Road Damage Detection and Application on Mobile

Trong Duy Nguyen*; Quang Duc Tran; Viet Tien Le; Kim Thanh Tran

Abstract: Currently, the research on detecting road damage is no longer strange. However, many studies only focus on the detection of the presence or absence of damage. When the road managers from the Ministry of Transport need to repair such damage, they need to know the type of damage clearly to take effective action. Hence, there was no consistent road damage data set available openly, leading to the absence of a benchmark for road damage detection. In this research data constitute 30,000 road images collected from India, Japan, and the Czech Republic automatically detecting road damages in these countries from the IEEE International Conference on Big Data 2020. We use state-of-the-art models (SOTA) in object detection to road damage obtaining results that rival older models. After this, Tensorflow Lite is used with a Yolov4-tiny backbone to deploy on mobile to detect road damage real time on Hoa Lac Hi-tech Park.

Keywords: road damage detection, object detection, yolo

1. Introduction

Roads infrastructure make a decisive contribution both directly and indirectly to the growth and development of the economy. According to the Ministry of Industry and Trade, transport's growth in 2021 approximately expanded. Roads will also greatly affect the lifespan, safety and well-being of humanity. Military absolutely needs roads as much as possible as well as no activities can be performed greatly without proper, live communication. Roads also are helpful in emergency cases. Moreover, the fee for road maintenance is quite cheap (in Viet Nam).

Fig 1. An image captured in Japan

At the moment, the road damage problem has significantly increased compared to the past's figure. In Japan, every year almost three thousand earthquakes appear. In Vietnam, by the unorganized way of transport, roads are being seriously damaged. It can be said that road conditions have, on average, recently been downgraded.

However this is the time that roads will probably need to be maintained. Road maintenance centers actually exist but their effectiveness seems not yet enough. A manual inspection is prone to personal error because to inspect carefully and conclude in detail, the agencies must be full of

experience and have an accumulated wisdom. Additionally, the inspection can become a struggle depending on how bad the weather is. The road inspection job is quite time-consuming.

The manual inspection must be improved by semi-automatic and ultimately automated methods because the network's existing pavements are aging while new ones are being added, making it impossible to meet the deadlines for assessments on time. Society demands another way, which should be faster, more accurate and less time-spending. Computer hardware now can process many types, many sizes of data. As the development of computer hardware, YOLOs technology in recent years has been highly developed , it is indeed superior, especially YOLO4, YOLO5. With a high standard dataset, models are pre-trained and based on those rates (mAP), through researching the results, the best fit model is chosen then.

2. Related work

2.1 Road Damage Detection

Road surface inspection is primarily based on visual observations by humans and quantitative analysis using expensive machines.

Among these, the visual inspection approach not only requires experienced road managers, but also is time consuming and expensive. Furthermore, visual inspection tends to be inconsistent and unsustainable, which increases the risk associated with aging road infrastructure. Considering these issues, municipalities lacking the required resources do not conduct infrastructure inspections appropriately and frequently, increasing the risk posed by deteriorating structures.

In contrast, quantitative determination based on large-scale inspection, like using a mobile measurement system (MMS)[1] or laser-scanning method [2] is also widely conducted. An MMS obtains highly accurate geospatial information using a moving vehicle; this system comprises a global positioning system (GPS) unit, an internal measurement unit, digital measurable images, a digital camera, a laser scanner, and an omnidirectional video recorder. Though quantitative inspection is highly accurate,it is considerably expensive to conduct such comprehensive inspections especially for small municipalities that lack the required financial resources.

Therefore, considering the above mentioned issues, several attempts have been made to develop a method for analyzing road properties by using a combination of recordings by in-vehicle cameras and image processing technology to more efficiently inspect a road surface. For example, a previous study proposed an automated asphalt pavement crack detection method using image processing techniques and a naive Bayes-based machine-learning approach [3]. In addition, a pothole-detection system using a commercial black-box camera has been previously proposed [4]. In recent times, it has become possible to quite accurately analyze the damage to road surfaces using deep neural networks[5][6][7].For instance, Zhang et al introduced CrackNet [7], which predicts class scores for all pixels. However, such road damage detection methods focus only on the determination of the existence of damage. Though some studies do classify the damage based on types for example,Zalama et al [8] classified damage types vertically and horizontally, and Akarsu et al [9] categorized damage into three types, namely, vertical, horizontal, and crocodile most studies primarily focus on classifying damages between a few types. Therefore, for a practical damage detection model for use by municipalities, it is necessary to clearly distinguish and detect different types of road damage; this is because, depending on the type of damage, the road administrator needs to follow different approaches to rectify the damage.

