electronics

MDPI

Article

Pothole Detection Using Image Enhancement GAN and Object Detection Network

Habeeb Salaudeen [1,2,*] and Erbuğ Çelebi [1,2]

[1] Department of Computer Engineering, Faculty of Engineering, Cyprus International University, Nicosia 99010, Cyprus; ecelebi@ciu.edu.tr
[2] Artificial Intelligence Application and Research Center, Cyprus International University, Nicosia 99010, Cyprus
* Correspondence: hksalaudeen@gmail.com

Abstract: Many datasets used to train artificial intelligence systems to recognize potholes, such as the challenging sequences for autonomous driving (CCSAD) and the Pacific Northwest road (PNW) datasets, do not produce satisfactory results. This is due to the fact that these datasets present complex but realistic scenarios of pothole detection tasks than popularly used datasets that achieve better results but do not effectively represents realistic pothole detection task. In remote sensing, super-resolution generative adversarial networks (GAN), such as enhanced super-resolution generative adversarial networks (ESRGAN), have been employed to mitigate the issues of small-object detection, which has shown remarkable performance in detecting small objects from low-quality images. Inspired by this success in remote sensing, we apply similar techniques with an ESRGAN super-resolution network to improve the image quality of road surfaces, and we use different object detection networks in the same pipeline to detect instances of potholes in the images. The architecture we propose consists of two main components: ESRGAN and a detection network. For the detection network, we employ both you only look once (YOLOv5) and EfficientDet networks. Comprehensive experiments on different pothole detection datasets show better performance for our method compared to similar state-of-the-art methods for pothole detection.

Keywords: pothole detection; small object detection; super-resolution; object detection; GAN; deep learning

Citation: Salaudeen, H.; Çelebi, E. Pothole Detection Using Image Enhancement GAN and Object Detection Network. *Electronics* **2022**, *11*, 1882. https://doi.org/10.3390/electronics11121882

Academic Editors: Ahmad Taher Azar, Anis Koubaa, Alaa Khamis, Ibrahim A. Hameed and Gabriella Casalino

Received: 28 April 2022
Accepted: 8 June 2022
Published: 15 June 2022

1. Introduction

1.1. Problem Description and Motivation

There are many applications for the detection of objects on the road, with some of the most promising occurring in autonomous driving [1,2], and surface defects that need to be reported to road repair ministries [3,4]. These applications are made possible through cameras that are mounted on moving vehicles. In order to address the challenges of detecting potholes from images and videos, there have been many methods proposed. These methods include processing images or videos captured with cameras from mobile phones [3], unmanned aerial vehicles (UAVs), and drones [5]. However, the methods do not reflect how pothole detection as an object detection problem can be perceived. Figure 1 shows on the left an image that was taken close up while standing over a pothole. This typically represents how most pothole datasets acquire data and what the state-of-the-art methods have used to train models to detect potholes. The image on the right shows a more realistic scenario of pothole instances captured from a moving vehicle, representing how pothole detection tasks should be perceived. When the methods are presented in a manner that reflects the problem well, the detection performance is not so good because of cases where the amount of noise present in images

or videos, most often at low resolution, causes small potholes to appear as insignificant objects that blend into the background. The datasets that present realistic representations of pothole detection problems include PNW [6] and CCSAD [7].

Figure 1. A more realistic task of pothole detection presented in the right column compared to an unrealistic instance.

When evaluating object detection methods' performance, researchers use datasets, such as ImageNet [8] and Microsoft Common Objects in Context (COCO) [9], containing objects that are relatively easy to detect. In addition, the objects often appear large in the images. However, some other objects captured from a distance often appear small, sometimes blending in with the background, and can be challenging to detect using popular object detectors [10]. For images containing these types of objects to be detected, researchers have found that high-resolution (HR) images offer more input features than low-resolution (LR) images as a result of the lack of input features for small objects [11–13].

In an attempt to improve the detection accuracy of the pothole object detection problems, researchers have proposed varieties of object detection methods [14–19] enhanced with super-resolution (SR) techniques that are employed to generate an enhanced image from a low-resolution image before performing object detection. In the field of remote sensing, where images are captured from a satellite and most often present the small object detection problem, several methods have been proposed based on super-resolution as well. SR techniques based on convolutional neural networks (CNN), such as single-image super-resolution convolution networks (SRCNN) [14] and accurate image super-resolution using very deep convolutional networks (VDSR) [15], have been proposed and show remarkable results in generating HR images and performing object detection. In addition to CNN-based methods, methods based on generative adversarial network (GAN) [16] have also been proposed. Super-resolution generative adversarial networks (SRGAN) [17], enhanced super-resolution generative adversarial networks (ESRGAN) [18], and end-to-end enhanced super-resolution generative adversarial networks (EESRGAN) [19] have demonstrated better performance in producing both realistic HR images and performing small object detection. These GAN-based models typically consist of generator and discriminator networks that are trained on a pair of LR and HR images, with the generator network generating HR images from the inputted LR images while the discriminator network tries to distinguish the real HR image from the generated HR image. The generator network eventually learns to produce HR images that are indistinguishable from the ground truth HR images, and the discriminator will not be able to distinguish between the images.

Another major challenge in detecting potholes on roads is the cost of the sensor devices used in such a process. Majorly, lidar sensors are exploited for 3D modeling of

the surrounding environment to detect obstacles and objects around the vehicle. A single lidar sensor can easily cost thousands of dollars. Cameras have been exploited as cheaper alternatives, but acquiring HR cameras that can capture high-quality images from a moving vehicle can also be expensive.

We have thus identified two main problems with detecting potholes from 2D images. First, the accuracy of object detection models can decline considerably when potholes appear as small objects at a distance compared to large objects. Second, LR cameras cannot provide good detection accuracy at a reasonable cost, while it is too costly to acquire HR cameras. Therefore, there is a need for a novel solution to improve the detection of potholes when they appear at a distance from LR images. To the best of our knowledge, no study has employed GAN-based super-resolution image enhancement and object detection algorithms to detect potholes when they appear in a captured image.

In this paper, we present a pipeline that combines an object detection network and a super-resolution network in order to detect potholes accurately from images. Recent research has shown that some state-of-the-art detectors can misclassify or totally miss objects that appear at a distance when trained on low-resolution images. Our proposed method can detect such instances of potholes that are at a distance. When we used super-resolution images to train our object detector to detect instances of potholes from images from different pothole detection datasets, namely the Sunny dataset, CCSAD, PNW, and Japan, the detector exhibited a more reliable detection performance than when using low-resolution images. We have combined several datasets to achieve this aim because there is no single benchmark dataset for pothole detection and most of the known datasets do not accurately represent the pothole detection scenario. This is a known challenge in pothole detection tasks. A comparison of our methods with similar studies using similar datasets indicates that our detector records a higher precision rate than comparative studies and has a reasonable recall value on the test set. The overall performance on the test datasets is also satisfactory. Section 5 provides detailed information about our results.

1.2. Contributions

The proposed methodology uses two components: the ESRGAN network [18] and an object detection network (YOLOv5 and EfficientDet Networks). This approach is inspired from work completed in the remote sensing field that employs super-resolution GANs to detect objects at different resolutions. We used ESRGAN to generate super-resolution images and trained an object detection network on these images.

The proposed strategy facilitates the detection of potholes from a distance accurately. Figure 2 shows the cases where some instances of potholes could not be detected from an LR image. The experimental results show that, using the state-of-the-art object detectors, the SR images used for training can significantly outperform the detectors trained on LR images both in accuracy and in detecting small potholes that appear in the frame. It is our hope that this study will expose the industry to the field of using HR images for the task of pothole detection. In this study, we also provide an overview of the state-of-the-art techniques used for the problem of pothole detection.

