Real-Time Deep Learning based Road Deterioration Detection for Smart Cities

Nusrat Mehajabin
Department of Electrical and Computer Engineering
The University of British Columbia
Vancouver, Canada
nusratm@ece.ubc.ca

Zhenchao Ma
Department of Electrical and Computer Engineering
The University of British Columbia
Vancouver, Canada
zhenchaoma@ece.ubc.ca

Yixiao Wang
Department of Electrical and Computer Engineering
The University of British Columbia
Vancouver, Canada
yixiaow@ece.ubc.ca

Hamid Reza Tohidypour
Department of Electrical and Computer Engineering
The University of British Columbia
Vancouver, Canada
htohidyp@ece.ubc.ca

Panos Nasiopoulos
Department of Electrical and Computer Engineering
The University of British Columbia
Vancouver, Canada
panosn@ece.ubc.ca

***Abstract*—Timely road condition inspection and maintenance are key components of infrastructure management for smart cities, as they reduce traffic congestion, accidents and repairing costs. Traditional road inspection methods that employ vibrations and/or laser scanning for detecting road deterioration use expensive equipment and dedicated municipality vehicles. Recently, computer vision techniques and artificial intelligence are emerging as alternative solutions to traditional approaches for road condition detection, offering more flexibility, higher accuracy, and overall lower cost. In this paper, we utilize convolutional neural network-based and vision transformer-based object detection models to accurately identify road deteriorations namely, potholes, cracks, and alligators. We compare four different state-of-the-art models in terms of detection accuracy and speed. Performance evaluations have shown that, on the same dataset the Swin Transformer model outperformed the other state-of-the-art methods by a substantial margin. With 74% detection accuracy, and 42 frames per second processing speed Swin Transformer exceled over EfficentDet, YOLOv4, and YOLOX. We also present a new comprehensive and balanced large-scale road condition dataset of 27,298 annotated images, captured by ordinary car cameras.**

***Index Terms*—Road deterioration, Object detection, Deep learning, YOLOv4, YOLOX, Swin Transformer, Transmission**

I. INTRODUCTION

As one of the most important topics in traffic engineering, road safety plays a pivotal role in safeguarding the lives of people using roads and highways in their day to day life. Road deterioration is considered one of the major contributing factors to traffic congestion and accidents, along with driving awareness, speeding and weather conditions [1]. Timely road repair is crucial in maintaining high quality roads, effectively preventing further deterioration and thus reducing road maintenance costs for municipalities. As road maintenance is a main component of city infrastructure management planning, municipalities have specific engineering divisions devoted to that task. Certified inspectors and engineering teams are responsible for frequently inspecting roads for damages, a task that is laborious, expensive, time consuming and prone to error. In addition to human observation, quantitative analysis using expensive equipment and specialized vehicles is another approach employed to inspect street conditions.

Due to the unsustainable and unscalable nature of the above-mentioned solutions, automatic road damage inspection has emerged as an important field of research. The literature includes road damage detection methods based on vibration [2], laser-scanning [3] and imaging [4] data analysis. These methods, however, have limitations as they either need specialized equipment that comes in contact with the road, or expensive scanning mechanisms attached to specialized vehicles. In recent times, computer vision aided techniques are being used to address these limitations, ranging from 3D imaging [5] [6] to remote sensing and artificial intelligence [7-10]. In [7] a unique deep learning network, known as CrackNet, is trained using 3D images to detect road deterioration. However, having access to 3D video feed requires specialized camera equipment and is not a practical solution. A machine learning-based method is presented in [9], which simply focuses on determining if there is damage on the road. One of the first most comprehensive, large scale, road damage datasets, which is captured in Japan, is presented in [11]. The same work introduces an object detection method that classifies these damages into eight different categories. A latest work on road deterioration is presented by Sadra et al., which uses an EfficientDet network for pavement crack detection [12]. Another recent work, [22] compares the effectiveness of YOLOv5 against Faster-RCNN [23] network and finds Faster-RCNN outperforms YOLOv5. However, the dataset used is unbalanced hence higher number of false positives encountered.

