Research Article

Deep Neural Network on Detection of Road Distress Using Mixture of Predicted and Observed Data

Nathan Lin Xiao[1], Wen Cheng[2]

[1]*Wuhan Yangtze International School, 10–1 Boxue Road, WEDZ, Wuhan, Hubei Province, China*
[2]*Civil Engineering Department, College of Engineering, California State Polytechnic University, Pomona, CA, USA*

ABSTRACT

Roadway distress detection is essential for ensuring a safe and comfortable driving environment. However, given the irregular shape, small area size, and occasionally very large number, of the road distress objects, it is often laborious to label the distress instances during the training process under the fully supervised algorithm. To address this issue, the study strives to apply semi-supervised learning for distress detection that claims to reduce the cost associated with the labeling process, while maintaining or even improving the learning accuracy in some situations. The research features three distinct backbones of Mask R-CNN models, Unmanned Aerial System imagery of two resolutions, three levels of pseudo-labeled data, eleven threshold values and two types of assessment (that is, in-resolution and out-of-resolution). The results demonstrate that semi-supervised Mask R-CNN models are effective in detecting road distress. Nonetheless, the sensitive analysis is recommended in the future research to identify the optimal pseudo ratio that could generate the highest prediction accuracy.

Keywords: Semi-supervised Learning; Unmanned Aerial System; Mask R-CNN Models; Data Label

INTRODUCTION

Roadway cracks are amid the most prevalently observed road surface degradations. Cracks not only reduce road pavement performance but also threaten traffic safety [1]. To ensure the quality of road pavement surface and to engender a safe environment for all roadway users, efficient and reliable detection and the maintenance of roadway cracks are very substantial. Early inspection and detection can help circumvent roadway damage and possible failure [2]. However, the traditional manual inspection highly depends on the inspector's engineering judgment and experience. It is prone to subjectivity since two inspectors can conclude different analyses for similar situations [3]. In addition, manual inspection is very time-consuming and labor-intensive [4-5]. Hence, there is an imperative need to utilize diverse techniques and better, reliable, and efficient roadway crack detection strategies.

To address the urgent need for concise roadway crack detection, there is considerable interest in developing efficient and reliable algorithms to automate object detection with the help of rapid advancement in technology, namely, computer graphics, deep learning,

Corresponding author: Nathan Lin Xiao, E-mail: nxiao25@students.wyischina.com.

Received June 12, 2024**; Accepted** June 22, 2024

and computer vision. Automation of crack detection leads to more objective and standardized rehabilitation decisions [6]. With an expeditious evolvement of image analysis, automated roadway crack detection has been widely explored over the past few decades [7-11]. In general, automated road crack detection methods benefit from various deep learning models of object detection, including single-stage models, two-stage object detection algorithms, and models based on encoder-decoder structure. Single-stage object detection models such as Single Shot MultiBox Detector (SSD) model [12-13] and You Only Look Once (YOLO) treat object detection as a simple regression problem by taking an image input and learning the class probabilities and bounding box coordinates [14]. On the other hand, two-stage object detection algorithms include Region-based Convolutional Neural Networks (R-CNN) [15], Faster R-CNN [16], and Mask R-CNN [17]. These models utilize a Region Proposal Network (RPN) to generate regions of interest (RoI) in the first stage and then send the region proposals down the pipeline for object classification and bounding-box regression. Compared to single-stage models, two-stage models reach higher levels of accuracy, but with longer computation time. Furthermore, models based on encoder-decoder structure, including U-Net [18], SegNet [19], Fully Convolutional Network (FCNs) [20], CrackSeg [21], use recurrent neural networks for sequence-to-sequence prediction problems. Encoder-decoder structure-based models tend to improve both efficiency and accuracy [22]. Another unique classification of object detection models includes fully supervised and semi-supervised models. Fully supervised learning is a subcategory of machine learning that utilizes fully labeled datasets to train its algorithms to classify data or predict outcomes accurately. As input data is fed into the model, the weight of the variables within the model are adjusted appropriately according to the cross-validation process [23]. This process eases training the model by providing a clear training dataset [24]. Due to these benefits, fully supervised object detection has been applied to a number of roadway crack detection models, including artificial neural networks [25], deep convolutional neural networks [26], and FCN [27]. Semi-supervised learning is similar to fully supervised learning. However, the major difference occurs in the data annotation. In semi-supervised learning, only a portion of the training data is labeled. The model is initially trained to predict the rest of the training dataset based on the information provided by the annotated data [28]. This method dramatically cuts down on the time required to annotate data for model training, which offers higher efficiency in model performance [29]. Due to these benefits, semi-supervised models have also seen some limited applications in transportation such as traffic incident detection [30], roadway crack detection [31], and roadway sign detection [32].

Similar to the large variety of CV in object detection, there is a wide range of CV methodologies for road crack detection from various perspectives. For starters, Prasanna et al. used computer vision techniques to detect and analyze cracks on a bridge by utilizing edge-detection based classification [33]. Another study by Yeum and Dyke also dedicated their work to automatically process and analyze a large volume of images of bridge cracks without controlling camera angles [34] A study by Zhang et al. trained a supervised deep neural network to classify features of each image patch into crack and non-crack in pavement images [35]. Additionally, Schmugge et al. (2017) adapted SegNet for crack segmentation in video frames and revealed that it could significantly improve the CNN-based method [36]. A study by Zhang et al. employed U-Net to process an image as a whole and generate a crack segmentation without pacifying and obtained an outstanding pixel-level accuracy [37]. Furthermore, a study by Singh & Shekhar, demonstrated that Mask R-CNN could be used to localize cracks and obtain their corresponding masks to extract other properties that are useful for the inspection [17]. Additionally, Attard et al. illustrated that a higher precision and recall value could be achieved through Mask R-CNN [38]. A study by Augustaukas and Lipnickas utilized U-Net convolutional neural network and its different layers for pixel-wise detection [39]. Moreover, a study by Zhou and Song demonstrated that heterogenous image fusion is a better alternative to image pre-processing [26]. A study by Yu et al adopted OTSU automatic threshold, guided filtering, and gamma image enhancement, then used the Zhang Suen skeleton extraction algorithm to extract crack skeletons [40]. The results demonstrated that the method was efficient and reliable. In addition to that, an ad hoc YOLOv2 was employed by Deng et al to detect concrete cracks from real-world images automatically [41]. The results demonstrated that the proposed mdoel outperformed Faster R-CNN in terms of both accuracy and inference speed. In recent studies, new deep learning models have also been proposed for crack detection, in particular, I-UNet [42], CrackU-Net [43], U-CliqueNet [44], SCHNet [45], U-HDN [46], and feature pyramid and hierarchical boosting network (FPHBN) [47].

Along with the multitude of methods used in roadway distress detection, many data collection methods were

also employed, primarily utilizing different camera types, including Pole cameras, wall-mounted cameras, and vehicle cameras, to name a few. In the field of remote sensing, applications of the unmanned aerial system (UAS) are becoming increasingly more popular. The primary advantage of UAS is its ability to carry a lightweight digital camera as well as its ability to have ample spatial coverage, which can be used to effectively and efficiently acquire imagery at unforeseen high resolution. UAS-acquired imagery can characterize the detailed spectral features of the objects, which would further improve the detection accuracy and overall performance of CV applications. UAS applications are rapidly growing in fields such as transportation engineering [48-49], construction engineering [50-51], surveying and mapping [52], hazard mapping [53] and surveillance [54-55]. Furthermore, UAS is safe and can mitigate risk factors to acquire information at a safe distance remotely. Using advanced UAS in combination with image technologies and efficient algorithms can help overcome current challenges and be an effective tool for road crack detection.

