance feature representation for pavement crack segmentation.

In this study, we propose a dual attention approach to capture semantic interdependencies in both spatial and channel dimensions for scale and type of input images. The dual attention mechanism achieves a fast focus on more important features and enhances the representativity of more relevant features for better classification performance. The dual attention approach enables modeling global context as well as multimodal features to improve classification performance for both small objects (e.g., cracks) and large objects (e.g., patches), which are in trade-off using other CNN models.

2.3. 3D Image Data in Pavement Assessment. Most of the existing deep learning studies were based on only intensity images using 2D imaging devices in transportation applications. With 2D intensity input images, CNNs suffer from some important limitations. The complexity of scenes, diversity of objects, background noises (stains, oil spills, and tire marks), and surrounding changes (light and shadow) make it difficult to distinguish foreground objects (defects) from the background (asphalt) in 2D images. With the advances in sensor technology, 3D imaging systems are available and increasingly employed by state and local transportation agencies for automated road condition assessment. A survey showed that 18 states in the U.S. adopted a 3D data collection system by 2017, and 17 states intended to utilize this technology by 2019 [1]. Different approaches have been studied for transportation applications such as GPR, LiDAR, Microsoft Kinect, and laser profilers [3]. In pavement applications, laser profilers are commonly used in surveying road roughness and megatexture (ASTM E950, ASTM E1926, and ISO 13473–5) [62–64]. Other techniques offer limitations such as relatively low resolution (in case of LiDAR) or low frequency (in case of Microsoft Kinect) to collect road surface profiles. The 3D laser imaging technique, such as Laser Crack Measurement System (LCMS) [65], is commercially available to collect high-resolution road surface profiles. This system utilizes surveying vehicles equipped with two laser imaging devices (left and right) and IMU. Using the 3D imaging system, intensity and range images can be acquired at speeds up to 100 km/h on on-road lanes with 4 m width under various lighting conditions. The 3D laser imaging technology has been used to evaluate crack [66, 67], pothole [68], raveling [69], rutting [70], joint [71], and texture [72]. Ghosh et al. [73] employed YOLO and Faster R–CNN to detect cracks in range images collected by the 3D imaging system. Yang et al. [74] utilized 3D laser technology to measure the growth of crack lengths when they are sealed and non-sealed to quantify the crack sealing benefit. Li et al. [28] proposed a CNN framework to classify range images into transverse cracks, longitudinal cracks, block cracks, and alligator cracks. Lang et al. [67] proposed a clustering-based algorithm to classify range images into the same categories of cracks as Li et al. [28]. Fei et al. [75] presented a deep CNN, called CrackNet-V, to segment cracks on asphalt range images. Li et al. [76] applied a filter-based method to segment cracks using 3D pavement images. Zhang et al. [77] proposed a recurrent neural network (RNN), called CrackNet-R, to detect pavement cracks at pixel-level in range images. Gui et al. [78] utilized laser-scanning 3D to detect pavement cracks by extracting hand-crafted features. Tsai and Chatterjee [68] proposed a threshold-based method to detect pavement potholes in range images collected by 3D laser technology. Zhang et al. [79] proposed a CNN-based architecture, called CrackNet to segment cracks in 3D pavement images. Zhang et al. [80] improved the crack segmentation results on 3D pavement images by proposing a deeper network, CrackNetII, in which the need for hand-crafted features is eliminated. Li et al. [81] presented a frequency analysis to detect pavement cracks from background texture in range images.

While there are existing studies using 3D laser imaging technology, they are limited to the use of either range or intensity images. In this study, we show that extracting features from both intensity and range (depth) images can significantly improve the CNN performance. We also show that by fusing intensity-specific and depth-specific features systematically, one can robustly and accurately classify not only cracks but also other pavement objects, including crack seals, patches, potholes, markers, manholes, and curbing in multiclass classification.

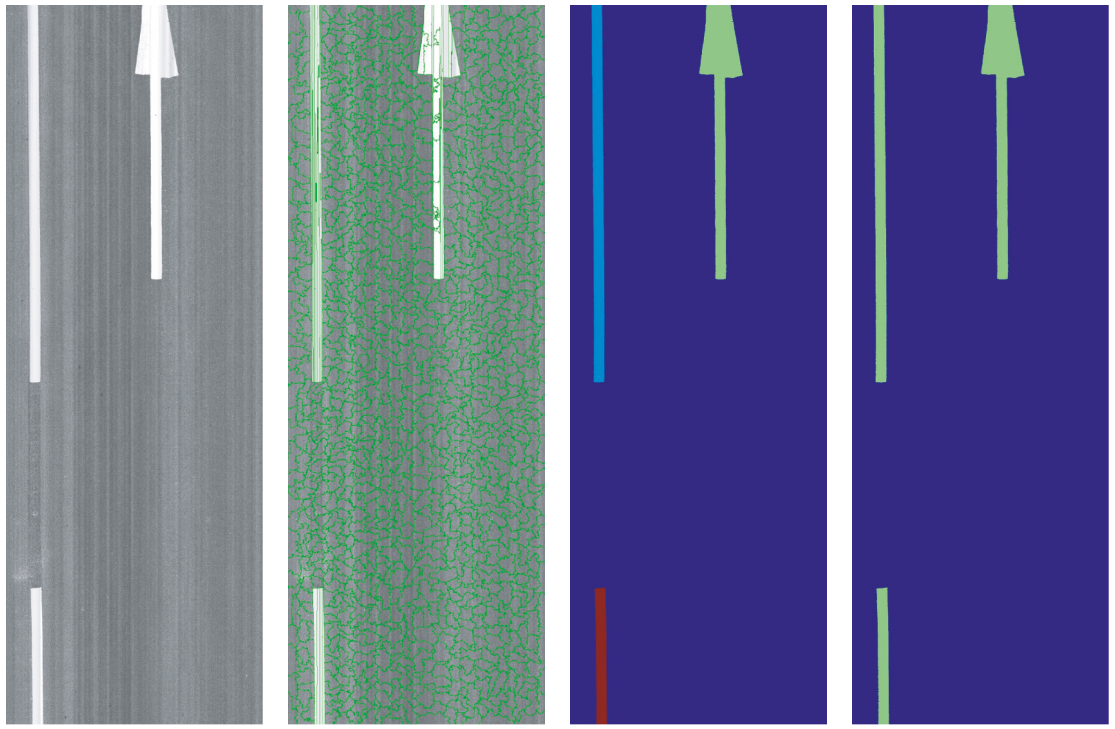
3. Data Preparation

3.1. Ground-Truth Labeling. The dataset used in this study contains 296 intensity images and the same number of range images with the size of 3700×10000 pixels spatial resolution of 1 mm/pixel. The gray-scale intensity and range images are collected by the 3D laser imaging device developed by Korea Institute of Civil Engineering and Building Technology (KICT) shown in Figure 1. The technical specifications of this device are provided in Table 1.

We provide pixel-level annotations of road objects for 8 categories, including 4 distress classes (crack, crack seal, patch, and pothole), 3 non distress classes (marker, manhole, and curbing), and 1 pavement class (asphalt) as the background. We annotate the intensity images using an in-house developed semiautomated software that makes the annotation process fast yet accurate. The annotation procedure is performed in two steps: (i) labeling area objects (all classes except for cracks) and (ii) labeling linear objects (i.e., cracks). To label area objects, the original image, shown in Figure 2(a), is grouped into homogeneous regions, called superpixels [82, 83]. As shown in Figure 2(b), superpixel segmentation preserves the edges and boundaries of objects. Therefore, superpixel-level labeling, rather than pixel-level labeling, can be performed, which reduces the labeling work significantly. To further facilitate the annotation process, an unsupervised mean shift clustering is applied, which groups the neighboring superpixels into a bigger cluster. The result of the superpixel clustering procedure is shown in Figure 2(c). Then, the human annotator can easily select the clusters that belong to the same object and label them. Also,

Table 1: Technical specifications of KICT 3D laser imaging device.

Scanning frequency	Transverse range	Lateral resolution (mm)	Vertical resolution (mm)	Data rate
5600 profiles per second	4 m (4096 points per profile)	1	0.5	10.4 Gb/km (720 Mb/km compressed)



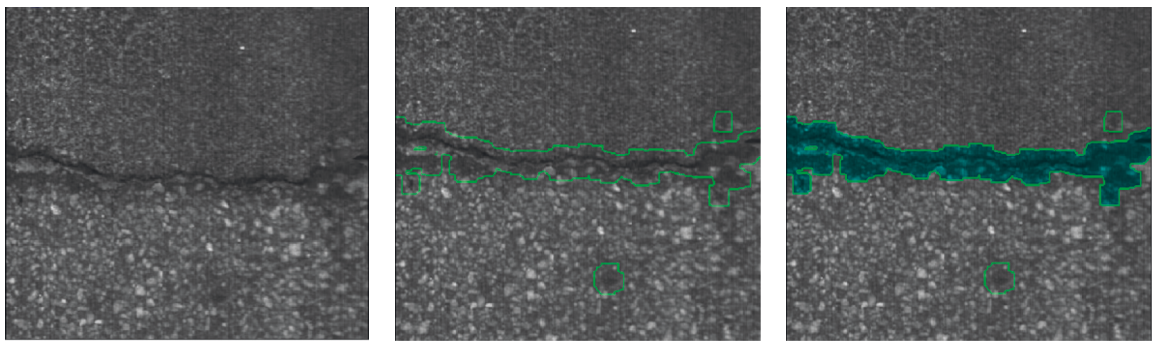
(a) (b) (c) (d)

Figure 2: Annotation procedures for areal objects. (a) Original image; (b) superpixel segmentation; (c) unsupervised mean shift clustering; (d) human correction of false clustering and classification.

the annotator is able to define new segments, which are missed by the clustering algorithm. Figure 2(d) demonstrates the final pixel-level labeling mask. Although the superpixel segmentation technique is beneficial for labeling area objects in the dataset, it is not effective for linear object labeling such as cracks. To label cracks, a morphological technique, called MorphLink-C, is employed to extract crack pixels in original images. MorphLink-C consists of a series of morphological operations, which is proposed by Wu et al. [84]. The original image in Figure 3(a) is a zoomed-in pavement image for better visualization of the existing crack. The cracks detected by MorphLink are shown in Figure 3(b) with the bounding boxes. Having the detected cracks, the human annotator can select the truly detected cracks within the image, as shown in Figure 3(c).

Figure 4 demonstrates the contents of different objects in the dataset. We observe that the population of road object pixels are highly imbalanced, for example, there are more than three million of asphalt pixels but only more than 4000 crack seal pixels in the dataset. Detecting objects with high variations in shape and size within a highly imbalanced dataset is a major challenge in pavement applications.