In this paper, we train and evaluate four different state-of-the-art object detection and classification networks to accurately and efficiently detect and classify three types of road deteriorations namely, cracks, potholes and alligators. The goal here is to assess the performance and generalizability of two single stage deep learning based object detection and classification networks namely, YOLOv4 and YOLOX and two two-stage networks namely EfficientDet and hierarchical vision transformer network known as Swin Transformer on a balanced dataset. Our objective is to utilize all human-driven and autonomous vehicles equipped with cameras to use our dedicated deep learning model designed for detecting road conditions and deteriorations and share this information with neighboring vehicles and municipalities. The end result will be avoidance of possible accidents and timely road maintenance, both aiming at reduced traffic and improved transportation. Performance evaluations identify the best method in terms of detection accuracy and inference (testing) time.

978-1-6654-6975-3/22/$31.00 ©2022 IEEE

The major contributions of this work are as follows:

- To the best of our knowledge, this is the first extensive comparison of road deterioration detection capabilities of four (EfficientDet, YOLOv4, YOLOX, and Swin Transformer) state-of-the-art object detection and classification models.
- We present the first ever annotated Canadian road deterioration condition dataset consisting of 27,298 images with three different deterioration classes.

The rest of this paper is organized as follows. In Section II, we present the new dataset along with the details of training and evaluation method. Section III presents the experiments conducted and discuss the performance of the different methods and discusses the results. Finally, in Section IV we conclude the paper with future directions.

II. Dataset and Networks

In this section, we present our balanced road deterioration dataset. There are three different kinds of road damages captured. For the object detection and classification network, initially, we considered six different widely used network architectures, namely YOLOv3, YOLOv4, YOLOX, Faster-RCNN, EfficientDet, and Swin Transformer and investigated their appropriateness for our task. Our preliminary tests showed that EfficientDet, YOLOv4, YOLOX and Swin Transformer outperformed the other networks. As a result, we chose these four to be our deep learning network architectures. The following subsections explain in detail the collection procedure of our new dataset, the labeling process, the distribution of the classes and the modifications we made to optimize the performance of the chosen networks for our task.

A. Data Collection and Labeling

Our dataset consists of a comprehensive number of videos captured by our team in the city of Vancouver, Canada. In total, our dataset contains 45 unique short video sequences. The videos capture different weather conditions (e.g., sunny, rainy, cloudy), illumination conditions (e.g., daytime, presence and absence of streetlights) and residential and freeway driving conditions. All videos are captured by cameras mounted on vehicles. After careful observation of all the videos and the types of deterioration present in them, we decided to narrow down the classes into three general road deterioration categories: potholes, cracks, and alligators. Through visual inspection, we chose 27,298 frames out of the captured videos and all the potholes, cracks and alligators in

Fig. 1. Examples of labeled frames of our dataset, showing potholes in red, cracks in yellow and alligators in green.

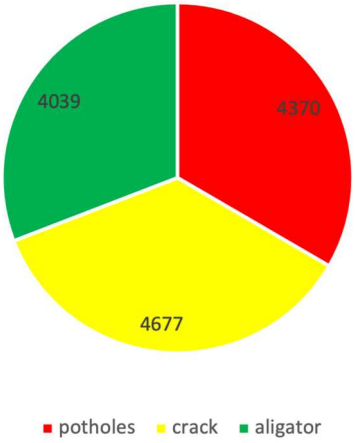

Fig. 2. Number of class instances appearing in our labeled frames

these frames were labeled using the Computer Vision Annotation Tool (CVAT) [14]. It is worth noting that all 30 fps were labelled. We came to that conclusion after visually checking the frames and realizing that they are significantly different due to the 40 km/h average car speed. Labelled data is in MS COCO [24] annotation format for ease of using with several different networks. Fig. 1 illustrates examples of labeled frames of our dataset, with potholes depicted in red, cracks in yellow and alligators in green.

