# Towards Autonomous Industrial Warehouse Inspection

Fahimeh Farahnakian, Lauri Koivunen, Tuomas Mäkilä, Jukka Heikkonen
*Department of Computing*
*University of Turku*
Turku, Finland
Email: fahfar, lamkoi, tusuma, jukhei@utu.fi

***Abstract*—In order to achieve autonomous warehouse inspection, a reliable rack monitoring and instant detection of rack is necessary. Damage detection is an essential task for pallet rack maintenance and it requires large amount of manual efforts. To address this problem, we employ deep learning to automatically detect damages with their per-pixel segmentation mask in this paper. We also investigated the impact of using related and unrelated data on the detection performance. For this purpose, we compared the performance of a detector when it is trained on: (1) the COCO dataset, (2) the ImageNet dataset and (3) a related task such as car damage dataset. Moreover, we evaluated the performance of the proposed detector based on different backbones. Experiments show that the detector with ResNet101 as feature extractor can achieve 93.45% accuracy in our real dataset. The code and dataset can be viewed at: https://gitlab.utu.fi/drone-warehouse/beam_defect_detection.git.**

***Index Terms*—Object detection, object segmentation, convolutional neural network, deep learning, warehouse damage detection**

## I. Introduction

The application of image capturing drones and computer vision techniques are overhauling how industries are conducting inspections.Warehouse pallet racks consist of hollow steel beams as the supporting structure for the inventory. Unfortunately, despite their surprising weight withstanding capacity they are easily damaged by daily operations from warehouse trucks or the like. A dent in the beam can quickly drop the load capacity of the shelves by a significant percentage. This can cause potentially catastrophic collapses of any loaded pallet racks and causing threat to humans operating the warehouse. Regular inspections should therefore be conducted, but this currently requires a lot of manpower. The inspectors have to walk through the warehouses and be able to inspect the correct parts from bottom shelf up to tens of meters high warehouse racks. Drones are already being used for these inspections, but only to aid the inspection as the drones are almost completely manually operated. Therefore, this calls for automatic and intelligent methods that can inspect pallet racks automatically. Moreover, these methods allow us to detect rack damages from the images or videos to speed up the inspection process and improve the accuracy simultaneously.

Inspired by the success of applying Convolutional Neural Networks (CNNs) in many number of computer vision problems including object classification [1], object detection [2], [3] and image segmentation [4], we use a CNN-based segmentation method which is called Mask-RCNN [4]. There are different version of CNN-based detectors such as Fast R-CNN [5], Faster RCNN [3], Single Shot multibox Detector (SSD) [6] and You Only Look Once (YOLO) [7]. Generally , the existing CNN-based detection methods can be categorized into two types: one-stage and two-stage object detection methods. One-stage methods are faster than two-stage methods since they do not use extra step for generating the candidates object locations. Although, two-stage object detection methods are more accurate. Mask R-CNN adopts the same two-stage methods, with an identical first stage.

In this paper, we employed Mask-RCNN to detect the damages and compute a pixel-wise mask for each detected damage in an image, taken from our dataset generated from pictures taken by Työtehoseura ry's SEMA approved racking inspectors from various Finnish warehouses. Damage detection task classifies damages and localizes each damage using a bounding box in the input image. As the rack damages are almost thin, using instance segmentation is more efficient than object detection (considering only bounding boxes). Therefore, we use object detection to classify each object and and localize each using bounding box, and instance segmentation to classify each pixel into a defined set of categories (e.g., damage and non-damage). To investigate how using related data can improve the detection performance, we collected the results when our proposed method use related and unrelated data. We also trained and compared our method with different backbones [8] (ResNet50 and ResNet101). We conduct experiments with real image data which was collected in a warehouse. The experiments show that we can get 9.78% more accuracy when the detector with backbone ResNet101 is trained on the related data (car damage [9]) comparing with unrelated data such as COCO dataset [10].

The paper is organized as follows. Section II describes the most significant related works. The proposed detection method is presented in Section III. Section IV shows the obtained results. Conclusions are described in Section V.