3.2. Data Preprocessing. In road surveying projects, the depth information in range images is often used to measure the macrostructure of pavement surface (ISO 13473–1) [64]. Although the depth resolution of the laser device on an absolute millimeter scale is important to determine the mean profile depth (MPD) in macrotexture surveying, a small variation in surface profile (e.g., crack depth) and low contrast in range images could be a disadvantage in road objects detection. To enhance the contrast, a histogram equalization (HE) can be applied to range images. HE enhances the contrast by effectively spreading out the most frequent intensity values (stretching out the intensity range of the image). It allows for areas with lower local contrast to obtain a higher contrast. In this study, Contrast Limited Adaptive Histogram Equalization (CLAHE) [85] is applied to a range of images. CLAHE differs from ordinary HE algorithms in two ways: (i) An adaptive HE computes several histograms, each corresponding to a small region of the image rather than computing the histogram for the entire image. Therefore, it improves the local contrast and edges in each region of the image. (ii) CLAHE sets a threshold to limit the contrast in each small region. The contrast limiting



(a) (b) (c)

Figure 3: Annotation procedures for linear objects. (a) Original image; (b) automatic crack detection by MorphLink technique; (c) human selection of truly detected cracks.

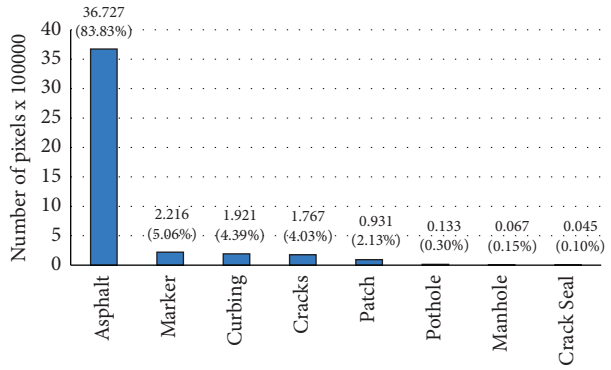


Figure 4: Number of pixels in our pavement classes.

procedure prevents the over-enhancement and amplification of noise in the image. Figure 5(a) shows a range image with cracks spreading all over the image. Also, the intensity distribution of the image and the cumulative distribution are presented for the range image as histogram and cdf, respectively. Figure 5(b) demonstrates the range image after using CLAHE enhancement and its corresponding histogram and cdf. We can see that the visibility of cracks is improved by redistributing the lightness values of the image without introducing noises to the image. Comparing the histograms before and after applying CLAHE to the image, the intensity range of the road image is expanded within the lower range (dark pixels 0–50) by redistribution of the values, as shown in Figure 5(b).

After the contrast enhancement of range images, we divide the original images into nonoverlapping 50×50 tiles to conduct multiclass classification experiments on pavement images. Then, each image tile is assigned to one of 8 categories of road objects. When a 50×50 tile has more than one class of pixels, the tile class is determined by a majority vote between the pixel number of nonbackground classes if exists, otherwise, the tile is classified as the background (asphalt). By aggregating the assigned classes for all tiles generated from an original image, a segmentation mask with a resolution of $50 \times 50\,\text{mm}^2$ can be produced. The reason for 50×50 tile generation comes from two sources: (i) Due to the large size of the original images (3700×10000), the segmentation task on the whole image is memory intensive and not practical; (ii) 50×50-pixel tiles, equivalent to $50 \times 50\,\text{mm}^2$, is small enough to contain only one pavement object for the classification task. Therefore, assembling the classification results into the whole image produces a segmentation mask with a high-resolution, which is satisfactory in pavement applications. Although having small input tiles results in high-resolution segmentation masks, it sacrifices the contextual information required from the deep networks to perform well. Due to the importance of contextual information for the classification task, we generate 250×250, and 500×500 tiles surrounding each 50×50 tile with the same center. Feeding multiscale tiles into the deep networks improves the classification performance of the smallest tile, which will be explained in Section 4.1.

4. Method

4.1. Dual Attention Convolutional Neural Network Architecture. The Dual Attention Convolutional Neural Network (DACNN), illustrated in Figure 6, is presented to classify pavement image tiles into one of the 8 existing classes in the dataset. The DACNN provides a systematic way of data fusion for heterogeneous input images including (i) intensity and range images (i.e., mode), and (ii) 50×50, 250×250, and 500×500 (i.e., scale), which is more effective than a simple feature concatenation. For this, the DACNN

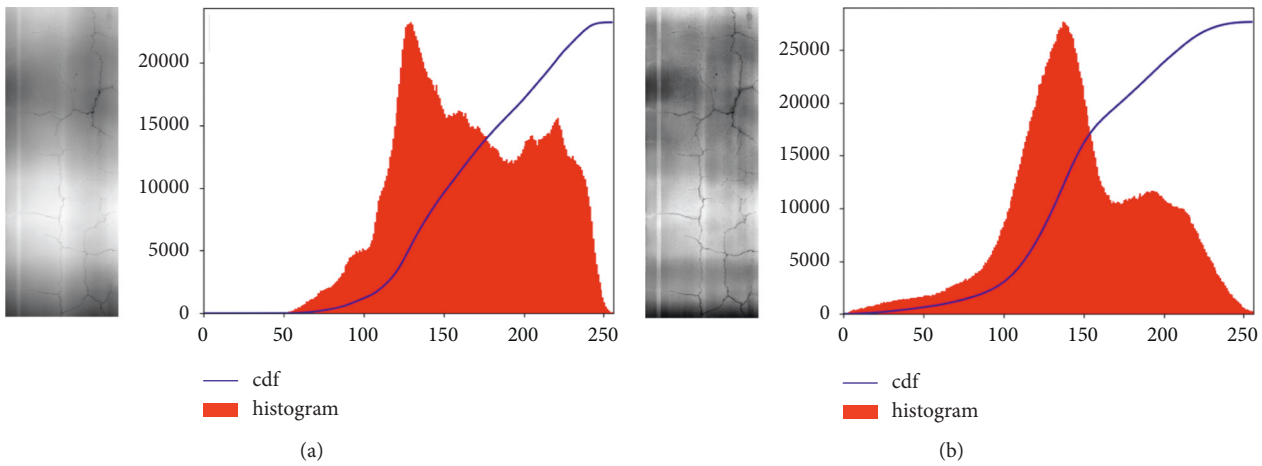


Figure 5: (a) Original range image and (b) CLAHE enhanced range image with corresponding histograms and cumulative histograms.

consists of two main streams of intensity and range modes, which are merged later by a mid-fusion strategy (i.e., mode-level attention module). Each mode steam consists of three scale streams to extract multiscale features, which are combined later using a mid-fusion strategy (i.e., scale-level attention module). The high-level architecture of the DACNN is shown in Figure 6.

Multiscale Input Tiles. Input tiles are extracted from the original intensity and range images at three scales, 50 × 50, 250 × 250, and 500 × 500. All the input tiles are resized to 50 × 50 before they are fed to the DACNN.

Feature Extraction (Scale). A conventional to combine multiscale multimodal input data is directly concatenating them at the input level. This approach has a disadvantage in that only similar patterns will be captured across the scales and modes. Instead of concatenating heterogeneous input data in an early fusion, we propose to feed input tiles to 6 separate CNNs to extract scale-specific and mode-specific features. Each CNN consists of three convolution layers with the filter numbers 32, 32, and 64, respectively. The filter size is 3 × 3 pixels for all convolution layers. Each convolution layer is then followed by a Batch Normalization layer and a rectified linear unit activation (ReLU), which are not shown in Figure 6 because of space limitation. It should be noted that up to this point the extracted feature maps are processed independently at each scale and mode level.

Mid-Fusion with Scale-Level Attention Module. The main idea of using multiscale input tiles is to allow features extracted from different levels of spatial context around the smallest tile (50 × 50) to contribute to the classifying decision. The level of contribution at each scale for different objects varies for different objects. For example, scale 1 is more informative for small objects (e.g., cracks), while scale 3 is more informative for classifying large objects (e.g., patches). Therefore, we use a scale-level attention module that decides how much attention to pay to scale-sensitive features. Unlike simple concatenation of multiscale features, the scale-level attention module weights the features from different input scales at each mode. The scale-level attention module consists of three convolution layers of 1 × 1 × 64, and one sigmoid layer to generate the weight scores for each scale. The generated score maps reflect the importance of scale-specific features at a specific position and scale for classifying the object in the tile.

Feature Extraction (Mode). After the mid-fusion with the scale-level attention module, the weighted feature maps get concatenated in intensity and range modes, separately. Then, they are passed through three convolution layers with the filter number of 128 and max-pooling layers. At this stage, the network is expected to extract more complex multiscale features in each mode. Depth-specific patterns can complement intensity patterns and help the overall model with this useful information.

Mid-Fusion with Mode-Level Attention Module. For the effective mid-fusion of complementary information of intensity and range data, we use a mode-level attention module that weights the mode-sensitive features extracted from intensity and range images, determining the contribution level of mode-sensitive features to the final classification output. In this way, the feature maps can be fused with different weights based on the contribution levels of road object classes, instead of being treated uniformly.

Feature Extraction (Classification). For each mode, the mode-level attention module outputs weight maps that are multiplied by the feature maps. The weighted feature maps get concatenated and passed through shared layers. Four convolution layers with the filter size of 256, 512, 512, and 1024 with two max-pooling layers are applied to extract higher-level multimodal features. Then the feature maps are flattened and passed to six fully-connected layers with the sizes 2048, 1024, 512, 256, 128, and 8.

*Classifier's Output.*The last fully-connected layer generates 8 numbers showing the probability of the 50 × 50 tile belonging to the 8 existing classes in the dataset. The higher the number is, the more probable the tile belongs to that specific pavement class. By assembling the predicted labels for the smallest tiles into the whole image,

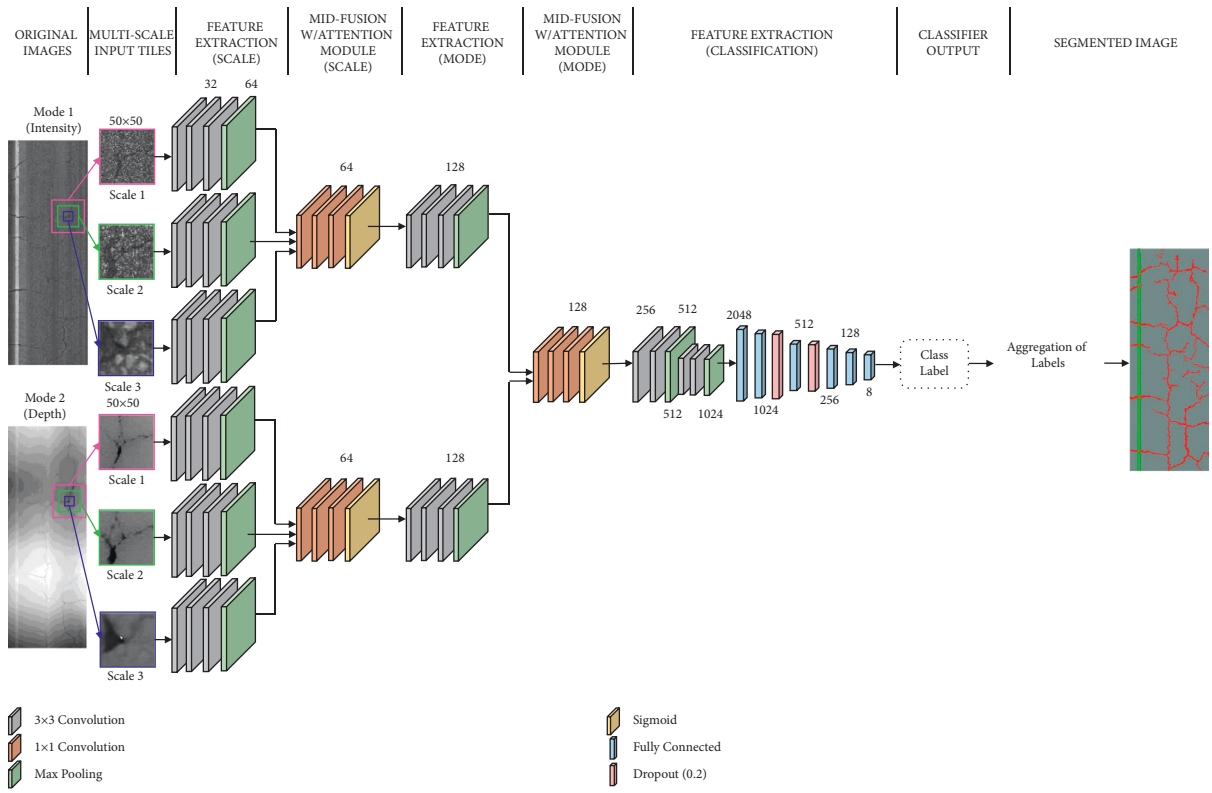


FIGURE 6: An overview of the DACNN. Range and intensity image tiles are generated at three scales to capture local and global information in each mode. The adaptive fusion of multiscale multimodal features is performed through scale-level and mode-level attention modules. The final class prediction for input tiles is assembled into the original image to create a mask.

the segmentation mask with the spatial resolution of $50 \times 50\,\text{mm}^2$ is created.