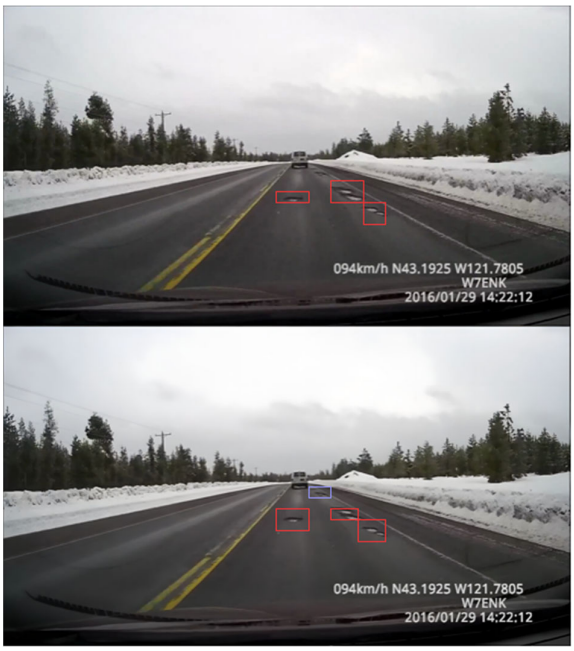

Figure 2. The first image shows detected instances on an LR image, and the second image shows detection on an SR image, which detected more instances of potholes, particularly an instance at a distance.

The paper is structured as follows: in Section 2, we review the existing literature on pothole detection techniques, image super-resolution techniques, and studies that have employed super-resolution along with object detection for small object detection tasks. In Section 3, we present our proposed methodology by showcasing the ESRGAN network and the object detection methods we have employed, such as YOLOv5 and EfficientNet. In Section 4, we provide the details of our experiments by providing information about the datasets we have used in training and testing our methods. In Section 5, we present comparative performance evaluation results derived from the experiments on different datasets and also a comparison with other studies that used similar datasets but on LR images. Finally, in Section 6, we conclude.

2. Related Works

In this section, we provide in-depth knowledge regarding the developing field of road surface anomalies detection. We discuss the topic within the context of other related areas and review the current research on the subject. This section presents the recent advancements in the fields related to this paper.

2.1. Pothole Object Detection

A variety of devices have been employed to collect data used in road surface anomalies detection. These devices include image acquisition devices, vibration-based sensors, and 3-D depth cameras. Object detection techniques often rely on image data captured by digital cameras [20,21] and depth cameras, thermal imaging technology, and lasers.

To extract the features of a pothole from images, convolutional-neural-network (CNN)-based techniques are more prevalent in this application. These models can accurately model non-linearity in patterns and perform automatic feature extraction on given images. In addition, they are desirable because of their robustness to filtering background noise and low contrast in road images [22]. CNNs have been successfully employed in many applications [1,3,5], but they are not effective in all scenarios. For

example, when the object to be detected is small relative to the image, or when high-resolution images are used to mitigate this problem, the computation required to process the data can be prohibitive. This is because CNNs consume a large amount of memory and computation time [23]. To address this, Chen et al. [23] suggest two workarounds to resize input images to the network or using image patches from HR images to train the network. The former workaround is a two-stage system in which a localization network (LCNN) is first employed to locate the pothole instance in the image and then a classification network based on part (PCNN) is utilized to determine the classes.

Salcedo et al. [4] recently proposed a series of deep learning models to develop a road maintenance prioritization system for India. The proposed models include UNet, which employs ResNet34 as the encoder (a neural network subcomponent), EfficientDet, and YOLOv5 on the Indian driving dataset (IDD). Another variation of the you only look once (YOLO) model has also been employed for the task of pothole detection. In a study by Silva et al. [24], the YOLOv4 algorithm was used to detect road damage on a custom dataset that provides an aerial view of roads from a flying drone. The accuracy of the YOLOv4 algorithm and its applicability in the context of identifying damages on highway roads was experimentally evaluated, with an accuracy of 95%.

Asphalt roads can be evaluated by creating 3D crack segmentation models. Guan et al. [25] employed a modified U-net architecture featuring a depth-wise separable convolution in an attempt to reduce the computational workload when working on a multi-view stereo imaging system that contains color images, depth images, and color-depth overlapped images of asphalt roads. The architecture produces a 3D crack segmentation model that considerably outperforms the benchmark models regarding both inference speed and accuracy.

Fan et al. [26] argued that approaches that have employed CNNs for road potholes are faced with challenges of annotating data to be used for training since deep learning models require a large amount of data. The authors thereby proposed a stereo vision-based road pothole detection dataset and an algorithm that is used to distinguish between damaged road and undamaged roads. The algorithm proposed derived inspiration from graph neural network, where the authors employed an additional CNN layer called the graph attention layer (GAL) to provide optimization for image feature representations for semantic segmentation.

Other methods besides deep learning—such as support vector machines (SVM) and nonlinear SVM—have been explored for extracting potholes from images. Gao et al. [27] employed texture features from grayscale images to train an SVM classifier to distinguish road potholes from cracks in the pavement.

In addition to the aforementioned machine-learning-based techniques, other approaches have been developed. Penghui et al. [28] used morphological processing in conjunction with geometric features from pavement images to detect pothole edges. Koch et al. [29] used histogram shape-based thresholding to detect defective regions in road surface images and subsequently applied morphological thinning and elliptic regression to deduce pothole shapes; texture features within these shapes were compared with those from surrounding non-pothole areas to determine if an actual pothole was present.

As previously mentioned, these proposed techniques produce good results on the test set, but they have not been trained and tested on realistic datasets of high complexity, such as those encountered in autonomous vehicles and unmanned aerial vehicles. Such models will likely underperform when applied to real-world scenarios.

2.2. Super-Resolution Techniques

Small object detection is commonly exploited in the remote sensing field, where researchers are often faced with small objects in the object categories, making the detection of these objects by state-of-the-art detectors challenging. As images are scaled

down by the generic detectors, such as SSD, Faster R-CNN, etc., the performance is reduced. Therefore, most of the proposed methods that use super-resolution images for small object detection are enormous in this field.

Enhanced deep SR network (EDSR) [30] introduces the idea of performing object detection on SR images in the remote sensing field for some of the popularly used architectures [19,31,32]. The ESRGAN [18] architecture improved on the existing super-resolution GAN networks to provide more realistic SR images. The authors employed residual-in-residual dense block (RRDB) with adversarial and perceptual loss to achieve this. The authors achieved a considerable improvement in a subsequent study regarding real-ESRGAN [33] with the use of only synthetic data with high-order degradation modeling, which were close to the real-world degradations.

SwinIR [34] addressed the issue of small object detection with SR data by proposing a transformer that had three parts: a shallow feature extraction step, a deep feature extraction step, and a high-quality image reconstruction step using the residual Swin transformer blocks (RSTB). This transformer produced good results on the DIV2K dataset and the Flickr2K dataset.

Zhang et al. [35] proposed a model called BSRGAN to address degradation issues of SR models that often affect the performance of such models. They proposed that BSRGAN uses random blue shuffle, down sampling, and noise degradation techniques to produce a more realistic degradation of LR images.

The dual regression network (DRN) [36] mapped LR images to HR ones and provided a corresponding degradation mapping function. The authors also found that their method achieved better performance in terms of PSNR (peak signal-to-noise ratio) and the number of parameters.

NLSN for non-local sparse network [37] uses a non-local sparse attention (NLSA) to address the problem of image SR. The method divides the input into hash buckets that contain relevant features, which prevents the network from providing noise or attention to areas of the image with less information during training.