## II. Related Work

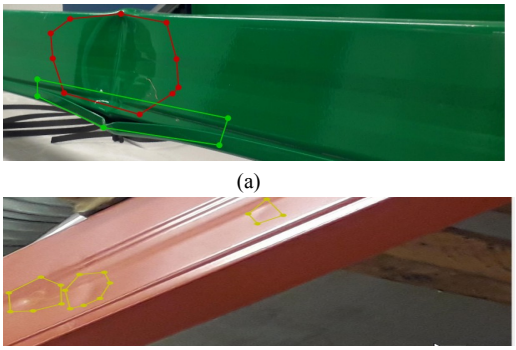
This section first provides a brief overview of some of the most well-known Convolutional Neural Networks (CNNs) used by the computer vision community for performing object detection and segmentation. In addition, we will review the literature of using deep learning for damage detection.

**Convolutional Neural Networks (CNNs):** in recent years CNNs [11] have achieved great improvements in various computer vision tasks as a widely used models of deep learning. In particular, the series of methods based on Region-based Convolutional Neural Network (R-CNN) [2] carry out object detection and instance segmentation significantly. R-CNN first identifies region proposals by Selective Search (SS) [12] and then classifies proposals into object categories via a CNN. A drawback of R-CNN is using the CNN for each region proposal, leading to time-consuming. To minimize the running time of R-CNN, Faster R-CNN [3] employs a CNN for region proposals instead of using SS. Faster R-CNN uses a region proposal network (RPN) to extract region proposals. Then, it employs a Region of Interest (RoI) pool layer to determine the bounding box coordinates and corresponding object's class. Generally, the region-based object detectors can be divided into two main groups: two-stage detectors (R-CNN [2], Faster R-CNN [3], R-FCN [13]) and one-stage detectors (SSD [6], YOLO [7]). Two-stage object detection methods have higher accuracy than one-stage detectors as they are considered hard examples which are poorly predicted by a model. However, they usually slower due to have an external module for generating the candidates target locations. One-stage object detection methods generate possible object proposals faster and simpler by eliminating classification operation and directly predicting bounding boxes [14].

An extension of Faster R-CNN for object instance segmentation is Mask R-CNN [4]. Compared to Faster R-CNN, Mask R-CNN can predicts a mask beside the object class and localization. Instance segmentation is more challenging than other computer vision tasks as it needs the correct detection of all objects in an image while also precisely segmenting each instance.

Mask R-CNN exceeds other instance segmentation approaches for instance Multi-task Network Cascades (MNC) [15] and Fully Convolutional Instance Segmentation (FCIS) [16] (MNC and FCIS won the COCO 2015 and 2016 segmentation challenges respectively). MNC employ a cascaded network for sharing convolutional features and also utilized RPN for performing instance segmentation [15]. FCIS is the first fully convolutiona solution for segmentation task which creates spurious edges [4]. Another interesting instance segmentation model is MaskLab [17], which is based on Faster R-CNN. MaskLab generates three outputs: bounding box detection, semantic segmentation and direction prediction.

Mask R-CNN performs instance segmentation into two steps: (1) gets the input image and generates the RoI proposals, and (2) classifies the proposals and generates the bounding boxes and corresponding a binary mask for each RoI. Mask R-CNN not only has a detection accuracy improvement, it also has a great advantages in detection of small objects [9]. It widely used in many real applications such as agriculture [18], medical image segmentation [19], construction [20], damage detection [9], [21] and etc. For this reason, we also motivated to use Mask R-CNN in this paper for damage detection in warehouses.



(a)

(b)

Fig. 1. Two examples of our dataset: (a) green and red polygons show tooth and dent damages and (b) yellow polygon shows mini-dent damage.

**Deep learning-based damage detection:** Although Deep Learning (DL) has been used in many real applications, there are only few studies on DL-based damage detection. Fully Convolutional Networks (FCN) has been proposed in [21] to detect two major types of damages, namely cracks and burn, in aircraft engine. Their results on the real images show that the proposed FCN can successfully extracts the damage regions with high accuracy. In [9], a vehicle-damage-detection segmentation algorithm based on Mask R-CNN has presented. To improvement the performance, they used a pre-trained model based on COCO dataset and fine-tuned the network parameters based on the dataset which are collected in this article. A road damage detection approach is proposed in [22] for road maintenance and decreasing the manual human efforts based on their own dataset collected. Their approach utilized RetinaNet [23] as a two-stage object detection method and trained RetinaNet with different backbones. Their results show that their model with VGG19 [24] as backbone can achieve the highest detection accuracy.