Effects of Range and Intensity Input Image Tiles. Range and intensity input images provide complementary information about road objects, which can improve object classification performance compared to intensity-only input images. Depth is a key feature for road object classification, such as cracks and potholes. These objects can be small or have a similar color and texture to the clean asphalt, and it makes them difficult to detect in gray-scale intensity images. However, they appear more clearly in range images due to their depth differences. Other pavement objects, such as markers, that have a distinct color or texture or do not have a significant depth can be easier to detect from intensity images. Figure 7 demonstrates the advantage of using intensity and range images over intensity images only containing markers, patches, and cracks.

4.2. Attention Modules. We design two types of attention modules as a mid-fusion strategy to adaptively aggregate multiscale multimodal features extracted from intensity and range image tiles. The mechanism of an attention module is to attend to relevant parts of input features, which is important for having a robust classification. The scale-level and mode-level attention modules enable the deep network to focus on visual representations that are more informative for the classification of the object in the input tile. Scale-level and mode-level modules incorporate both spatial and channel-wise attention into the network.

As illustrated in Figure 8(a), the scale-level attention module generates the score maps ($\mathbf{S}^m$) with the dimension of $C \times H \times W$ for each scale, where $m \in \{1, 2, 3\}$ is the scale number, C is the number of channels, W is the width, and H is the height of the input features ($\mathbf{F}^m$). The weighted feature maps, $\mathbf{F}^m$, are generated by the inner product of:

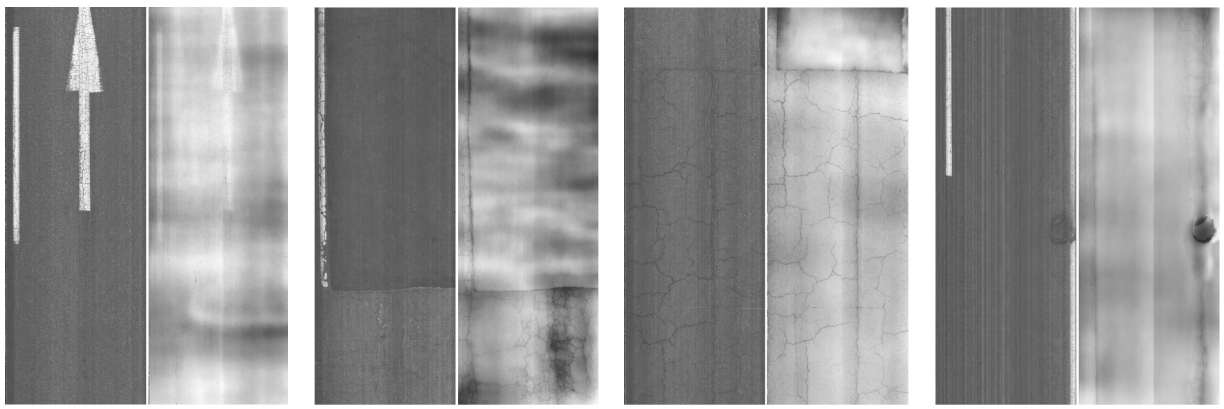
$$\mathbf{F}^m = \mathbf{F}^m . \mathbf{S}^m, \tag{1}$$

or

$$f^m_{w,h,c} = s^m_{w,h,c} . f^m_{w,h,c}, \tag{2}$$

where $\tilde{f}^m_{w,h,c}$ is the weighted feature at the spatial position (w, h) for the channel number c at the scale m; and $s_{w,h,c}$ is the score corresponding to the input feature $f^m_{w,h,c}$ at the spatial position (w, h) for the channel number c at the scale m. The attention module assigns a score between 0 and 1 to the feature maps of each scale in each channel and spatial position. Therefore, each element in the feature map $x_{w,h,c}$ is revised to $x_{w,h,c}$, in which scale, channel, and spatial information is considered. This module not only localizes the object spatially but also selects the most discriminative channel.

The mechanism of the mode-level attention module, shown in Figure 8(b), is similar to the scale-level one. In this module, the shared module among the modes generates the score maps for each mode to focus on the most



(a) (b) (c) (d)

Figure 7: Illustration of pavement objects in intensity and range images: (a) markers, (b) marker, patch, and cracks, (c) patch and cracks, and (d) marker and pothole.

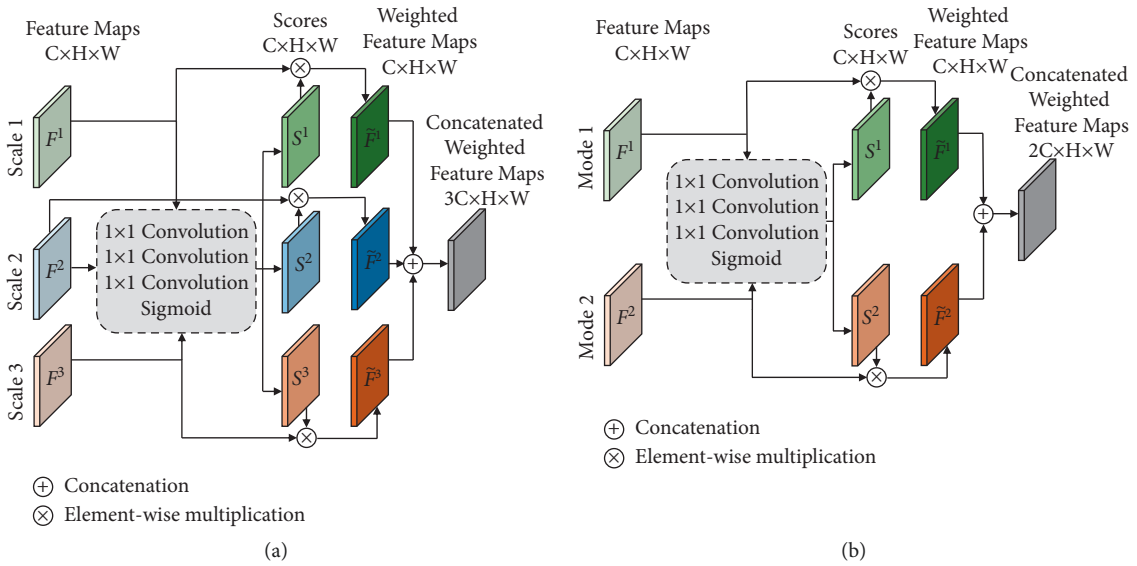


Figure 8: The details of (a) scale-level attention module and (b) mode-level attention module.

discriminative part of visual representations. The attention module assigns higher weights to the channel and regions of the mode features that are more relevant and informative for the classification step of that particular object.

4.3. Implementation Details. We train the classifiers in a fully supervised manner. The Adam optimizer with a learning rate of $\alpha = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1e{-}8$ is used, where β_1 and β_2 are exponential decay rates, and ε is a constant for numerical stability. The Adam optimizer inherits the advantages of other optimization algorithms, including the momentum feature of SGD and the adaptive learning feature of AdaDelta. The Adam optimizer also provides faster computation time and requires fewer parameters for tuning. The networks are trained for 800 epochs with a mini-batch size of 200. In each epoch, the network uses 60,000 random tiles out of more than 6 million tiles in the training dataset. The model with the best performance on loss for the validation dataset is selected as the model used in the testing mode. The training is conducted on an NVIDIA TitanX GPU with a memory configuration of 12 GB. The codes are implemented in Python 3.7.3 and TensorFlow 1.14.0.

5. Experiments

5.1. Baseline Models with Single-Scale Input Images and Results. Four different baseline classifiers, widely used in pavement applications, are trained to classify pavement image tiles into one of the existing 8 classes in the dataset. The deep CNNs compared in this study can be divided into three categories. (i) VGGNet was proposed by Simonyan and Zisserman [30] for ImageNet challenge 2014. The main idea behind VGGNet is to use filters with a small size (3×3), decreasing the number of parameters, and stack more of them to achieve the same receptive field as if a larger filter

were used. VGG16 and VGG19 have a total number of 16 and 19 convolutional and fully-connected layers, respectively. The deep architecture of VGGs is proved beneficial for image classification tasks. However, the gradient vanishing problem has appeared with the deeper architectures. (ii) ResNet proposed by He et al. [33] for ImageNet challenge 2015, alleviates the gradient vanishing problem by introducing skip-connections so that the input in each layer is passed to the next layer. Using identity skip-connections as well as batch normalization allows for training deep networks. ResNet50 has a total number of 50 convolutional and fully-connected layers. (iii) DenseNet proposed by Huang et al. [86] in 2017, extends ResNet's idea by including skip-connections from all previous layers. The dense concatenation to all subsequent layers preserves the features in preceding layers and allows for the classification of images in a wide range of scales. DenseNet121 has a total of 121 convolutional and is fully connected.

Figure 9 shows an overview of the deep networks used for pavement object classification in this study. The classifiers are trained with only intensity input tiles as well as intensity and range input tiles to evaluate the effect of exploiting depth information along with intensity information. As shown in Figure 9(a), 50×50 image tiles are generated and are concatenated as a 3-channel image to train the deep networks with only intensity images. When training the networks with both intensity and range images, as shown in Figure 9(b), 50×50 image tiles of each mode are concatenated at the input level as a 2-channel image (early fusion) and fed to the network.

Table 2 summarizes the results for all classifiers using (i) only intensity and (ii) intensity and range input pavement tiles. The performance of each classifier is evaluated on each pavement object and on average in terms of precision, recall, and F-score.

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP}, \\ \text{Recall} &= \frac{TP}{TP + FN}, \\ F * \text{score} &= \frac{2TP}{2TP + FP + FN}, \end{aligned} \tag{3}$$

where TP, FP, and FN are true positives, false positives, and false negatives, respectively. The precision determines how many of positive predictions are really positive, while the recall shows the ability of the network in predicting all the relevant instances. The F-score is a harmonic mean of precision and recall that is a useful measure to find the balance between these two metrics. The results show that using both range and intensity images improves the performance of all classifiers in terms of overall precision, recall, and F-score.

