mathematics

MDPI

Article

VPP: Visual Pollution Prediction Framework Based on a Deep Active Learning Approach Using Public Road Images

Mohammad AlElaiwi [1], Mugahed A. Al-antari [2], Hafiz Farooq Ahmad [1], Areeba Azhar [3], Badar Almarri [1] and Jamil Hussain [4,*]

[1] Computer Science Department, College of Computer Sciences and Information Technology (CCSIT), King Faisal University, P.O. Box 400, Al-Ahsa 31982, Saudi Arabia
[2] Department of Artificial Intelligence, College of Software & Convergence Technology, Daeyang AI Center, Sejong University, Seoul 05006, Republic of Korea
[3] Department of Mathematics, College of Natural & Agricultural Sciences, University of California-Riverside (UCR), Riverside, CA 92521, USA
[4] Department of Data Science, College of Software & Convergence Technology, Daeyang AI Center, Sejong University, Seoul 05006, Republic of Korea
* Correspondence: jamil@sejong.ac.kr; Tel.: +82-2-3408-3180

Abstract: Visual pollution (VP) is the deterioration or disruption of natural and man-made landscapes that ruins the aesthetic appeal of an area. It also refers to physical elements that limit the movability of people on public roads, such as excavation barriers, potholes, and dilapidated sidewalks. In this paper, an end-to-end visual pollution prediction (VPP) framework based on a deep active learning (DAL) approach is proposed to simultaneously detect and classify visual pollutants from whole public road images. The proposed framework is architected around the following steps: real VP dataset collection, pre-processing, a DAL approach for automatic data annotation, data splitting as well as augmentation, and simultaneous VP detection and classification. This framework is designed to predict VP localization and classify it into three categories: excavation barriers, potholes, and dilapidated sidewalks. A real dataset with 34,460 VP images was collected from various regions across the Kingdom of Saudi Arabia (KSA) via the Ministry of Municipal and Rural Affairs and Housing (MOMRAH), and this was used to develop and fine-tune the proposed artificial intelligence (AI) framework via the use of five AI predictors: MobileNetSSDv2, EfficientDet, Faster RCNN, Detectron2, and YOLO. The proposed VPP-based YOLO framework outperforms competitor AI predictors with superior prediction performance at **89**% precision, 88% recall, 89% F1-score, and 93% mAP. The DAL approach plays a crucial role in automatically annotating the VP images and supporting the VPP framework to improve prediction performance by 18% precision, 27% recall, and 25% mAP. The proposed VPP framework is able to simultaneously detect and classify distinct visual pollutants from annotated images via the DAL strategy. This technique is applicable for real-time monitoring applications.

Keywords: AI-based visual pollution prediction (VPP); deep active learning (DAL); deep learning; simultaneous VP detection and classification

MSC: 68T45

check for updates

Citation: AlElaiwi, M.; Al-antari, M.A.; Ahmad, H.F.; Azhar, A.; Almarri, B.; Hussain, J. VPP: Visual Pollution Prediction Framework Based on a Deep Active Learning Approach Using Public Road Images. *Mathematics* **2023**, *11*, 186. https://doi.org/10.3390/math11010186

Academic Editors: Ezequiel López-Rubio, Esteban Palomo and Enrique Domínguez

Received: 25 November 2022
Revised: 23 December 2022
Accepted: 24 December 2022
Published: 29 December 2022

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the beginning of 2018, the Kingdom of Saudi Arabia (KSA) launched the Quality of Life (QoL) project under the Saudi Vision 2030 framework, contingent on the usage of advanced AI technology to improve the quality of life of its residents by establishing a more comfortable environment for their contemporary lifestyles. The program aims to increase inhabitant engagement with numerous social and cultural activities based on entertainment, culture, tourism, sports, and other sectors able to nurture an increased

quality of life. Heightened participation in such activities is predicted to have a positive economic and social impact by allowing for the establishment of numerous jobs and a diverse range of activities being made available to Saudi residents [1]. As such, the current standing of Saudi cities could be elevated to make them among the world's most livable cities [2]. The community targeted for this program consists of individuals residing within the boundary of Saudi Arabia, including, but not limited to, citizens, residents, visitors, and tourists. As an integral part of the KSA 2030 vision, a strategic economic and social reform framework, municipalities across thirteen provincial regions in Saudi Arabia have launched intensive remedial policies in an effort to secure high living standards for residents of the Kingdom. As we know, the continuation of expansive and invasive anthropogenic influences on the natural environment endangers all living organisms. As defined by the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES), the five significant ecosystem propulsors and biodiversity losers in dire need of swift and effective change are (1) climate change, (2) direct exploitation, (3) pollution, (4) biological invasions, and (5) sea-use change [3]. As such, they are reversing the significant environmental damage engendered by man, which is unequivocally the primary source of discussion in influential environmental discourses throughout recent history.

As defined by toxicity-based literature, pollution is an offshoot of industrial and economic progression, with acute consequences for the environment and its inhabitants. Abiotic drivers, or the non-living components of an ecosystem, precipitated through human activity result in inexhaustible levels of pollution released into both untouched and man-made ecosystems. The degree of such drivers is, of course, in direct relation to the distance between natural and urban areas—an interval that continues to diminish as the demand for ecosystem inputs increases in direct correlation with the growth of human populations. Although much research on air, water, and land pollution exists, sensory pollution, or human-induced stimuli that interfere with the senses, is a relatively unevaluated phenomenon with severe repercussions. This pollution of "disconnection" has recently evolved to include visual pollution (VP)—disturbances or obstructions in the natural environment. Visual pollutants are the final benefactors to multimodal environmental deterioration when examined alongside other forms of sensory pollution in urban environments. Visual pollution (VP), as detailed in this study, refers specifically to disruptive presences that limit visual ability on public roads, with an emphasis on excavation barriers, potholes, and dilapidated sidewalks.

Visual pollution appears in digital images with varying irregular shapes, colors, and sizes, as observed in Figure 1. This particular form of pollution is a relatively recent concern when considering the current plethora of contaminants habitually spotlighted in the academic literature [4]. Several factors, however, have driven an upsurge in visual pollutants; the incessant construction of new buildings, the inevitable deterioration of asphalt roads as well as sidewalks, and even weather conditions, for instance, are directly connected to the rise in VP.

It is important to adhere to and follow government rules for the construction of buildings or any other civil works in neighborhoods to minimize the occurrence of visual pollution. In an effort to mitigate the adverse effects of such disagreeable elements, the government of Saudi Arabia has launched several field campaigns that manually inspect the country for visual pollutants and alert all construction protocol violators to swiftly rectify any virulent activities in order to avoid disciplinary action [1]. However, this non-automatic process is highly time-consuming, economically unfeasible, and mentally as well as physically draining for employees. As such, our team endeavors to architect AI technological processes applicable to real-time investigations of three distinct visual pollutants: (1) excavation barriers, (2) potholes, and (3) dilapidated sidewalks. Identifying and predicting VP, in particular, can be achieved by training convolutional neural networks (CNNs) with various layers of artificial neurons in the context of image recognition and vision computing [5]. Prior to You Only Look Once (YOLO) [6], all multi-stage object detectors (R-CNN, Fast R-CNN, Faster R-CNN, and others) that exhibited state-of-the-art

(SOTA) accuracy used regions to localize targets rather than assessing whole input images. The YOLO architecture consists of a single neural network with a neck, a head, and a backbone, with varying numbers of outputs in the head. When applied to real-time data, these algorithms can also trade-off between accuracy and speed, resulting in unreliable models. On the other hand, YOLO is a series of contemporary object detection models that predicts bounding boxes and classification probabilities from complete images in single evaluations. Because of its speed, precision, more robust network architecture, and efficient training method, this model eventually superseded most traditional SOTA algorithms [7]. In recent years the YOLO detection series has been proven to be a great resource for cutting-edge real-time object detection, as well as to have a significant amount of financial potential. To use YOLOv5 for visual pollution detection we would need to train a YOLOv5 model to recognize specific types of visual pollution, such as excavation barriers, potholes, or dilapidated sidewalks. This could be done by collecting a dataset of images that contain these types of visual pollution and using them to train the model. Once the model has been trained, it can be used to detect and classify visual pollution in new images or video frames. One potential application of this approach could be to use YOLOv5 for the automated monitoring of public spaces for visual pollution, such as streets, parks, or sidewalks. This could help identify areas where intervention is needed to address visual pollution and improve an environment's appearance. It could also be used to monitor the effectiveness of efforts to reduce visual pollution over time. Due to its high speed and performance, we employed the YOLO architecture as an objective backbone detector for the current study.