B. Data Distribution

It is advised when training object detection networks, to use a balanced dataset of labeled and empty (no label present) frames [15]. Based on that recommendation, we randomly removed 5164 empty frames – frames which did not have any of the 3 classes from our training and validation dataset, resulting in 13,679 labeled frames and equal number of empty frames, for a total of 27,298 frames. Fig. 2 shows the number of class instances appearing in the labeled frames. We observe that in the 13,679 instances, the three classes (potholes, cracks and alligators) are almost equally distributed (4370 potholes, 4677 cracks and 4039 alligators). It is worth mentioning that class balance is one of the most influential factors for the accurate predictive performance of classifiers. In the context of road deterioration detection, an imbalanced dataset has shown to cause models to predict indistinguishable data as the majority class [11].

C. Training the Deep Learning Networks

As we mentioned above, we chose YOLOv4, YOLOX and Swin Transformer as our classification and object detection deep learning networks. And EfficientDet to compare our results with [12]. The main reasons for these choices are the following unique properties of each network. Both YOLOv4

TABLE I. VALIDATION RESULTS FOR EACH CATEGORY FOR THE IOU OF 0.50 AND 0.75 WITH OUR YOLOV4 MODEL

Class Name	**Threshold$_{0.5}$**	**Threshold$_{0.75}$**	**AP$_{0.5}$**	**AP$_{0.75}$**
Potholes	TP = 3455 FP = 364	TP = 2861 FP = 958	98.05%	72.81%
Cracks	TP = 3230 FP = 935	TP = 1448 FP = 2717	88.66%	24.92%
Alligators	TP = 2991 FP = 619	TP = 1726 FP = 1884	94.49%	37.17%

TABLE II. DISTRIBUTION OF THE THREE CLASSES FOR TRAINING AND VALIDATION

Class Name	**Categories Statistics**		
	Training Data	***Validation Data***	***Total***
Potholes	3980	390	4370
Cracks	4227	450	4677
Alligators	3670	369	4039

and YOLOX are one-stage object detection networks. However, the former is anchor based whereas the latter is anchor free. As they are single-stage they are faster and proven to yield accurate results in real-time. On the other hand, Swin Transformer is a two-stage vision transformer computed using hierarchical shifted windows and is known for its linear computational complexity with respect to image size. Therefore, the selection of networks ensure variety in complexity and speed.

As a first step, we trained YOLOv4 starting with the pretrained weights in the darknet framework (CSPDarknet-53) [15]. The anchor boxes are generated by clustering the dimensions of the ground truth boxes from the original dataset, to find the most common shapes/sizes. Since the anchor box in YOLOv4 is content-dependent, before training we had to recalculate the anchor box for our implementation. However, for YOLOX no such recalculation was necessary as it is an anchor free model. Swin Transformer like YOLOX did not require any preprocessing. The training time for YOLOv4 and YOLOX were 38 and 36.5 hours respectively.

We used the input size of 608×608 to be able to detect small objects, as smaller sizes would reduce the accuracy of the network. We set the batch size to 64 for YOLOv4 and YOLOX and 16 for Swin Transformer, the learning rate to 0.0001, and changed the number of filters in the layers related to the number of classes accordingly, which in our case is 3. Training time for Swin Transformer was 66 hours. Our dataset of 27,298 frames was divided into 90% for training and 10% for validation. Table I shows the distribution of the three classes for training and validation. We trained the networks using the Tesla V100 SXM2 GPU, with 32GB of HBM2 memory and 640 tensor cores available by a state-of-the-art computing cluster [16].