In more detail, we compare the baseline models' performances for different classes when they are trained with intensity-only images and intensity-range images. To interpret the results, we divide the classes into two categories: (i) the pavement objects having a height difference with adjacent pixels including crack, crack seal, pothole, manhole, and patch; (ii) pavement objects having no significant height difference with adjacent pixels including marker, curbing, and asphalt. Using range-intensity input images improved the performance of VGG16, VGG19, ResNet50, and DenseNet121 on the first category of objects, including crack, crack seal, pothole, manhole, and patch, on average by 18.8%, 20.6%, 11.9%, and 14.5% in terms of F-score. The average improvement of the baseline models on crack, crack seal, patch, pothole, and manhole are 12.6%, 22.5%, 21.6%, 22.1%, and 3.6% in terms of F-score. The lower improvement of manhole classification compared to the other four objects comes from the fact that manholes have distinct shapes and textures in intensity images. Therefore, providing range data as complementary information to the network has a milder effect. Incorporating range images into the network barely changes the performance of baseline models on the classification of pavement objects in the second category. In fact, the range image of marking, curbing, and asphalt provide no extra information to the networks for the classification task.

Providing depth information to the DACNN improves the classification results on the first category of objects by 3.2% in terms of F-score. In more detail, utilizing range-intensity images increases the performance of the DACNN on the classification of crack, crack seal, patch, pothole, and manhole by 2.4%, 7.8%, 1.2%, 2.6%, and 2.3% in terms of F-score, respectively. The improvement of DACNN performance by adding depth information is less than such improvement in baseline models. This is because of the high performance of the trained DACNN with intensity-only images which creates less capacity for improvements. As shown in Table 2, the average F-score for DACNN with intensity-only images is 92.9% while the number for VGG16, VGG19, ResNet50, and DenseNet121 is 59.9%, 59.9%, 62%, and 63.4%, respectively. The DACNN also outperforms VGG16, VGG19, ResNet50, and DenseNet121 on average by 23.3%, 22%, 25.4%, and 22.4%, respectively, in terms of F-score when the networks are trained with range-intensity input data. The significant improvement of DACNN classification performance over the baseline models comes from encoding contextual information to the network and adaptively fusing the features through the attention modules. In section (5.2), we show that the performance of baseline models improves by providing multiscale input tiles to the networks. However, DACNN still outperforms those models by having an effective fusion strategy for combining multiscale multimodal features.

Figure 10 demonstrates sample segmentation at a spatial resolution of $50 \times 50\,\text{mm}^2$ for different algorithms when trained with intensity-only and intensity-range pavement tiles. It can be seen DACNN achieves the best results by extracting a robust representation of range and intensity images. In more detail, we can see that cracks at the top left corner of the image are identified better when the depth information is encoded into all the networks. Range data provide more distinctive features helping the networks to distinguish between foreground and background when intensity values are not distinctive.

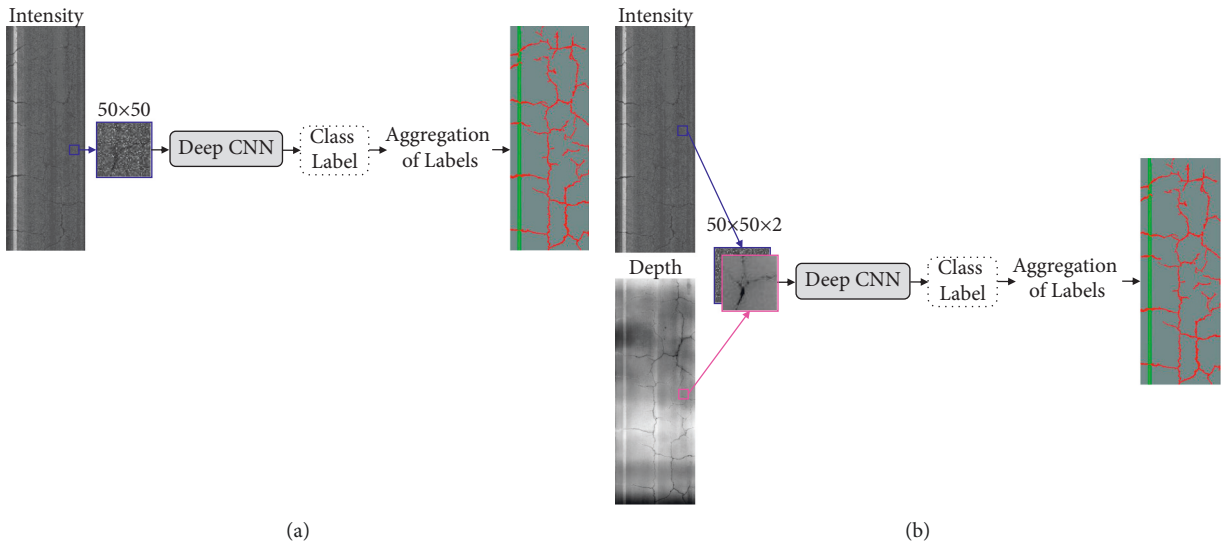


Figure 9: An overview of baseline classifiers trained with single-scale (a) intensity images, and (b) intensity and range images.

Table 2: Comparison of deep CNNs for classification of pavement objects using single-scale intensity and range input tiles.

Metric	Method	Input image	Crack	Crack seal	Patch	Pothole	Marker	Manhole	Curbing	Asphalt	Avg
Precision	VGG16	Intensity	0.660	0.533	0.676	0.529	0.923	0.850	0.947	0.945	0.758
		Intensity + Range	0.732	0.684	0.849	0.637	0.944	0.875	0.956	0.959	0.830
	VGG19	Intensity	0.593	0.577	0.653	0.636	0.932	0.840	0.944	0.95	0.766
		Intensity + Range	0.670	0.707	0.823	0.570	0.940	0.887	0.955	0.963	0.814
	ResNet50	Intensity	0.637	0.702	0.661	0.554	0.928	0.831	0.945	0.949	0.776
		Intensity + Range	0.679	0.777	0.793	0.593	0.931	0.879	0.954	0.959	0.821
	DenseNet121	Intensity	0.647	0.529	0.688	0.543	0.928	0.850	0.946	0.949	0.760
		Intensity + Range	0.737	0.875	0.789	0.608	0.935	0.895	0.952	0.961	0.844
	DACNN (ours)	Intensity	0.864	0.897	0.965	0.947	0.966	0.919	0.983	0.986	0.941
		Intensity + Range	**0.887**	**0.942**	**0.972**	**0.953**	**0.971**	**0.965**	**0.993**	**0.987**	**0.959**
Recall	VGG16	Intensity	0.256	0.018	0.345	0.253	0.917	0.646	0.985	0.988	0.551
		Intensity + Range	0.438	0.241	0.563	0.463	0.908	0.669	0.988	0.985	0.657
	VGG19	Intensity	0.373	0.033	0.348	0.139	0.904	0.669	0.991	0.981	0.555
		Intensity + Range	0.480	0.209	0.596	0.571	0.920	0.727	0.986	0.983	0.684
	ResNet50	Intensity	0.33	0.073	0.364	0.271	0.911	0.618	0.990	0.984	0.568
		Intensity + Range	0.453	0.163	0.533	0.450	0.925	0.645	0.988	0.984	0.643
	DenseNet121	Intensity	0.324	0.100	0.363	0.319	0.912	0.655	0.990	0.984	0.581
		Intensity + Range	0.431	0.218	0.648	0.528	0.926	0.676	0.992	0.985	0.676
	DACNN (ours)	Intensity	0.780	0.837	0.942	0.909	0.957	**0.937**	**0.990**	0.991	0.918
		Intensity + Range	**0.805**	**0.947**	**0.958**	**0.956**	**0.966**	**0.937**	**0.990**	**0.993**	**0.944**
F-score	VGG16	Intensity	0.369	0.034	0.457	0.343	0.920	0.734	0.966	0.966	0.599
		Intensity + Range	0.548	0.356	0.677	0.536	0.926	0.758	0.972	0.972	0.718
	VGG19	Intensity	0.458	0.063	0.454	0.223	0.918	0.745	0.967	0.965	0.599
		Intensity + Range	0.560	0.323	0.691	0.602	0.930	0.799	0.970	0.973	0.731
	ResNet50	Intensity	0.434	0.133	0.469	0.364	0.919	0.709	0.967	0.966	0.620
		Intensity + Range	0.543	0.269	0.638	0.512	0.928	0.744	0.971	0.971	0.697
	DenseNet121	Intensity	0.432	0.169	0.475	0.402	0.920	0.740	0.967	0.966	0.634
		Intensity + Range	0.544	0.349	0.712	0.566	0.930	0.770	0.971	0.973	0.727
	DACNN	Intensity	0.820	0.866	0.953	0.928	0.961	0.928	0.986	0.988	0.929
		Intensity + Range	**0.844**	**0.944**	**0.965**	**0.954**	**0.969**	**0.951**	**0.992**	**0.990**	**0.951**

5.2. Baseline Models with Multiscale Input Images and Results. Figure 11 shows an overview of the deep networks trained with multiscale input tiles to classify pavement objects. The multiscale image tiles are generated at three scales, 50 × 50, 250 × 250, and 500 × 500, for each mode of intensity and depth. As shown in Figure 11(a), multiscale tiles are concatenated as a 3-channel image to train the deep networks with only intensity images. When training the networks with