Figure 1. Examples of the three categories central to this study: (**a**,**b**) represent the barrier category, (**c**,**d**) illustrate the sidewalk category, and (**e**,**f**) depict the pothole category. All RGB images were collected from Saudi Arabia.

The objective of this study is to assist government organizations with AI-based technology that automatically predicts and recognizes visual pollution without user intervention.

The major contributions of this work are summarized as follows:

- The proposed AI-based real-time visual pollution prediction (VPP) aims to simultaneously detect and categorize visual pollution (VP) from color images.
- An end-to-end AI-based framework is trained and evaluated using a private dataset in a multi-class classification scenario to simultaneously predict various pollutants.
- A new private VP dataset is collected by the Ministry of Municipal and Rural Affairs and Housing (MOMRAH), Saudi Arabia. This dataset has various VP classes and is called the MOMRAH benchmark dataset: excavation barriers, potholes, and dilapidated sidewalks.
- Deep active learning (DAL) supports MOMRAH experts in automatically annotating the VP dataset for multiple tasks: detection with a bounding box and classification with a class label. The annotation process is conducted at an object level, not just at an image level. This is because some images carry multiple and different objects at once.
- A comprehensive training process is conducted to optimize and select the optimal solution for the proposed VPP. We perform various emerging AI predictors, which are MobileNetSSDv2, EfficientDet, Faster RCNN, Detectron2, YOLO-v7, and YOLOv5.
- An ablation or adaptation study is conducted to check the reliability of the proposed AI-based VPP framework when unseen images from different sources are used.

The rest of this paper is organized as follows: A review of the contemporary literature relevant to this study is presented in Section 2. Technical details of the proposed VPP framework are presented in Section 3. The results of the experimental study are reported and discussed in Section 4. Finally, Section 5 presents our conclusions.

2. Related Works

The concept of visual pollution (VP) was identified in the mid-twentieth century, alongside ongoing investigations of the malicious nature of air and water pollution. Contrary to the plethora of academic literature concentrated on air and water pollutants, however, is VP, a relatively unexplored issue essential to providing comfortable living environments in a modernizing world. Initially, researchers defined VP as the impairment of a region's visual quality caused by unnecessary advertisements and signage [8]. Lately, however, this concept has been expanded to include any element that results in landscape-based chaos; a myriad of factors, including perpetual construction, the inevitable demise of asphalt roads, erosion, and even a lack of commitment by residents in following garbage management protocols all coincide with the current interpretation of VP [9]. Exposure to VP has also been proven to beget several adverse mental and physical consequences. According to research on the effect of VP on human physiology and psychology, the absence of VP can reduce the perception of pain by increasing cortisol production in the body [1]. Recent emphasis has been placed on managing visual pollutants via identification-based software, such as a geographic information system (GIS), through which methods of cartographic visualization can be adopted in mapping and, correspondingly, reducing VP [10]. Simultaneously, Delphie and ordering weighing methods have also been used in the academic literature to manipulate a number of visual pollutants [10].

In addition, the analytical hierarchy process (AHP) is considered a multi-criteria decision-making technique for dealing with subjective and numerous contradictory criteria for investigating the effects of VP [8]. Artificial intelligence (AI) technology has recently garnered attention in several research fields, including medicine and healthcare [11–14], weather forecasting [15], energy control systems [16], army studies [17], and air as well as water pollution prediction [10]. The colossal success AI technology has had in such topics makes it highly effective in tackling various practical issues [13,18]. Deep learning feature extraction, in particular, is key in architecting a convolutional neural network (CNN) able to predict any feature-based anomaly. Figure 2 follows a contemporary timeline of advanced AI-based techniques used for object detection mentioned in [19].

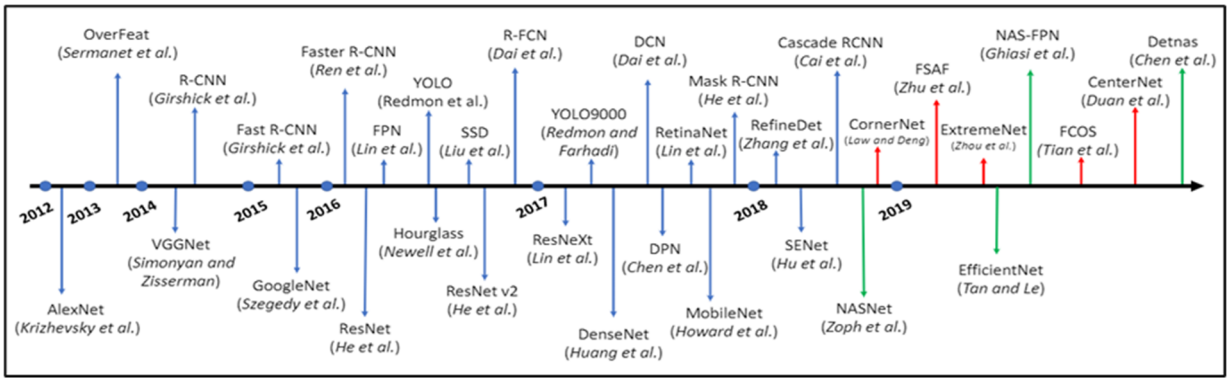


Figure 2. State-of-the-art AI-based object detection techniques [19].

In 2011, Koch et al. presented a two-stage method to detect fissures in images of asphalt roads [20]. Firstly, the image is segmented into two defect and non-defect regions, and any potential pothole shapes are determined via geometric characteristics based on said defect regions. The textural characteristics of the extracted regions are then compared with the textures of the remaining normal regions. If the textures of the defect regions are coarser and grainier than the normal surface, the region is classified as a pothole. An accuracy of 86%, precision of 82%, and recall of 86% were observed. N. Ahmed et al., on the other hand, used a deep convolutional network made up of five convoluting and max-pooling layers to classify VP into four categories: billboards and signage, network and communication towers, telephone and communication wires, and street litter [5]. They collected a dataset of 200 images per category from the Google Images search engine and achieved 95% training accuracy and 85% validation accuracy in their results. Shu et al. adopted a similar deep learning technique via the YOLOv5 model to detect pavement cracks from a dataset of 400 street view images in multiple Chinese cities [21]. A detection accuracy of 70% with a speed detection ability of 152 ms was encountered in identifying cracks in both paved and non-paved street images. Yang et al. proposed a more contemporary detection methodology based around a feature pyramid and hierarchical boosting network (FPHBN) to detect fissures [22]. This method can integrate contextual information from low-level and high-level features in a feature pyramid to generate accurate maps for fissure detection. They achieved an acceptable average intersection over union (AIU) of 0.079, but the execution time for a single image was high—approximately 0.259 s. An ensemble learning methodology was used by Liu et al. to visually detect smoke in an effort to reduce the air pollution produced by industrial factories [10]. Three different CNN architectures with five, eight, and eleven convolutional as well as pooling layers were trained separately using two different visual smoke datasets. Smoke was then detected via the ensemble majority voting strategy. The average detection results over two different datasets were obtained with an overall accuracy of 97.05%, precision of 99.86%, recall of 96.16%, and an F1-measure of 97.97%. Wakil et al. developed a visual pollution assessment (VPA) tool for predicting VP in an urban environment in Pakistan [23]. Their proposed VPA tool has assisted regulators in assessing and charting VP consistently and objectively, while also providing policymakers with an empirical basis for gathering evidence, hence facilitating evidence-based and evidence-driven policies that are likely to have a significant impact, especially in developing countries. In 2021, Wakil et al. presented a web-based spatial decision support system (SDSS) to facilitate stakeholders (i.e., development control authorities, advertisers, billboard owners, and the public) in balancing the optimal positioning of billboards under current governing regulations [24]. The SDSS system has been functional in identifying urban hot spots and exploring suitable sites for new billboards, therefore assisting advertising agencies, urban authorities, and city councils in better planning and managing existing billboard locations to optimize revenue and improve urban aesthetics [24]. Chmielewski et al. proposed a methodological framework for the measurement of VP using tangential view landscape metrics accompanied by statistically significant

proofs [9]. The visible area metrics were found to be highly sensitive VP indicators; the maximum visible distance metrics provided evidence for the destructive effect of outdoor advertisements (OAs) on view corridors [9]. In this paper, an end-to-end deep learning predictor is adopted, trained, and evaluated based on real datasets generated from the KSA. The proposed prediction framework aims to simultaneously detect and classify visual pollutants in three categories: excavation barriers, potholes, and dilapidated sidewalks.