2.3. Super-Resolution Based Object Detectors

For object detection tasks, both training and inference are affected by the size of the objects. The existing detectors work well with medium-to-large-sized objects but struggle when detecting small-sized objects (objects occupying less than 5% of the overall image size or objects with dimensions in a few pixels). This is because small objects are often indistinguishable from the features of other classes or the background, thereby leading to lower accuracy for the detector.

One technique for improving detector accuracy has been to use data augmentation to oversample small objects of interest, thus increasing the possibility that the small objects will overlap with the prediction [38]. However, this technique has proven to decrease accuracy on other objects in the dataset by reducing the overall amount of training data available for those objects. Another technique proposed for improving detector accuracy is training on both small and large objects of multiple resolutions [39].

YOLOv3 [40] is an object detection system that uses the feature pyramid network (FPN) to quickly provide users with the location of objects in a specific field of view. The system has had great success at detecting small objects due to its ability to detect and locate them without having to perform multiple scans of the same area. One of the significant improvements this network provides is the addition of a new classifier that enables the system to track objects during different stages of their movements, which allows YOLOv3 to locate smaller objects more effectively. However, the network lacks significantly when it comes to processing time. To further improve the performance of small object detection, different modifications have been made to the architecture.

To further improve the performance of YOLOv3 on small object detection and processing speed, Chang et al. [41] proposed amendments to the structure of the network. First, the authors proposed using the K-means algorithm using the width and

height of the object's bounding box to obtain appropriate anchor boxes for the objects of interest in a dataset to mitigate the challenge of the objects having different sizes. This modification provides faster network training since the generated anchor boxes are now much closer to the dataset objects.

Lv et al. [42] proposed optimizing the loss function of the YOLOv3 by changing the default loss function L2 and classification loss cross-entropy to GIoU (generalized intersection-over-union) loss function and focal loss, respectively, due to the lack of robustness of the L2 loss function and vulnerabilities, such as the model being sensitive to examples with significant errors and, while trying to adjust, sacrificing example values, with small mistakes. To this end, the GIoU loss function, a variation of the IoU loss function, is proposed to provide a general improvement for the YOLOv3 network

In studies by Bashir and Wang [11] and Courtrai et al. [43], SR networks were used to increase the spatial resolution of LR datasets before feeding the SR images to detector networks for actual detection tasks. Such SR networks have been exploited in recent studies to scale LR images for 2× and 4× scale factors, resulting in remarkable results. In recent years, image generation models that produce single or a pair of images have been widely used for visual representation. Examples include single-image super-resolution (SISR) [44] using a single input; Ferdous et al. [45] used a generative adversarial network (GAN) to produce SR images and SSD to perform object detection on the images; Rabbi et al. [19] combined ESRGAN [18] and EEGAN [46] to develop their own integrated end-to-end small object detection network; Wang et al. [47] proposed a multi-class cyclic GAN with residual feature aggregation (RFA), which is based on both image SR and object detection. The proposed method replaced conventional residual blocks with RFA-based blocks and concatenated the features of the images to improve the performance of the network.

3. Materials and Methods

In this paper, we present an efficient architecture for the detection of small objects in the context of pothole detection on road images. Toward this goal, we propose a network consisting of two separately trained deep neural network modules: ERSGAN (enhanced super-resolution generative adversarial networks), which is used to upscale the images by 4× while producing HR images from LR images, and an object detection network for detecting instances of potholes. Figure 2 shows the proposed architecture. The ERSGAN is first used to upscale the images in our dataset 4 times their original scale, while the object detection network is trained on the HR images.

3.1. Super-Resolution with ESRGAN

The SR network based on ESRGAN architecture was designed to generally improve the perceptual quality of super-resolution images. ESRGAN typically employs the basic architecture of SRResNet [17] with few adjustments made to the discriminator network and the perceptual loss for produce a better performance.

To improve the quality of the generated images, primarily texture, two main changes were introduced to the architecture of the SRGAN. First, all BN layers were removed and replaced by a new residual-in-residual dense block (RRDB), which combines both dense connections and multi-level residual networks within a single module.

From the architecture of the SRGAN, the batch normalization (BN) layers were removed from the residual block as depicted in Figure 3. This was conducted because the removal of BN layers from architectures have produced better results in PSNR tasks, such as deblurring [48] and super-resolution [30], since BN layers perform normalization of features with mean and variance in a given batch during the training process and even use an estimate of mean and variance of the entire training dataset during testing; therefore, the statistics for training and testing datasets significantly vary, thereby introducing undesirable artifacts and also limiting the generalization ability. Therefore,

to achieve more stable training and performance of the network, the BN layers were removed. It was also observed that the procedure improved the generalization ability of the network and also caused a reduction in the computational workload.

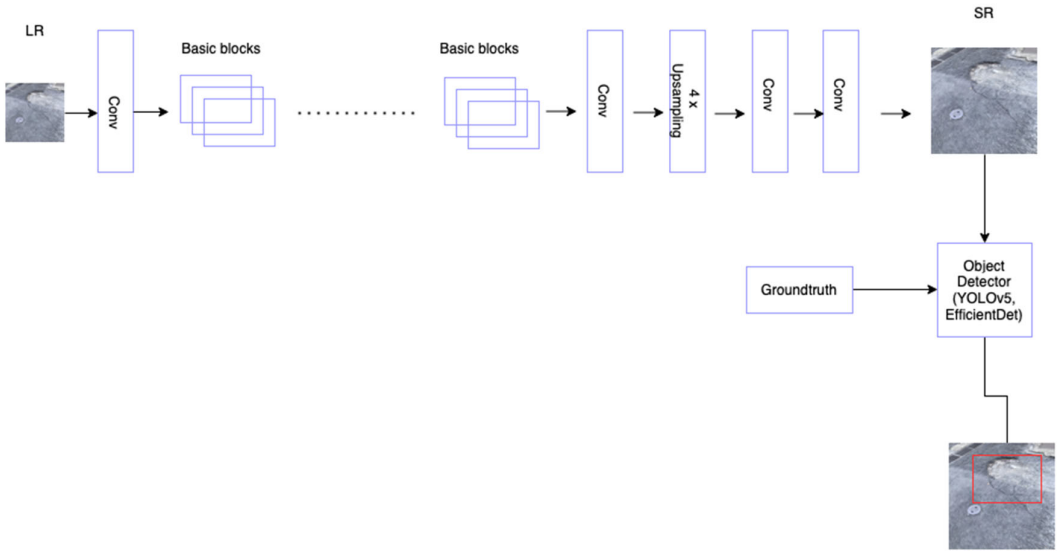

Figure 3. The proposed end-to-end architecture consisting of the ESRGAN super-resolution architecture and object detector.

While retaining the architectural design of the SRGAN, RRDB is proposed as the basic block, with the intention to use more layers and connections to improve the performance of the network. It features a residual-in-residual structure and uses dense blocks in the main connections.

In addition to these changes made to the SRGAN architecture, the authors also exploited a different training technique to achieve a better performance: residual scaling, which scales down the residuals through the multiplication of a given constant value between 0 and 1 before adding them to the main path as an attempt to establish stability and the usage of smaller initialization, which makes training easier since the parameter variance will become smaller.