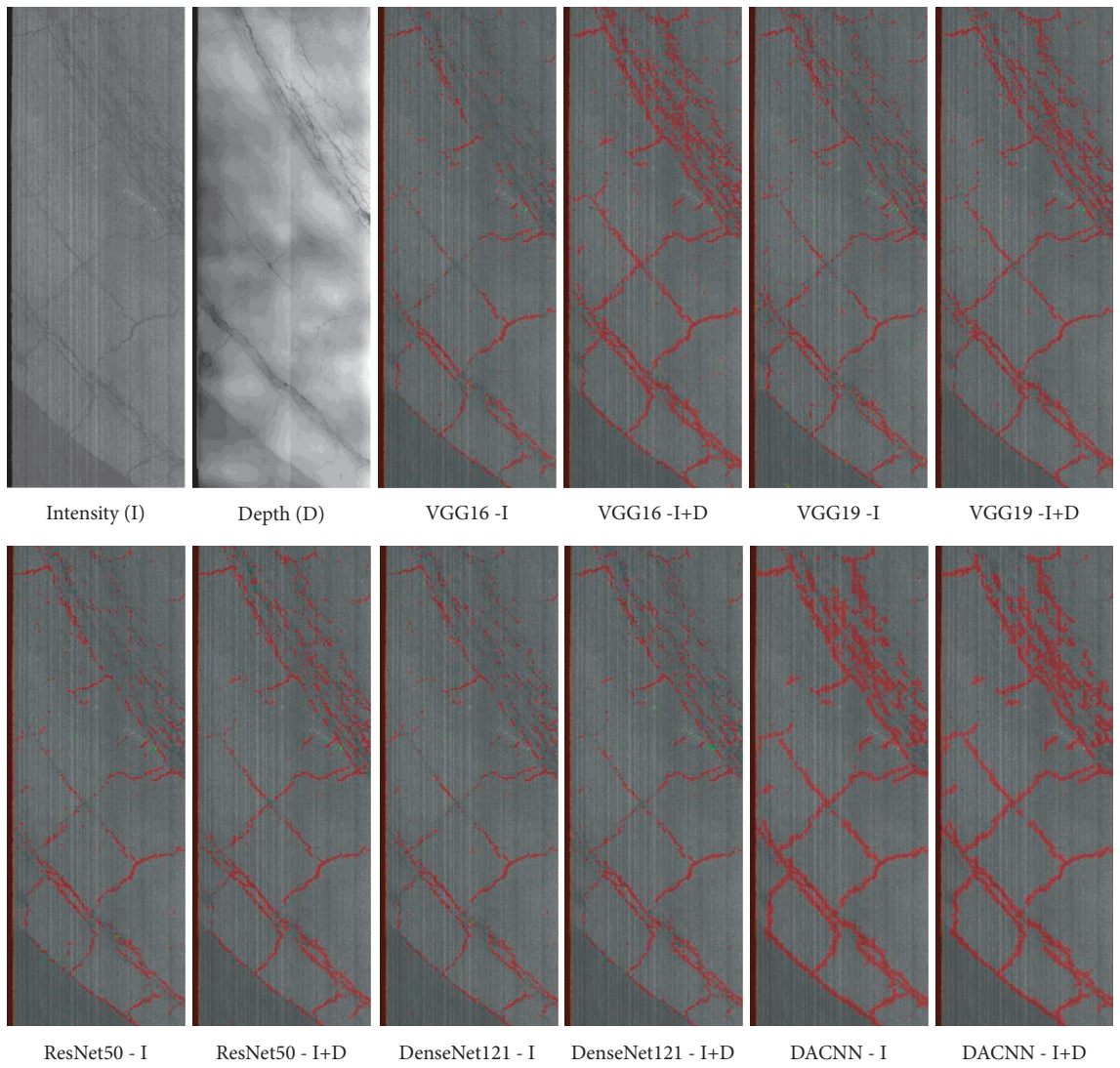


Figure 10: Classification results of road cracks using different algorithms trained with intensity-only and intensity-range images. Segmentation masks are created by aggregating classification results of 50 × 50 tiles.

both intensity and range images, as shown in Figure 11(b), the 3-channel image of each mode are merged at the input level (early fusion) and fed to the network.

Table 3 summarizes the performance of baseline models on the classification of 8 pavement classes in terms of precision, recall, and F-score. Comparing the results with the single-scale version of the networks, incorporating the contextual information into the networks improves the average F-score of VGG16, VGG19, ResNet50, and DenseNet121 by 28.3%, 29.3%, 24.3%, and 24.4%, respectively, when trained with intensity-only images. Furthermore, extracting depth features along with intensity features increases the average F-score of the VGG16, VGG19, ResNet50, and DenseNet121 by 4.1%, 3.4%, 4%, and 5.1%, respectively.

Although encoding the contextual information and incorporating the depth data into the network significantly enhances the performance of the baseline models, the DACNN classifies the objects more robustly by having an effective mid-fusion strategy. The DACNN outperforms VGG16, VGG19, ResNet50, and DenseNet121 trained with multiscale multimodal features by 2.8%, 2.5%, 4.8%, and 2.2%, respectively, on average in terms of F-score. More specifically, the DACNN improves the crack classification (as one of the most important distress types in pavement condition assessment) by 8.8%, 7.2%, 8.7%, and 7% in terms of F-score compared to VGG16, VGG19, ResNet50, and DenseNet121, respectively. This demonstrates the effectiveness of attention modules for pavement object classification.

6. Discussion

6.1. Qualitative and Quantitative Analysis of DACNN. One of the most important comparison metrics to evaluate the performance of multiclass classification models is their capability to distinguish between classes. AUC (Area under the Curve) of ROC (Receiver Operating Characteristics) is

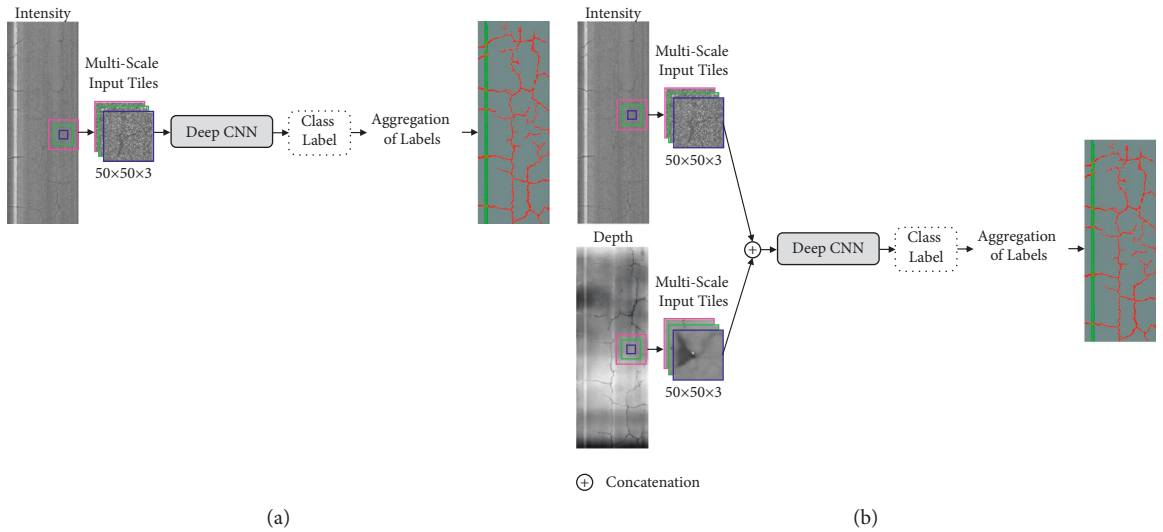


Figure 11: An overview of baseline classifiers trained with (a) multiscale intensity images and (b) multiscale intensity and range images.

Table 3: Comparison of deep CNNs for classification of pavement objects using multiscale intensity and range input tiles.

Metric	Method	Input image	Crack	Crack seal	Patch	Pothole	Marker	Manhole	Curbing	Asphalt	Avg
Precision	M-VGG16	Intensity	0.755	0.850	0.874	0.865	0.949	0.930	0.982	0.977	0.898
		Intensity + Range	0.824	0.966	0.929	0.920	0.958	0.933	0.990	0.982	0.938
	M-VGG19	Intensity	0.775	0.883	0.859	0.902	0.959	0.959	0.984	0.977	0.912
		Intensity + Range	0.786	0.904	0.92	0.914	0.959	0.910	0.986	0.985	0.921
	M-ResNet50	Intensity	0.772	0.894	0.798	0.885	0.952	0.952	0.976	0.975	0.901
		Intensity + Range	0.781	0.940	0.932	0.898	0.949	0.917	0.980	0.984	0.923
	M-DenseNet121	Intensity	0.756	0.839	0.869	0.883	0.960	0.934	0.981	0.975	0.900
		Intensity + Range	0.811	0.889	0.972	0.920	0.961	0.933	0.983	0.984	0.932
	DACNN (ours)	Intensity	0.864	0.897	0.965	0.947	0.966	0.919	0.983	0.986	0.941
		Intensity + Range	**0.887**	**0.942**	**0.972**	**0.953**	**0.971**	**0.965**	**0.993**	**0.987**	**0.959**
Recall	M-VGG16	Intensity	0.640	0.771	0.838	0.854	0.965	0.910	0.985	0.984	0.868
		Intensity + Range	0.699	0.831	0.954	0.919	0.967	0.946	0.988	0.989	0.912
	M-VGG19	Intensity	0.632	0.773	0.883	0.873	0.959	0.906	0.985	0.986	0.875
		Intensity + Range	0.758	0.920	0.952	0.920	0.961	0.966	0.989	0.985	0.931
	M-ResNet50	Intensity	0.606	0.617	0.853	0.775	0.961	0.897	0.985	0.984	0.835
		Intensity + Range	0.735	0.693	0.918	0.891	0.970	0.927	0.989	0.985	0.889
	M-DenseNet121	Intensity	0.636	0.766	0.824	0.810	0.952	0.906	0.984	0.985	0.858
		Intensity + Range	0.741	0.929	0.917	0.943	0.962	0.946	0.990	0.988	0.927
	DACNN (ours)	Intensity	0.780	0.837	0.942	0.909	0.957	**0.937**	**0.990**	0.991	0.918
		Intensity + Range	**0.805**	**0.947**	**0.958**	**0.956**	**0.966**	**0.937**	**0.990**	**0.993**	**0.944**
F-score	M-VGG16	Intensity	0.693	0.808	0.856	0.859	0.958	0.920	0.984	0.981	0.882
		Intensity + Range	0.756	0.893	0.941	0.920	0.961	0.940	0.989	0.985	0.923
	M-VGG19	Intensity	0.696	0.824	0.871	0.887	0.959	0.931	0.985	0.982	0.892
		Intensity + Range	0.772	0.912	0.936	0.917	0.960	0.937	0.988	0.985	0.926
	M-ResNet50	Intensity	0.679	0.730	0.825	0.827	0.956	0.924	0.980	0.980	0.863
		Intensity + Range	0.757	0.797	0.925	0.895	0.960	0.922	0.984	0.985	0.903
	M-DenseNet121	Intensity	0.691	0.801	0.846	0.845	0.956	0.919	0.982	0.980	0.878
		Intensity + Range	0.774	0.908	0.944	0.932	0.962	0.940	0.986	0.986	0.929
	DACNN	Intensity	0.820	0.866	0.953	0.928	0.961	0.928	0.986	0.988	0.929
		Intensity + Range	**0.844**	**0.944**	**0.965**	**0.954**	**0.969**	**0.951**	**0.992**	**0.990**	**0.951**

a measure of how strongly the classifier separates the classes. Higher the AUC, the better the model is capable of predicting true classes. To evaluate the DACNN performance, ROC curves for all investigated methods are plotted in Figure 12. Comparing the AUC values, DACNN demonstrates a stronger ability to separate classes while predicting the pavement objects.