3. Materials and Methods

The schematic diagram of the proposed VPP framework is demonstrated in Figure 3. The proposed framework is able to directly predict the VP objects from the whole input image without user interactions and interventions. This is key to developing a rapid framework for real-time predicting purposes. Accurate and rapid object prediction is crucial for real-time AI applications. In this paper, a comprehensive experimental study is conducted and compares the performances of several object detection methods: MobileNetSSDv2 [25], EfficientDet [26], Faster R-CNN [27], Detectron2 [28], and YOLO [6,12–14,29,30]. The best predictor is selected to achieve the best prediction performance in a compact structure, which is determined to be YOLOv5. Thus, the proposed VPP has a compact, lightweight deep structure and could predict even multiple objects at once. Generally, deep learning detectors consist of two main parts: the CNN-based backbone used for deep feature extraction, the head predictor used to predict the class type, and the bounding box coordinators for the objects [6]. Recently, deep learning detectors are developed by inserting some different deep layers between the backbone and the head, and this part is called the neck network [6]. The VPP has a deep learning backbone for extracting deep high-level features based on the concept of deep learning convolutional networks. Indeed, many deep networks in the literature are used and have their capabilities for deep feature extraction proven, such as VGG [31], ResNet [32], DenseNet [33], Swin Transformer [34], CSP with SPP [35], and others.

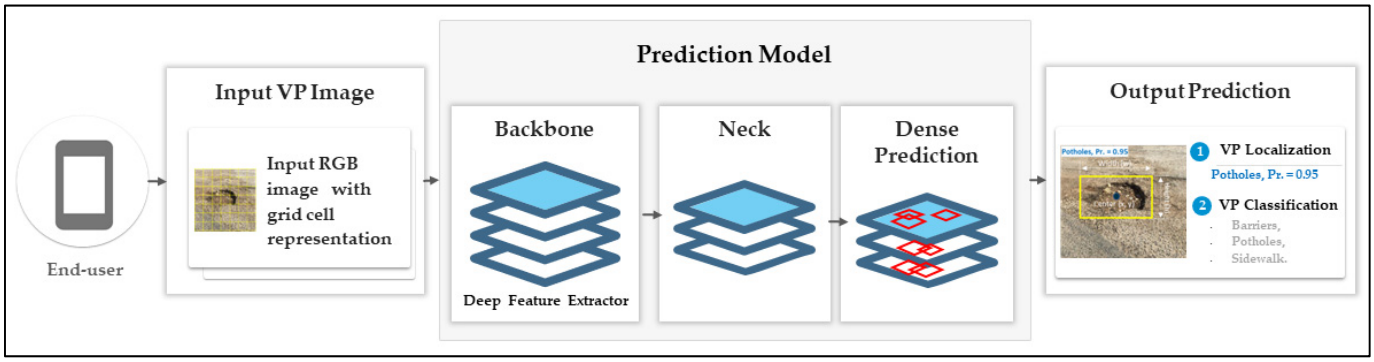
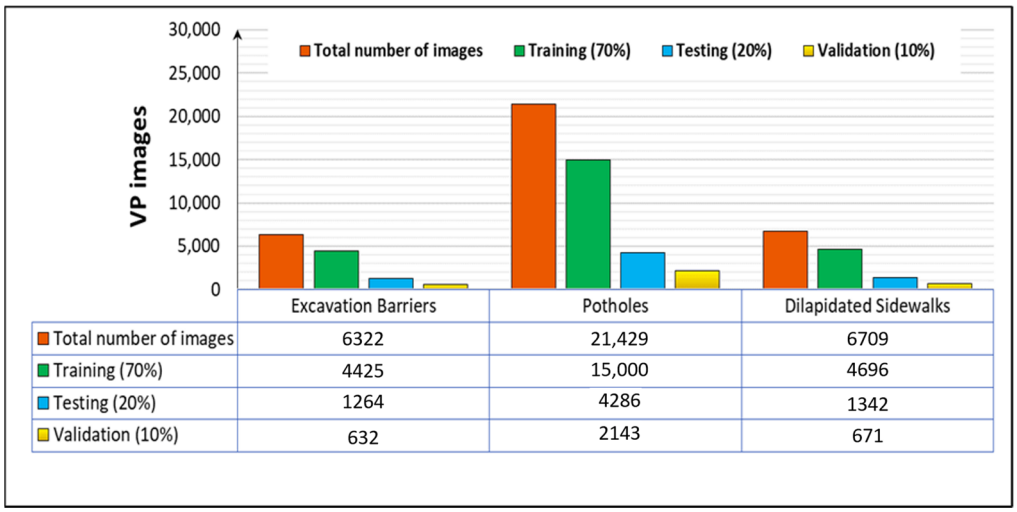


Figure 3. Abstract view of the proposed visual pollution predictor (VPP) framework based on the YOLOv5 predictor. The process consists of three steps: (**a**) feeding an input image, (**b**) using a YOLO prediction model, and (**c**) outputting a prediction with localization and classification.

The neck network is then a key link between the backbone and heads, and is designed to better use the extracted deep features via the backbone network. It includes several bottom-up and top-down deep learning paths for reprocessing and rationally using the extracted features from the backbone network. Here, the output has multiple predictors, or factors, for detection and classification tasks. Afterward, predictor layers are used to predict the object's existing probabilities. For detection, the bounding box predictors are the center coordinators (x,y), width (w), and height (h). For classification, different neurons are assigned to predict a VP object's type to be a barrier, pothole, or sidewalk. All predictors are stored in a tensor of prediction, as shown in Figure 3.

3.1. Visual Pollution Real Dataset: MOMRAH VP Dataset

The dataset is collected from different regions in the Kingdom of Saudi Arabia (KSA) via the Ministry of Municipal and Rural Affairs and Housing (MOMRAH) as a part of a visual pollution campaign to improve Saudi Arabia's urban landscape. To collect this dataset, Saudi citizens and expatriates are requested to take pictures of visual pollutants by using their smartphones and upload them to the government-created Balady mobile application [4]. Our team received official permission from Saudi Arabia's MOMRAH to use the collected data for this study. The VP real dataset is called the MOMRAH VP dataset, and it has 34,460 RGB images for three different classes, which are excavation barriers, potholes, and dilapidated sidewalks. The MOMRAH dataset is publicly published to enrich the research domain with a new VP image dataset [36]. The data distribution over three different classes is shown in Figure 4. Fortunately, some images have more than one object, and this helps to increase the number of training object ROIs. Thus, the total number of object ROIs per class are recorded to be 8417 for excavation barriers, 25,975 for potholes, and 7412 for dilapidated sidewalks. Unfortunately, this dataset lacks annotation labels for both detection and classification tasks since it is collected for the first time as a raw dataset. To annotate all of the images for detection (i.e., bounding box) and classification (i.e., classification label) tasks, a deep active learning strategy is used, where the initial 1200 VP images (i.e., 400 images per class) are manually annotated by four experts. The DAL strategy of the data annotation is presented in Section 3.3.



	Excavation Barriers	Potholes	Dilapidated Sidewalks
Total number of images	6322	21,429	6709
Training (70%)	4425	15,000	4696
Testing (20%)	1264	4286	1342
Validation (10%)	632	2143	671

Figure 4. Visual pollution real dataset (i.e., the MOMRAH VP dataset) distribution over three different classes: excavation barriers, potholes, and dilapidated sidewalks. The dataset per class is split into 70% for training, 10% for validation, and 20% for testing.

3.2. Data Pre-Processing

The following pre-processing steps are performed to prepare the dataset for fine-tuning the deep learning models within the proposed framework: irrelevant image removal, normalization, resizing, and data splitting. Experts investigate the raw RGB images in the MOMRAH VP dataset in-depth, and irrelevant, inaccurate, or unreliable images related to the visual pollution topic are immediately excluded. Some examples of irrelevant and excluded images are depicted in Figure 5. Since the normalization process could improve the overall prediction performance, the VP images are normalized to bring their intensity into the range of [0, 255] [13,18]. Meanwhile, all images are resized using bi-cubic interpolation to scale their intensity pixels into the same range of 460×600.