3.1.1. Relativistic Discriminator

In addition to the improvements that have been covered so far, which are mainly completed in the generator network of the architecture, the authors also proposed the enhancement of the discriminator network based on the relativistic GAN [49], which differs from the standard discriminator D used in SRGAN. The relativistic discriminator D_{Ra} attempts to deduce the probability of a real image is relatively more realistic than a generated image by the generator. This differs from the SRGAN discriminator, which only estimates the probability of an input image x is real. Thus, the authors deduced the discriminator loss and the adversarial loss for the generator network as:

$$L_D^{Ra} = -E_{x_r}\left[log\left(D_{Ra}(x_r, x_f)\right)\right] - E_{x_f}\left[log\left(1 - D_{Ra}(x_f, x_r)\right)\right]. \quad (1)$$

$$L_G^{Ra} = -E_{x_r}\left[log\left(1 - D_{Ra}(x_r, x_f)\right)\right] - E_{x_f}\left[log\left(D_{Ra}(x_f, x_r)\right)\right]. \quad (2)$$

E_{x_f} represents for the operation of taking average for all the generated data in a given mini-batch. The generated image derived from the input LR image is represented as x_f and x_r for the real SR image. Both real and generated data gradients provide benefit to the generator during adversarial training.

3.1.2. Perceptual Loss

SRGAN proposed a loss, which constrains the features after the activation. As for perceptual loss L_{percep}, the constraints on features are completed before activation. The standard perceptual loss proposed by Johnson et al. [50] can be enabled on the activation layers of a pre-trained deep neural network with the minimal distance between the two activated features. As for ESRGAN, the authors proposed to use the features before the activation layers to overcome the drawbacks of the original perceptual loss. The identified drawbacks include: sparse activated features which provides weak supervision and lead to lower performance, and inconsistent reconstructed brightness of SR images when compared to the ground truth image. The loss for the generator network is thus given as:

$$L_G = L_{percep} + \lambda L_G^{R_a} + \eta L_1 \quad (3)$$

The content loss L_1 is used to evaluate the 1-norm distance between the ground truth image and the recovered image, while λ and η are coefficients for balancing the different loss terms.

Finally, to prevent unwanted noise in the result, the authors proposed a network interpolation strategy that trains a PSNR-oriented network and a fine-tuned GAN-base network, with both network parameters interpolated to derive an interpolated model G with the following parameters: θ_G^{INTERP}, θ_G^{PSNR}, and θ_G^{GAN}:

$$\theta_G^{INTERP} = (1 - \alpha)\theta_G^{PSNR} + \alpha\theta_G^{GAN} \quad (4)$$

This allows the network to produce good results without the introduction of artefacts and a continuously balanced perceptual quality through training. In the case of potholes, the network showcased its ability to remove significant noise from the images. In addition to upscaling the images to 4× the original size, deblurring was adequate and the edge information of the pothole shapes is enhanced as well; see Figure 4.

Figure 4. On the left is the SR generated by ESRGAN (4× scale) compared with the LR image on the right. It can be observed that the SR images provide more feature representation of the objects on scene than the LR images.

3.2. Object Detection

We proposed both YOLOv5 and EfficientDet-D1 architectures for the object detection tasks. We will briefly discuss the architectures of the networks in the following sections.

3.2.1. You Only Look Once (YOLOv5)

The YOLOv5 [51] architecture is a little bit different from the previous YOLO version. While other previous versions use Darknet, the new version uses PyTorch and CSPDarknet53 [52] as the backbone network.

YOLOv5 is an effective, fast, and efficient object detection and classification model. It is a single-stage object detection network. It uses the focal loss function, which is used in classification tasks, and runs a convolutional feature extractor on top of a backbone architecture that is pretrained on ImageNet. It then passes the extracted features to a detection head composed of two subnetworks. One subnetwork outputs bounding box coordinates, while the other outputs class probabilities for each bounding box. This architecture is unique because it uses both classification and regression losses to train its model rather than using a single loss function like most other object detection networks. It also uses the same anchor boxes across all scales of images during training so that it can produce consistent results regardless of input size.

The backbone network is an important part of the architecture as it is used to solve the vanishing gradient problem that used to exist in large backbones and is also used to integrate gradient change into the feature map, which tends to reduce the inference speed and model size while improving the accuracy.

It also uses a path aggregation network (PANet) [53] as the neck of the network as a technique to boost the information flow. The PANet uses a newer type of feature pyramid network (FPN) [54] with a large number of bottom-up and top-down layers to improve the propagation of low-level features in the model, thus improving the localization in lower layers and enhancing the localization accuracy of the network.

YOLOv5 employs the same head architecture as both the YOLOv4 [55] and YOLOv3 [41] architectures. The YOLO layers generate three different outputs of feature maps to achieve multi-scale prediction in an attempt to enhance the prediction of small to large objects. The generated feature maps are fed into the backbone network for feature extraction and to the PANet for feature fusion. To calculate the loss, focal loss or binary cross-entropy with logits loss is used. The loss is calculated based on the bounding box regression score, objectness score, and class probability score.

3.2.2. EfficientDet

The architecture of EfficientDet [56] was conceived from the evaluation of several object detection architectures with the intention to optimize several areas and improve the general efficiency. The evaluations made by the authors include the examination of backbone network, feature fusion, and class/box network. To develop an efficient multi-scale feature fusion, weighted bi-directional feature pyramid network (BiFPN) was proposed to replace commonly used architecture that often contributes to unequal fusion. The BiFPN employs learnable weights to deduce the importance of each input feature and continuously apply bottom-up and top-down multi-scale feature fusion.

The authors [56] also proposed a compound scaling method for the backbone that can effectively scale up resolution, depth, and width for the backbone, feature, and class prediction networks. Therefore, combining the backbone and the BiFPN, the EfficientDet network could achieve better accuracy on object detection tasks while using fewer numbers of parameters than other object detectors.

The general architecture features an ImageNet-pretrained EfficientNets, which serves as the backbone of the network; the BiFPN is the feature network that continuously applies bottom-up and top-down bidirectional feature fusion, which is then read into the class and box network to predict object class and the bounding box.

3.3. Training

The proposed architecture is trained in a way where the ESRGAN network is used to generate the SR from the input LR image and then the bounding boxes of the object of interest are scaled accordingly. Afterwards, the newly generated image is fed into the object detectors for training on the task of pothole detection.

4. Experiments

As previously mentioned, the training of the architecture was completed in a separate manner from the ESRGAN network used to generate SR images and the object detectors trained on the SR images generated. The pre-trained model of the ESRGAN that has been trained on the DIV2K dataset of HR images usually used for image restoration tasks was employed.

For the EfficientDet model, we employed adaptive learning rate and learning rate scheduling mechanism, which work together to reduce the learning rate to the defined schedule. The operation is based on cosine decay schedule. The schedule periodically applies the cosine decay function to the optimizer at a given step. We employed a learning rate base of 8×10^{-2}, with the warmup learning rate set to 0.001 and interval set to every 2500 steps. The training consists of a total of 40,000 steps and a batch size of 8. We used the momentum optimizer to update the entire architecture weights until it converges.

As for the YOLOv5 architecture, in order to match with the size of the EfficientDet network, the larger size of YOLOv5 called YOLOv5l was used. The version has 49.0 mAP on the COCO dataset, which is a decent performance. An initial learning rate of 0.01 was set and the final one cycle learning rate set to 0.2. SGD momentum optimizer was employed, which updates at every 3 epochs and has a momentum of 0.8 and initial bias value of 0.1. The training consists of a total of 200 epochs with a batch size of 16.

We implemented the EfficientNet architecture with the Tensorflow object detection framework and the YOLOv5 with the PyTorch framework and both trained and tested using NVIDIA P-100 GPU with 16 GB on the Google Colaboratory platform (pro version). The training for the YOLOv5 architecture took about 10 h to complete, while for the EfficientDet, it took about 12 h to complete.