D. Performance Metrics

In order to evaluate the networks, we use mean Average Precision (mAP), which is obtained by taking the combined weighted average of the average precision (AP) [17-19] for all category tests. In image recognition and classification, mAP is a common and efficient metric for evaluating models [10], calculated as follows:

$$mAP = \frac{1}{|Q_R|}\sum_{q \in Q_R} AP(q) \qquad (1)$$

where Q_R refers to the number of validation sets and AP is the average precision. Other than object recognition accuracy, the success of our approach relies on the ability of the network to localize each class. For this purpose, we use Intersection over Union (IoU), which is a measure of the degree of intersection of two detection frames (for target detection) [20] [21]. IoU is a reliable approach for measuring the amount of overlap between two bounding boxes or segmentation masks and is given by the following equation:

$$IoU = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \quad (2)$$

where B_p and B_{gt} represent the predicted and the ground-truth bounding box, respectively. More precisely, the IoU value represents the overlap area between the predicted bounding box and the real bounding box. A large IoU value indicates that the prediction is accurate and the parameters of the predicted bounding box can better represent the position of the target object in the image.

We chose the IoU threshold of 0.50, which is considered a good threshold for the predicted bounding box. Fig. 3 shows the Average Loss and mAP@0.5 (mAP with the IoU threshold of 0.50) for training the YOLOv4 darknet using our custom dataset with input image size of 608×608 pixels. It took our network approximately 13 hours to achieve the best weights at acceptable loss of 0.70 and mAP@0.5 levels of 94%.

In order to further evaluate the accuracy performance of the model and to make sure the 0.50 is an acceptable threshold choice for our task, we examined different Intersections over Union (IoU) thresholds ranging from 0.25 to 0.75. Our results showed that threshold 0.50 led to the best trade-off between Average Precision and localization. Table I shows the true positive (TP), false positive (FP) values, and Average Precision accuracy of each class for threshold values of 0.50 and 0.75 for YOLOv4. Similar results were observed for YOLOX and Swin Transformer. We observe that for threshold 0.50, the Average Precision of detecting potholes is the highest at 98.05%, while the lowest precision of 88.66% is corresponds to cracks. We also observe from Table I that as we increase the threshold from 0.50 to 0.75, the number of TPs drops while the FPs increases for all 3 classes, the accuracy (AP) for each class decreases significantly. This drop is much worse for the case of cracks and alligators. This is expected, as potholes are very distinct compared to cracks and alligators that often look alike. Recall that in our implementation, the main objective is to diagnose the presence of these deteriorations and not their accurate position in a frame, so the threshold of 50% for the Intersection on Union is more than acceptable for our task.

III. EVALUATION AND DISCUSSION

To better evaluate the performance of the three models, we decided to compare them with the latest state-of-the-art deep learning method presented in [12]. To this end, we had to retrain the EfficientDet network used in [12] using our custom dataset and modify it to detect the three classes of potholes, cracks and alligators. During training the mean and standard deviation is calculated over our dataset and used for normalization.

TABLE III. MODEL PERFORMANCE DETAILS WITH DIFFERENT THRESHOLD ON OUR CUSTOM DATASET

Model	Backbone	$mAP_{0.50}$	Inference Time
EfficientDet	Efficient-B0	52.70%	53.1ms/18fps
YOLOv4	CSPDarknet-53	66.32%	24.3ms/41fps
YOLOX	Yolov3 with Darknet-53	70.11%	18.7ms/53fps
Swin T	Mask RCNN	74.00%	23.5ms/42fps

A. Experiment Setup

The machine used for the experiments had two Intel Silver 4216 Cascade Lake 2.1GHz CPU, 40GB RAM, with Linux Operating system, and 99.4% of CPU was utilized with 30-70% memory efficiency. A Tesla V100 SXM2-32GB GPU was used. The EfficientDet model D0 with the efficient-B0 as backbone was used and the input size for this specific version of EfficientDet is fixed to 512×512. Backbone networks used for the other networks are presented in Table III.