Furthermore, the application of deep learning for road surface damage identification has been proposed by few studies, for example, studies by Maeda et al [6] and Zhang et al [4]. However, the method proposed by Maeda et al [6], which uses 256 × 256 pixel images, identifies the damaged road surfaces, but does not classify them into different types. In addition, the method of

Zhang et al [4] identifies whether damage occurred exclusively using a 99 × 99 patch obtained from a 3264 × 2448 pixel image. Further, a 256 × 256 pixel damage classifier is applied using a sliding window approach [10] for 5,888 × 3,584 pixel images in order to detect cracks on the concrete surface [11]. In these studies, classification methods are applied to input images and damage is detected. Recently, it has been reported that object detection using end-to-end deep learning is more accurate and has a faster processing speed than using a combination of classification methods. As an example of a method using end-to-end deep learning performing better than traditional methods, white line detection based on end-to-end deep learning using OverFeat [12] outperformed a previously proposed empirical method [13]. However, to the best of our knowledge, no example of the application of an end-to-end deep learning method for road damage detection exists. It is important to note that classification refers to labeling an image rather than an object, whereas detection means assigning an image a label and identifying the object's coordinates as exemplified by the ImageNet competition [14].

Therefore, considering this, we apply the end-to-end object detection method based on deep learning to the road surface damage detection problem, and verify its detection accuracy and processing speed. In particular, we examine whether we can detect eight classes of road damage by applying state-of-the-art object detection methods (discussed later in 2.2). Although many excellent methods have been proposed, such as segmentation of cracks on the concrete surface [15][16], our research uses an object detection method.

3.Methodology

3.1 Data Collection

Damage Type			Detail	Class Name
Crack	Linear Crack	Longitudinal	Wheel mark part	D00
			Construction joint part	D01
		Lateral	Equal interval	D10
			Construction joint part	D11
	Alligator Crack		Partial pavement, overall pavement	D20
Other Corruption			Rutting, bump, pothole, separation	D40
			White line blur	D43
			Cross walk blur	D44

Table 1. Road damage types in our dataset and their definitions.

The image dataset for road damage identification was provided by the IEEE BigData Cup Challenge 2020, and it includes three countries: Czech Republic, India, and Japan. A smartphone application collects each image in JPEG format. For training the road damage detector, a total of 26,620 images were gathered with a resolution of 720 x 720 pixels in the Czech Republic, 3,595 images in Czech, and 9,892 images in India, and 10,533 images with a resolution of 600 × 600 pixels in Fig(2). In addition, for the images from Japan, numerous additional damage categories were included to maintain consistency with prior versions of the dataset, the details of which may be seen in the research paper []. It's also worth mentioning that the dataset includes images of roads that don't have the aforementioned damage on the pavement surface.These images of uncracked roads have been included to help models created to identify road failures detect false positives.

Fig 2. Sample Images from India,Czech,Japan.

The collected images were annotated in PASCAL VOC format [ ] using the labelImg tool. The annotations include marking the road damage label and location in the image. The data is labeled in the PASCAL VOC [] format using the associated XML format(Fig 3). It contains information on the road damage label as well as the coordinates of where it appears in the image. The four categories of damage discussed before are mostly covered by the damage labels.