4.1. Datasets

It is important to use datasets that present the real scenario of detecting potholes in the wild, which typically involves a camera mounted on a moving vehicle capturing images at a distance. Dhiman and Klette [57] established the lack of proper datasets and benchmark dataset for the task of pothole detection and proposed to combine a number of datasets to mitigate these challenges. In our research, we follow the same ideology.

1. CCSAD—Guzmán et al. [7] presented a dataset named challenging sequences for autonomous driving (CCSAD), which consists of captured video at 20 fps using two Basler Scout scA1300-32fm firewire greyscale cameras from a moving vehicle on the street of Mexico. The dataset presents instances of potholes on the road surface amongst other objects. The entire dataset is divided into four segments, colonial town streets, urban streets, avenues and small roads, and tunnel networks. It is a very large dataset of about 500 GB consisting of calibrated and rectified pairs of sterol images, videos, and meta-data for each of the segments. The image resolution of the dataset is 1096 × 822.
2. Japan—The Japan dataset has been widely used in road damage detection competitions and some research work. It contains about 163 k images of roads with dimensions of 600 × 600 collected across Japan. Different categories of road damages are presented in the dataset, including cracks and potholes; however, few instances of potholes are presented in the dataset. We have selected from the few available images with instances of potholes contained in them.
3. Sunny—This dataset presents several images of pothole instances, mostly small-sized and at a distance. The image resolution is 3680 × 2760 captured with a GoPro camera mounted on the vehicle.
4. PNW—The dataset [6] is a YouTube video recorded on a Pacific Northwest highway during the winter season. The dataset presents a realistic pothole detection problem with roads that have been dilapidated by melting snow and rainwater. The vehicle from which the video is recorded has a typical speed range of 45 km/h to 90 km/h. Images of dimension 1280 × 720 were extracted from the video frames.

We used about 1300 images for the training dataset, which comprise SR images from the CCSAD's Urban Sequence 1 dataset, the Japan dataset, and the PNW dataset collectively; see Figure 5. The Sunny dataset was also used in the mix, but the super-resolution pipeline was not performed on it because the images come at a higher resolution. The validation dataset contains about 188 images, representing about 12% of the entire data, to monitor the performance of the model during training. The testing set contains 81 images, which are used to evaluate the performance of the models.

Figure 5. The Japan dataset. Generated SR image is shown on the left and the corresponding LR image on the right.

4.2. Evaluation Metrics for Detection

Both YOLOv5 and EfficentDet networks give output as bounding boxed with the corresponding classes. In our case, we our have one class (pothole) and the rest as background class. To evaluate the performance of our proposed methods, we have employed precision and recall.

To determine the values of the proposed evaluation metrics, we can calculate the values with true positives (TP), which are the set of corrected detected objects, false positives (FP)—set of wrongly detected objects, false negatives (FN)—set of objects that are not detected by the detector. Therefore, the precision is given as the as the ratio of true positives to all the other predicted objects (Equation (5)), and the recall is given as the ratio of detected objects to the number of all objects in the dataset (Equation (6)). We set the IoU to vary between 0.5 to 0.95 for a more generalized evaluation.

The output of our detection models is the bounding boxes along with the single pothole class, which is the standard for most object detection networks. To evaluate the results of our experiments, we used average precision (AP) at a given calculated intersection over union (IoU). Precision and recall are thus used in computing AP.

We can measure the error in predicted bounding boxes by comparing them to their true locations. A box is considered to be correctly predicted if it overlaps with a ground truth box. This overlap is calculated as the overlap between the detected and ground truth box (IoU). If we consider all boxes with an IoU greater than a given value as true positives, while the remaining boxes as false positives, we can compute the precision at the given IoU. Moreover, if we vary the IoU between values 0.5 and 0.95 at a given step of 0.05, we can combine these precision values at each step to compute average precision (AP) at IoU = 0.5:0.95.

To obtain a single representative performance, we use mean average precision (mAP) as our evaluation metric. We evaluate performance by considering results at both IoU = 0.5 and IoU = 0.5:0.95 since this is the range of overlap between the object detections

and ground truth objects during training. It is calculated by measuring the average distance between instances of the ground truth and predicted bounding boxes and then dividing that by the number of ground truth instances.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

5. Results

5.1. Detection with SR and LR Images

To determine the performance of the detectors, we trained both proposed object detectors on the SR images generated by the ESRGAN network. We also used LR images of the same datasets to train and test the models. The YOLOv5l and EfficientDet-D1 models were used, trained on SR and LR images, respectively.

Table 1 shows the results of the object detectors in terms of the mAP on each of the datasets employed. The EfficientDet network achieved up to 10.6% AP on the datasets, while the YOLOv5l network also achieved up to 12% AP on the datasets when the LR images were used for training and testing. These results are considerably lower than the SR images. The detection results for the models trained on the SR images are evident in the table. We have achieved up to 32% AP on the dataset for the YOLOv5 model and 26% AP with the EfficientDetD1 model.

Table 1. Detection on super-resolution images and low-resolution images for both models. The AP (average precision) values are calculated using 10 different IoUs from 0.5 to 0.95 with 0.05 step intervals.

Model	Image Resolution	Test Results (mAP at IoU = 0.5:0.95)	Test Results (mAP at IoU = 0.5)	Test Results (Recall)	Test Results (Precision)
ESRGAN + EfficientDet	LR	10.6%	20%	30%	53%
ESRGAN + YOLOv5	LR	12%	30%	41%	60%
ESRGAN + EfficientDet	SR	26%	39%	66.77%	100%
ESRGAN + YOLOv5	SR	32%	46%	70%	97.60%

5.2. Discussion

We have used IoU = 0.5 to determine and calculate the recall. From the experiments, it can be noted that the methods recorded higher precision than the recall value. In its essence, the methods can detect instances of potholes better, including the tiny instances, but might not detect all the instances of potholes. The reason for the lower recall value can be related to the misclassification that occurs in the training and testing datasets.

Due to the nature of the datasets used for training and testing, errors were recorded during testing. It was observed that there were considerably large numbers of mislabels in the Japan dataset, where instances of cracks were labeled as potholes or plain road surfaces were labeled as potholes. In addition, manhole covers were also mislabeled as pothole instances. This can be connected to the low resolution of the images during annotation and the multi-class property of the Japan dataset. It is also worth mentioning that the Sunny dataset contains instances of a great deal of visual occlusions, where the potholes are almost invisible even to humans. This contributed to the detection error encountered during the experiments. Therefore, a better result is obtained from the PNW and CCSAD datasets compared to the Japan dataset. However, the Japan dataset is almost 50% of the entire dataset used.

The effect of the low recall value recorded can reduce the accuracy of the pothole detection system. This means that the detection system may not catch all the negative

instances that it should be catching, which could result in an increased amount of false positives. Thus, the performance of the detector is considerably reduced such that it may not be able to detect all the positive cases that it should be detecting. While it will correctly detect pothole instances in an image, it has a higher probability of not detecting all the pothole instances in frame.

The recall value could have been significantly improved if we had relabeled the datasets accordingly. This will reduce the mislabeling that exists in the datasets, therefore improving the overall performance of the target detectors. Moreover, collecting a new, independent, and realistic pothole database will considerably resolve the encountered issue.

Apart from the characteristics of the potholes in the dataset, computing resources also contributed to the detection errors. To train on the SR images, we resized their input to object detectors. However, since we were not able to utilize the full capability of the SR images and train with more images and more training time, our performance suffered. The experimental results clearly indicate that the evaluation metrics will significantly improve with a well-labeled and clean dataset where pothole instances are clearly labeled.