Figure 13 shows segmentation samples of DACNN generated by integrating classified pavement tiles. The corresponding heatmaps for the pavement classes are also demonstrated for qualitative comparisons. A hotter color

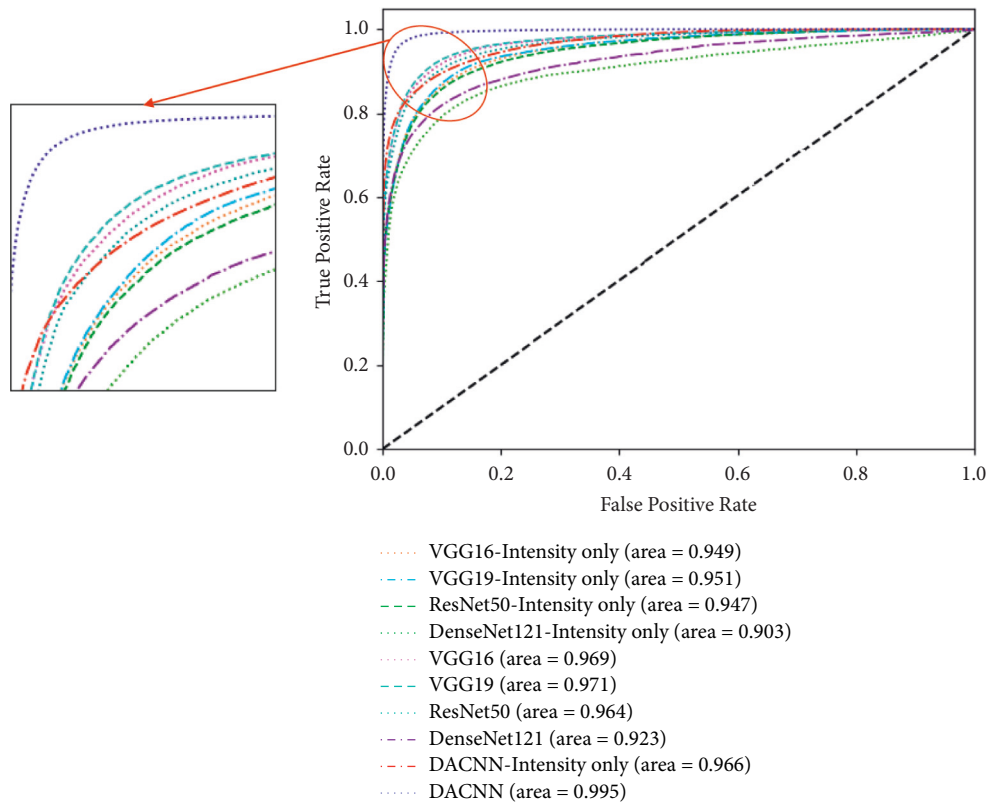


Figure 12: Receiver Operating Characteristic (ROC) curves. The presented DACNN achieves the highest area under the curve (AUC).

means a greater probability that the pixels belong to the corresponding class. The heatmaps reveal that the DACNN predicts the pavement object robustly with a strong separation from the rest of the objects.

Figure 14 visualizes the performance of the classifiers in terms of TP, TN, FP, and FN. Having the networks' predictions, we are able to analyze their performance in more detail. Especially in pavement applications, we care about not only increasing TPs but also decreasing FNs and FPs simultaneously. The reason is coming from: (i) having a high FN means that positive distresses are missed leading to an underestimation for road condition assessment, which is dangerous for safety considerations; (ii) having a high FP means that pavement tiles are misclassified as distresses leading to an overestimation, which is not cost-efficient for road assessment. As we can see in Figure 14, DACNN not only increases TPs but also significantly reduces FPs and FNs compared to all other methods. Other than DACNN which presents the best results, encoding depth information into all other networks also increases TPs and reduces FPs and FNs. For the pavement objects with a more distinctive representation in range images including cracks, crack seals, patches, potholes, and manholes, the improvements are more significant after combining the range data with intensity images. Figure 14 shows that DACNN generates the largest number of FPs and FNs for the crack classification. The reason mainly comes from the low contrast between cracks and the background within pavement images. Figure 15 demonstrates examples of DACNN predictions with FPs and FNs on crack classification.

6.2. Contrast Enhancement. As described in section 3.2, a histogram equalization technique, CLAHE, is employed to adjust the intensity values and improve the contrast in range images. CLAHE is a modified version of adaptive histogram equalization that limits the contrast to avoid over-amplification and noises in the images. Cliplimit value is the threshold defined to apply a limit over the image contrast. In this study, we conducted a grid search to optimize this hyperparameter for DACNN algorithm. Table 4 summarizes the DACNN performance while using different cliplimit values. Considering the F-score values, cliplimit = 4 is used as the threshold value for CLAHE.

6.3. Computational Cost. We compare the computational cost of investigated algorithms in this study in two cases: (i) The networks are trained with only intensity input tiles; (ii) The networks are trained with both intensity and range input images. This way, we can examine how encoding depth information to the networks affects the computational costs. To highlight the trade-off between performance and speed, our proposed method, DACNN is also compared to the baseline approaches. Table 5 summarizes the computational costs for different classification approaches used in this study, in terms of the number of trainable variables, training time per epoch, and inference time for 100 batches. While the first column presents the costs for intensity-only trained networks, including VGG16, VGG19, ResNet50, and DenseNet121, the second column presents the costs for the same networks trained with both intensity and range images.

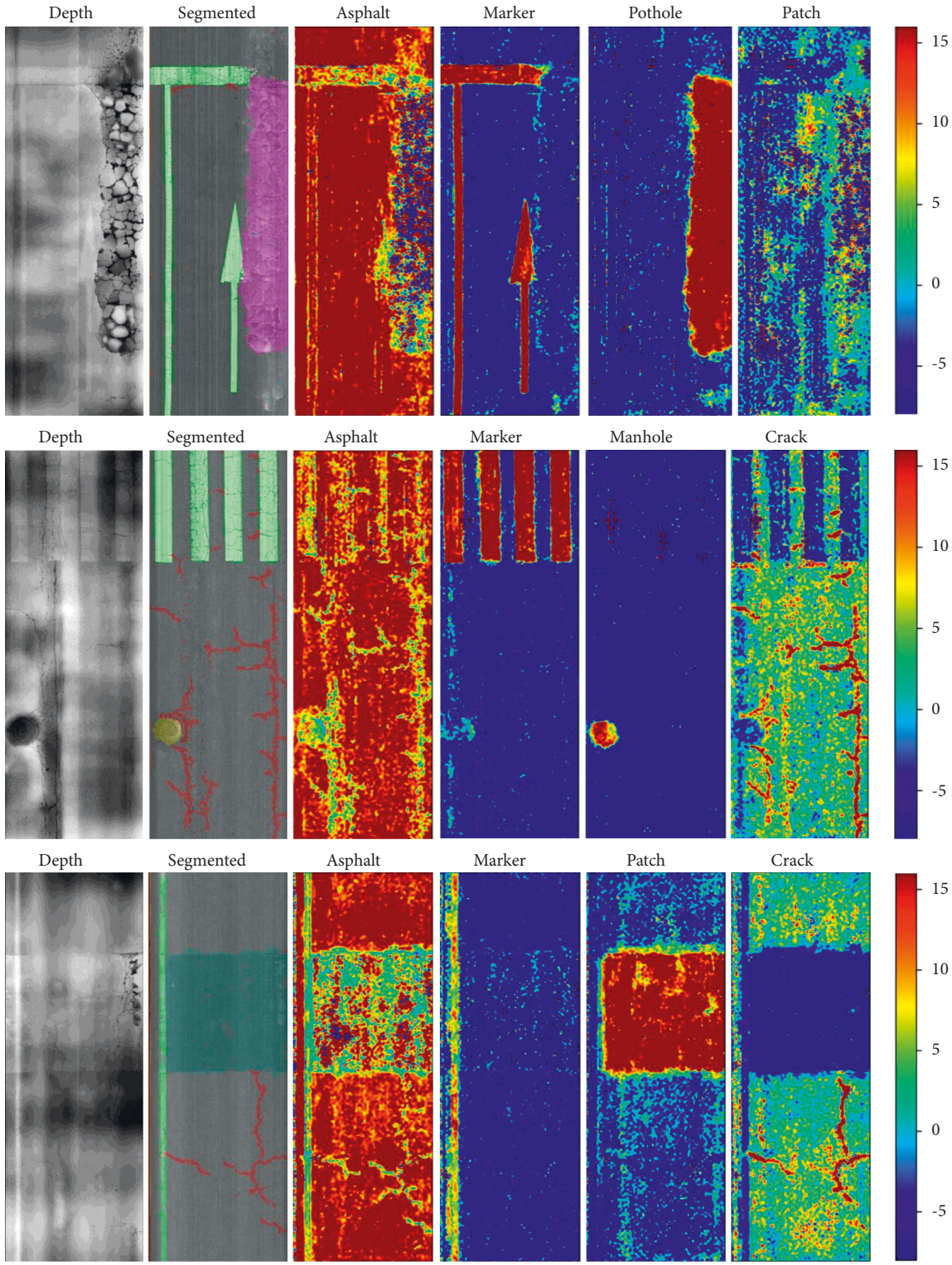


Figure 13: Classification results of the DACNN. Hotter colors mean a greater probability that the pixels belong to the specified class.

Comparing the first two columns reveals that the extra computational costs brought by encoding depth information to the baseline models were almost negligible. However, the average F-score increased by 16.5% for objects with discriminative features in the range of images (crack, crack seal, pothole, manhole, and patch). The third column shows the computational costs for DACNN when the depth branch is removed, and the last column shows the cost for DACNN trained with both intensity and range images. It can be concluded that by providing a limited extra source of computations, we can improve the classification results. Training with intensity-only, DACNN enhances the

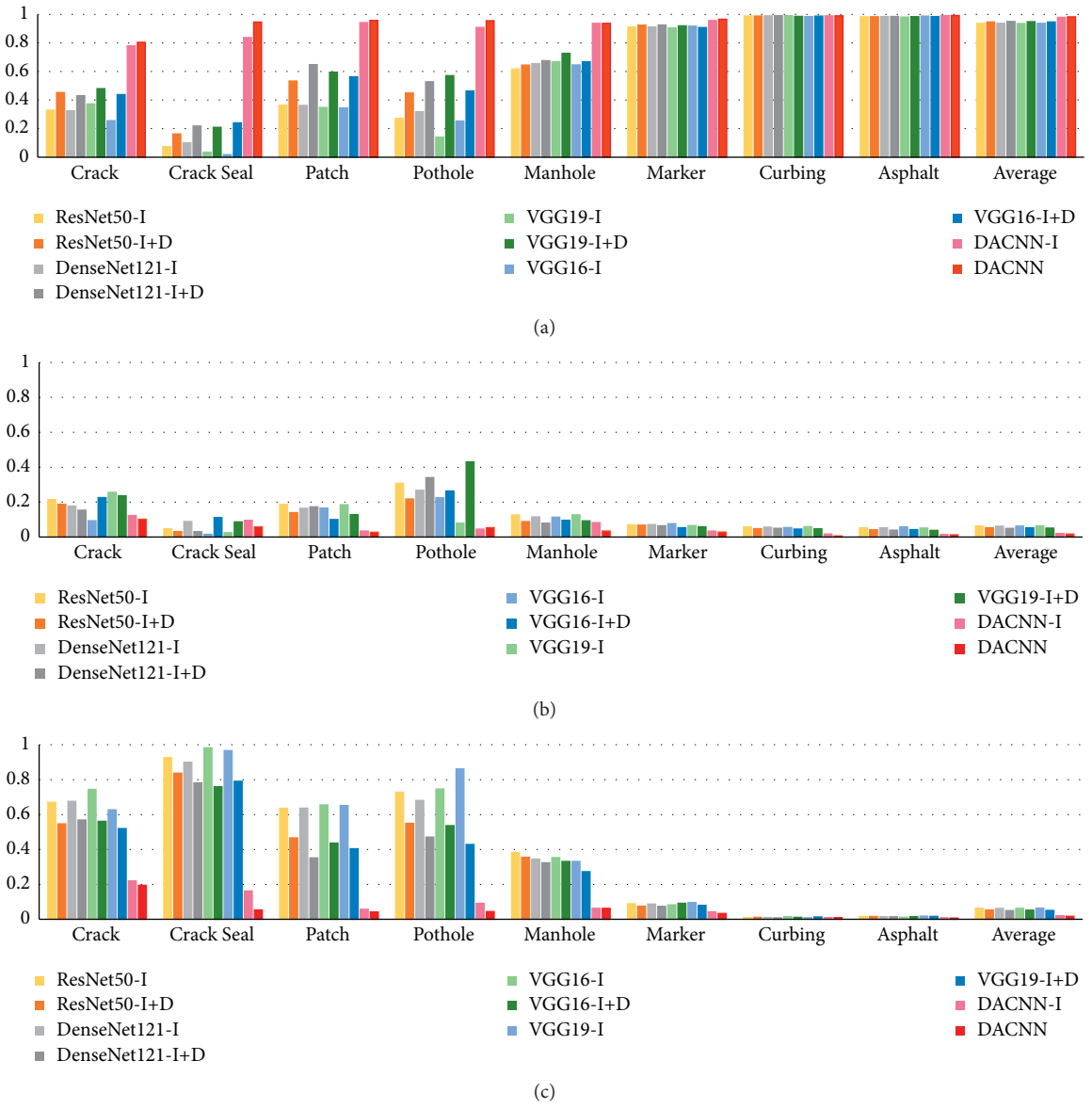


Figure 14: Normalized (a) TP, (b) FP, and (c) FN of each class using different algorithms.