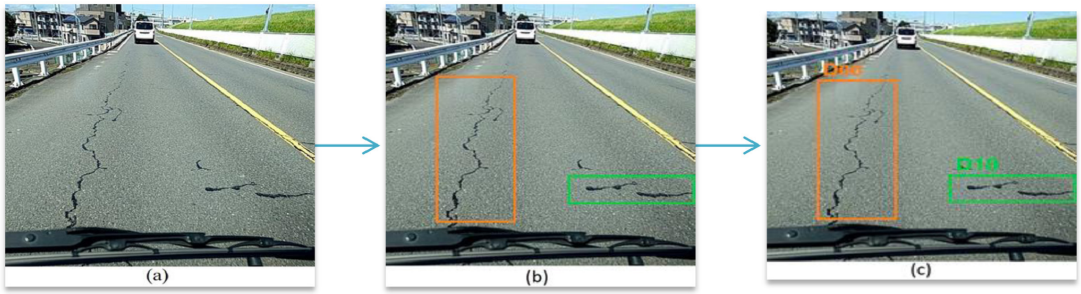


Fig 3 . Annotation Pipeline (a) original image, (b) image with bounding boxes, (c) final annotated image containing bounding boxes and class labels.

3.2 Object Detection Model

3.1.1 Yolov4 Scaled

YOLOv4 uses the Cross Stage Parietal (CSP) [24] Networks architecture as a feature extractor. It adds to the DarkNet-53 architecture the concept of cross-stage partial connection. This separates the input feature maps of a so-called dense block into two parts. The first part does not pass through the dense block and directly composes the input of the next transition layer. The second part passes through the dense block and then joins the first part at the input of the transition layer. [25] In this paper the author developed a model scaling technique based on YOLOv4 and ideas from EfficientNet proposed scaled-YOLOv4. They redesign YOLOv4 and propose YOLOv4-CSP, and then based on YOLOv4-CSP developed scaled-YOLOv4.In scaled-YOLOv4 they suggest the upper and lower bounds of linear scaling issues up/down

models and respectively analyzed the that need to be paid attention to in model scaling for small models and large models are YOLOv4 tiny and YOLOv4 large. Thus, scaled-YOLOv4 can achieve the best trade-off between speed and accuracy.

3.1.2 YOLOX

YOLOX [26] is a single-stage object detector that makes several modifications to YOLOv3 with a DarkNet-53 backbone. Specifically, the author replaced YOLO's head with a decoupled one and greatly appreciated the converging speed. It contains a 1 × 1 conv layer to reduce the channel dimension, followed by two parallel branches with two 3 × 3 conv layers respectively. With Yolox, the author has replaced Anchor-based detector with Anchor-free detector. Anchor-free mechanism significantly reduces the number of design parameters which need heuristic tuning and many tricks involved for good performance, making the detector, especially its training and decoding phase, considerably simpler.They add Mosaic and MixUp into our augmentation strategies to boost YOLOX's performance .Author used SimOTA Advanced label assignment not only reduces the training time but also avoids additional solver hyperparameters in SinkhornKnopp algorithm.Besides DarkNet53, They also test YOLOX on other backbones with different sizes, where YOLOX achieves consistent improvements against all the respective counterparts.

3.1.3 YOLOR

YOLOR (You only learn one) [27] also aims to make use of implicit and explicit knowledge and encode them as one (Fig 4).

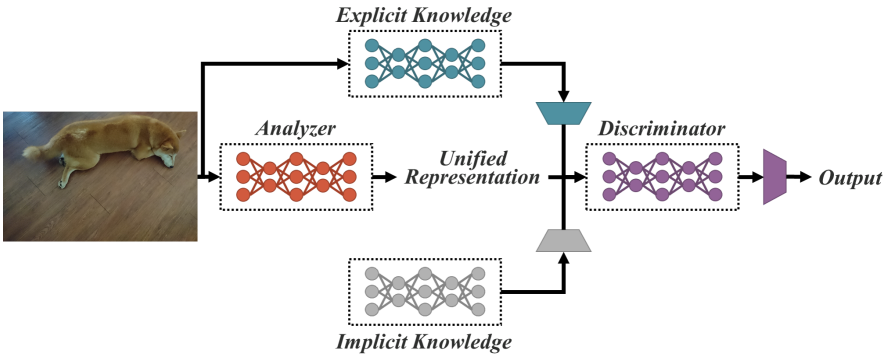


Fig 4. YOLOR concept with implicit and explicit knowledge-based multi-task learning