The ESRGAN + YOLOv5 and the ESRGAN + EfficientDet models are used to detect instances of potholes in any given SR image, and the results are satisfactory as it can be observed in Figure 6 that the ESRGAN + YOLOv5 model correctly detects instances of potholes in an image even when they appear small or at a distance within the frame. Figure 6 also shows the LR images input into the ESRGAN network and the corresponding SR image output. The image enhancement provided by the ESRGAN network has helped the detectors to obtain higher AP values by making the images visually good enough to identify the objects easily. It is evident from the figure that the visual quality of the generated SR images is quite good compared to the corresponding LR images, with both detectors detecting cases of small potholes. Few misclassifications were also experienced, even with the varying lighting conditions. What can be considered as a misclassification is in the cases where cracks, hand-hole covers, and a black refuse nylon on the road were detected as potholes, as shown in Figure 7.

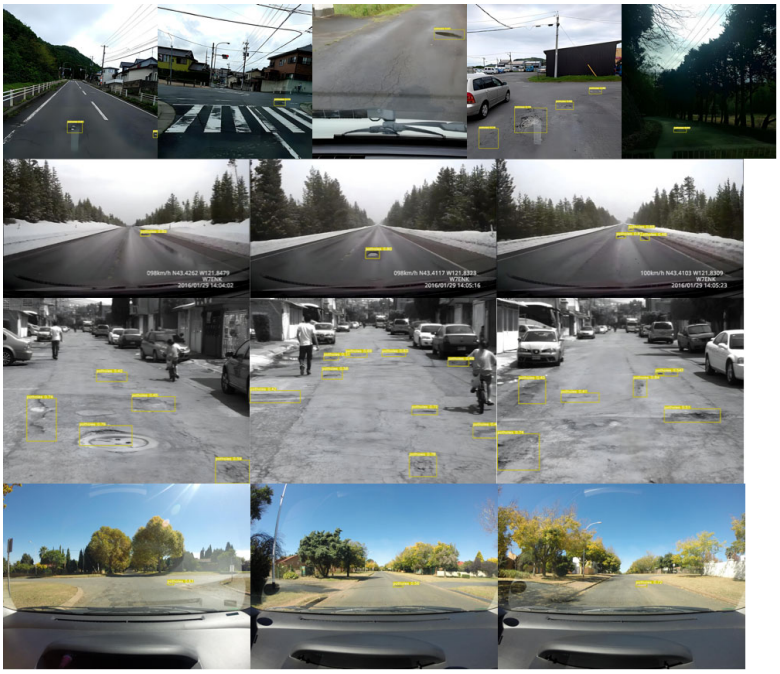

Figure 6. Detected potholes from the validation set using the ESRGAN + YOLOv5 method, with each row showing images from the different datasets.

Figure 7. Misclassifications experienced during validation, where a manhole cover, crack, and black nylon are classified as pothole instance.

To measure the performance of the models and compare with the state-of-the-art results, we generally employed the precision and recall as common classification measures. The metrics were previously defined and explained. We compared the results of both proposed object detectors and they produced good results on detecting pothole instances and almost identical values.

We completed the evaluation on a per dataset basis for the test set and later evaluated collectively for each model, as presented in Table 1. Table 2 presents the results for the selected few frames from the CCSAD urban sequences 1, which compares the detected potholes with the ground truth for the images. It also shows the comparison with the Dhiman and Klette [57] results on similar frames from the same dataset in an attempt to compare our results with other studies. The table shows that both our proposed methods significantly outperform the compared study in terms of precision but with a lower recall value, which can be related to the mislabeling issues previously identified. The ESRGAN + YOLOv5 model also performed better than the ESRGAN + EfficientDet model.

Table 2. Comparative evaluation of the proposed ESRGAN + YOLOv5 and ESRGAN + EfficientDet with Dhiman and Klette [57] best-performing model (LM1) on selected frames from the CCSAD dataset.

Method	Mean Precision (%)	Mean Recall (%)
ESRGAN + EfficientDet	100	71.5
ESRGAN + YOLOv5	100	72.2
LM1	89.9	92.8

Table 3 presents the results for a few selected frames from the PNW dataset since the PNW dataset accurately represents the scenario at hand, with a fast-moving vehicle and realistic road damage challenges. Both models misclassify a black trash bag on the road

as a pothole because of the dark color property and spherical shape of the bag, as can be seen in Figure 5. Both models showed great potentials in identifying realistic potholes on the road surface from a moving vehicle and at a distance, even when the holes are filled with water, as also illustrated by the images in the same figure. On the Sunny dataset, the YOLOv5 model has an overall precision of 100% and recall of 56.7%. On the Japan dataset, the EfficientDet model has an overall precision of 81.25% and recall of 65.8% while the YOLOv5 model has an overall precision of 86.3% and recall of 61.6% (see Table 4).

Table 3. Comparative evaluation of the proposed ESRGAN + YOLOv5 and ESRGAN + EfficientDet with Dhiman and Klette [57] best-performing model (LM1) on extracted frames from the PNW dataset.

Method	Mean Precision (%)	Mean Recall (%)
ESRGAN + EfficientDet	100	63
ESRGAN + YOLOv5	92.5	86.1
LM1	88.6	85.05

Table 4. Precision and recall on each dataset in the test set for both models.

Dataset	ESRGAN + YOLOv5		ESRGAN + EfficientDet	
	Mean Precision (%)	Mean Recall (%)	Mean Precision (%)	Mean Recall (%)
CCSAD	100	72.2	100	71.5
Sunny	100	57	60	34
PNW	92.5	86.1	100	63
Japan	86.3	61.58	81.25	65.8

The overall precision and recall value for all 81 images in our testing dataset is 100% precision and 66.77% recall for the EfficientDet model and 97.60% precision and 70% recall for the YOLOv5 model.

We also performed a comparative analysis with several studies that have used similar datasets for the task of pothole detection The results are compared with ours in Table 5. Dhiman and Klette [57] employed the same datasets and Mask R-CNN model in their studies.We have done this not to compare our methods with theirs but to show that, even when a segmentation model is used, our proposed methods with super-resolution can perform better in some instances with object detection. Our results generally provided better precision but lower recall than theirs, with their best method recording an overall precision of 88% and overall recall of 84%, while our best model (both models have very close overall results) with YOLOv5 has an overall precision of 97.60% and recall of 70%. Moreover, Darapaneni et al. [58] and Kortmann et al. [59] also employed the Japan dataset for the purpose of pothole detection. Table 4 compared our results side by side with theirs, with our methods performing better than both studies in terms of precision and recall.

Table 5. Comparative evaluation of the proposed ESRGAN + YOLOv5 and ESRGAN + EfficientDet with Darapaneni et al. [58] and Kortmann et al. [59] on the Japan dataset.

Author	Method	Mean Precision (%)	Mean Recall (%)
Darapaneni et al. [58]	YOLOv3	60	50
Darapaneni et al. [58]	YOLOv4	90	11
Darapaneni et al. [58]	YOLOv5	40	40
Kortmann et al. [59]	FRCNN	68.56	54.02
Our method	ESRGAN + EfficientDet	81.25	65.85
Our method	ESRGAN + YOLOv5	86.3	61.58

The proposed method outperforms the existing approaches in detecting smaller potholes in the frame or at a distance and also boosts the general accuracy of detection. Especially with challenging datasets, our approach could accurately detect pothole instances where other common methods could not perform well on these datasets. To this end, our method provides a means of obtaining accurate detection of potholes from low-quality imaging devices and in complex, unpredictable scenarios. The success of our proposed method, with the detection performance improving significantly with the use of SR images on the dataset and the state-of-the-art detectors, can miss potholes that are smaller in size.