## III. Methodology

In this section, we first describe the collected real image dataset. Then, we present the proposed workflow for performing damage detection on our image dataset. Finally, we briefly discuss the proposed parameters for training the proposed model and the methodology was tested on our image dataset.

### *A. Data collection and pre-processing*

First of all, we require to collect the data and reprocess it to train our model. Since there is no publicly available warehouse pallet damages dataset, we utilize drone operating on a warehouse to collect images on damages. We obtain a set

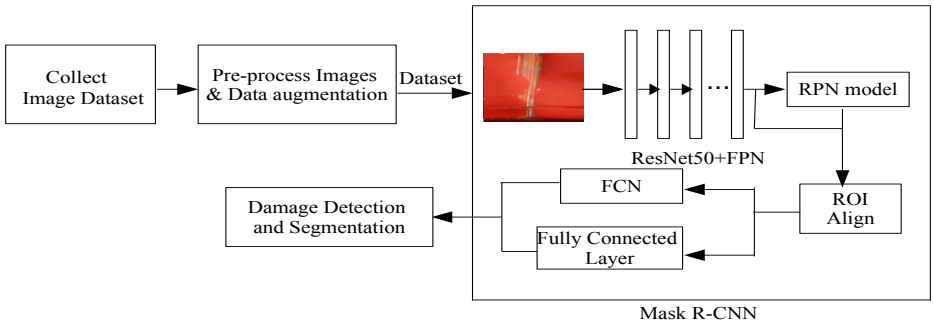



Fig. 2. The overall workflow of the proposed rack damage-detection framework.

of 75 images consists of three main damage types in racks such as dent, mini-dent and tooth. The dataset contains different shapes and colors of racks and various light conditions for model generalization. Figure 1 shows two example damages of the dataset.

Data augmentation is an efficient way to avoid overfitting and has has been widely used in [25], [26]. It generates more images from the original 75 images by a number of random transformations [27] by applying rotation, swirl, cropping, vertical flip and horizontal flip.

We randomly divided our dataset into two datasets after augmentation: training (629 images) and validation (85 images). In order to evaluate the performance of the proposed method, we collected an additional 85 images as a test dataset. All images have a size of $1024 \times 1024$ pixels. Images were annotated by VGG Image Annotator (VIA) tool[1]. We saved the annotations in a JSON file. Each mask is a set of polygon points. Figure 1 shows two example images with polygon region annotation that is manually generated with VIA tool. The value of pixels corresponding to damage inside the bounding polygons are 1 and the rest of the pixels value is 0 regarded as background.

### B. General Workflow

We employed the state-of-the-art Mask R-CNN algorithm to develop racking damage detection and segmenting framework for enhancing performance of autonomous warehouse inspection operation. Mask R-CNN can detect objects in an image and also generate mask for each object. In fact, it combines object detection and semantic segmentation. Figure 2 illustrates the general workflow for defect detection of racks based on Mask R-CNN. The workflow consists of four main steps as follows:

1) The input image is processed by the deep Residual Network (ResNet) [8] for extracting and obtaining corresponding feature maps. Moreover, the Feature Pyramid Network (FPN) [28] is used to extract multiple-scale information in order to improve representation of objects at multiple scales.
2) Region Proposal Network (RPN) uses the features maps for obtaining a large number of the Region Of Interests (ROIs). The softmax classifier is applied for binary classification of foreground and background and achieved more accurate candidate regions, and removing duplicate the ROIs by Non-Maximum Suppression (NMS). NMS selects the ROI with the highest scores. After that, the final proposals (ROIs) pass to the next step.
3) RoIAlign [4] creates a feature map for the selected ROI from the previous step.
4) In the last step, there are two branches: one branch is a fully connected layer for object classification and the other is Full Convolution Network (FCN) for generating pixel segmentation (mask).