B. Results and Discussion

The EfficientDet achieved an acceptable loss of 0.39 and mAP@0.5 levels of 52.70% on the validation images for the best weight sets, using our custom dataset. Similar to the validation stage for the other networks, we examined different Intersections over Union thresholds ranging from 0.25 to 0.75 for EfficientDet. Our results again verified, threshold 0.50 led to the best trade-off between Average Precision and localization for EfficientDet. Thus, we decided to use this threshold for the test stage of the networks. We tested the YOLOv4, YOLOX, Swin Transformer and the EfficientDet with 529 frames of an unseen images. Table III shows the mean average precision performance of detecting potholes, cracks and alligators by the models, using our dataset. From the table we see that, YOLOv4 achieves an mAP of 66.32% for the threshold of 0.5, which is 13.62% higher than that achieved by the state-of-the-art EfficientDet. Swin Transformer achieves 74% testing mAP which is +7.68 mAP compared to YOLOv4. We also observe that the inference (test) time for YOLOv4 is 24.3 ms per frame, while that of EfficientDet is 53.1 ms per frame. This shows that YOLOv4 achieves high detection accuracy in real-time (41.2 frames/s), while EfficientDet has lower accuracy and can process only 18.8 frames per second approximately (much lower than the 30fps that corresponds to real-time). Swin Transformer achieves even higher detection speed of 23.5 ms per frame leading up to 42 frames/s. The best performance in terms of latency is however, by the YOLOX model. This model has an mAP@0.5 of 70% with detection speed of 53 frames/s. From our analysis of the network, it is evident that Swin Transformer performs better compared to all the other popular object detection and classification models in terms of accuracy. Though Swin Transformer processes 11 frames less than YOLOX per second, Swin Transformer still achieves real time performance. We also report the two best models' (YOLOX and Swin Transformer) complexities in terms of number of parameters in Table. IV. We observe that the Swin model has significantly fewer parameters than YOLOX while using a larger input resolution. In conclusion, Swin is a more suitable solution for on board deep learning units of vehicles offering higher accuracy, more input flexibility and real-time performance.

In Fig. 4 we present the road damage detection results with prediction confidence from the three tested models for some of the video sequences. Fig. 4 (a), (b) illustrates the results from YOLOX, (c), (d) from Swin Transformer and (e), (f) from YOLOv4.

IV. CONCLUSIONS

In this paper, we employed three state-of-the-art object detection and classification network to solve the road deterioration detection problem accurately and in real-time. To this end, we captured a new comprehensive large-scale dataset using ordinary car cameras and annotated the captured

Fig. 3. Examples of road deterioration detection using (a), (b) YOLOX (c), (d) Swin Transformer and (e), (f) YOLOv4

TABLE IV. MODEL COMPLEXITY COMPARISON

Model	# Parameters	Input Size
YOLOX-x	99.00M	640×640
Swin Transformer-s	66.07M	1333×800

data for potholes, cracks and alligators. Our system is designed to be scalable and sustainable as it works with any vehicle equipped with a camera and GPS. Evaluation results have shown that the Swin Transformer model outperformed the other methods in terms of accuracy, and model complexity while being comparable in terms of latency.

Future work involves the design of an end-to-end solution, involving deployment of efficient data exchange and intelligent collaboration between the multiple levels of edge nodes such as vehicles and road side units as well as cloud servers.

ACKNOWLEDGMENT

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC – PG 11R12450), and TELUS (PG 11R10321). This research was enabled in part by support provided by WestGrid (www.westgrid.ca) and Compute Canada (www.computecanada.ca).

REFERENCES

[1] E. Chung, O. Ohtani, H. Warita, M. Kuwahara, and H. Morita (2005), "Effect of rain on travel demand and traffic accidents," in Proc. 8th Int. IEEE Conf. Intell. Transp. Syst., Sep. 13–16.

[2] P. M. Harikrishnan and V. P. Gopi (2017), "Vehicle vibration signal processing for road surface monitoring, " in IEEE Sens. J., vol. 17, no. 16, pp. 5192-5197, 15 Aug.15.

[3] X. Yu and E. Salari (2011), "Pavement pothole detection and severity measurement using laser imaging," in Proc. IEEE Int. Conf. Electro Inform. Technol., May.