Upon reviewing the above studies, it is clear that existing data and methodologies for roadway distress detection, although demonstrating some favorable results, can be improved further. Mobile Measurement System (MMS) collects existing data, primarily through vehicle-mounted digital cameras or low-cost smartphone cameras, is often time-consuming, and provides minimal spatial coverage as the collected data is often from the driver's point of view. In addition, the majority of the ongoing methodologies are based on single-stage detectors. Such methods are limited to producing the bounding box, and thus fail to provide geometric information such as dimensions, orientations, and pixel-wise segmentation of roadway cracking. Furthermore, limited studies have explored the semi-supervised learning approach for object detection. Consequently, it is necessary to provide additional research to further expand the overall understanding of CV's applications in road distress recognition. To this end, the present study naturally extends one recent study by applying the semi-supervised algorithm to the Mask R-CNN models to detect and segment road cracks. Some appealing features are worth mentioning [56]. First, three distinct backbone models are evaluated and compared using average precision (AP) scores as the primary evaluation criterion. Second, image augmentation techniques are applied to UAS imagery to avoid overfitting. Third, different levels of unlabeled data are experimented along with eleven thresholds for comprehensive assessment.

METHODOLOGY

Data Collection Process

The study area is located in the city of San Dimas (latitude: 34° 06' 45.75" N, longitude: 117° 49' 21.35" W) covering an area of approximately 3,000 m^2. The site is a paved residential street that varies in ground elevation by 3.6 m having a length of approximately 215 m. Two aerial surveys were performed consisting of a low- and high-resolution optical sensor. On April 15, 2020, the low-resolution (4:3 aspect ratio: 4,000 × 3,000) dataset acquired 101 images using a DJI Phantom 4 UAS with a 12.4-megapixel camera. On May 16, 2020 using a DJI Phantom 4 Pro v2.0 UAS with a 20-megapixel camera, the high-resolution (4:3 aspect ratio: 4,864 × 3,648) dataset collected 324 images. The UASs flew at about 37 m above ground level collecting images at nadir with a forward overlap of 95%, a sidelap of 90%, and a flight speed of 1.8 m/s for a flight mission that lasted about 5 minutes and 35 seconds. The flights were performed using an autopilot flight path produced in DroneDeploy [57]. For consistency, both aerial surveys followed the same framework. On average, there are approximately 198 and 129 road surface distress instances per image for higher- and lower-resolution, respectively.

Data Annotation

The popular VIA Annotation Software [58], a lightweight and standalone web application annotation tool, was utilized for data annotation. Since the damage of the roadway is relatively small in such high-resolution images, the authors zoomed in at the scale of 7 times to 8 times the original scale to have a better look of the street. During annotation, each instance was cut if there was another road damage intersecting it or it is the end of that road surface defect instance. A cropped portion of a sample input annotated image is shown in Figure 1 (captured at the scale 4x).

Semi-Supervised Learning

Deep learning networks demonstrated their capabilities on a wide range of supervised-learning tasks on extensive collections of labeled data such as ImageNet or COCO dataset. These deep models usually require a large amounts of labeled training data in order to provide remarkable performance on these tasks. However, the process of annotating these datasets is often difficult, costly, and time-consuming. For the field of computer vision, visual data can be acquired relatively easy, yet only a small portion of collected data are annotated,

which leaves a significant number of dataset samples unlabeled. Under such conditions, semi-supervised learning (SSL) has emerged to resolve the lack of labeled data issue and opened an exciting new research pathway in deep learning. SSL is the middle ground between supervised and unsupervised learning. The principle of semi-supervised learning allows the models to take advantage of both labeled data and an arbitrary amount of unlabeled data for the training process in order to gain more understanding of the population. However, SSL is very sensitive. In other words, it is only applicable under certain conditions as it might also lead to the degradation problem in model's performance. SSL contains 4 simple steps. Initially, deep learning models are trained on the manually annotated training data. Then, the trained weights will be then loaded to these models to generate road distress detection on unlabeled portion of the dataset, which is also known as the pseudo-label. The arbitrary amount of newly created pseudo-label data is combined with the original labeled data resulting in the combined training dataset. Those deep learning models are then re-trained on this combined dataset. The ratio of the pseudo-label data to the total labeled data is also noted as pseudo-ratio. Overall, the semi-supervised learning pipeline on road surface damage detection using Mask R-CNN variants is illustrated in Figure 2.

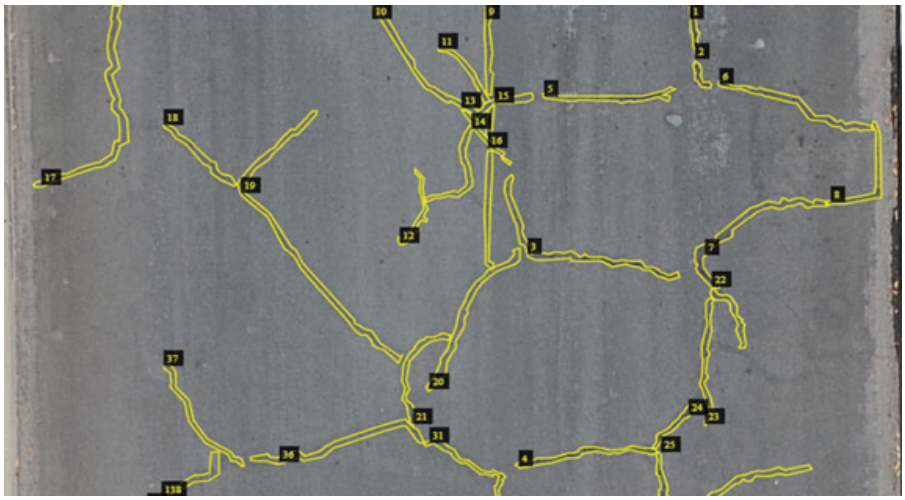

Figure 1. Sample Annotation of Roadway Damage using VGG Image Annotator (VIA).

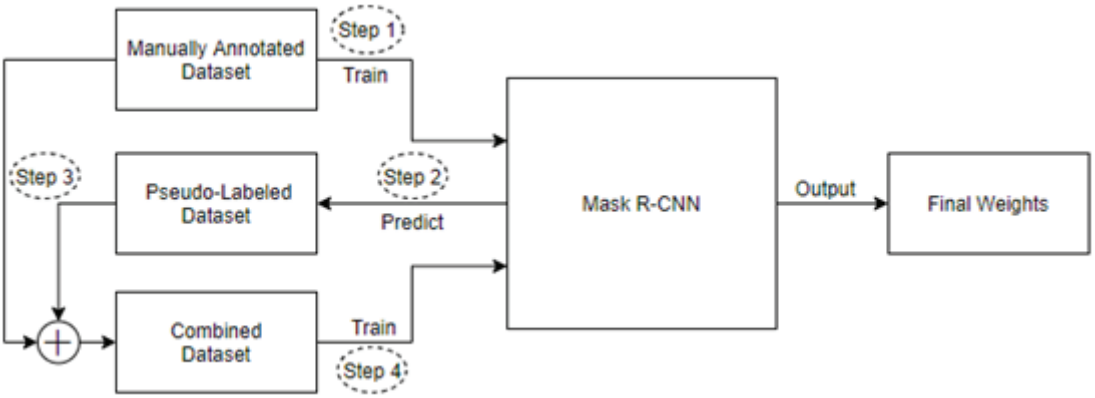

Figure 2. Semi-Supervised Learning Pipeline for Road Damage Detection.