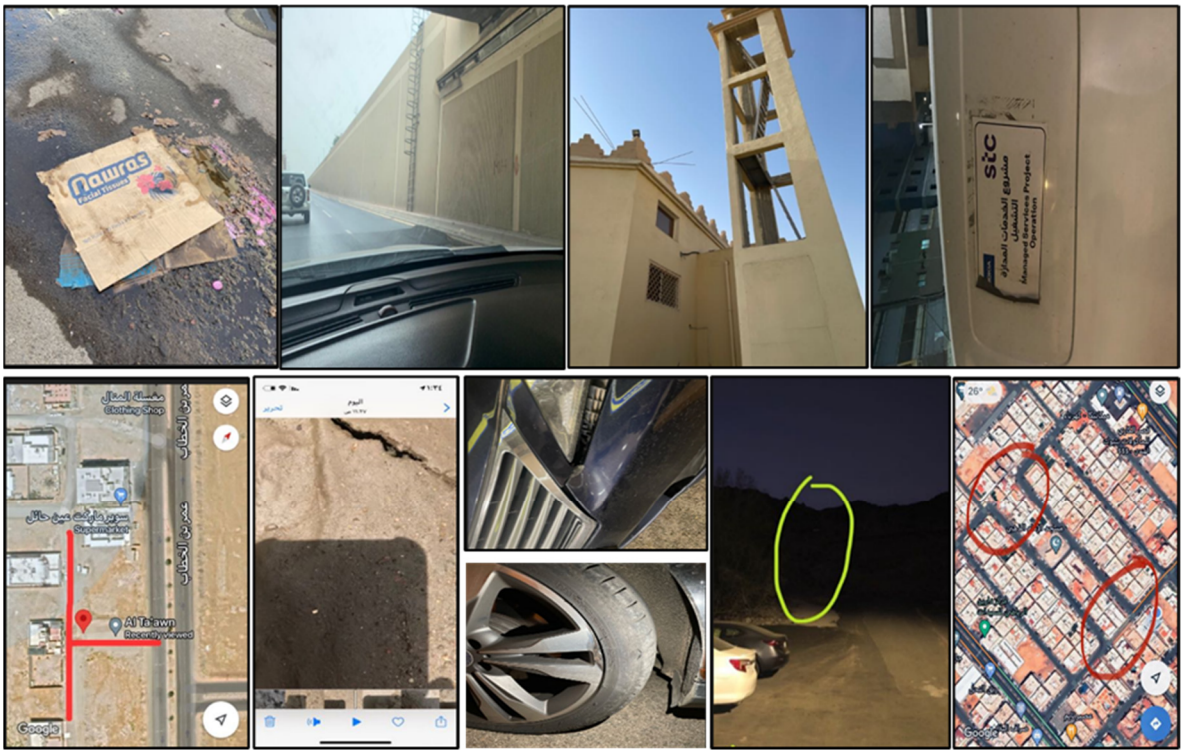


Figure 5. Some examples of the irrelevant images that are excluded during the pre-processing step.

3.3. Deep Active Learning (DAL) for Automatic Data Annotation

Active learning provides an effective method for people to help annotate data, as participants only need to inspect the data they are interested in, while a learning algorithm can automatically adaptively choose and prioritize other data for annotation. Data annotation is especially expensive for object detection tasks. Each object detection frame typically has tens of thousands of pixels, and annotators have to label them manually with boxes around the objects. Annotation can be as simple as drawing a bounding box, but is still highly time-consuming. In addition to the costs, monitoring and controlling the quality of the annotations are more challenging. To summarize, human-in-the-loop may be necessary for general object detection systems, but it is expensive and more difficult in regard to controlling the quality of annotations.

Active learning uses annotated data to reduce the amount of work required to accomplish a target performance. It is used for object classification, image segmentation, and activity recognition. Active learning begins by training a baseline model using a small, labeled dataset, which is then applied to an unlabeled dataset. It estimates, for each unlabeled sample, whether this sample contains essential information that the baseline model has not yet learned by using various query selection strategies (random, uncertainty (entropy), and more). Once the samples containing the most important information have been identified and labeled by the trained model and verified by a human, they can be added to the initial training dataset to train a new model that is anticipated to perform better.

Several different strategies can be used for active learning. One common strategy is called "query by committee," which involves training a committee of multiple models on the available labeled data and then having each model make predictions on the unlabeled data. The model then selects the data points on which the models disagree the most and requests labels for those points in order to resolve the disagreement. This method, query by committee, can be effective because it allows the model to focus on the most informative and uncertain data points, leading to faster and more efficient learning. Another common strategy is "uncertainty sampling", which involves selecting data points for which the model is least certain of the correct label. In this method, data points with the highest

entropy (a measure of uncertainty) or data points closest to the model's decision boundary are selected. Other active learning strategies can be used, such as "representative sampling", in which the model selects data points that are representative of the overall distribution of the data, or "variance reduction", in which the model selects data points that are expected to have the most significant impact on reducing the variance in the model's predictions. In this work, we employed the representative sampling technique using the visual similarity algorithm provided by the Voxel51 brain module.

The proposed deep learning VPP framework is developed to detect and classify the VP objects into three classes: excavation barriers, potholes, and dilapidated sidewalks. To train and develop such a VPP framework, all images in the dataset must be annotated for detection and classification tasks. For the classification task, all images are annotated by four experts in the ministry of MOMRAH by providing an associated class label for each image. For the detection task, a detection label must be represented as a bounding box to surround the whole object (i.e., ROI) inside the image with the coordinators of the start point (x_1, y_1), end point (x_2, y_2), width (w), and height (h), as shown in Figure 3a**.** To perform this labeling, four experts are requested in parallel to manually annotate the best and most clear 400 images from each class by using the CVAT toolbox [37]. Since the labeling process is challenging and time-consuming, the deep active learning (DAL) strategy is mainly involved and used to automatically annotate the rest of the VP images. The primary process of the DAL strategy is depicted in Figure 6. The deep active learning strategy is performed with the following steps: First, we select the best clear 400 images from each class, and four experts become involved to manually annotate the object localization by using the CVAT toolbox. Second, the best deep learning detector model is selected to be trained based on the annotated small dataset (i.e., 400 images per class). Third, the trained DL model is used to test the most relevant and similar images among the remaining unlabeled ones. Fourth, based on the query strategy, the most relevant and exciting samples are selected via the visual similarity approach to be checked by expert-in-the-loop. The selection procedure is usually carried out by checking the high similarity among the initial samples in the first round and the remaining unlabeled ones. The high-similarity instances are selected to be systematically verified and reviewed by an expert. Indeed, the experts interact with machine-in-the-loop to check, modify, and confirm the automated labeling process. The experts have to check that all of the images received some label boxes and manually adjust the boxes' locations and class labels, add some other boxes for the unseeing objects, or even delete the wrong detected boxes. Fifth, after the experts complete the labeling correction process for the first round, the AI model is retrained again using the new trusted labeled images (i.e., 400 + new confirmed subset). Finally, the VP images with lower similarity ratios that could not be labeled in the first round are used as a testing set for the second round of the DAL cycle. This way, the automatic DAL process is repeated until the stopping criteria are satisfied by correctly annotating all of the VP images.

Figure 7 shows some examples of the deep active learning procedure for annotating the images and building a benchmark dataset. Once the DAL process is completed and a benchmark dataset is built, the images per class are randomly split into three different sets: 70% for training, 10% for validation, and 20% for testing. The training and validation sets are used to train and fine-tune the AI models, while the evaluation strategy is performed using the isolated testing set.