### C. Implementation Details

We implemented the Mask R-CNN method using an open-source package built on Keras and Tensorflow developed by team of Mask R-CNN on Github [29]. The backbone of Mask R-CNN is ResNet [8] with 50 and 101 layers based on FPN network. The backbone strides is [4, 8, 16, 32, 64]. We use a learning rate of 0.0001 with mini-batch size of 2 images. Momentum and weight decay values are 0.9 and 0.0001, respectively. For training the model, Adam optimization is utilized with 10 epochs and 100 training steps per epoch. The RPN anchors span five scales and three aspect ratios [0.5, 1, 2], following [28].

## IV. Experimental Results

We evaluated our method when it uses the pre-trained weights on (1) COCO dataset [10], (2) ImageNet dataset [30] and (3) related dataset (car damage [9]) instead of building the model from scratch. Using the weights of a trained model trained on another dataset and fine tuning the weights for the proposed task is called transfer learning. Transfer learning can decrease the number of images which are required to train a model.

Figure 3 represents the architecture of Mask R-CNN for damage detection and segmentation. It shows the process of segmentation on an input image according to the described workflow of Mask R-CNN in Section III (B). As shown in

[1]https://www.robots.ox.ac.uk/ vgg/software/via/via.html

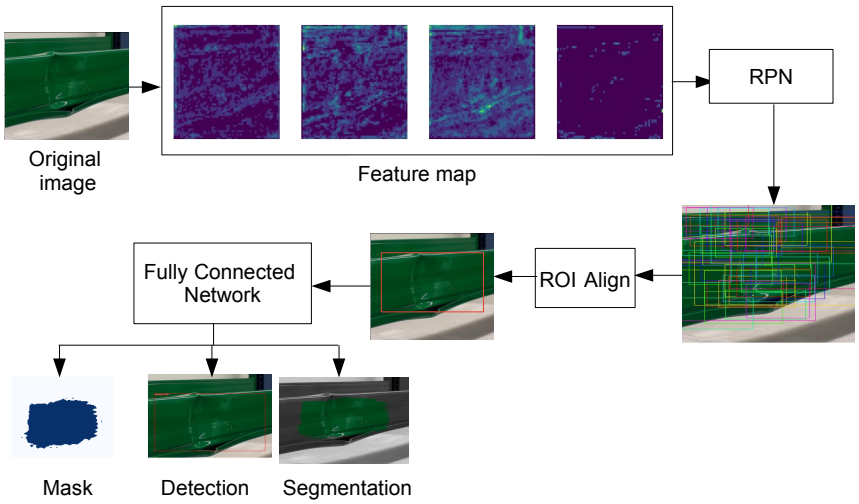



Fig. 3. Mask R-CNN architecture for damage detection and segmentation.

Figure 3 the input image passes through CNN backbone to create feature map. The feature map becomes the input of RPN. Then, a lot of boxes (anchors) and correspond scores are generated by RPN over the image. After refining boxes around detected objects, a binary mask is generated for the final ROI.

Table I presents the performance of the damage detection for the test dataset based on two proposed backbones: ResNet50 and ResNet101. When the Intersection-Over-Union (IoU) of predicted and real bounding boxes exceeds $x\%$ for an object, a detection is counted as a true positive. Based on two values of 50 and 70 for $x$, we calculate the Average precision (AP): $AP_{50}$ and $AP_{70}$. The best results are highlighted in bold. Based on the results three observations can be, made. **First**, our method can achieve the best performance (93.45%) when it uses ResNet101 as backbone. **Second**, the proposed method can learn more efficient features when it is pre-trained on a related task such as car damage dataset. The $AP_{70}$ of ResNet50 and ResNet101 is 7.07% and 9.78% increased when using the car damage dataset compared to the COCO dataset, respectively. **Third**, the AP value for ResNet101 on car damage dataset is 3.71% decreased when the IoU threshold set to 70% from 50%.