Mask R-CNN

Mask R-CNN is a simple framework for object classification, detection, and instance segmentation, which can effectively recognize objects in an image while maintaining high-quality segmentation mask for each predicted instance. Object recognition has become fundamental visual tasks in the field of computer vision. In recent years, numerous deep learning object recognition models have been implemented on various open-source platforms. Mask R-CNN is also supported in Detectron2 framework, which is an open-source software system by Facebook AI Research (FAIR) implementing most state-of-the-art algorithms for object detection and classification task. Mask R-CNN expands from Faster R-CNN, which adopts two stages. The first stage extracts multi-scale features from the input image and generates the anchors, also known as the proposal bounding box. The second stage refines these anchors into bounding boxes with the according category and generates the segmentation masks. The general network structure of Mask R-CNN algorithm is demonstrated in Figure 3.

Backbone Network

Detectron2 supports variations of backbone as the feature exactor in the form of base_network-head format. The base networks are mainly variants of residual architecture: ResNet50 (R-50), ResNet101 (R-101), ResNeXt101 (X-101-32x8d). These base networks will be combined with another 3 different types of head known as FPN, C4, and DC5.

a) ResNet

In recent years, neural networks have become deeper as the complexity of the machine learning tasks become more difficult and complex. Therefore, many state-of-the-art networks have increased from solely a few layers such as AlexNet to over hundred convolutional layers with very complicated architecture structure. Due to the increment in the number of layers, more in-depth information regarding the input data will be extracted and learned through the deep learning algorithms resulting in better performance on predicting more complex functions. However, as the neural network becomes deeper, it suffers the huge training barrier, which is also known as vanishing gradients. Specifically, during the gradient descent process, back-propagation from the final layer to the initial layer is computed. Therefore, if the gradients are small, multiple multiplication between the weight matrix through each layer can either exponentially decrease to zero, or exponentially explode to a very large value. With the help of normalization through initialization or intermediate layer, deep neural network might be able to converge normally. However, when these deep networks are able to start converging to the minima, they may

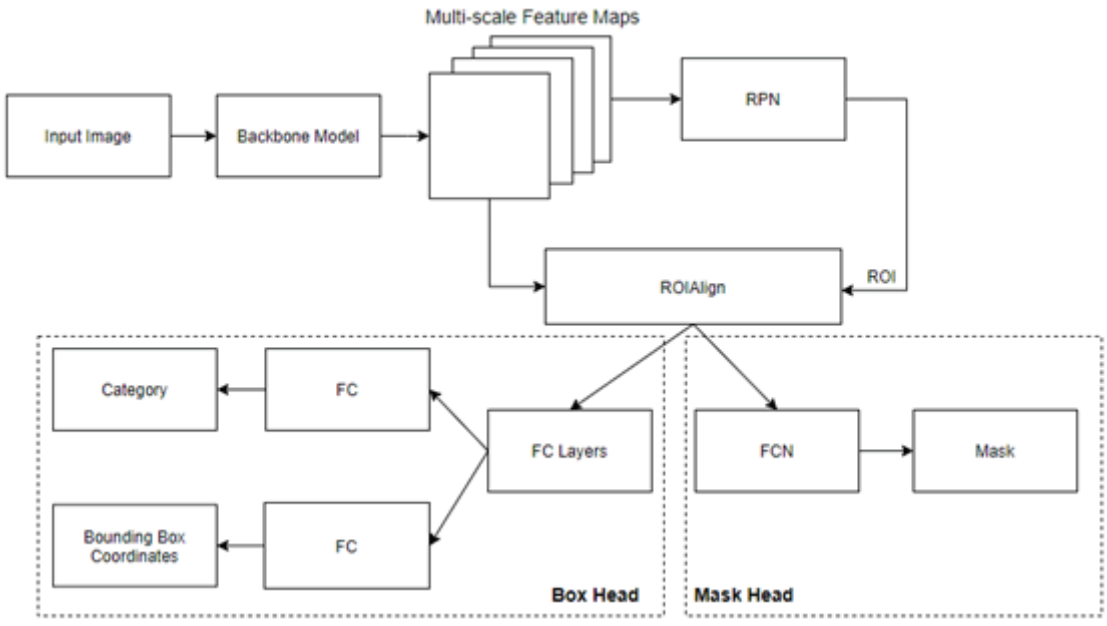

Figure 3. The Structure Diagram of Mask R-CNN Algorithm.

undergo degradation problem, meanig that the model's accuracy gets saturated with increment in network depth. In other words, the network is able to predict the dataset's function before reaching to the final layer. These additional layers are redundant and might make the model fail to learn the identity function to carry out the result to the output from the layer that the model already learned everything. As a result, deep residual network was proposed trying to solve this problem by introducing the new residual mapping function defined as:

$$h(x) = f(x) + x \quad (1)$$

Where h(x) is the residual mapping function in term of input x and f(x) is the initial mapping function which is also in term of input x.

Residual networks are able to better optimize the new residual mapping compared to the original mapping alone. With this new residual mapping function, residual network introduced shortcut path (or, skip connection), allowing the input information to flow through from layer to layer easily without going through any convolutional layers. In addition, regularization will skip through these additional new layers if they do not contribute to the model performance. In this case, the residual mapping h(x) will simply act as an identity function to carry out the output of the previous layer. As the result, the model prediction ability will not be affected from them. In residual network, there are two main types of blocks depending on whether the input and output's dimensions are similar or different, which is illustrated in Figure 4. The Identity Block is the standard block in ResNet variants where the input and output have the same dimension. Meanwhile, the architecture utilizes Convolutional Block when the dimensions do not match up.

This residual block or identity block is the fundamental component in ResNet architecture. ResNet architecture is then constructed from the combination of different residual blocks. Specifically, both ResNet-50 and ResNet-101, as mentioned above, have the same five-stage structure. The first stage only consists of the basic stem block, which is the combination of one convolutional neural network, followed by batch normalization layer, a non-linear action function, and a max pooling layer. Stages 2 to 5 is stack of convolutional blocks and identity blocks. In total, ResNet-50 and ResNet-101 have 50 layers and 101 layers, respectively. Their only difference is the number of convolutional blocks in stage 4, which is 5 blocks for R-50 and 22 blocks for R-101.

b) ResNeXt

ResNeXt architecture further enhances the advantages of the original ResNet architecture by introducing the 'next' dimension (also called cardinality). Building upon the principle of repeating same topology building blocks, ResNeXt expands the cardinality dimension by splitting the original space into subspaces with the same topology. These lower-dimensional representation embeddings will be transformed with arbitrary function T_i. The output from the transformation function of these subspaces will be then aggregated using summation. ResNeXt still maintains the parameters complexity of the building