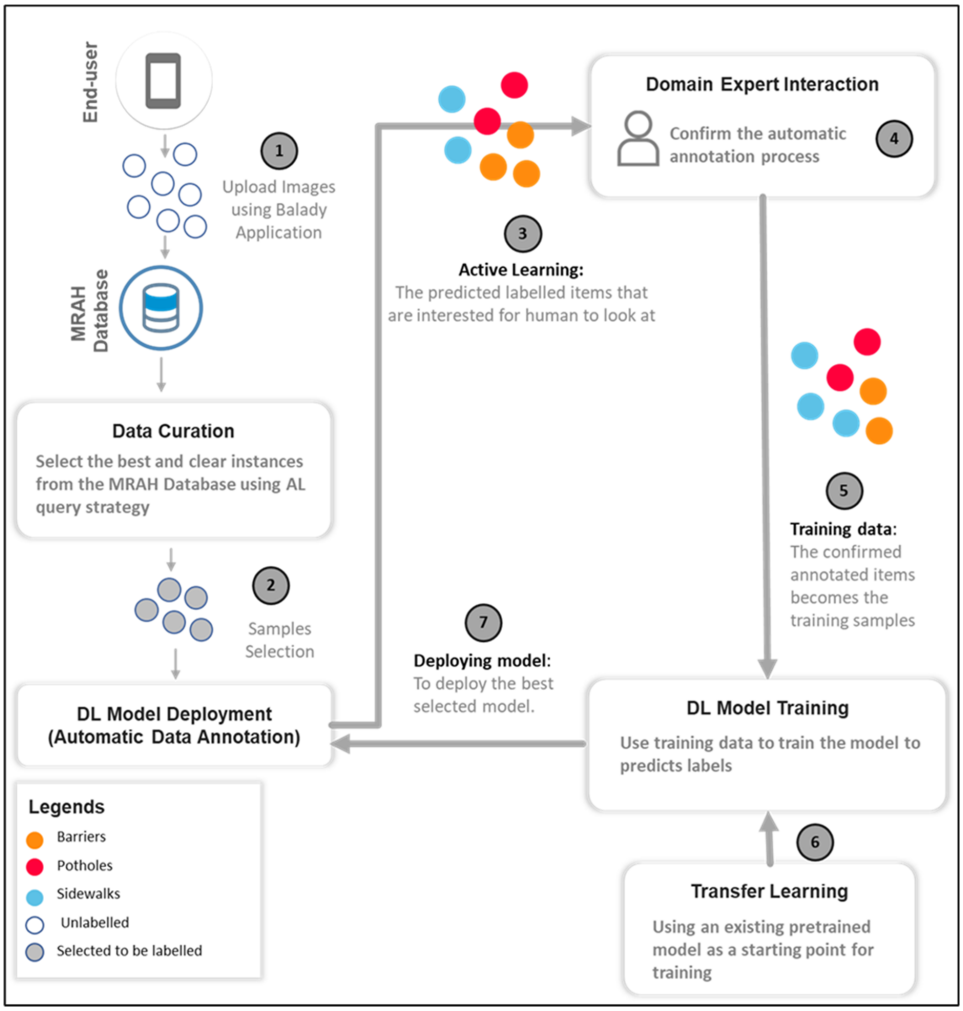


Figure 6. Deep active learning (DAL) strategy for the automatic data annotation process.

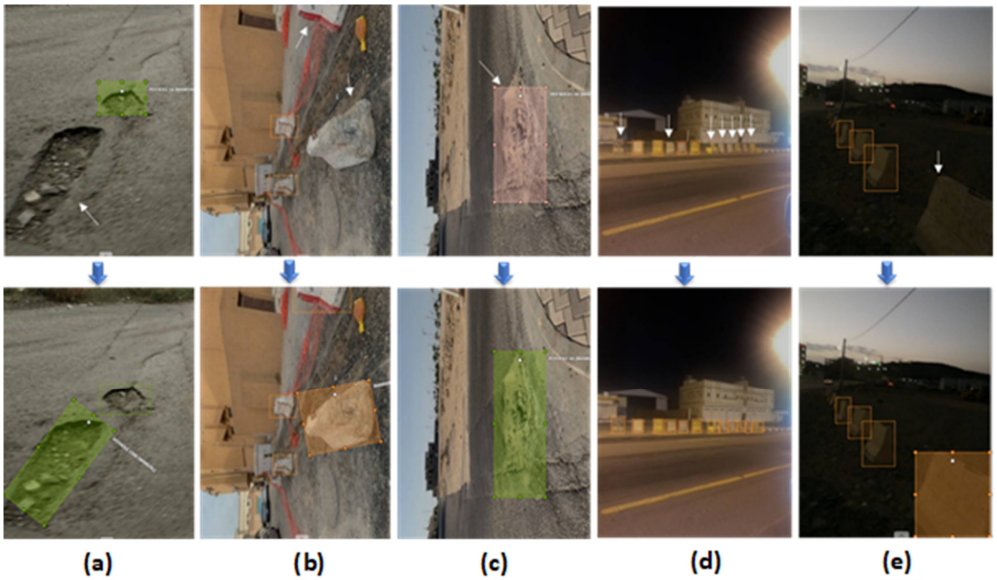


Figure 7. Some examples of the deep active learning (DAL) procedure for image annotation. The first row shows the automatic annotation via a machine during the first round, while the second row depicts the same images but with an expert's interventions and label corrections. Examples from the three categories of excavation barriers, potholes, and dilapidated sidewalks are shown in numbers (**a**–**e**).

3.4. Training Data Enlargement via an Augmentation Strategy

Training data augmentation is a well-proven technique used to enlarge the number of training images for model generalization, avoid over-fitting, and solve the class imbalance problem [38]. To effectively fine-tune deep learning models, a large number of images is required [12,39]. The effectiveness augmentation strategy is mainly used to expand the nature of the dataset. Thus, the deep learning model could be more robust due to the varying image conditions. Augmentation based on the image photometric and/or geometric distortions is recently used to increase the number of training images [6]. For photometric distortion, we imperially adjust the images' hue, saturation, and value by 0.015, 0.7, and 0.4, respectively. For geometric distortion, 0.9 random scaling, 0.1 translation, and 0.5 rotation lift-right are used. Moreover, the recent augmentation methods of Mosaic and MixUp are used with probabilities of 1 and 0.1, respectively [6]. Finally, a total VP training augmented dataset of 41,804 images is generated to fulfill the requirements of deep learning models: 8417 excavation barriers, 25,975 potholes, and 7412 dilapidated sidewalks.

3.5. The Concept of VP Object Detection—VPP-Based YOLO

The AI-based deep learning method "You Only Look Once (YOLO)" has different architectures, such as YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Basically, all versions of YOLOv5 use the deep learning architecture of the cross-stage partial network (CSP) Darknet with spatial pyramid pooling (SPP) layers [35] as a backbone, a path aggregation network (PANet) [40] as a neck, and head detectors [41]. The difference among these versions basically depends on the number of feature extraction modules and the size as well as number of the convolution kernels at each specific location inside the deep network [6]. The schematic diagram of the YOLOv5 is depicted in Figure 8. We select the YOLO predictor since it has an excellent reputation as a one-stage detector with very high prediction speed [6,13,42]. Indeed, the YOLO predictor is mainly used as a detection method regression methodology. It can handle whole input images and predict both object localization as well as object classification type [41–43]. As shown in Figure 8, YOLOv5 consists of backbone, neck, and detector networks, or head predictors, representing the final prediction output.

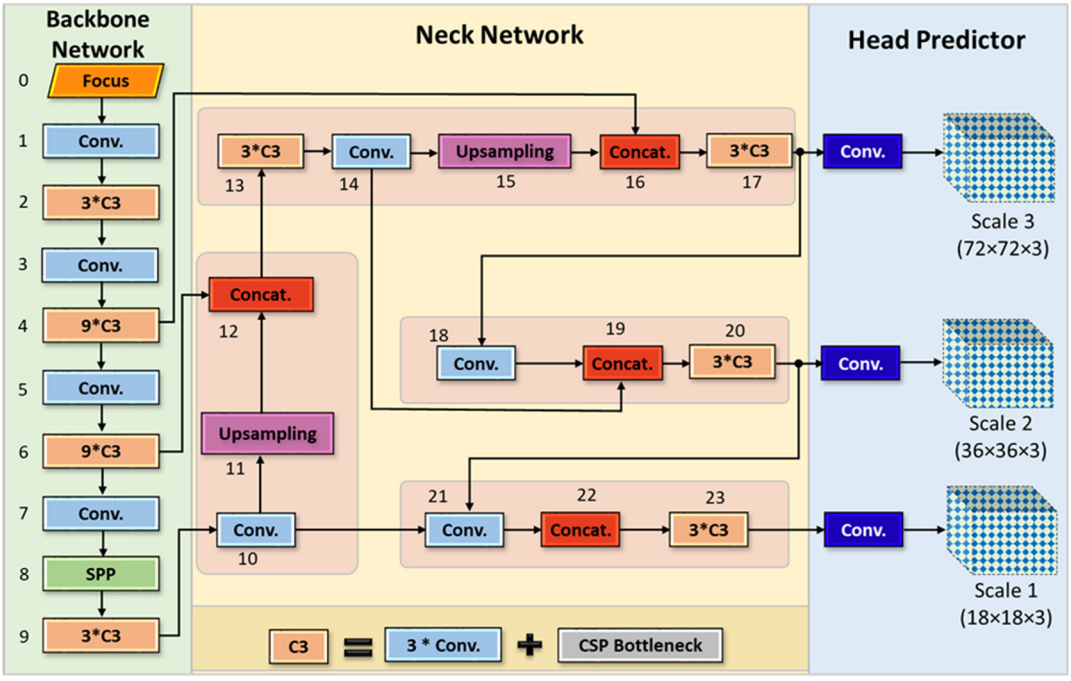


Figure 8. Schematic diagram of VPP-based YOLOv5 to detect and classify the real VP public road images. The * indicates the convolutional process.