[4] P. J. Chun, K. Hashimoto, N. Kataoka, N. Kuramoto, and M. Ohga (2015), "Asphalt pavement crack detection using image processing and naive bayes based machine learning approach," Journal of Japan Society of Civil Engineers, Ser. E1 (Pavement Engineering), vol. 70, no. 3, pp. I1–I8.

[5] S. Ryu, T. Kim and Y. Kim (2015), "Image-based pothole detection system for its service and road management system," Math. Probl. Eng., vol. 2015.

[6] G. Vitor, D. Lima, A. Victorino, and J. Ferreira (2013), "A 2D/3D vision based approach applied to road detection in urban environments," in Proc. IEEE IV, Jun..

[7] A. Zhang, K. C. P. Wang, B. Li, E. Yang, X. Dai, Y. Peng, Y. Fei, Y. Liu, J. Q. Li, and C. Chen. 2017. "Automated pixel-level pavement

crack detection on 3D asphalt surfaces using a deep-learning network," Computer-Aided Civil and Infrastructure Engineering 32, 10 (2017), pp. 805-819.

[8] E. Schnebele, B. Tanyu, G. Cervone, and N. Waters (2015), "Review of remote sensing methodologies for pavement management and assessment," Eur. Transp. Res. Rev., vol. 7, no. 2, pp. 1–19.

[9] Q. Shi, X. Liu, and X. Li (2017), "Road detection from remote sensing images by generative adversarial networks," IEEE Access, vol. 6, pp. 25486–25494.

[10] Y. Fei, K. C. P. Wang, A. Zhang, C, Chen, J. Q. Li, Y. Liu, G. Yang, and B. Li (2017b), "Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network," J. Comput.-Aided Civil Infrastruct. Eng., vol. 32, no. 10, pp. 805–819.

[11] B. Akarsu, M. Karakose, K. Parlak, A. Erhan, and A. Sarimaden (2016), "A fast and adaptive road defect detection approach using computer vision with real time implementation," IJAMEC, vol. 4, no. 1, pp. 290–295.

[12] N. Sadra, M. Naddaf-Sh, A. R. Kashani, and H. Zargarzadeh (2020), "An efficient and scalable deep learning approach for road damage detection," In Proc. IEEE Int. Conf. Big Data, Dec..

[13] A. Bochkovskiy, C. Wang, and H. M. Liao (2020), "Yolov4: Optimal speed and accuracy of object detection." In Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit, Jun..

[14] OpenVINO Toolkit, 2020. GitHub repository, https://github.com/openvinotoolkit/cvat

[15] YOLOv4 darknet. GitHub repository, https://github.com/AlexeyAB/darknet.

[16] Compute Canada state-of-the-art advanced research computing network. Available from: https://www.computecanada.ca.

[17] M. Everingham, and J. Winn. 2012. The pascal visual object classes challenge 2012 (voc2012) development kit. Pattern Analysis, Statistical Modelling and Computational Learning. Tech. Rep 8 (2012).

[18] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick (2015), "Microsoft COCO: Common Objects in Context," in Proc. ECCV..

[19] M. Everingham, L. V. Gool, C. KI Williams, J. Winn, and A. Zisserman (2010), "The pascal visual object classes (voc) challenge. International journal of computer vision," 88, 2 (2010), pp. 303-338.

[20] E. Bochinski, T. Senst, and T. Sikora (2018), "Extending IOU based multi-object tracking by visual information," In Proc of the 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 1-6.

[21] L. Tychsen-Smith, and L. Petersson (2018), "Improving object localization with fitness nms and bounded iou loss In Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit, IEEE, 6877-6885.

[22] R. Vishwakarma, and V. Ravigopal (2020), "Cnn model & tuning for global road damage detection." In Proc *IEEE International Conference on Big Data (Big Data)*. IEEE, 5609-5615.

[23] Ren, S., He, K., Girshick, R., and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

[24] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." *European conference on computer vision*. Springer, Cham, 2014.