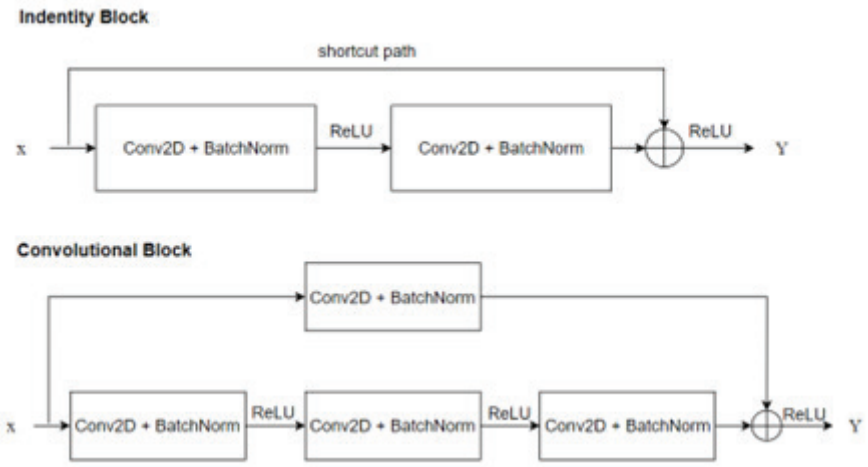

Figure 4. Convolutional Block in ResNet.

block when compared with its ResNet counterpart. Specifically, in this study, the parameters complexity between R-101 and X-101 still remain the same. Overall, with the introduction of cardinality, it is believed to be more effective in training the deep residual models to adapt to new datasets.

c) FPN, C4, DC5

After the input image is fed through the different deep residual base networks to extract features, these output features are fused with feature pyramid network (FPN) in order to obtain stronger semantic feature maps, which contribute directly to the ability to accurately predict road surface damage instances. As the input image is forwarded deeper through the deep residual network, the spatial resolution deceases and more high-level features are detected, leading to richer semantic information for deeper layer. However, the information regarding the targets such as locations is no longer precise as information loses due to the up-sampling and down-sampling layers. In order to obtain both high resolution and rich semantic information, top-down pathway is added alongside with the deep residual network. In top-down pathway, lateral connection is applied between the reconstructed layers from upsampling and the corresponding feature maps from residual network to help better predict objects' location. Aside from FPN, Detectron2 also supports the experimental head C4 and DC5. C4 is the baseline model implemented in the Faster R-CNN paper utilizing a ResNet conv4 backbone with conv5 head. Meanwhile, DC5 incorporates a ResNet conv5 backbone with dilations in conv5, a standard convolutional layer and fully connected heads for mask and bounding box prediction accordingly, which is proposed in Deformable ConvNet paper [59].

Region Proposal Network (RPN)

RPN is a unique model module that includes both regressors and classifiers. Using sliding mechanism, RPN scans through different-scale strong semantically features maps outputted from backbone network and proposes region of interest (ROI) that may contain objects. RPN generates multiple anchors with the anchor point placed in the middle of the sliding window at different sizes and scales. These proposals will be then fed through the classifier and the regressor. The classifier is responsible for providing the probability of objects that are present within the proposed region. On the other hand, the regressor will refine these ROIs and output the bounding box coordinates. By utilizing anchors, the model is translational invariant. In other words, if the input image is translated using various transformations such as rotation, resize, brightness level, there is no variance on the output.

ROIAlign

Those regions of interest outputted from RPN are then mapped to multi-level features maps from backbone model using ROIAlign to extract the corresponding features resulting in a set of shared feature maps, which are subsequently sent to the fully connected layers and fully convolutional network for object classification and mask segmentation task, respectively. The original ROI Pooling method from Faster R-CNN introduced a lot of quantization operations to map the generated proposals to retrieve integer value x and y coordinate. However, as the ROIs are not aligned with the original grid of the feature maps, ROI Pooling suffers from misalignment issue leading to the lower performance due to the loss of lots of useful information. This misalignment issue might not affect the classification and detection ability of the model. Yet, since mask head requires more fine-grained alignment, the problem adversely affects the process of generating precise pixel-level segmentation. Thus, Mask R-CNN adopts new pooling layer ROIAlign to further enhance the mapping of the proposed ROIs to the multi-scale semantic feature maps. The problem of harsh quantization of ROI Pooling layer is resolved by applying ROIAlign layer with bilinear interpolation algorithm on the aligned ROI. Initially, the ROIAlign layer traverses each region of interests and keeps the floating-point number for the ROI position unquantized. These proposals are divided into k x k cells with unquantized boundary. Within each cell, four fixed-value coordinates are computed using bilinear interpolation, which are then aggregated with either max pooling or average pooling operation. The result of ROIAlign layer is fixed-size ROIs with no quantization error.

Loss Function

Mask R-CNN is trained based on multi-task loss function that is a combination of classification, localization loss (previously introduced in Faster R-CNN), and segmentation mask loss as illustrated in Equation 2:

$$L = L_{cls} + L_{box} + L_{mask} \quad (2)$$

Where L_{cls}, L_{bbox}, L_{mask} are classification loss, bounding box loss and mask segmentation loss accordingly. L_{cls} and L_{bbox} are further divided into corresponding loss inside RPN and Box Head module defined in Equation 3 and Equation 4.

$$L_{cls} = L_{cls\_rpn} + L_{cls\_boxhead} \quad (3)$$

$$L_{box} = L_{box\_rpn} + L_{box\_boxhead} \quad (4)$$

Where L_{cls_rpn} is the loss for anchor binary classifier; L_{box_rpn} is the bounding box regression loss of RPN; $L_{box\_boxhead}$ represents the loss of the classifier within the box head of Mask R-CNN; and $L_{box\_boxhead}$ denotes the loss for bounding box refinement. The detailed calculations for each items are shown below.

- Loss function for classification task:

$$L_{cls\_rpn} = L_{cls\_boxhead} = \frac{1}{N_{cls}} \sum_i L_{cls} (p_i, p_i^*) \quad (5)$$

Where $L_{cls} (p_i, p_i^*)$ is log loss function between the predicted probability of an anchor i having an object (p_i), and the binary ground truth label of anchor i having an object (p_i^*) can be defined as:

$$L_{cls}(p_i, p_i^*) = - p_i^* log(p_i) - (1 - p_i^*) log(1-p_i) \quad (6)$$

- Loss function for bounding box regression task:

$$L_{box\_rpn} = L_{box\_boxhead} = \frac{\lambda}{N_{box}} \sum_i p_i^* L_1^{smooth} (t_i - t_i^*) \quad (7)$$

Where L_1^{smooth} is the smooth L1 loss function between the predicted four parameterized coordinates t_i and ground truth coordinates t_i^*.

- L_mask is defined as average binary cross-entropy loss for pixel-level segmentation task:

$$L_{mask} = -\frac{1}{m^2} \sum_{1 \le i,j \le m} [y_{ij} \log (\hat{y}_{ij}^k) + (1 - y_{ij}) \log (1 - \hat{y}_{ij}^k) \quad (8)$$

Where y_{ij} is the label for the pixel at position i, j in the region of size m x m, $\hat{y}_{ij}^k$ is the predicted label for the same pixel for the ground-truth class k.

Training

Hyperparameters

All the deep learning models used in the experiments were trained on a high-performance computing cluster. Each model was trained on a single node using 2 NVIDIA Tesla P100 graphics card machine which has 64 GB memory each. All the Mask R-CNN models utilized the weights pre-trained on MS-COCO dataset to initialize the training process, which were then fine-tuned for the road surface damage detection task. Specifically, for this research, only 3 different backbone models R50, R101, and X101 are investigated. Some important hyper-parameters are modified to better adapt with the roadway distress dataset summarizing in Table 1.