Qualitative prediction results of the proposed method on the test dataset are shown in Figure 4 and Figure 5 for ResNet50 and ResNet101, respectively. These figures show four examples of images from our test dataset. They illustrate the segmentation and detection results for our model. The first row is the original image and the subsequent rows show the results when the model used a pre-trained model based on COCO, ImageNet and car damage datasets, respectively. The model output consists of bounding boxes, the probability of detection and masks. The results show that the pixel-wise segmentation is improved when the model is pre-trained on car damage (last row in each image) compared to the COCO and ImageNet datasets.

TABLE I
OBJECT DETECTION RESULTS (%AP) BASED ON DIFFERENT PRE-TRAINED DATASET AND BACKBONES FOR THE TEST DATASET.

| Backbone | Pre-trained dataset | $AP_{50}$ | $AP_{70}$ |
|---|---|---|---|
| **ResNet50** | COCO | 81.03 | 82.98 |
| | ImageNet | 82.41 | 86.12 |
| | Car damage | 86.43 | 90.05 |
| **ResNet101** | COCO | 82.63 | 83.67 |
| | ImageNet | 85.63 | 89.54 |
| | Car damage | **89.74** | **93.45** |

## V. CONCLUSION

This paper presents a pallet rack damage detection method which uses Mask-RCNN as a famous convolutional neural network method for both detection and segmentation tasks. This method detects individual damages in warehouse racks and obtain pixel-wise mask for each detected damage in an image. To evaluate our method, we have collected a real image dataset of different type of pallet rack damages in warehouse. We have compared the performance of our method based on two backbone: ResNet50 and ResNet101. In addition, the results have collected when the method is pretrained on a unrelated dataset (COCO and ImageNet) and related dataset (car damage). The experimental results show that this method can achieve 93.45% detection accuracy when is pre-trained on a related task and used ResNet101 as a backbone structure.

## ACKNOWLEDGEMENTS

We acknowledge the help of Työtehoseura ry for their invaluable warehouse inspection expertise and for their fault picture collection allowing us to build the dataset used on this publication. We also gratefully acknowledge the Helsinki-Uusimaa Regional Council for funding.