Capturing agents do not have to employ expensive sensors, such as LIDAR or HD cameras, to be able to obtain a precise detection of pothole instances on asphalt surfaces.

However, due to the super-resolution technique used, more processing power is needed on-board to perform real-time processing, thereby increasing the computational cost. The lesser the computing resources are, the longer the inference time to detect the pothole instances.

6. Conclusions

Potholes significantly contribute to road accidents all over the world and are also culprits in causing wear-and-tire on vehicles. In this study, we have proposed a technique based on the super-resolution of images to mitigate the overlooked realistic characteristics of pothole detection tasks. We proposed the GAN-based ERSGAN network to achieve super-resolution from low-resolution images and two object detectors, YOLOv5 and EfficientDet networks, to perform the task of pothole detection from the SR images. The proposed method provides several enhancements over the state-of-the-art research in pothole detection, such as the accurate identification of potholes in a challenging environment, accurate detection of potholes that appear smaller in the image and at a distance, tackling the issue of small object detection, and resolving the issue of detection from 2D images generated by inexpensive equipment. Both object detection techniques produced similar results in pothole detection on the test set by recording good precision and recall values. Both methods are easy to implement, but the YOLOv5 method provides a faster training and inference speed than the EfficientDet method.

In addition, when compared to the state-of-the-art methods in pothole detection on similar datasets, the proposed method significantly outperforms the other methods, especially with a higher precision but with a lower recall value than the state-of-the-art method that employed instance segmentation. Overall, the mAP values for the proposed methods are significantly higher than the results obtained from the models trained on LR images.

The results we have presented here no doubt shed more light on the task of small object detection and how super-resolution images are used to mitigate the challenges it poses. While super-resolution GANs are popularly used in the field of remote sensing for object detection, other domains with similar challenges have not fully utilized the benefits it presents. This study thereby tries to bridge these gaps in different areas of research and is not limited to the task of pothole detection. In future works, we intend to focus on employing end-to-end training for both the super-resolution network and the object detection network. In addition, our work is focused on developing light-weight super-resolution networks to significantly reduce the inference time and employ lightweight semantic segmentation networks to detect all objects in frame and on the road surface. While there is no standard benchmark dataset for a complex pothole detection dataset, we intend to collect and release such a dataset that will realistically represent potholes on road surfaces for autonomous vehicles.

Author Contributions: Conceptualization, H.S. and E.Ç.; methodology, H.S and E.Ç.; software, H.S.; validation, H.S. and E.Ç.; formal analysis, E.Ç.; investigation, H.S.; resources, H.S.; data curation, H.S.; writing-original draft preparation, H.S.; writing-review and editing, H.S. and E.Ç.; visualization, H.S.; supervision, E.Ç.; project administration, E.Ç.; funding acquisition, H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dewangan, D.K.; Sahu, S.P. PotNet: Pothole Detection for Autonomous Vehicle System using Convolutional Neural Network. *Electron. Lett.* **2020**, *57*, 53–56.
2. Kavith, R.; Nivetha, S. Pothole and Object Detection for an Autonomous Vehicle Using YOLO. In Proceedings of the 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 6–8 May 2021; pp. 1585–1589. https://doi.org/10.1109/ICICCS51141.2021.9432186.
3. Patra, S.; Middya, A.I.; Roy, S. PotSpot: Participatory Sensing Based Monitoring System for Pothole Detection using Deep Learning. *Multimed. Tools Appl.* **2021**, *80*, 25171–25195. https://doi.org/10.1007/s11042-021-10874-4.
4. Salcedo, E.; Jaber, M.; Requena Carrión, J. A Novel Road Maintenance Prioritisation System Based on Computer Vision and Crowdsourced Reporting. *J. Sens. Actuator Netw.* **2022**, *11*, 15. https://doi.org/10.3390/jsan11010015.
5. Junqing, Z.; Jingtao, Z.; Tao, M.; Xiaoming, H.; Weiguang, Z.; Yang, Z. Pavement Distress Detection using Convolutional Neural Networks with Images captured via UAV. *Autom. Constr.* **2022**, *133*, 103991. https://doi.org/10.1016/j.autcon.2021.103991.
6. PNW Dataset. Available online: www.youtube.com/watch?v=BQo87tGRM74 (accessed on 23 January 2022).
7. Guzmán, R.; Hayet, J.; Klette, R. Towards Ubiquitous Autonomous Driving: The CCSAD Dataset. In *Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29–31 2011, Proceedings, Part II*; Springer: Berlin/Heidelberg, Germany, 2011.
8. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *arXiv* **2014**, arXiv:1409.0575.
9. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common Objects in Context. *arXiv* **2015**, arXiv:1405.0312.
10. Yang, L.; Peng, S.; Nickolas, W.; Yi, S. A survey and Performance Evaluation of Deep Learning Methods for Small Object Detection. *Expert Syst. Appl.* **2021**, *172*, 114602. https://doi.org/10.1016/j.eswa.2021.114602.
11. Bashir, S.M.A.; Wang, Y. Small Object Detection in Remote Sensing Images with Residual Feature Aggregation-Based Super-Resolution and Object Detector Network. *Remote Sens.* **2021**, *13*, 1854. https://doi.org/10.3390/rs13091854.
12. Haris, M.; Shakhnarovich, G.; Ukita, N. Task-Driven Super Resolution: Object Detection in Low-Resolution Images. In *Neural Information Processing. 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12 2021, Proceedings, Part VI*; Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N., Eds.; Springer: Cham, Switzerland, 2021; p. 1516. https://doi.org/10.1007/978-3-030-92307-5_45.
13. Luo, Y.; Cao, X.; Zhang, J.; Cao, X.; Guo, J.; Shen, H.; Wang, T.; Feng, Q. CE-FPN: Enhancing Channel Information for Object detection. *arXiv* **2022**, arXiv:2103.10643.
14. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. https://doi.org/10.1109/tpami.2015.2439281.
15. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. https://doi.org/10.1109/cvpr.2016.182.
16. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
17. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–27 July 2017. https://doi.org/10.1109/cvpr.2017.19.
18. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Loy, C.C. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *Computer Vision—ECCV 2018 Workshops: Munich, Germany, September 8–14, 2018, Proceedings, Part III*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 63–79. https://doi.org/10.1007/978-3-030-11021-5_5.