In order to adapt to different augmented images, YOLOv5 has the capability to integrate an adaptive anchor frame calculation on the input images. Thus, YOLO could automatically initialize the anchor frame size when the input images are changed and fed to the deep networks [42]. CSP and SPP are utilized for extracting deep feature maps using multiple convolutional and pooling layers for the backbone network. In fact, the CSP network is used to accelerate the learning process, while SPP is used to extract deep features from different scales of the specific feature maps. Both CSP and SPP networks are used to increase the prediction accuracy compared to older versions of YOLO [35]. Indeed, many deep networks in the literature are used and have proven their capabilities for deep feature extraction, such as VGG [31], ResNet [32], DenseNet [33], and Swin Transformer [34]. The feature pyramid deep learning structures of the feature pyramid network (FPN) and the pixel aggregation network (PAN) are consecutively used for the neck network. The FPN conveys the strongest semantic deep features from the top to the lower feature maps. Simultaneously, the PAN is used to convey the strong localization of deep features from lower to higher feature maps. Indeed, both deep learning networks are jointly utilized to strengthen the extracted feature. Thus, the detection performance is increased due to the benefits of both the FPN and PAN. For the final detection procedure, the head predictor is utilized to detect the final target objects with different feature maps' sizes [6]. The head output is mainly designed to detect the final object localization and predict the object type inside the inputted whole image.

3.5.1. Hyperparameters' Evolution

In deep learning, hyperparameters are parameters set prior to formal training. Appropriate hyperparameters could enhance a model's performance. The YOLOv5 algorithm had 23 hyperparameters that were primarily used to set the learning rate, loss function, data enhancement parameters, and others. It was necessary to retrain the appropriate hyperparameters, since all of the data in this study were significantly different from those of the public dataset. YOLOv5 was able to perform hyperparameter optimization by using a genetic algorithm that primarily employed mutation to produce offspring based on the optimal combination of all predecessors, with a probability of 0.90 and a standard deviation of 0.20. In this study, 320 generations of iterative training were set, and the model's F1 and mAP were used to evaluate and determine the optimal hyperparameters. The optimality of the corresponding hyperparameters is denoted by the maximum value of the fitness function in the evolutionary process.

3.5.2. Transfer Learning

Transfer learning, a popular technique in deep learning, could improve the efficiency and robustness of the model training. Typically, external convolutional networks are employed primarily for extracting generic features and concentrating on individual recognition, such as color, shape, and edges. Deeper networks place a greater emphasis on learning task-specific characteristics, primarily for classifying targets. Through the characteristics of transfer learning, the detection algorithm utilized the pre-trained weight during training, eliminating the need for random initialization. This training method could decrease the model's search space and increase training efficiency. The YOLOv5 algorithm utilized the pre-trained weight from the COCO dataset, which contained 1.2 million targets in 80 categories. Although the pre-training weight contained many general features, the COCO dataset differed significantly from this study's recognition target. Therefore, it was necessary to determine if transfer learning could detect potholes, sidewalks, and barrier detection by using the model's mAP.

3.6. Experimental Setting

For training, the strategy of multi-scale training is used to learn prediction across different resolutions of the inputted VP images [40]. Moreover, a mini-batch size of 32 and a number of epochs of 100 are utilized for training and validating the proposed AI models.

A stochastic gradient descent (SGD) optimizer is used with an initial learning rate of 0.01, a final one-cycle learning rate of 0.1, a momentum of 0.937, a weight decay of 5×10^{-4}, warmup epochs of 3, a warmup momentum of 0.8, and a warmup initial bias learning rate of 0.1. The predicted box loss gain, class loss gain, and object loss gain are designed to be 0.05, 0.3, and 0.7, respectively. Moreover, the IoU training threshold and anchor-multiple thresholds are adjusted to be 0.2 and 4, respectively.

3.7. Implementation Environment

The comprehensive experimental study is achieved by using a PC with the following specifications: an Intel(R) Core(TM) i7-10700KF CPU @ 3.80GHz, 32.0 GB of RAM, six CPUs, and one NVIDIA GeForce RTX 3060 GPU.

3.8. Evaluation Strategy

We used the standard evaluation parameters regarding training loss, validation loss, precision, recall, and mean average precision (mAP). The loss of YOLOv5 was used to evaluate the inconsistency between the model prediction results and the ground truths, and it was composed of three components: bounding box loss, object loss, and classification loss. In order to prevent the under-fitting or over-fitting of the VPP model, training loss (loss of the training set) and validation loss (loss of the validation set) would be observed during the training process to obtain the optimal detection model. The mAP metric comprises the product of the accuracy and recall of the detected bounding boxes and ranges from 0 to 1, with higher values denoting superior performance. The mAP may represent the model's global detection performance, especially in comparison to F1. The mAP can be obtained by calculating the area under the corresponding precision–recall curve, which is the standard metric for evaluating an object detection algorithm. In evaluating an object detection algorithm, the mAP is frequently used as the primary performance metric. Based on the principle of IoU, the mAP is an excellent indicator of the network's sensitivity. IoU is the ratio of the overlap area between the ground truth and its predicted areas to the union area. Precision and recall are calculated using true positive (TP), false positive (FP), true negative (TN), and false negative (FN) based on the multi-class confusion matrix. The weighted average of precision and recall are utilized to calculate the F1-score (F1).

4. Experimental Results and Discussion

The experiment of this study is conducted via three evaluation scenarios. First, the dataset initially labeled by experts (i.e., 400 images per class) is used to select the best prediction AI model for our proposed VPP framework. The best AI model is also tested and verified with various activation functions to achieve the best prediction performance. Simultaneously, the trainable hyperparameters of the selected model are carefully optimized via different initialization strategies. Second, once the deep learning model is selected and optimized, the deep active learning (DAL) strategy is used to automatically annotate the remaining raw VP images in our private MOMRAH dataset. Finally, the proposed VPP framework is trained and evaluated using the big data of the labeled VP images over three trails. Meanwhile, the prediction performance of the VPP framework is directly compared with that of other state-of-the-art prediction models using the same MOMRAH dataset.

4.1. The Optimization Results of the Proposed AI-Based VPP Framework

A comprehensive experimental study is conducted to optimize the capability of the proposed AI-based VPP framework for selecting the best solution that leads to optimal prediction performance. To perform this study, the initial curated benchmark dataset (i.e., 400 images per class) annotated by the experts is used. We sequentially investigate three factors that could support the proposed framework, providing better prediction performance. First, various depth and width deep learning networks are investigated using four different YOLO architectures, which are YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. This is to select the optimal version of the YOLO detector that could achieve

the best evaluation performance. Second, once the optimal YOLO version is selected, six activation functions are used and investigated: LeakyReLU, ReLU, Sigmoid, Mish, SiLU, and Tanh. Finally, we investigate three different initialization methods for the trainable hyperparameters of the best AI predictor selected in the first step. All of the experimental results regarding this optimization strategy are presented in the following sections.

4.1.1. Evaluation Results Based on the Various YOLO Structures' Depth and Width

By evaluating four various YOLO networks, we find that YOLOv5x achieves the best prediction performance and outperforms the other architectures. This could be due to its largest convolutional deep learning structure compared with the smaller versions (i.e., YOLOv5s, YOLOv5m, and YOLOv5l), since it is known that deeper and wider deep learning models can achieve better performance. To achieve this finding, all deep learning YOLO predictors are separately trained and tuned using the initial curated dataset, which consists of 400 VP images of each class (i.e., excavation barriers, potholes, and dilapidated sidewalks). All models are trained using the same training settings of 250 epochs and the default hyperparameter initialization method. Figure 9 depicts the optimized loss function performance over 250 epochs during the training time of all of the deep learning models. It is shown that all versions of the YOLO detectors could learn well and achieve better loss values by increasing the number of epochs. YOLOv5x is optimized well, achieving the lowest loss function compared with the other YOLO versions, while YOLOv5s is fine-tuned and achieves the lowest performance in terms of all of the loss functions, as shown in Figure 9. Figure 10 shows the evaluation metrics of precision, recall, and mAP, which were recorded for the same training settings of the four versions of the YOLO detectors. It is clearly shown that all of the evaluation metrics during the training time improve with an increase in the training epochs. This means that the deep learning detectors learned well without any over-fitting to the seen training data.