Explicit knowledge can be called any data that is immediately understood so that certain conclusions can be drawn ,and can be obtained using a shallow neural network. Implicit knowledge can be defined as detailed feature extraction from given data while raw feature extraction can be acquired using deep neural networks. The introduction of Explicit knowledge along with Implicit knowledge in this model leads to better performance in various tasks . The architecture of the YOLOR model proposes a single neural network to perform multiple tasks such as feature allocation, prediction refinement, and multi-task learning. Multi-task learning includes performing tasks such as object detection, multi-label image classification, feature embedding. Feature embedding refers to the process of extracting features and classifying them based on the properties of these features. All these features work with the help of 7 layers of complexity present in the ant YOLOR structure with maxpool class.

3.1.4 EfficientDet

Mixing Tan, et al researched and found that systematically balancing: network depth, width, and resolution can result in better performance for Convolutional Neural Networks. The author optimized the balance of all dimensions of network width, depth and resolution during ConvNet scaling and has created a model with the name EfficientNet [21]. The EfficientNet-B7 model achieved SOTA in CNN model accuracy at that time.

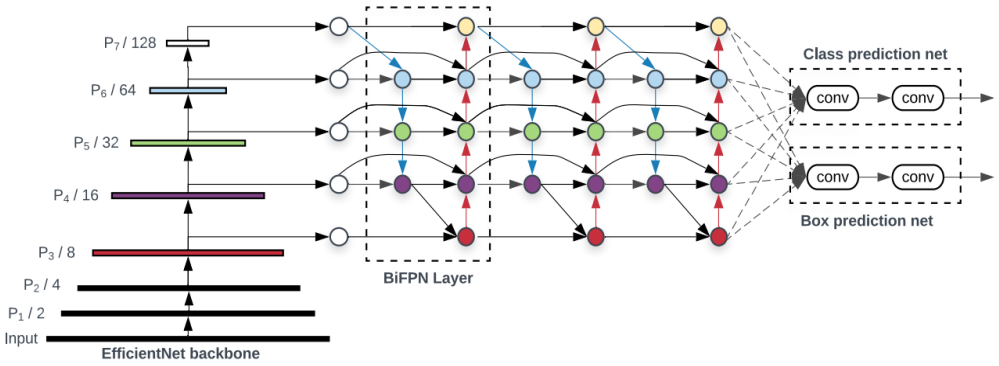


Fig 7. EfficientDet architecture

From previous work on the EfficientNet network, the authors studied an Object Detection model named EfficientDet [22].This model introduces BiFPN which is a two-way weighted feature network whose weights can be learned to determine the importance of different input features, applying multi-scale feature matching top-down and bottom up. Combining the two points: EfficentNet backbone and BiFPN above resulting in EfficientDet(Fig 7), a new family of object detectors. These models have significant better accuracy and efficiency across a wide spectrum of resource constraints.

3.1.5 DETR

As we all know, one of the features of One Stage Object Detection algorithms is the use of anchor boxes (like YOLO, SSD, Retina Net, etc.). This method has some problems like imbalance in labels (no. The number of anchor boxes is too large while the number of objects in an image is small, which makes empty-object labels a lot more than object labels). And the way is usually solved by the Non-maximum suppression algorithm (NMS). But in general, the implementation of models using anchor boxes is not easy.