19. Rabbi, J.; Ray, N.; Schubert, M.; Chowdhury, S.; Chao, D. Small-object Detection in Remote Sensing Images with End-to-end Edge-enhanced GAN and Object Detector Network. *Remote Sens.* **2020**, *12*, 1432. https://doi.org/10.3390/rs12091432.
20. Kamal, K.; et al. Performance Assessment of Kinect as a Sensor for Pothole Imaging and Metrology. *Int. J. Pavement Eng*. **2016**, *19*, 565–576.
21. Li, S.; Yuan, C.; Liu, D.; Cai, H. Integrated Processing of Image and GPR Data for Automated Pothole Detection. *J. Comput. Civ. Eng.* **2016**, *30*, 04016015.
22. Sha, A.; Tong, Z.; Gao, J. Recognition and measurement of pavement disasters based on convolutional neural networks. *China J. Highw. Transp.* **2018**, *31*, 1–10.
23. Chen, H.; Yao, M.; Gu, Q. Pothole detection using Location-aware Convolutional Neural Networks. *Int. J. Mach. Learn. Cybern.* **2020**, *11*, 899–911.
24. Silva, L.A.; Sanchez San Blas, H.; Peral García, D.; Sales Mendes, A.; Villarubia González, G. An Architectural Multi-Agent System for a Pavement Monitoring System with Pothole Recognition in UAV Images. *Sensors* **2020**, *20*, 6205. https://doi.org/10.3390/s20216205.
25. Jinchao, G.; Xu, Y.; Ling, D.; Xiaoyun, C.; Vincent, C.S.; Lee, C.J. Automated Pixel-level Pavement Distress Detection based on Stereo Vision and Deep Learning. *Autom. Constr.* **2021**, *129*, 103788. https://doi.org/10.1016/j.autcon.2021.103788.
26. Fan, R.; Wang, H.; Wang, Y.; Liu, M.; Pitas, I. Graph Attention Layer Evolves Semantic Segmentation for Road Pothole Detection: A Benchmark and Algorithms. *IEEE Trans. Image Process.* **2021**, *30*, 8144–8154. https://doi.org/10.1109/TIP.2021.3112316.
27. Gao, M.; Wang, X.; Zhu, S.; Guan, P. Detection and Segmentation of Cement Concrete Pavement Pothole Based on Image Processing Technology. *Math. Probl. Eng.* **2020**, *2020*, 1360832.
28. Wang, P.; Hu, Y.; Dai, Y.; Tian, M. Asphalt Pavement Pothole Detection and Segmentation Based on Wavelet Energy Field. *Math. Probl. Eng.* **2017**, *2017*, 1604130.
29. Koch, C.; Brilakis, I. Pothole Detection in Asphalt Pavement Images. *Adv. Eng. Inf.* **2011**, *25*, 507–515.
30. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced Deep Residual Networks for Single Image Super-Resolution. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–27 July 2017; pp. 136–144. https://doi.org/10.1109/CVPRW.2017.151.
31. Shermeyer, J.; Van Etten, A. The Effects of Super-resolution on Object Detection Performance in Satellite Imagery. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019. https://doi.org/10.1109/CVPRW.2019.00184.
32. Wei, Z.; Liu, Y. Deep Intelligent Neural Network for Medical Geographic Small-target Intelligent Satellite Image Super-resolution. *J. Imaging Sci. Technol.* **2021**, *65*, art00008.
33. Wang, X.; Xie, L.; Dong, C.; Shan, Y. Real-ESRGAN: Training Real-World Blind Super-Resolution With Pure Synthetic Data. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 11–17 October 2021; pp. 1905–1914.
34. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. SwinIR: Image Restoration Using Swin Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 11–17 October 2021.
35. Zhang, K.; Liang, J.; Van Gool, L.; Timofte, R. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 11–17 October 2021.
36. Guo, Y.; Chen, J.; Wang, J.; Chen, Q.; Cao, J.; Deng, Z.; Tan, M. Closed-loop matters: Dual Regression Networks for Single Image Super-resolution. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 18–20 June 2020; pp. 5406–5415. https://doi.org/10.1109/CVPR42600.2020.00545.
37. Mei, Y.; Fan, Y.; Zhou, Y. Image Super-Resolution With Non-Local Sparse Attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 11–17 October 2021; pp. 3517–3526.
38. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for Small Object Detection. In Proceedings of the 9th International Conference on Advances in Computing and Information Technology, Sydney, Australia, 21–22 December 2019; pp. 119–133. https://doi.org/10.5121/csit.2019.91713.
39. Park, D.; Ramanan, D.; Fowlkes, C. Multiresolution Models for Object Detection. In Proceedings of the European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010; pp. 241–254. https://doi.org/10.1007/978-3-642-15561-1_18.
40. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. https://doi.org/10.48550/arXiv.1804.02767.
41. Chang, L.; Chen, Y.-T.; Wang, J.-H.; Chang, Y.-L. Modified Yolov3 for Ship Detection with Visible and Infrared Images. *Electronics* **2022**, *11*, 739. https://doi.org/10.3390/electonics11050739.
42. Lv, N.; Xiao, J.; Qiao, Y. Object Detection Algorithm for Surface Defects Based on a Novel YOLOv3 Model. *Processes* **2022**, *10*, 701. https://doi.org/10.3390/pr10040701.
43. Courtrai, L.; Pham, M.T.; Lefèvre, S. Small Object Detection in Remote Sensing Images based on Super-resolution with Auxiliary Generative Adversarial Networks. *Remote Sens.* **2020**, *12*, 3152. https://doi.org/10.3390/rs12193152.
44. Hui, Z.; Li, J.; Gao, X.; Wang, X. Progressive Perception-oriented Network for Single Image Super-resolution. *Inf. Sci.* **2021,** *546*, 769–786. https://doi.org/10.1016/j.ins.2020.08.114.

45. Ferdous, S.N.; Mostofa, M.; Nasrabadi, N. Super Resolution-assisted Deep Aerial Vehicle Detection. In Proceedings of the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, Baltimore, MD, USA, 15–17 May 2019; p. 1100617. https://doi.org/10.1117/12.2519045.
46. Jiang, K.; Wang, Z.; Yi, P.; Wang, G.; Lu, T.; Jiang, J. Edge-Enhanced GAN for Remote Sensing Image Superresolution. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5799–5812. https://doi.org/10.1109/TGRS.2019.2902431.
47. Wang, Y.; Bashir, S.M.A.; Khan, M.; Ullah, Q.; Wang, R.; Song, Y.; Guo, Z.; Niu, Y. Remote Sensing Image Super-resolution and Object Detection: Benchmark and State of the Art. *Expert Syst. Appl.* **2022**, *197*, 116793. https://doi.org/10.1016/j.eswa.2022.116793.
48. Nah, S.; Kim, T.H.; Lee, K.M. Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–27 July 2017.
49. Jolicoeur-Martineau, A. The Relativistic Discriminator: A Key Element missing from Standard Gan. *arXiv* **2018**, arXiv:1807.00734.
50. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-time Style Transfer and Super-resolution. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherland, 8–16 October 2016.
51. Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; NanoCode012; Kwon, Y.; TaoXie; Fang, J.; imyhxy; Michael, K.; et al. ultralytics/yolov5: v6.1-TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference (v6.1). *Zenodo* **2022**. https://doi.org/10.5281/zenodo.6222936.
52. Wang, C.; Liao, H.Y.; Yeh, I.H. CSPNET: A New Backbone that can Enhance Learning Capability of CNN. *arXiv* **2019,** arXiv:1911.11929.
53. Shu, L.; Lu, Q.; Haifang, Q.; Jianping, S.; Jiaya, J. Path Aggregation Network for Instance Segmentation. *arXiv* **2018**, arXiv:1803.01534.
54. Tsung-Yi, L.; Piotr, D.; Ross, G.; Kaiming, H.; Bharath, H.; Serge, B. Feature Pyramid Networks for Object Detection. *arXiv* **2016**, arXiv:1612.03144.
55. Bochkovskiy, A.; Wang, C.; Liao, H.Y. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
56. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
57. Dhiman, A.; Klette, R. Pothole Detection Using Computer Vision and Learning. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 3536–3550. https://doi.org/10.1109/TITS.2019.2931297.
58. Darapaneni, A.; et al. Pothole Detection Using Advanced Neural Networks. In Proceedings of the IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 27–30 October 2021; pp. 0567–0572. https://doi.org/10.1109/IEMCON53756.2021.9623237.
59. Kortmann, F.; Talits, K.; Fassmeyer, P.; Warnecke, A.; Meier, N.; Heger, J.; Drews, P.; Funk, B. Detecting various Road Damage Types in Global Countries Utilizing Faster R-cnn. In Proceedings of the IEEE International Conference on Big Data, Atlanta, GA, USA, 10–13 December 2020; pp. 5563–5571. https://doi.org/10.1109/BigData50022.2020.9378245.