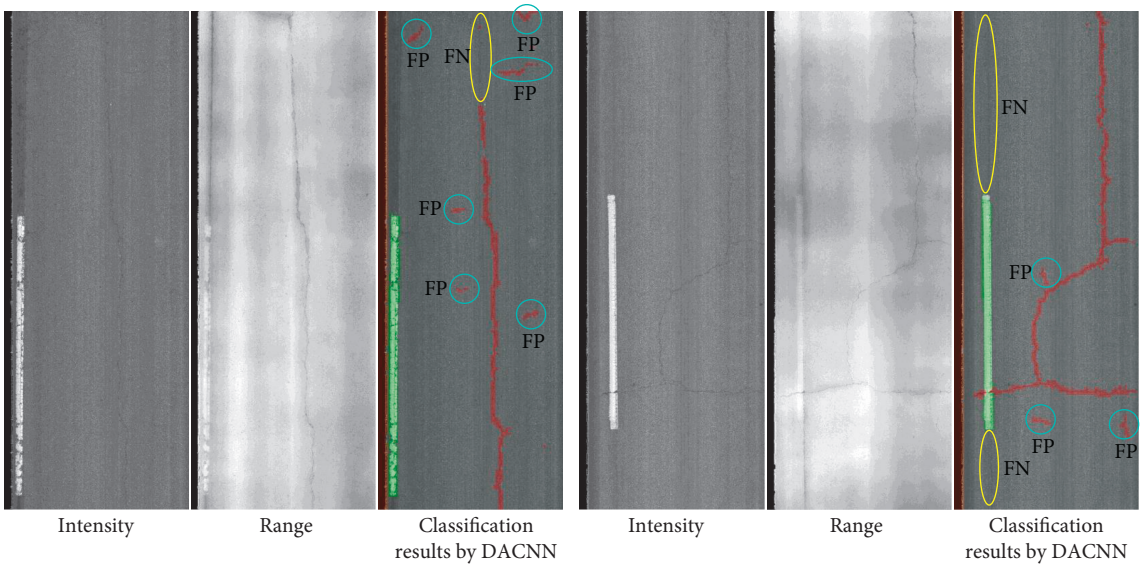


Figure 15: Examples of DACNN predictions with FPs and FNs on crack classification.

Table 4: Effect of different threshold values for histogram equalization on DACNN results.

Metric	Cliplimit = 2	Cliplimit = 3	Cliplimit = 4	Cliplimit = 5
Precision	0.924	0.973	**0.959**	0.967
Recall	0.913	0.878	**0.944**	0.906
F-score	0.918	0.923	**0.951**	0.936

Table 5: Comparison of computational costs for different classification approaches.

Computational costs	Baselines (intensity-only)	Baselines (intensity-range)	DACNN (intensity-only)	DACNN (intensity-range)
Number of parameters	51 M	51 M	61 M	63 M
Training time/epoch	67.1 s	82.6 s	175.6 s	247.5 s
Inference time/100 batches	4.2 s	4.3 s	8.7 s	10.2 s

classification results by capturing contextual information by 31.6% in F-score compared to the baseline methods (first vs. third column). Training with both intensity and range, DACNN improves the classification results by an adaptive fusion strategy by 23.3% in F-score compared to the baseline methods (second vs. fourth column). It should be noted that DACNN is not developed with the goal of having a real-time classification. In most practices, automated assessments of road conditions are performed offline where accuracy and robustness are the most important factors.

7. Conclusions

A deep learning-based model termed DACNN is presented to improve the performance of multiclass classification for road objects. Both intensity and range images are fed to the DACNN to enrich the image representation learned by the network. Discriminant feature representations obtained by encoding range images help the network to capture complex topology and to handle noises and illumination variances. Furthermore, feeding multiscale input images into the DACNN enables the network to catch both local and global fields of view, which is beneficial for classifying pavement objects with various sizes and shapes. We designed dual attention modules as an effective way to fuse scale-specific and mode-specific features to model the semantic interdependencies in spatial and channel dimensions. The position attention selectively aggregates the feature at each position by a weighted sum of the features at all positions, and channel attention selectively emphasizes interdependent channel maps by integrating associated features among all channel maps. This way, the network learns better the relevant content for each specific object at each scale and mode contributing to more precise classification results.

The effectiveness and feasibility of the DACNN were compared with four baseline CNN models. The comparison results showed that the DACNN outperforms all compared CNNs. The results also showed that encoding depth information into the networks improves the classification results of VGG16, VGG19, ResNet50, DenseNet121, and the DACNN by 11.9%, 13.2%, 7.7%, 9.3%, and 2.2% in terms of averaged F-score, respectively, compared to when these models are trained with intensity-only images. The classification improvements are even more significant for pavement objects that are distinctive in range images by having height differences with neighboring pixels. For example, incorporating depth data with intensity information improves the crack classification by 17.9%, 10.2%, 10.9%, 11.2%, and 2.4% in terms of averaged F-score in VGG16, VGG19, ResNet50, DenseNet121, and the DACNN, respectively. In addition to encoding depth data, DACNN yields more improvements by capturing global context through multiscale input tiles, as well as focusing on the most important feature representations through attention modules. The DACNN outperforms VGG16, VGG19, ResNet50, and DenseNet121 by 23.3%, 22%, 25.4%, and 22.4%, respectively, in terms of averaged F-score, while they are all trained with range-intensity tiles.

Although the developed DACNN achieves great performance in pavement object classification, some limitations still exist in our model. Therefore, extra effort is required to make our model more practical and effective. Firstly, our model classifies 50×50 pavement tiles into different categories. Although $50 \times 50\ mm^2$ spatial resolution is acceptable in most road surveys, a pixel-level segmentation is required for some pavement applications such as crack width measurements. Secondly, quantifying the severity of pavement distresses is of necessity for road condition assessment, but it cannot be obtained directly from our model. Lastly, self-attention mechanisms capturing long-range dependencies in the network can be explored for further improvements. Furthermore, one can conduct hyperparameter studies for the training of the network and provide quantitative comparisons.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no potential conflicts of interest.

Acknowledgments

This project was partially supported by Korea Institute of Civil Engineering and Building Technology (KICT). Data Transfer Solution (DTS) partially helped in the preparation of the ground-truth dataset used in this study.

References

[1] K. A. Zimmerman, "Pavement management systems: Putting data to work," vol. 20-05, 2017, https://www.trb.org/Publications/Blurbs/175607.aspx.

[2] S. Mathavan, K. Kamal, and M. Rahman, "A review of three-dimensional imaging technologies for pavement distress detection and measurements," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2353–2362, 2015.

[3] T. B. Coenen and A. Golroo, "A review on automated pavement distress detection methods," *Cogent Engineering*, vol. 4, no. 1, Article ID 1374822, 2017.

[4] W. Cao, Q. Liu, and Z. He, "Review of pavement defect detection methods," *IEEE Access*, vol. 8, pp. 14531–14544, 2020.

[5] A. Ragnoli, M. R. De Blasiis, and A. Di Benedetto, "Pavement distress detection methods: a review," *Infrastructure*, vol. 3, no. 4, 2018.

[6] N. S. P. Peraka and K. P. Biligiri, "Pavement asset management systems and technologies: a review," *Automation in Construction*, vol. 119, Article ID 103336, 2020.

[7] H. S. Munawar, A. W. A. Hammad, A. Haddad, C. A. P. Soares, and S. T. Waller, "Image-based crack detection methods: a review," *Infrastructure*, vol. 6, no. 8, p. 115, 2021.

[8] Y.-A. Hsieh and Y. J. Tsai, "Machine learning for crack detection: review and model performance comparison," *Journal of Computing in Civil Engineering*, vol. 34, no. 5, Article ID 04020038, 2020.

[9] K. Gopalakrishnan, "Deep learning in data-driven pavement image analysis and automated distress detection: a review," *Data*, vol. 3, no. 3, p. 28, 2018.

[10] M. Pak and S. Kim, "A review of deep learning in image recognition," in *Proceedings of the 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*, pp. 1–3, IEEE, Kuta Bali, Indonesia, August 2017.

[11] S. Zhou and W. Song, "Deep learning-based roadway crack classification using laser-scanned range images: a comparative study on hyperparameter selection," *Automation in Construction*, vol. 114, Article ID 103171, 2020.

[12] X. Wu, D. Sahoo, and S. C. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, 2020.

[13] A. Abdollahi and B. Pradhan, "Integrating semantic edges and segmentation information for building extraction from aerial images using unet," *Machine Learning with Applications*, vol. 6, Article ID 100194, 2021.

[14] J. Tang and Y. Gu, "Automatic crack detection and segmentation using a hybrid algorithm for road distress analysis," in *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 3026–3030, IEEE, Manchester, UK, October 2013.

[15] A. Ayenu-Prah and N. Attoh-Okine, "Evaluating pavement cracks with bidimensional empirical mode decomposition," *EURASIP Journal on Applied Signal Processing*, vol. 2008, pp. 861701–861707, 2008.

[16] S. Chambon, P. Subirats, and J. Dumoulin, "Introduction of a wavelet transform based on 2d matched filter in a Markov random field for fine structure extraction: application on road crack detection," *Image Processing: Machine Vision Applications II*, vol. 7251, p. 72510A, 2009.

[17] L. Sun and Z. Qian, "Multi-scale wavelet transform filtering of non-uniform pavement surface image background for automated pavement distress identification," *Measurement*, vol. 86, pp. 26–40, 2016.

[18] Y. Hu and C.-x. Zhao, "A novel lbp based methods for pavement crack detection," *Journal of pattern Recognition research*, vol. 5, no. 1, pp. 140–147, 2010.

[19] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[20] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational Intelligence and Neuroscience*, vol. 2018, 2018.

[21] J. Chai, H. Zeng, A. Li, and E. W. Ngai, "Deep learning in computer vision: a critical review of emerging techniques and application scenarios," *Machine Learning with Applications*, vol. 6, Article ID 100134, 2021.