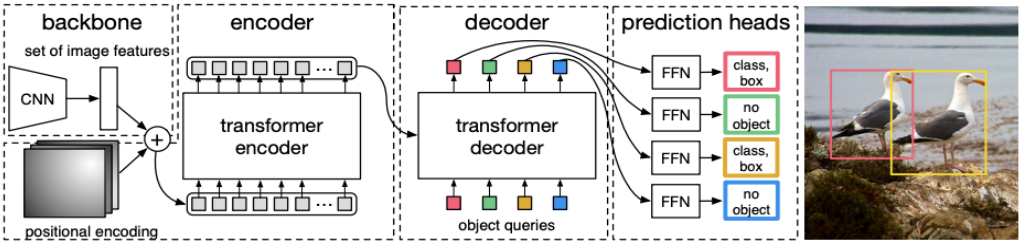


Fig 8 . DETR architecture

DETR was born with an anchor-free object detection approach that is easy to implement, solves the problem of imbalance classes and does not need NMS inference [23]. First, about the labeling mechanism, assuming the model outputs a certain number of boxes (and corresponding classes), how do we know which boxes will predict ground true objects? The OD transformer uses a Hungarian algorithm to solve this assignment problem. During training, the author used bipartite matching to uniquely assign predictions with ground truth boxes. Prediction with no match should yield a “no object” class prediction. Second, how can an OD transformer not use NMS, this means that different boxes must have each other's information to avoid duplicate predictions. The author takes the idea of attention mechanism Fig (8) with Objects queries of n

predictions that the model will output for each input image. After the backbone model extracts features of the image, those features will be passed through conv 11 to bring the channel size to model_dimension (512) and then flatten to include in the encoder . Then, the output of the encoder becomes memories and along with spatial positional embeddings is fed to the decoder multihead encoder-decoder attention at each layer of the decoder. In the decoder, slot predictions are initialized randomly. In each decoder layer, the slots will pay attention to each other, then the slots will pay attention to the encoder output, this is each slot prediction will pay attention to which region and what features on the input image, from there, giving prediction for objects in that region. Thus, the slots have each other's information to avoid overlapping predictions, and there will be no need to use NMS during prediction.

3.2 Mobile Application

In general, vehicles designed specifically for road in-spection are expensive. Meanwhile, mobile devices such as smartphones have made remarkable progress in recent years, and examples of road inspection using smartphone sensors are increasingly common. Using a smartphone is advantageous insofar as it is possible to inspect the road surface cheaply and exhaustively. For example, In [29] proposed a method to measure the flatness of a road using the accelerom-eter of a smartphone installed in a car. And [30] proposed a method that visualizes (on a map) potholes detection by smartphone sensors.