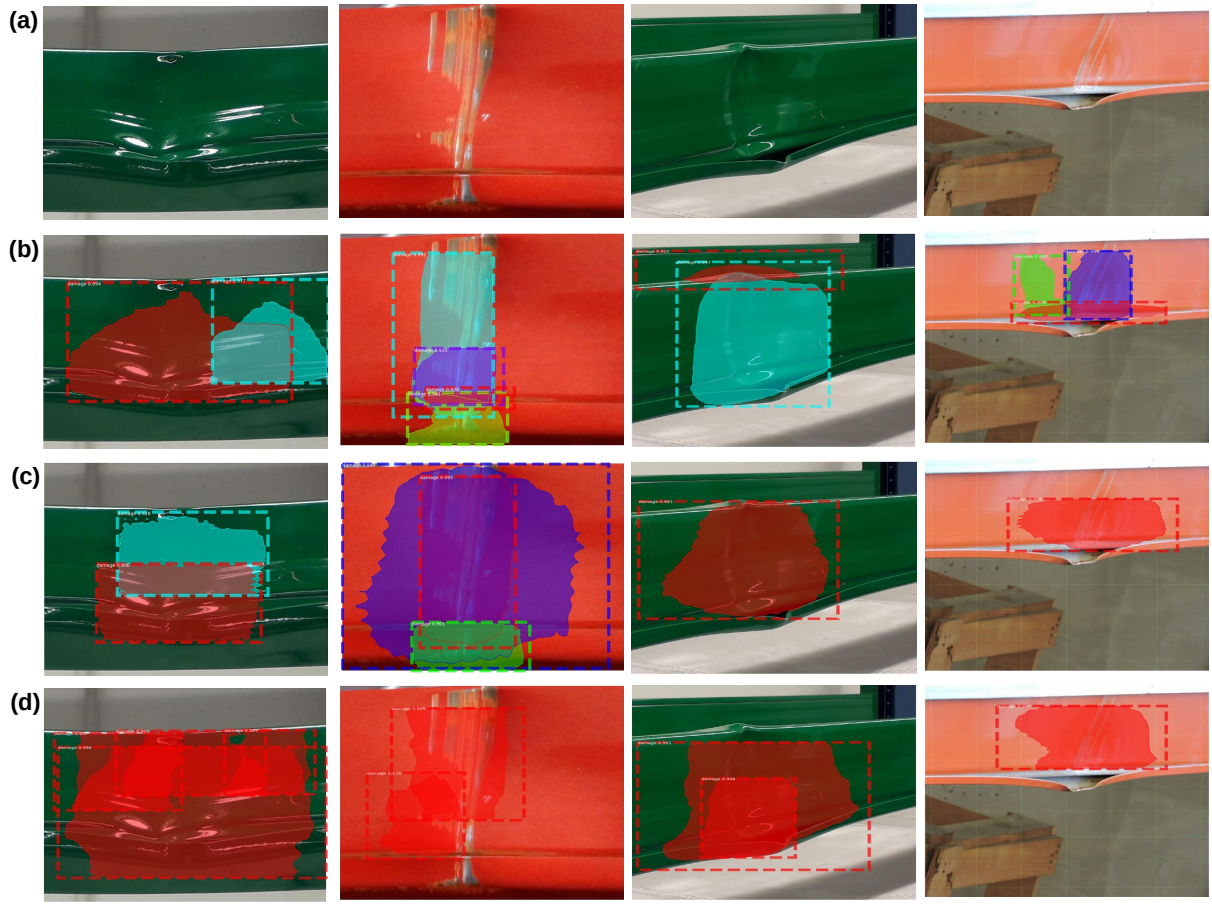



Fig. 4. The final prediction results and the color splash of Mask-RCNN based on **ResNet50** backbone. (a) Original input RGB images. (b) Detection results of the model when is pre-trained on COCO dataset. (c) Detection results of the model when is pre-trained on ImageNet dataset. (d) Detection results of the model when is pre-trained on car damage dataset

.

## REFERENCES


[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[2] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, June 2017.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[5] Ross B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015.

[6] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[7] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[9] Qinghui Zhang, Xianing Chang, and Shanfeng Bian. Vehicle-damage-detection segmentation algorithm based on improved mask rcnn. *IEEE Access*, PP:1–1, 01 2020.

[10] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[11] Yoshua Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009.

[12] J. R. Uijlings, K. E. Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *Int. J. Comput. Vision*, 104(2):154–171, September 2013.

[13] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks, 2016.

[14] Fahimeh Farahnakian and Jukka Heikkonen. Rgb-depth fusion framework for object detection in autonomous vehicles. In *2020 14th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 1–6, 2020.

[15] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. *CoRR*, abs/1512.04412, 2015.

[16] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. *CoRR*, abs/1611.07709, 2016.

[17] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. *CoRR*, abs/1712.04837, 2017.

[18] Cheng Shuhong, Zhang Shijun, and Zhang Dianfan. Water quality monitoring method based on feedback self correcting dense connected convolution network. *Neurocomputing*, 349:301 – 313, 2019.

[19] Ke Wang, Xun Zhang, Leyun Pan, Caixia Cheng, Antonia Dimitrakopoulou-Strauss, Yueping Li, and Nie Zhe. Multi-path dilated residual network for nuclei segmentation and detection. *Cells*, 8:499,