Augmentations

To make the models more robust, augmentations were applied to the input image by modifying the DataLoader from Detectron2 in order to avoid overfitting. However, the model only applies random horizontal and vertical flipping transformation with the probability of 0.6 in the second time training of the deep residual networks, because the graphical computing resources would run out of memory as more instances are added to the training

Table 1. Summary of Important Hyperparameters in Detectron2 for Road Damage Detection Task

Hyper-parameter name	Detectron2's parameter name	Value
Warmup iteration	cfg.SOLVER.WARMUP_ITERS	2000
Base learning rate	cfg.SOLVER.BASE_LR	0.001
Training iteration	cfg.SOLVER.MAX_ITER	30000
Checkpoint period	cfg.SOLVER.CHECKPOINT_PERIOD	10000
Number of classes	cfg.MODEL.ROI_HEADS.NUM_CLASSES	1
Batch size per image	cfg.MODEL.ROI_HEADS.BATCH_SIZE_PER_IMAGE	128
Anchor sizes	cfg.MODEL.ANCHOR_GENERATOR.SIZES	(4, 8, 16, 32, 64)
Anchor aspect ratio	cfg.MODEL.ANCHOR_GENERATOR.ASPECT_RATIOS	(0.5, 1.0, 2.0)
Input image's width	cfg.INPUT.MIN_SIZE_TRAIN, cfg.INPUT.MIN_SIZE_TEST	600
Input image's height	cfg.INPUT.MAX_SIZE_TRAIN, cfg.INPUT.MAX_SIZE_TEST	800

dataset. However, during the first-time training, besides horizontal and vertical flipping, the input data were also augmented with random brightness with the intensity within the range of 0.3 to 1.8, in order to make the training data to be more diverse in each training iteration.

Evaluation Criteria

To determine the effectiveness of variants of Mask R-CNN models in the study, the average precision (AP) score was evaluated. In COCO dataset or those datasets having similar format, mean average precision (mAP) often refers to the average precision. In general, the precision is defined by Equation 2:

$$p = \frac{TP}{TP+FP} + \frac{TP}{total\ detections} \tag{9}$$

To further understand the AP score, recall is necessary to be computed along with precision, which is shown in Equation 3:

$$r = \frac{TP}{TP+FN} + \frac{TP}{total\ groundtruths} \tag{10}$$

AP score can then be calculated as the area under the precision and recall curve. The corresponding expression is shown as below:

$$AP_T = \int_0^1 p(r)dr \tag{11}$$

Where p(r) is the precision function in term of recall. T is the IoU threshold. AP_T is the average precision score at a specific threshold T.

RESULTS

Inference

The inference was done using the Detectron2 framework. Table 2 shows the average time taken to detect road distress instances per image for the different

Table 2. Detection Time for Different Backbone Model with Pseudo Ratio Average over 11 Threshold Values from 0.20 to 0.70 with Step Size of 0.05

Models	Pseudo Ratio	Resolutions for Training	AIS on L(s)	AIS on H(s)
mask_rcnn_R_50_FPN_3x	0.75	H	12.05	14.49
		L	11.33	14.51
mask_rcnn_R_50_FPN_3x	0.5	H	12.53	14.10
		L	9.23	1.42
mask_rcnn_R_50_FPN_3x	0.25	H	11.33	13.60
		L	6.84	8.34
mask_rcnn_R_101_FPN_3x	0.75	H	12.22	14.82
		L	10.70	13.70
mask_rcnn_R_101_FPN_3x	0.50	H	13.32	14.80
		L	8.39	10.03
mask_rcnn_R_101_FPN_3x	0.25	H	8.48	10.45
		L	6.74	8.40
mask_rcnn_X_101_32x8d_FPN_3x	0.75	H	12.19	14.37
		L	10.17	11.60
mask_rcnn_X_101_32x8d_FPN_3x	0.5	H	12.12	13.08
		L	7.72	8.72
mask_rcnn_X_101_32x8d_FPN_3x	0.25	H	6.53	8.42
		L	5.69	6.88

Notes: 1. AIS is Average Inference Speed; 2. H represents high resolution and L represents low resolution; 3. Pseudo ratio is defined as the ratio of the pseudo-label data to the total labeled data.

scenarios of backbone, pseudo ratio and image resolution.

As shown in Table 3, it is clear that the models take more time to predict road surface defects based on high-resolution pictures in most cases. With the speed of 1.8 m/s, the drone can travel approximately a distance of 6.04 m and 13.26 m while processing the images for detection purpose. Therefore, before the UAS is flies to the next area for next frame capture, the detection system is done with prediction of the present picture. In other words, the Mask R-CNN satisfies the real-time prediction requirement. The sample detection of road distress instances predicted for one high resolution imagery is shown in Figure 5.

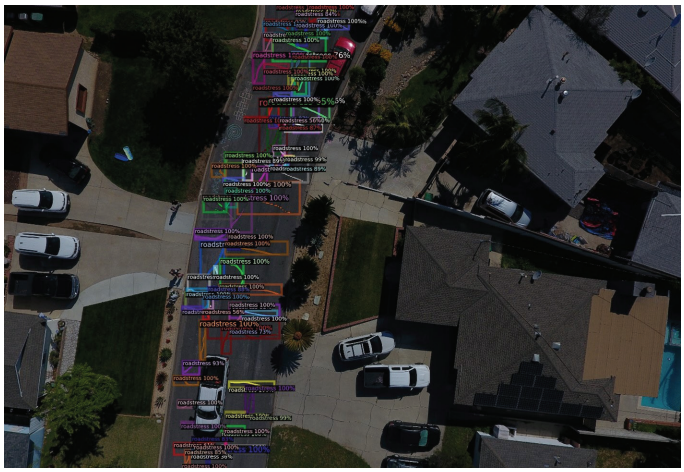

Figure 5. The Sample of Road Distress Instances Detected for Higher Resolution Image.

Performance Evaluation

For comprehensive assessment of the model performance, the AP scores of the detectors were reported for various scenarios. First, both in-resolution (i.e., training and validating the models based on the data of the same level of resolution, high or low) and out-of-resolution (or, model training and validation based on the different levels of resolutions) evaluations were conducted. Second, three levels of pseudo ratio (that is, 0.25 or low, 0.50 or medium, 0.75 or high) were employed. Three, eleven threshold values were explored that range from 0.2 to 0.7, with the increment of 0.05. Finally, as mentioned before, three categories of backbones were chosen and tested that contain R_50, R_101 and X_101. Overall, there are 198 pairs of AP scores observed. To identify the reliable relationship among the detection performance (based on both in-resolution and out-of-resolution) and the influential factors including pseudo ratio, threshold value, input image resolution and the backbone model, the statistical analysis was conducted in two ways. First, the association was performed between detection performance and each of the individual factors via Analysis of variance (ANOVA), paired t-test, and pairwise correlation analysis. Second, a joint model was developed that estimated the above two performance scores with all contributing factors being considered at the same time. The common trends illustrated in both statistical analyses are anticipated to yield more reliable conclusions with greater confidence.