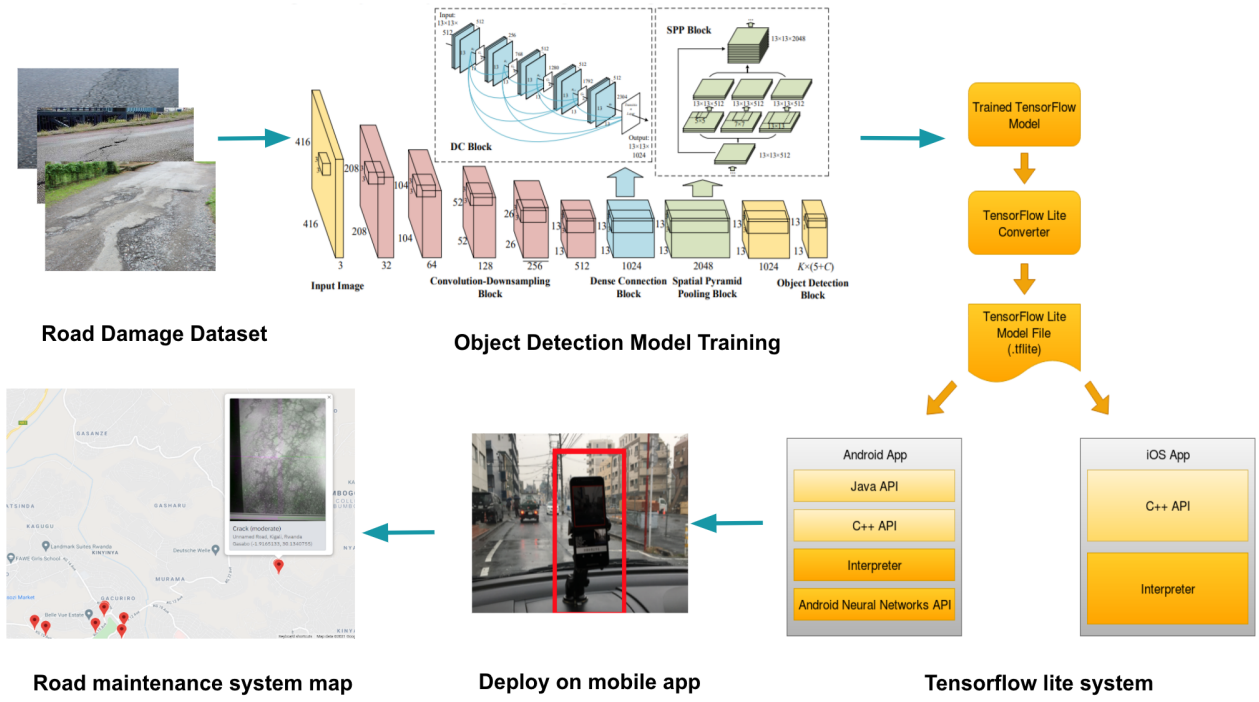


Fig 9. Steps to implement Mobile Application method

Therefore, we demon-strate that using end-to-end object detection systems is feasible for processing such images. The steps are shown in (Fig9) : Step 1) Road Damage Dataset collection .2) Use Object Detection model to train and evaluate dataset. 3) Tensorflow lite system to convert trained model to suitable format for mobile.4) Deploy mobile application. 5) Resend information when the app detects damaged roads to the road maintenance system.

4. Results

In this work we use the AP50 metric to evaluate each model for each pattern. We applied the Object Detection SOTA models (Scaled YOLOv4, YOLOX, YOLOR, EfficientNet, DETR) to test and compare on each class.

To evaluate the model, we use the AP_{50}metric, which is one of the widely used metrics in the field of object detection. When the Intersection Over Union (IOU) is 0.5, the average precision is AP@0.5IOU. The approach of evaluating objective detection methods is known as intersection over union. IOU requires a ground-truth bounding box, which is the pattern's labeled bounding box region. IOU additionally requires a third bounding box, which is the model's projected bounding box. The amount of overlap of two bounding boxes is divided by the area of the union to calculate the IOU score.

$$Precision=\frac{tb}{tb+fb}$$

$$Recall=\frac{tb}{tb+fn}$$

Where:

- tb:True positives mean patterns is recognized as patterns correctly.
- fp:False positives mean those are not patterns but detected as patterns.
- fn:False negatives mean pattens is not recognized by the model.

Value of AP is calculated as an area under the precision-recall curve calculated using the formula:

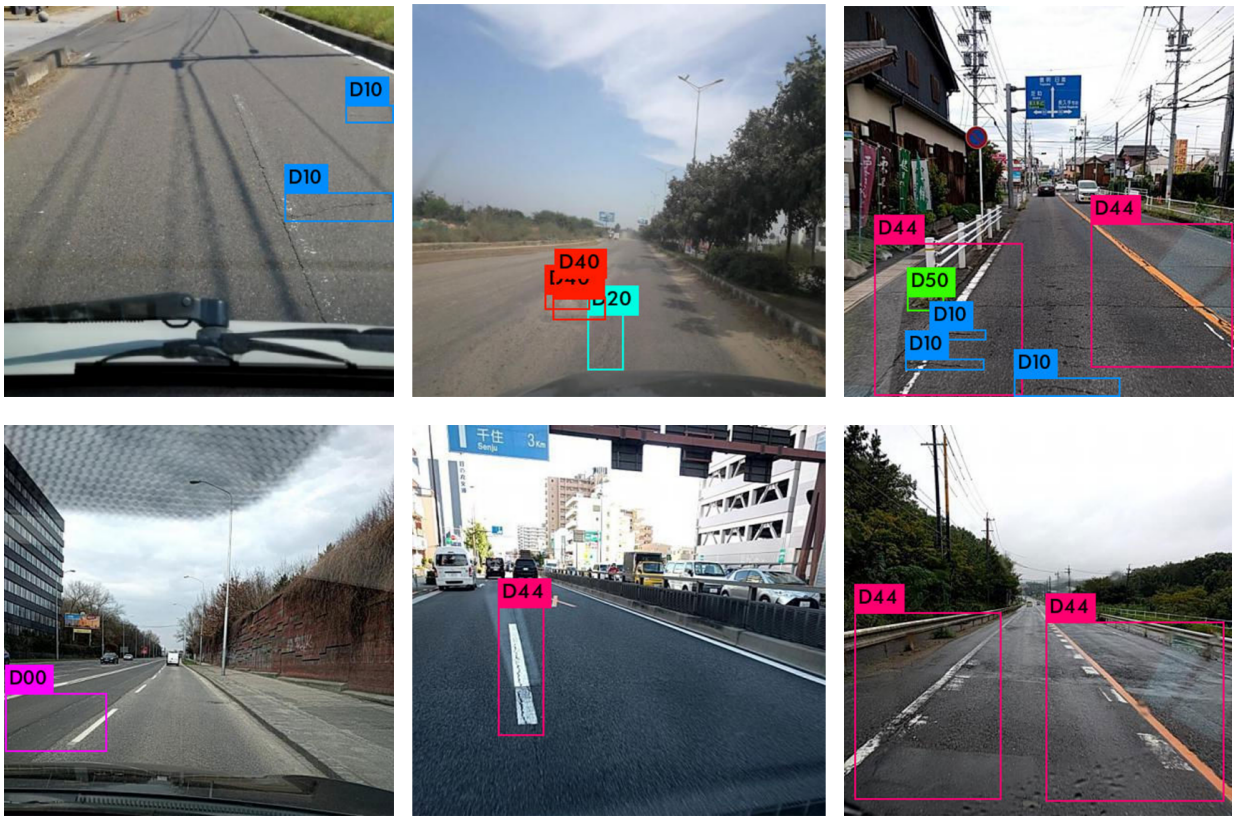
$$\int_0^1 p(r)dr$$

With p mean precision, r mean recall @0.5IOU indicates that if the IOU score is more than 0.5, the sample is positive; otherwise, it is negative. When IOU is 0.5,AP_{50} displays the model's accuracy.

Model(AP50)/ class	**D00**	**D01**	**D10**	**D11**	**D20**	**D40**	**D43**	**D44**
YOLOV4-S	79.4	78.6	78.1	80.9	73.3	72.7	77.8	78.4
YOLOX	95.4	**97.3**	**98.9**	96.3	97.2	**98.7**	**98.9**	95.6
YOLOR	93.9	89.7	96.2	**97.8**	**97.8**	97.9	92.4	95.3
EfficientDet	**98.2**	96.0	91.8	95.8	95.6	97.5	95.5	96.7
DERT	93.9	96.4	95.3	94.6	97.6	94.1	96.3	**97.2**

Table 2 .Although Efficient is the best at D00,in D44 best is the DERT model. Overall, the YOLOX and YOLOR yields better results. Next, the mAP of Scaled YOLOv4 is 77.4%. That means detection by smartphone sensors gives acceptable results.

As discussed in the Road damage detection section, identifying patterns with low cost will be useful in Road surface inspection. Therefore, for a practical damage detection model for use by municipalities, help the road administrator to follow different approaches to rectify the damage.



Figures 11.Detected samples using Scaled YOLOv4.

5. Conclusion

In this work, we used a comprehensive data collection for identifying and categorizing road damage. We gathered about thirty thousand road pictures in Japan at most and also in India and Czech. These road damage images were subsequently visually verified, divided into eight classes, and almost all of them were annotated before being made available as a training data set and test set. Based on the findings, throughed training with several randoms, finally it went to YOLOv4 tiny as the surprisingly unexpected time-training but also effective.

6.References

[1] KOKUSAI KOGYO CO., L. (2016). Mms(mobile measurement system).

[2] Yu, X. and Salari, E. (2011). Pavement pothole detection and severity measurement using laser imaging. In Electro/Information Technology (EIT), 2011 IEEE International Conference on, pages 1–5. IEEE

[3] Chun, P.-j., Hashimoto, K., Kataoka, N., Kuramoto, N., and Ohga, M. (2015). Asphalt pavement crack detection using image processing and na¨ıve bayes based machine learning approach. Journal of Japan Society of Civil Engineers, Ser. E1 (Pavement Engineering), 70(3).

[4] Jo, Y. and Ryu, S. (2015). Pothole detection system using a black-box camera. Sensors, 15(11):29316–29331

[5] Zhang, L., Yang, F., Zhang, Y. D., and Zhu, Y. J. (2016). Road crack detection using deep convolutional neural networks. In Image Processing (ICIP), 2016 IEEE International Conference on, pages 3708–3712. IEEE.

