Improvements of YoloV3 for road damage detection

Qinglang Wang, Jingchao Mao, Xu Zhai, Jie Gui, Wenjie Shen and Ye Liu*

College of Automation & College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, 210023, China

*Corresponding author’s e-mail: yeliu@njupt.edu.cn

Abstract. Automatic detecting road damage from images is a challenging problem. Recent advances in deep learning based detectors offer a powerful tool for resolving this problem. However, these deep learning detectors are designed for detecting generic objects, the specific characteristics of road damage are not considered in these methods. We propose in this paper an improved method based on YoloV3 that takes in consideration the slenderness and tininess nature of the road damage, which require low-level and detailed description. Our method fuses the low-level feature with high level feature to enhance the description power of the network, and also an improvement to the loss function further boosts the detection performance. Experimental results have demonstrated the effectiveness of proposed method.

1. Introduction
Inspecting the road to find road damages is crucial for road maintenance, however, this task costs a lot of manpower. Using computer vision techniques especially deep learning techniques [1][2][4][5][6][7] to automatically detect road damages from images captured from a video camera is promising. However automatic detection from image is a non-trivial task due to:

1) Most of the road damages are slender cracks which is significant different from conventional objects in public detection datasets such as MS COCO.

2) The road damages may be long and discontinuous thus the detector may fail to predict the correct bounding box that covers the complete damage area.

Based on the observations that most of the road damages are long and slender cracks, we integrate the low-level features to the detection head of YoloV3. These low-level features are from the very beginning of the network backbone, which although much less semantic than high-level features captured the visual detail of the road cracks. Then based on observations a large portion of the detection failures are due to incomplete bounding box prediction in the direction of crack extends, we propose an improved loss function which pays special attention to the direction of road crack extension. Experimental results have demonstrated that the improvements can effectively boost the performance of detection.

2. Method

2.1. YoloV3 detector
We chose YoloV3 as the baseline detector. YoloV3[7] is the improved version of YOLO series [4][6] deep learning object detector. YoloV3 introduced a backbone network named Darknet53 which contains multiple residual blocks like Resnet[3]. Different from its previous versions, YoloV3 incorporates the anchoring mechanism which is originated from Fast RCNN[1] and Faster RCNN[2], and it introduces
the FPN[10] mechanism into the network, which integrates different level of features. The network structure is illustrated in Figure 1 (without the red dashed arrows). Although FPN can to some extent fuse low-level and high-level features, the fused feature maps are from relatively high-level part of the network and the feature maps from earlier layers are ignored. Although these feature maps contain less semantic information than high-level features, the positional information and visual detail is better preserved which is helpful for problems like road damage detection.

![Figure 1. Network structure of YoloV3 and our improvements (red arrow).](image)

2.2. Improvements for road damage detection

Figure 2 gives two examples of road damages, the width of the crack may be only several pixels, features from high-level layers may lose the visual details of these road damages.

Based on the above analysis, we modify the network as follows: the output feature map of the 2nd convolutional layer is fused with the output of the 1st residual block via feature concatenation. Since their sizes are different, we adopt the mean pooling with stride 2 to down-sample the feature map. And the fused feature map is sent to the three detection branches of YoloV3, where it is down-sampled (if necessary) and fused with original feature maps via feature concatenation. This is illustrated in the red arrows in Figure 1. In this way, the features from early layers are fused with high-level features which encode both semantic information and visual details.
2.3. crack direction-aware loss

Our second improvement is based on the observation that original YoloV3 detector tends to predict incomplete bounding boxes along the direction road crack extends. The loss of YoloV3 consists of three parts: a localization loss $l_{bbox}$, an objectness loss: $l_{obj}$ and a classification loss: $l_{class}$. The final loss is the weighted sum of the three losses. We are interested in the localization loss, which measures the coordinate deviation and shape differences between predicted bounding box and ground truth.

$$l_{bbox} = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \sum_{k=0}^{1} \sum_{l=0}^{1} (2 - w_i \times \hat{h}_i) [D_{center} + D_{wh}]$$

$$D_{center} = (x_i - \bar{x})^2 + (y_i - \bar{y})^2$$

$$D_{wh} = (w_i - \hat{w}_i)^2 + (h_i - \hat{h}_i)^2$$

We want to take advantage of the direction of road cracks, however this kind of information is not contained in the bounding box labels. Fortunately, the class label contains an approximated direction, for example class D00 and D01 are longitudinal road damages, D10 and D11 are lateral road damages as shown in Figure 2. We add a factor $\alpha$ to width and height differences:

$$D_{wh} = \alpha(w_i - \hat{w}_i)^2 + (1 - \alpha)(h_i - \hat{h}_i)^2$$

For class D00 and D01, we set $\alpha$ to a value smaller than 0.5, which means difference in y direction is playing a more important role in loss than that in x direction. And for class D10 and D11, we set $\alpha$ to a value greater than 0.5, which means difference in x direction is paid more attention to. And for other classes, it is set to 0.5. In this way, we force the network to pay attention to the x-boundary for lateral cracks and to the y-boundary for longitudinal road cracks.

3. Experiments

We evaluate our method on the Road Damage Dataset[9], which is collected with a smartphone in car in seven cities in Japan. The dataset contains 9053 images and 15435 annotated bounding boxes. Among
them, 7240 images are chosen for train and 1813 images for testing. Following the evaluation protocol on the COCO [8] dataset, we use mAP50 as the metrics to measure the precision. Specifically, mAP50 is computed as average precision of all categories at the IoU threshold 0.5.

We compare our method with original YoloV3, the result in Table 1 have shown that our method has made a considerable improvement in terms of mAP50 compared with YoloV3. In order to evaluate the role of the two improvements, we also evaluate the method with single improvement: YoloV3 with low-level feature fusion (LF) and YoloV3 with direction-aware loss (DAL), the results are listed in Table 1, which demonstrates the effectiveness of each improvement. Figure 3 gives some examples from the detection results of proposed method.

### Table 1. Evaluation results of proposed method compared with YoloV3.

| Method              | mAP50  |
|---------------------|--------|
| YoloV3              | 0.496  |
| Proposed with LF    | 0.502  |
| Proposed with DAL   | 0.509  |
| Proposed with LF+DAL| 0.523  |

![Detection results of proposed method.](image)

4. **Conclusion**

We have proposed in this paper a method which make two improvements to YoloV3 for better detection of road damages. These two improvements take advantage of the visual characteristics of the road damage, and experiments have demonstrated the effectiveness of the proposed method.

### References

[1] Girshick, R., 2015. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[2] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[3] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[4] 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 computer vision and pattern recognition (pp. 779-788).
[5] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y. and Berg, A.C., 2016, October. Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

[6] Redmon, J. and Farhadi, A., 2017. YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).

[7] Redmon, J. and Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[8] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[9] Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T. and Omata, H., 2018. Road damage detection and classification using deep neural networks with smartphone images. Computer-Aided Civil and Infrastructure Engineering, 33(12), pp.1127-1141.

[10] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).