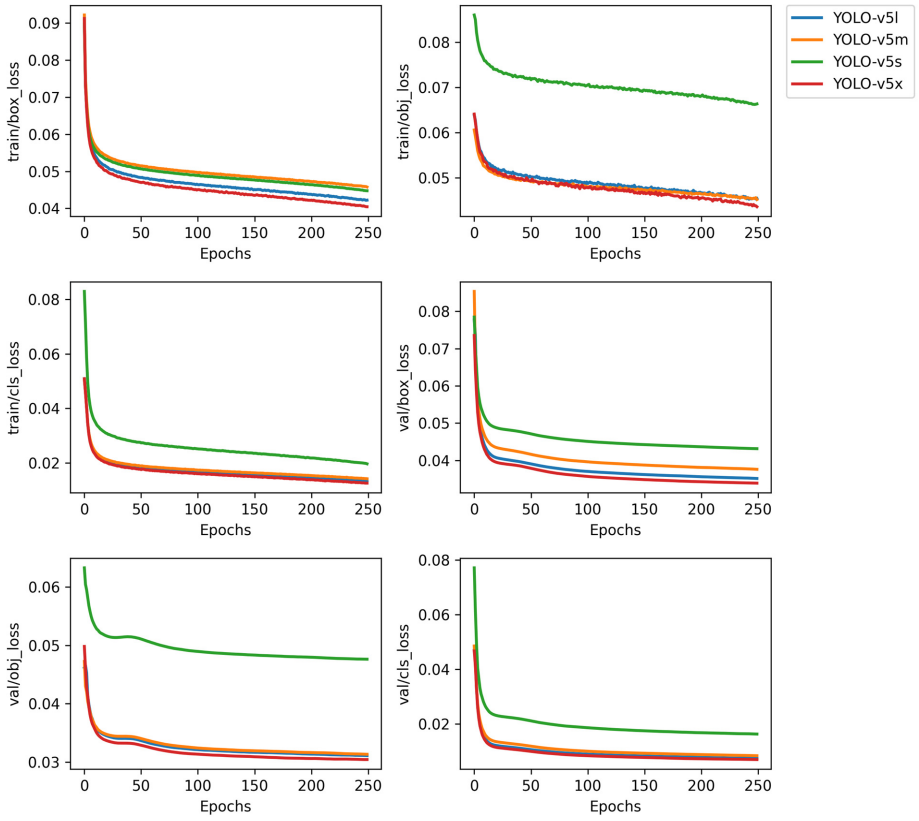


Figure 9. Training parameter optimization results of the proposed VPP framework based on various deep learning YOLO structures (i.e., YOLOv5s, YOLOv5m, or YOLOv5l) in terms of train/valid detected box, object, and cls loss functions.

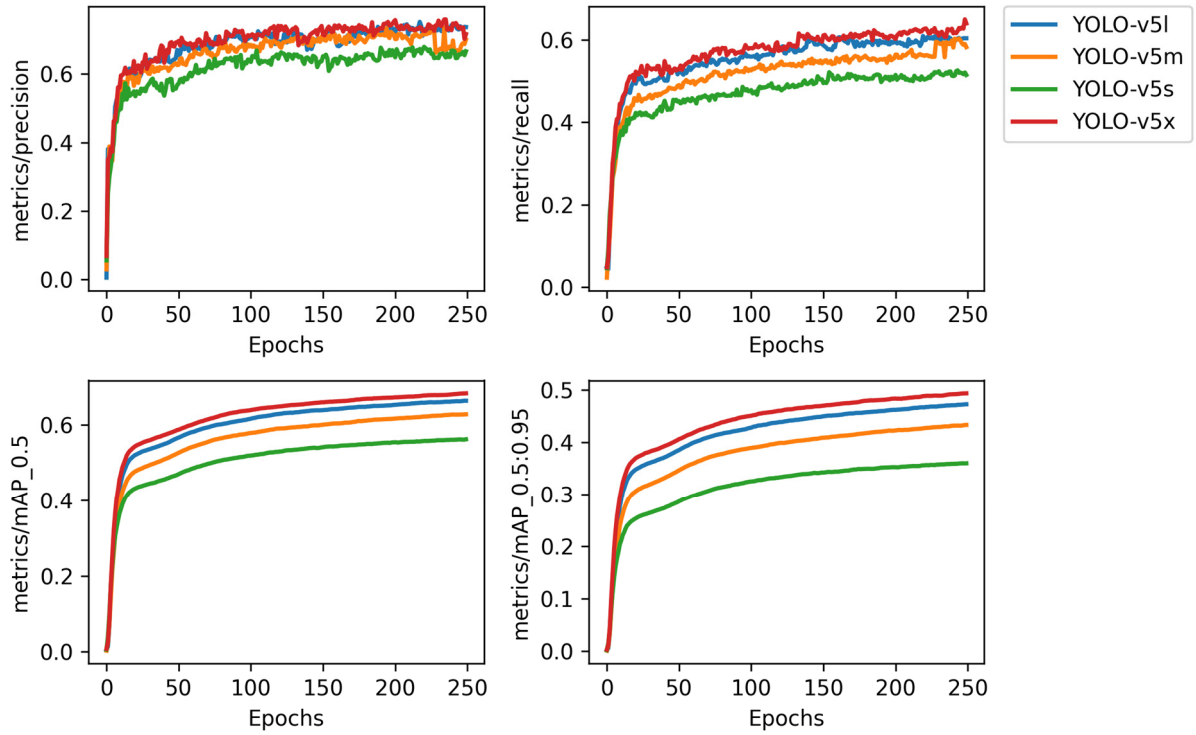


Figure 10. Evaluation performance of the proposed VPP framework based on various deep learning YOLO structures (i.e., YOLOv5s, YOLOv5m, or YOLOv5l) in terms of precision, recall, and mAP.

The best detection performance of all varieties of the YOLO detectors over three trials are presented in Table 1. It is obviously shown that YOLOv5x could achieve the best prediction performance, with 70% precision, 62% recall, an F1-score 66%, and 67% mAP. On the other hand, YOLOv5x is the heaviest deep learning model, with a model size of 169.26 MB and 88,453,800 trainable parameters. This means that its volume and number of parameters could characterize the model's complexity, requiring more GPU memory and a long time for fine-tuning all of the parameters. In contrast, YOLOv5s has the smallest deep learning architecture, with a model size of 14.08 MB and 72,318 parameters. Comparing the aforementioned experiments, it is clear that YOLOv5x is a superior deep-learning model that could achieve the best prediction performance over three classes of potholes, sidewalks, and barrier detection.

Table 1. The evaluation performance of all versions of the YOLO detectors as an average over three trails.

AI Model	Precision	Recall	F1-Score	mAP
YOLOv5s	0.65	0.50	0.57	0.55
YOLOv5m	0.69	0.57	0.62	0.61
YOLOv5l	0.72	0.59	0.65	0.65
YOLOv5x	**0.70**	**0.62**	**0.66**	**0.67**

4.1.2. Evaluation Results of the Best YOLO Candidate with Various Activation Functions

Once YOLOv5x is selected as the best candidate for the proposed VPP framework, we conduct another optimization study to select the optimal activation function that could support YOLOv5x, achieving better prediction results. Six activation functions are used to achieve this goal: LeakyReLU, ReLU, Sigmoid, Mish, SiLU, and Tanh. YOLOv5x is separately trained and evaluated six times according to each activation function. Meanwhile, YOLOv5x is fine-tuned using the initial curated dataset over 250 epochs with the default hyperparameter initialization strategy. The training and evaluation results over 250 epochs using all of the activation functions are compared, as shown in Figures 11 and 12.