[22] A. Alfarrarjeh, D. Trivedi, S. H. Kim, and C. Shahabi, "A deep learning approach for road damage detection from smartphone images," in *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data)*, pp. 5201–5204, IEEE, Seattle, WA, USA, December 2018.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[24] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, "Road damage detection and classification using deep neural networks with smartphone images," *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 12, pp. 1127–1141, 2018.

[25] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in *Proceedings of the European Conference on Computer Vision*, pp. 21–37, Springer, Germany, September 2016.

[26] L. Song and X. Wang, "Faster region convolutional neural network for automated pavement distress detection," *Road Materials and Pavement Design*, vol. 22, no. 1, pp. 23–41, 2021.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.

[28] B. Li, K. C. P. Wang, A. Zhang, E. Yang, and G. Wang, "Automatic classification of pavement crack using deep convolutional neural network," *International Journal of Pavement Engineering*, vol. 21, no. 4, pp. 457–463, 2020.

[29] K. Gopalakrishnan, S. K. Khaitan, A. Choudhary, and A. Agrawal, "Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection," *Construction and Building Materials*, vol. 157, pp. 322–330, 2017.

[30] K. Simonyan and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, 2014, https://arxiv.org/abs/1409.1556.

[31] S. L. H. Lau, E. K. P. Chong, X. Yang, and X. Wang, "Automated pavement crack segmentation using u-net-based

convolutional neural network," *IEEE Access*, vol. 8, pp. 114892–114899, 2020.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, Germany, November 2015.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[34] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[35] T. Chen, Z. Cai, X. Zhao et al., "Pavement crack detection and recognition using the architecture of segnet," *Journal of Industrial Information Integration*, vol. 18, Article ID 100144, 2020.

[36] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," 2014, https://arxiv.org/abs/1409.0473?source.

[37] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[38] Y. Tian, W. Hu, H. Jiang, and J. Wu, "Densely connected attentional pyramid residual network for human pose estimation," *Neurocomputing*, vol. 347, pp. 13–23, 2019.

[39] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "Adding attentiveness to the neurons in recurrent neural networks," in *Proceedings of the European Conference on Computer Vision*, pp. 135–151, (ECCV), Germany, October 2018.

[40] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: a neural network for large vocabulary conversational speech recognition," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4964, IEEE, Shanghai, China, March 2016.

[41] M. Sperber, J. Niehues, G. Neubig, S. Stüker, and A. Waibel, "Self-attentional acoustic models," 2018, https://arxiv.org/abs/1803.09519.

[42] K. Xu, J. Ba, R. Kiros et al., "Show, attend and tell: neural image caption generation with visual attention," in *International Conference on Machine Learning*, pp. 2048–2057, PMLR, Breckenridge, Colorado, USA, 2015.

[43] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 375–383, San Juan, PR, USA, 2017.

[44] S. Wang, L. Hu, L. Cao, X. Huang, D. Lian, and W. Liu, "Attention-based transactional context embedding for next-item recommendation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[45] H. Ying, F. Zhuang, F. Zhang et al., "Sequential recommender system based on hierarchical attention network," in *IJCAI International Joint Conference on Artificial Intelligence,* Stockholmsmässan, 2018.

[46] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2204–2212, 2014.

[47] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 2017–2025, 2015.

[48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[49] M. Stollenga, J. Masci, F. Gomez, and J. Schmidhuber, "Deep Networks with Internal Selective Attention through Feedback Connections," *Advances in neural information processing systems*, vol. 27, 2014.

[50] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, Salt Lake City, UT, USA, June 2018.

[51] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3286–3295, Seoul, Korea (South), October 2019.

[52] W. Song, G. Jia, D. Jia, and H. Zhu, "Automatic pavement crack detection and classification using multiscale feature attention network," *IEEE Access*, vol. 7, pp. 171001–171012, 2019.

[53] H. Wan, L. Gao, M. Su, Q. Sun, and L. Huang, "Attention-based convolutional neural network for pavement crack detection," *Advances in Materials Science and Engineering*, vol. 2021, Article ID 5520515, 13 pages, 2021.

[54] W. Qiao, Q. Liu, X. Wu, B. Ma, and G. Li, "Automatic pixel-level pavement crack recognition using a deep feature aggregation segmentation network with a scse attention mechanism module," *Sensors*, vol. 21, no. 9, p. 2902, 2021.

[55] W. Wang and C. Su, "Convolutional neural network-based pavement crack segmentation using pyramid attention network," *IEEE Access*, vol. 8, pp. 206 548–206 558, 2020.

[56] E. Eslami and H.-B. Yun, "Attention-based multi-scale convolutional neural network (a+ mcnn) for multi-class classification in road images," *Sensors*, vol. 21, no. 15, p. 5137, 2021.

[57] Q. Zhou, Z. Qu, and C. Cao, "Mixed pooling and richer attention feature fusion for crack detection," *Pattern Recognition Letters*, vol. 145, pp. 96–102, 2021.

[58] Z. Qu, W. Chen, S.-Y. Wang, T.-M. Yi, and L. Liu, "A crack detection algorithm for concrete pavement based on attention mechanism and multi-features fusion," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2021.

[59] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2net: A New Multi-Scale Backbone Architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, 2019.

[60] Y. Pan, G. Zhang, and L. Zhang, "A spatial-channel hierarchical deep learning network for pixel-level automated crack detection," *Automation in Construction*, vol. 119, Article ID 103357, 2020.

[61] H. Liu, X. Miao, C. Mertz, C. Xu, and H. Kong, "Crackformer: transformer network for fine-grained crack detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3783–3792, Montreal, BC, Canada, October 2021.

[62] A E1926-08, "Standard Practice for Computing International Roughness index of Roads from Longitudinal Profile Measurements," *West Conshohocken, PA: ASTM International*, 2015.

[63] ASTM, "Standard Test Method for Measuring the Longitudinal Profile of Traveled Surfaces with an Accelerometer

Established Inertial Profiling Reference," *Annual books of ASTM standards*, vol. 4, pp. 703–707, 1997.

[64] I O for Standardization, *Characterization of Pavement Texture by Use of Surface Profiles: Determination of Megatexture*, International Organization for Standardization, 2009.

[65] J. Laurent, D. Lefebvre, and E. Samson, "Development of a new 3d transverse laser profiling system for the automatic measurement of road cracks," in *Symposium on Pavement Surface Characteristics*, vol. 6th, Portoroz, Slovenia, 2008.

[66] J. Guan, X. Yang, L. Ding, X. Cheng, V. C. Lee, and C. Jin, "Automated pixel-level pavement distress detection based on stereo vision and deep learning," *Automation in Construction*, vol. 129, Article ID 103788, 2021.

[67] H. Lang, J. J. Lu, Y. Lou, and S. Chen, "Pavement cracking detection and classification based on 3d image using multi-scale clustering model," *Journal of Computing in Civil Engineering*, vol. 34, no. 5, Article ID 04020034, 2020.

[68] Y. C. J. Tsai and A. Chatterjee, "Pothole detection and classification using 3d technology and watershed method," *Journal of Computing in Civil Engineering*, vol. 32, no. 2, Article ID 04017078, 2018.

[69] Y.-C. J. Tsai, Y. Zhao, B. Pop-Stefanov, and A. Chatterjee, "Automatically detect and classify asphalt pavement raveling severity using 3d technology and machine learning," *International Journal of Pavement Research and Technology*, vol. 14, no. 4, pp. 487–495, 2021.

[70] Y. J. Tsai, Z. Wang, and F. Li, "Assessment of rut depth measurement accuracy of point-based rut bar systems using emerging 3d line laser imaging technology," *Journal of Marine Science and Technology*, vol. 23, no. 3, p. 8, 2015.

[71] Y. J. Tsai, Y. Wu, C. Ai, and E. Pitts, "Critical assessment of measuring concrete joint faulting using 3d continuous pavement profile data," *Journal of Transportation Engineering*, vol. 138, no. 11, pp. 1291–1296, 2012.

[72] F. Hong and Y. R. Huang, "Measurement and characterization of asphalt pavement surface macrotexture using three dimensional laser scanning technology," *Journal of Testing and Evaluation*, vol. 42, no. 4, pp. 20130147–20130890, 2014.

[73] R. Ghosh and O. Smadi, "Automated Detection and Classification of Pavement Distresses Using 3d Pavement Surface Images and Deep Learning," *Transportation Research Record*, vol. 2675, pp. 1359–1374, 2021.

[74] Z. Yang, X. Zhang, Y. Tsai, and Z. Wang, "Quantitative Assessments of Crack Sealing Benefits by 3d Laser Technology," *Transportation Research Record*, vol. 2675, pp. 103–116, 2021.

[75] Y. Fei, K. C. P. Wang, A. Zhang et al., "Pixel-level cracking detection on 3d asphalt pavement images through deep-learning-based cracknet-v," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 1, pp. 273–284, 2020.

[76] B. Li, K. C. P. Wang, A. Zhang, Y. Fei, and G. Sollazzo, "Automatic segmentation and enhancement of pavement cracks based on 3d pavement images," *Journal of Advanced Transportation*, vol. 2019, Article ID 1813763, pp. 1–9,, 2019.

[77] A. Zhang, K. C. P. Wang, Y. Fei et al., "Automated pixel-level pavement crack detection on 3d asphalt surfaces with a recurrent neural network," *Computer-Aided Civil and Infrastructure Engineering*, vol. 34, no. 3, pp. 213–229, 2019.

[78] R. Gui, X. Xu, D. Zhang, and F. Pu, "Object-based crack detection and attribute extraction from laser-scanning 3d profile data," *IEEE Access*, vol. 7, pp. 172728–172743, 2019.

[79] A. Zhang, K. C. P. Wang, B. Li et al., "Automated pixel-level pavement crack detection on 3d asphalt surfaces using a deep-learning network," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 10, pp. 805–819, 2017.

[80] A. Zhang, K. C. P. Wang, Y. Fei et al., "Deep learning–based fully automated pavement crack detection on 3d asphalt surfaces with an improved cracknet," *Journal of Computing in Civil Engineering*, vol. 32, no. 5, Article ID 04018041, 2018.

[81] Q. Li, D. Zhang, Q. Zou, and H. Lin, "3d laser imaging and sparse points grouping for pavement crack detection," in *Proceedings of the 25th European Signal Processing Conference (EUSIPCO)*, pp. 2036–2040, IEEE, Kos, Greece, September 2017.

[82] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy rate superpixel segmentation," in *Proceedings of the CVPR 2011*, pp. 2097–2104, IEEE, Colorado Springs, CO, USA, June 2011.

[83] W. Sultani, S. Mokhtari, and H.-B. Yun, "Automatic pavement object detection using superpixel segmentation combined with conditional random field," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 7, pp. 2076–2085, 2018.

[84] L. Wu, S. Mokhtari, A. Nazef, B. Nam, and H.-B. Yun, "Improvement of crack-detection accuracy using a novel crack defragmentation technique in image-based road assessment," *Journal of Computing in Civil Engineering*, vol. 30, no. 1, Article ID 04014118, 2016.

[85] S. K. Shome and S. R. K. Vadali, "Enhancement of diabetic retinopathy imagery using contrast limited adaptive histogram equalization," *International Journal of Computer Science and Information Technologies*, vol. 2, no. 6, pp. 2694–2699, 2011.

[86] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, Honolulu, HI, USA, July 2017.