First, the authors is interested in identifying the relationship between categorical variables (or, backbone

Table 3. Result of ANOVA Analysis between AP Scores and Categorical Variables

Resolution Type	Scenarios	Df	Sum of Square	Mean of Square	F Value	Pr(>F)
In Resolution	Backbone Model	2	1849	924.6	1.089	0.339
	Residuals	195	165634	849.4		
Out of Resolution	Backbone Model	2	1222	611	0.489	0.614
	Residuals	195	243701	1250		
In Resolution	**Level of Pseudo Ratio**	**2**	**35766**	**17883**	**26.480**	**6.7E-11**
	Residuals	**195**	**131717**	**675**		
Out of Resolution	**Level of Pseudo Ratio**	**2**	**11847**	**5924**	**4.956**	**0.008**
	Residuals	**195**	**233075**	**1195**		

Notes: 1. Df represents the degree of freedom; 2. Pr(>F) represents the statistical significance of two groups relying on F value represents the statistical significance of two groups relying on F value; 3. The bold text indicates the significant difference with the p-value being less than 0.05.

types and pseudo ratio levels) and AP scores. The popular ANOVA test was conducted with the detailed results being shown in Table 3. The degree of freedom, which is 2 for all cases, since backbone and pseudo ratio have three categories. The p-value Pr(>F) represents the statistical significance of two groups relying on F value. Usually, 0.05 is a critical value to identify their correlation. If p-value is less than 0.05, these two groups are supposed to be statistically correlated with each other. In review of Table 3, it is known that backbone models seem to exert no statistically significant influence on the two AP scores. However, the level of pseudo ratio demonstrates statistically significant impact on both in resolution and out-of-resolution AP scores. Scrutiny of the 195 pairs of scores indicates the lowest pseudo ratio (i.e., 0.25) yields the largest average AP scores, while the highest ratio of 0.75 yields the least average AP scores. Given the benefit of the semi-supervised algorithm in enhancing the data sample size for training and possibly the associated prediction accuracy, it is highly recommended to perform a sensitive analysis in the future for the ideal pseudo ratio that generates the highest AP scores at different levels of sample size.

Similar procedure was followed to determine the relationship between imagery resolution and AP score. To perform the analysis, the paired t-test was approached as resolution has only two categories (high and low) and images are created from the same location. The t-test results in Table 4 clearly show the statistically significant difference exists (the p-value less than 0.05) among resolution and AP score. The findings indicate that high-resolution have better detection performance than the low resolution.

In addition to the relationship between AP scores and the categorical variables, the study also explores the relationship between AP scores and the numerical variables such as Threshold value. Even though the pseudo ratio was treated as the categorical level in the ANOVA test, it can be used as numerical value for detailed correlation analysis. Hence, the Pearson's correlation coefficients for the pair of AP scores and both threshold and pseudo ratio were calculated. The detailed correlation results are presented in scatterplot matrices in Figure 6. As illustrated in the plots, the distributions of the variables are shown on the diagonal, the upper triangle of the matrix represents the value of correlation coefficient, and the lower triangle of the matrix illustrates the bivariate scatterplots. In review of Figure 6, it is known that both threshold and pseudo

Table 4. Result of T-Test between Resolution and AP Score

Items	Degree of freedom	t-value	Mean of the difference	p-value
In_Resolution_AP	98	20.425	44.123	<2.2E-16
Out_of_Resolution_AP	53	-51.307	-66.428	<2.2E-16

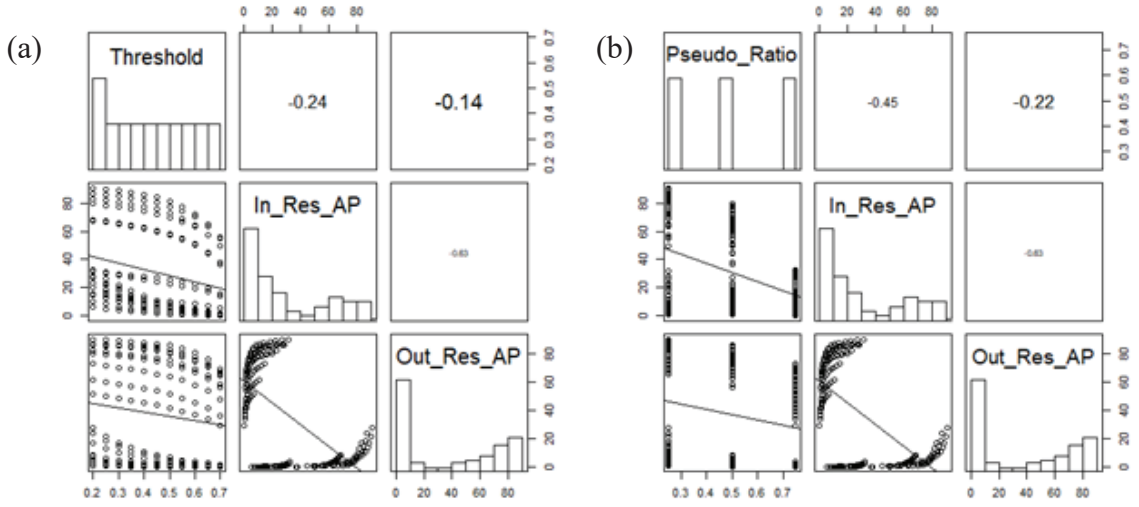

Figure 6. Scatter Plot and Correlation Matrix. Note: The font size associated with the correlation coefficients indicates the statistical significance. The larger the font size, the more statistically significant the coefficient.

ratio appear to have a negative correlation with the AP scores, which signifies that the increase of the threshold values or Pseudo Ratio would reduce the AP scores for both cases of in-sample and out-of-sample detections. The explanation is that the stricter criterion (or, higher threshold) would reduce the number of TPs, which also tend to be decreased with more unlabeled data being used in the training process. However, for pseudo ratio, given the relatively small coefficient magnitude for out of resolution (-0.22) and the small statistical significance for in resolution (or, the smaller font size associated with -0.45), it is expected that the pseudo ratio is not strictly linearly related with the detection performance. Such phenomenon suggests the need to further explore the actual relationship between pseudo ratio and AP scores, in hopes of identifying the optimum pseudo ratio yielding the highest detection accuracy.

Despite that the relationship between detection performance and individual factors is explored using various tests, the "true" influence of each factor on the AP scores with the co-existence of other variables is still unknown. To address this issue, the present study employed a bivariate random effect model to capture the associated relationship while considering the unobserved heterogeneities shared by the two AP scores. The detailed model results are shown in Table 5. As illustrated, some of the results in the joint model are consistent with previous tests related with individual factors. Threshold value and pseudo ratio are again negatively correlated with both AP scores. Interestingly, the relationship between imagery resolution and AP scores seems to be inconsistent with previous result. Based the results of t-test, high image resolution tends to have statistically better results than low resolution for both cases of AP score. Nonetheless, the joint model demonstrates that low resolution has statistically larger AP score than the base resolution in the case of out of resolution identification. A possible explanation may be that the lower resolution image dataset needs less iterations for the model training process, which leads to much better performance when using the same backbone models and threshold values. In addition, different than the ANOVA test results, the backbone of X_101 appears to be significantly better than the base of R_50 for both AP scores, while R_101 turned out be inferior to R_50 in the case of out of resolution assessment. The above phenomena signify the importance of considering all factors together simultaneously.