[6] Maeda, H., Sekimoto, Y., and Seto, T. (2016). Lightweight road manager: smartphone-based automatic determination of road damage status by deep neural network. In Proceedings of the 5th ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, pages 37–45. ACM

[7] Zhang, A., Wang, K. C., Li, B., Yang, E., Dai, X., Peng, Y., Fei, Y., Liu, Y., Li, J. Q., and Chen, C. (2017). Automated pixel-level pavement crack detection on 3d asphalt surfaces using a deep-learning network. Computer-Aided Civil and Infrastructure Engineering, 32(10):805–819.)
[8] Zalama, E., Gomez-Garc ´ ´ıa-Bermejo, J., Medina, R., and Llamas, J. (2014). Road crack detection using visual features extracted by gabor filters. Computer-Aided Civil and Infrastructure Engineering, 29(5):342–358.
[9] Akarsu, B., KARAKOSE, M., PARLAK, K., Erhan, A., ¨ and SARIMADEN, A. (2016). A fast and adaptive road defect detection approach using computer vision with real time implementation
[10] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence, 32(9):1627–1645.
[11] Cha, Y.-J., Choi, W., and Buy¨ uk¨ ozt ¨ urk, O. (2017). Deep ¨ learning-based crack damage detection using convolutional neural networks. Computer-Aided Civil and Infrastructure Engineering, 32(5):361–378.
[12] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.
[13] Huval, B., Wang, T., Tandon, S., Kiske, J., Song, W., Pazhayampallil, J., Andriluka, M., Rajpurkar, P., Migimatsu, T., Cheng-Yue, R., et al. (2015). An empirical evaluation of deep learning on highway driving. arXiv preprint arXiv:1504.01716.
[14] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 248–255. IEEE.
[15] O'Byrne, M., Ghosh, B., Schoefs, F., and Pakrashi, V. (2014). Regionally enhanced multiphase segmentation technique for damaged surfaces. Computer-Aided Civil and Infrastructure Engineering, 29(9):644–658
[16] Nishikawa, T., Yoshida, J., Sugiyama, T., and Fujino, Y. (2012). Concrete crack detection by multiple sequential image filtering. Computer-Aided Civil and Infrastructure Engineering, 27(1):29–47.
[17] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on comput
[18] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, pages 1440–1448.
[19] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Compute
[20] Redmon, J. and Farhadi, A. (2016). Yolo9000: better, faster, stronger. ArXiv:1612.08242.
[21] MixingTan,QuocV.Le,2020.EfficientDet:Scalable And Efficient Object Detection. ArXiv:1911.09070
[22] Nicolas Carrion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier,Alexander Kirillov, SergeyZagoruyko,2020.End-to-End Object Detection with Transformers .ArXiv:2005.12872v3.
[23] Aleksey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao 2020.YOLOv4: Optimal Speed and Accuracy of Object Detection.In: ArXiv preprint .ArXiv:2004.10934
[24] Chien-Yao Wang, Alexey Bochkovskiy, Hong-Yuan Mark Liao 2020.Scaled-YOLOv4: Scaling Cross Stage Partial Network .In: ArXiv:2011.08036
[25] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, Jian Sun. 2021 YOLOX: Exceeding YOLO Series in 2021".In: ArXiv:2107.08430
[26] Chien-Yao Wang,I-Hau Yeh,Hong-Yuan Mark Liao 2021.You Only Learn One Representation: Unified Network for Multiple Tasks.ArXiv:2105.04206
[27] Mingxing Tan, Quoc V. Le 2019,EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,ArXiv:1905.11946
[28] Buttlar, W. G. & Islam, M. S. (2014). Integration of Smart-Phone-Based Pavement Roughness Data Collection Tool with Asset Management System.USDOT region ,Regional University Transportation Center, NEXTRANS Center, West Lafayette, IN
[29] Casas-Avellaneda,D.A.&Lopez-Parra, J. F. (2016), Detection and localization of potholes in roadways using smart-phones, Dyna,83(195), 156–62.