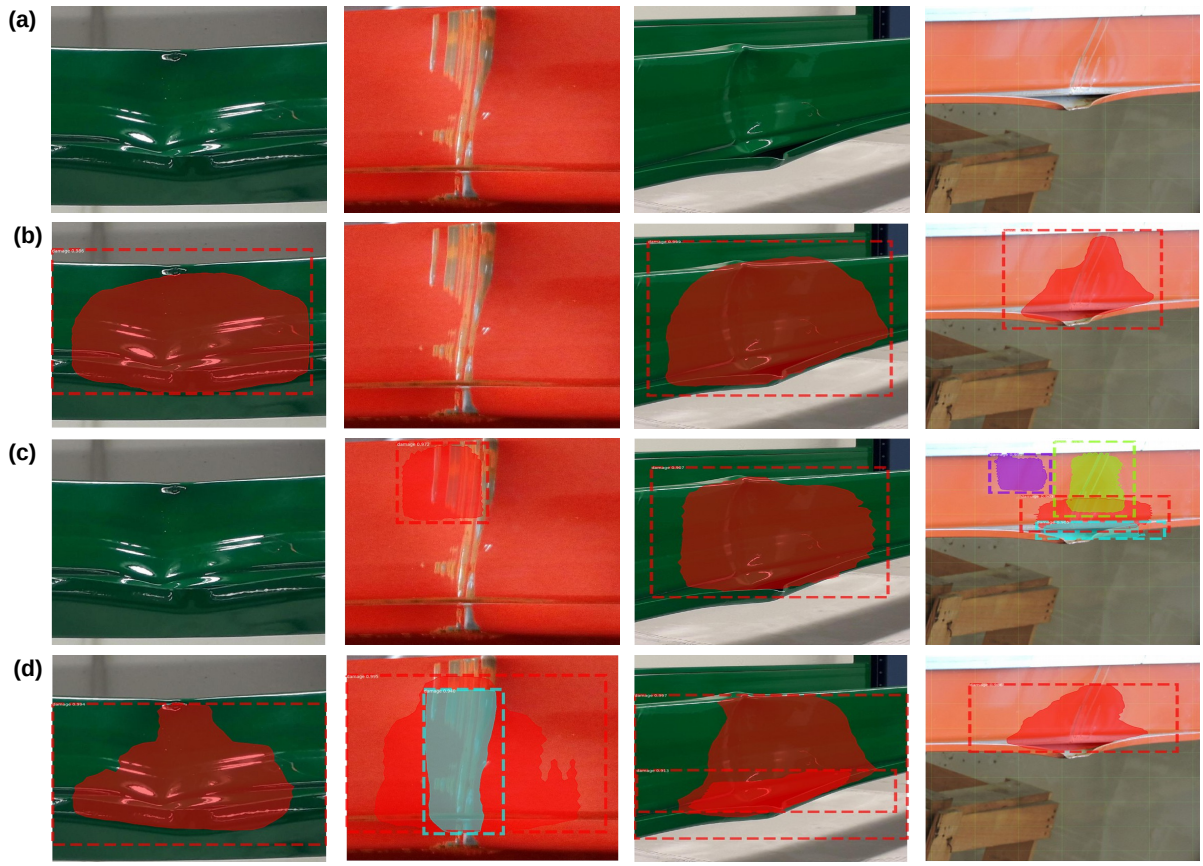



Fig. 5. The final prediction results and the color splash of Mask-RCNN based on **ResNet101** backbone. (a) Original input RGB images. (b) Detection results of the model when is pre-trained on COCO dataset. (c) Detection results of the model when is pre-trained on ImageNet dataset. (d) Detection results of the model when is pre-trained on car damage dataset

.

05 2019.

[20] Jirui Yang, Luyan Ji, Xiurui Geng, Xue Yang, and Yongchao Zhao. Building detection in high spatial resolution remote sensing imagery with the u-rotation detection network. *International Journal of Remote Sensing*, 40:1–23, 03 2019.

[21] Zejiang Shen, Xili Wan, Feng Ye, Xinjie Guan, and Shuwen Liu. Deep learning based framework for automatic damage detection in aircraft engine borescope inspection. pages 1005–1010, 02 2019.

[22] L. Ale, N. Zhang, and L. Li. Road damage detection using retinanet. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5197–5200, 2018.

[23] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[25] F. Farahnakian, M. Haghbayan, J. Poikonen, M. Laurinen, P. Nevalainen, and J. Heikkonen. Object detection based on multi-sensor proposal fusion in maritime environment. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 971–976, 2018.

[26] M. Haghbayan, F. Farahnakian, J. Poikonen, M. Laurinen, P. Nevalainen, J. Plosila, and J. Heikkonen. An efficient multi-sensor fusion approach for object detection in maritime environments. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2163–2170, 2018.

[27] Sebastien C. Wong, Adam Gatt, Victor Stamatescu, and Mark D. McDonnell. Understanding data augmentation for classification: when to warp? *CoRR*, abs/1609.08764, 2016.

[28] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016.

[29] Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.