CONCLUSIONS

Roadway distress detection is important for generating a safe and comfortable driving environment. However, compared to other roadway objects, it is more time-consuming to detect the roadway distress due to its relatively small size of the and sometimes the large number of instances under some conditions. Therefore,

Table 5. Joint Model Parameter Estimates

Variables		β1 (In Resolution AP)		β2 (Out of Resolution AP)	
		Mean	**SD**	**Mean**	**SD**
Fixed Effects					
(Intercept)		**143.75**	**4.234**	**-173.583**	**4.896**
Backbone Model	mask_rcnn_R_50_FPN_3x (Base)				
	mask_rcnn_R_101_FPN_3x	-3.305	2.045	**-2.465**	**1.213**
	mask_rcnn_X_101_32x8d_FPN_3x	**4.164**	**2.045**	**3.565**	**1.213**
Resolution	High (base)				
	Low	**-43.273**	**1.670**	**66.109**	**0.990**
Threshold Value	**-40.113**	**5.220**	**-30.724**	**3.117**	
Pseudo Ratio	**-61.224**	**4.068**	**-37.570**	**2.419**	
Random Effects					
Observation. ID	**5.326**	**3.457**	**4.366**	**3.014**	

Notes: 1. SD represents standard deviation; 2. The bold cells represent the variables which are statistically significant.

there is a high demand for easy, efficient, and ideally cheap approaches to recognizing the surface damage on the roadway. The previous research (56) demonstrated the great potential of the freely available Mask R-CNN models in detecting the road distress. However, the authors recommended to enhance the model performance by using more data for model-training. Given the irregular shape, small area size, and occasionally very large number, of the road distress objects, it often becomes arduous to annotate the required huge number of distress instances for the supervised learning during the training process. As a special instance of weak supervision, the semi-supervised learning, via combining the unlabeled data with some amount of labeled data, can help reduce the cost associated with the labeling process, while maintaining or even improving the learning accuracy. Albeit with the documented advantages, semi-supervised learning has seen limited applications in transportation field. To this end, the authors extended the previous research (56) by applying the semi-supervised learning algorithm to detect the roadway distress, whose fully labeled training sets are often infeasible because of the high labeling cost resulting from the supervised learning. Compared with the similar studies in the literature, the present research further expands the toolset for road distress detection with the following contributions and features.

1. Both low and high resolution UAS imagery data were collected by DJI Phantom 4 with a 12.4-megapixel camera and DJI Phantom 4 Pro v2.0 UAS with a 20-megapixel camera, respectively.
2. Three levels of pseudo radio (that is, 0.25 or low, 0.5 or medium, 0.75 or high) were experimented along with eleven thresholds and three backbone types for the comprehensive comparison of model performance.
3. Detailed statistical analysis was performed on the raw results to identify the relationships between AP scores and individual factors, and the co-influence of a set of factors on the detection performance.

In review of the detailed results, the following major conclusions were drawn:

1. Semi-supervised Mask R-CNN models appear to be effective in detecting road distress in most cases. This finding indicates the possibility of a high detection accuracy with low labeling cost, which is ideal for roadway distress given the nature of this type of object.
2. Higher thresholds would lead to lower AP scores as the stricter detection standard tends to remove more false positives.
3. For imagery resolution, the high resolution outperforms the low counterpart, if no other influential factors are considered. However, with the existence of other covariates, the low- resolution yields surprisingly better performance in case of out-of-resolution assessment. It is possible that the smaller low-resolution data may require less iterations during the training process.
4. Overall, the pseudo ratio tends to have a negative relationship with the model performance. However, the correlation between pseudo ratio and the AP score is not statistically significant for in-resolution assessment, and the correlation coefficient magnitude for the case of out-of-resolution is somewhat small. Such fact implies the ratio of the unlabeled data is not strictly linearly related with the detection performance, and there is a need to perform the sensitive analysis to identify the optimal pseudo ratio that could generate the highest prediction accuracy.

Albeit with unique contributions and exciting research findings, the present work can be further enhanced in different ways. First, only three pseudo ratios were explored that demonstrated non-significant correlation with the detection performance. It is highly recommended that more pseudo ratios be analyzed due to the sensitiveness of the number of unlabeled data to the prediction accuracy. Second, three Mask R-CNN models were experimented. More models based on different backbones or other advanced CV tool such as YOLOv4 could be approached for more reliable results. Third, instead of the simple binary detection regarding whether there is a road distress or not, the study can be expanded to further classify distress objects into different categories in terms of dimension or orientation, which may be more insightful to the roadway professionals. Finally, more detailed information about the study (e.g., code) could be obtained from the dedicated website: https://github.com/KossBoii/SS_Roadstress_Detection.git.

REFERENCES

1. Fan S, Wang H, Zhu H, Sun W. Evaluation of self-healing performance of asphalt concrete for low-temperature fracture using semicircular bending test. Journal of Materials in Civil Engineering. 2018.30(9):04018218.

2. Dhital D, Lee JR. A fully non-contact ultrasonic propagation imaging system for closed surface crack evaluation. Experimental mechanics. 2012.52(8):1111–22.
3. Meignen D, Bernadet M, Briand H. One application of neural networks for detection of defects using video data bases: identification of road distresses. In: Database and Expert Systems Applications 8th International Conference, DEXA'97 Proceedings. IEEE; 1997. p. 459–64.
4. Mohan A, Poobal S. Crack detection using image processing: A critical review and analysis. Alexandria Engineering Journal. 2018;57(2):787–98.
5. Shi X, Hansen G, Mills M, Jungwirth S, Zhang Y. Preserving the value of highway maintenance equipment against roadway deicers: a case study and preliminary cost benefit analysis. Anti-Corrosion Methods and Materials. 2015. 63 (1), 1-8
6. Cheng HD, Miyojim M. Novel system for automatic pavement distress detection. Journal of Computing in Civil Engineering. 1998;12(3):145–52.
7. Petrou M, Kittler J, Song KY. Automatic surface crack detection on textured materials. Journal of materials processing technology. 1996;56(1–4):158–67.
8. Huang Y, Xu B. Automatic inspection of pavement cracking distress. Journal of Electronic Imaging. 2006;15(1):013017.
9. Gavilán M, Balcones D, Marcos O, Llorca DF, Sotelo MA, Parra I, et al. Adaptive road crack detection system by pavement classification. Sensors. 2011;11(10):9628–57.
10. Zou Q, Cao Y, Li Q, Mao Q, Wang S. CrackTree: Automatic crack detection from pavement images. Pattern Recognition Letters. 2012;33(3):227–38.
11. Salman M, Mathavan S, Kamal K, Rahman M. Pavement crack detection using the Gabor filter. In: 16th international IEEE conference on intelligent transportation systems (ITSC 2013. IEEE; 2013. p. 2039–44.
12. Maeda H, Sekimoto Y, Seto T, Kashiyama T, Omata H. Road damage detection using deep neural networks with images captured through a smartphone. 2018.
13. Zhang L, Shen J, Zhu B. A research on an improved Unet-based concrete crack detection algorithm. Structural Health Monitoring. 2020;1475921720940068.
14. Alfarrarjeh A, Trivedi D, Kim SH, Shahabi C. A deep learning approach for road damage detection from smartphone images. In: 2018 IEEE International Conference on Big Data (Big Data. IEEE; 2018. p. 5201–4.
15. Chun C, Ryu SK. Road surface damage detection using fully convolutional neural networks and semi-supervised learning. Sensors. 2019;19(24):5501.
16. Arman MS, Hasan MM, Sadia F, Shakir AK, Sarker K, Himu FA. Detection and classification of road damage using R-CNN and faster R-CNN: a deep learning approach. In: International Conference on Cyber Security and Computer Science. Cham: Springer; 2020. p. 730–41.
17. Singh J, Shekhar S. Road damage detection and classification in smartphone captured images using mask r-cnn. 2018. arXiv:1811.04535
18. Cheng J, Xiong W, Chen W, Gu Y, Li Y. Pixel-level crack detection using U-net. In: TENCON 2018-2018 IEEE Region 10 Conference. IEEE; 2018. p. 0462–6.
19. Chen T, Cai Z, Zhao X, Chen C, Liang X, Zou T, et al. Pavement crack detection and recognition using the architecture of segNet. Journal of Industrial Information Integration. 2020.18:100144.
20. Dung CV. Autonomous concrete crack detection using deep fully convolutional neural network. Automation in Construction. 2019;99:52–8.
21. Song W, Jia G, Zhu H, Jia D, Gao L. Automated pavement crack damage detection using deep multiscale convolutional features. Journal of Advanced Transportation. 2020;
22. Bang S, Park S, Kim H, Kim H. Encoder–decoder network for pixel-level road crack detection in black-box images. Computer-Aided Civil and Infrastructure Engineering. 2019;34(8):713–27.
23. Collingwood L, Wilkerson J. Tradeoffs in accuracy and efficiency in supervised learning methods. Journal of Information Technology & Politics. 2012.9(3):298–318.
24. Xu G, Song Z, Sun Z, Ku C, Yang Z, Liu C, et al. Camel: A weakly supervised learning framework for histopathology image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019. p. 10682–91.
25. Sari Y, Prakoso PB, Baskara AR. Application of neural network method for road crack detection. Telkomnika. 2020;18(4):1962–7.
26. Zhou S, Song W. Deep learning–based roadway crack classification with heterogeneous image data fusion. Structural Health Monitoring. 2021.20(3):1274–93.
27. Sharma S. A data-driven approach for pavement surface distress classification. 2020.
28. Brefeld U, Scheffer T. Semi-supervised learning for structured output variables. In: Proceedings of the 23rd international conference on Machine learning. 2006. p. 145–52.
29. Chen T, Kornblith S, Swersky K, Norouzi M, Hinton G. Big self-supervised models are strong semi-supervised learners. 2020.
30. Chakraborty P, Sharma A, Hegde C. Freeway traffic incident detection from cameras: A semi-supervised learning approach. 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE; 2018. p. 1840–5.
31. Wang W, Su C. Semi-supervised semantic segmentation network for surface crack detection. Automation in Construction. 2021;128:103786.
32. He Z, Nan F, Li X, Lee SJ, Yang Y. Traffic sign recognition by combining global and local features based on semi-supervised classification. IET Intelligent Transport Systems. 2020;14(5):323–30.
33. Prasanna P, Dana KJ, Gucunski N, Basily BB, La HM, Lim

RS, et al. Automated crack detection on concrete bridges. IEEE Transactions on automation science and engineering. 2014;13(2):591–9.
34. Yeum CM, Dyke SJ. Vision-based automated crack detection for bridge inspection. Computer-Aided Civil and Infrastructure Engineering. 2015;30(10):759–70.
35. Zhang R, Zheng Y, Mak TWC, Yu R, Wong SH, Lau JY, et al. Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain. IEEE journal of biomedical and health informatics. 2016;21(1):41–7.
36. Schmugge SJ, Rice L, Lindberg J, Grizziy R, Joffey C, Shin MC. Crack segmentation by leveraging multiple frames of varying illumination. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE; 2017. p. 1045–53.
37. Zhang K, Musha Y, Yang B. Improvements Based on Feature Fusion Single Shot Multibox Detector. Design Engineering; 2020. 414–420 p.
38. Attard L, Debono CJ, Valentino G, Castro M, Masi A, Scibile L. Automatic crack detection using mask r-cnn. 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA). IEEE; 2019. p. 152–7.
39. Augustaukas R, Lipnickas A. Pixel-wise road pavement defects detection using U-net deep neural network. 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). IEEE; 2019. p. 468–71.
40. Yu X, Wang X, Da X, Zhao J. Crack Detection Algorithm of Complex Bridge Based on Image Process. In: CICTP 2020. 2020. p. 1341–53.
41. Deng J, Xuan X, Wang W, Li Z, Yao H, Wang Z. A review of research on object detection based on deep learning. Journal of Physics: Conference Series. 2020 Nov;1684(1):012028.
42. Wang L, Ye Y. Computer vision-based road crack detection using an improved I-UNet convolutional networks. In: 2020 Chinese Control and Decision Conference (CCDC). IEEE; 2020. p. 539–43.
43. Huyan J, Li W, Tighe S, Xu Z, Zhai J. CrackU-net: a novel deep convolutional neural network for pixelwise pavement crack detection. Structural Control and Health Monitoring. 2020;27(8):2551.
44. Li G, Ma B, He S, Ren X, Liu Q. Automatic tunnel crack detection based on u-net and a convolutional neural network with alternately updated clique. Sensors. 2020;20(3):717.
45. Pan Y, Zhang G, Zhang L. A spatial-channel hierarchical deep learning network for pixel-level automated crack detection. Automation in Construction. 2020;119:103357.
46. Fan Z, Li C, Chen Y, Wei J, Loprencipe G, Chen X, et al. Automatic crack detection on road pavements using encoder-decoder architecture. Materials. 2020;13(13):2960.
47. Yang F, Zhang L, Yu S, Prokhorov D, Mei X, Ling H. Feature pyramid and hierarchical boosting network for pavement crack detection. IEEE Transactions on Intelligent Transportation Systems. 2019;21(4):1525–35.
48. Barmpounakis EN, Vlahogianni EI, Golias JC. Unmanned Aerial Aircraft Systems for transportation engineering: Current practice and future challenges. International Journal of Transportation Science and Technology. 2016;5(3):111–22.
49. Sutheerakul C, Kronprasert N, Kaewmoracharoen M, Pichayapan P. Application of unmanned aerial vehicles to pedestrian traffic monitoring and management for shopping streets. Transportation Research Procedia. 2017;25:1717–34.
50. Melo RRS, Costa DB, Álvares JS, Irizarry J. Applicability of unmanned aerial system (UAS) for safety inspection on construction sites. Safety science. 2017;98:174–85.
51. Zhou S, Gheisari M. Unmanned aerial system applications in construction: a systematic review. Construction Innovation; 2018.
52. Fitzpatrick BP. Unmanned Aerial Systems for Surveying and Mapping: Cost Comparison of UAS versus Traditional Methods of Data Acquisition (Doctoral dissertation), University of Southern California; 2016.
53. Suh J, Choi Y. Mapping hazardous mining-induced sinkhole subsidence using unmanned aerial vehicle (drone) photogrammetry. Environmental Earth Sciences. 2017;76(4):144.
54. Puri A. A survey of unmanned aerial vehicles (UAV) for traffic surveillance. Department of computer science and engineering, University of South Florida; 2005. 1–29 p.
55. Haddal CC, Gertler J. Homeland security: Unmanned aerial vehicles and border surveillance. Library of Congress Washington DC Congressional Research Service; 2010.
56. Truong LNH, Mora OE, Cheng W, Tang H, Singh M. Deep Learning to Detect Road Distress from Unmanned Aerial System Imagery. Transportation Research Record. 2021;03611981211004973.
57. Yulianandha Mabrur, A. Analisis pemanfaatan opensource dronedeploy dalam proses mozaik foto udara (UAV). Pawon: Jurnal Arsitektur, 2019. 3(02), 79-92.
58. Dutta A, Zisserman A. The VIA annotation software for images, audio and video. In: Proceedings of the 27th ACM international conference on multimedia. 2019. p. 2276–9.
59. Haque MF, Lim HY, Kang DS. Object detection based on vgg with resnet network. 2019 International Conference on Electronics, Information, and Communication (ICEIC). IEEE; 2019. p. 1–3.