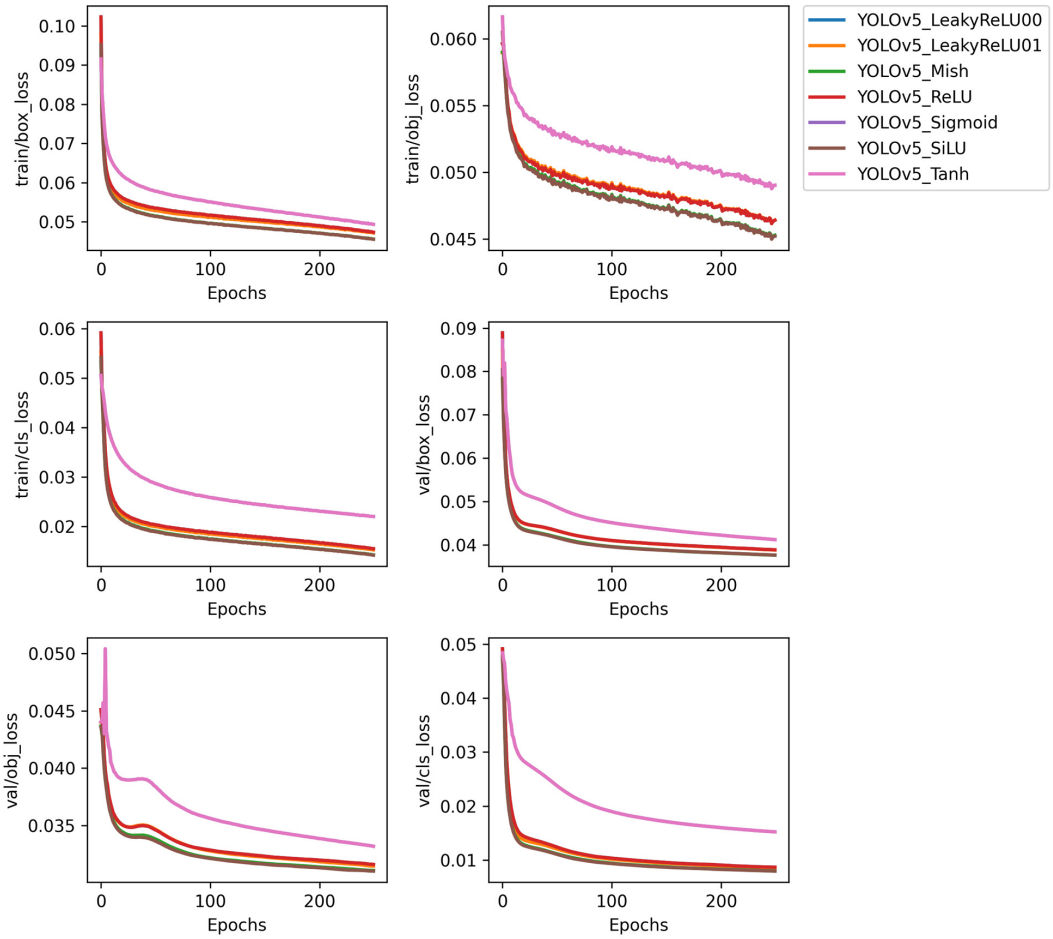


Figure 11. Training and validation loss functions of the proposed VPP framework based on the best candidate of the selected YOLOv5x over 250 epochs. The deep learning YOLOv5x is separately trained using six different activation functions: LeakyReLU, ReLU, Sigmoid, Mish, SiLU, and Tanh.

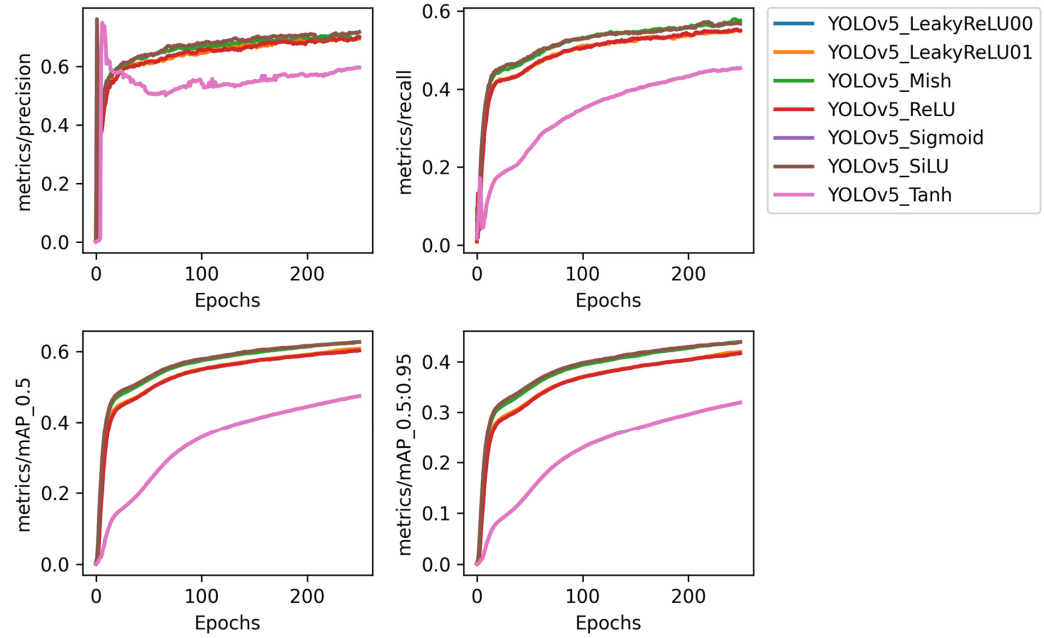


Figure 12. Evaluation prediction performance of the proposed VPP framework using different activation functions of LeakyReLU, ReLU, Sigmoid, Mish, SiLU, and Tanh.

From the empirical results, the selected activation functions of LeakyReLU, Sigmoid, Mish, and SiLU could similarly support the YOLO model to achieve better prediction performance than Tanh and ReLu. The worst evaluation performance is achieved using the Tanh activation function. To conclude, we choose to use the Mish default activation function for conducting the rest of our experiments in this study.

4.1.3. Influence of Hyperparameter Optimization on Prediction Performance

To further improve the prediction performance of YOLOv5x, an additional experimental study is conducted to investigate the most efficient training hyperparameter initialization strategy. The YOLOv5x model is separately trained using three different hyper-parameters and initialization strategies, which are hyp.scratch-low (https://github.com/ultralytics/yolov5/blob/2da2466168116a9fa81f4acab744dc9fe8f90cac/data/hyps/hyp.scratch-low.yaml (accessed on 23 June 2022)), hyp.scratch-med (https://github.com/ultralytics/yolov5/blob/2da2466168116a9fa81f4acab744dc9fe8f90cac/data/hyps/hyp.scratch-med.yaml (accessed on 23 June 2022)), and hyp.scratch-high (https://github.com/ultralytics/yolov5/blob/2da2466168116a9fa81f4acab744dc9fe8f90cac/data/hyps/hyp.scratch-high.yaml (accessed on 23 June 2022)). YOLOv5 has around 30 hyperparameters utilized for a variety of training configurations. These values are specified in *.yaml files located in the/data directory. Better initial predictions will provide better ultimate outcomes; thus, it is essential to establish these parameters correctly before evolving. The same training settings and deep learning YOLOv5x structure are used for each instance of training. By conducting this study, the training and validation loss function values could be reduced with the best evolved hyperparameters that can also support YOLOv5x to achieve better prediction performance results. Figures 13 and 14 depict the training evaluation results of YOLOv5x using various hyperparameters and initialization strategies.

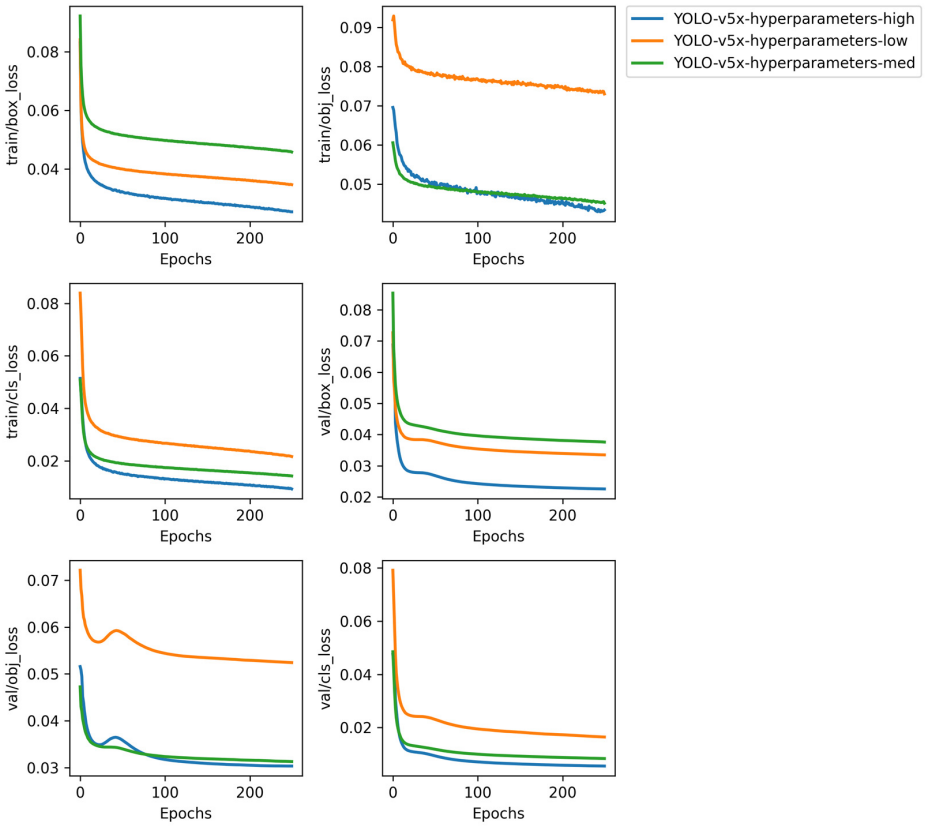


Figure 13. Training and validation loss functions of the proposed VPP framework over 250 epochs. The deep learning YOLOv5x is separately trained using three different hyperparameters and initialization strategies: hyp.scratch-low, hyp.scratch-med, and hyp.scratch-high.

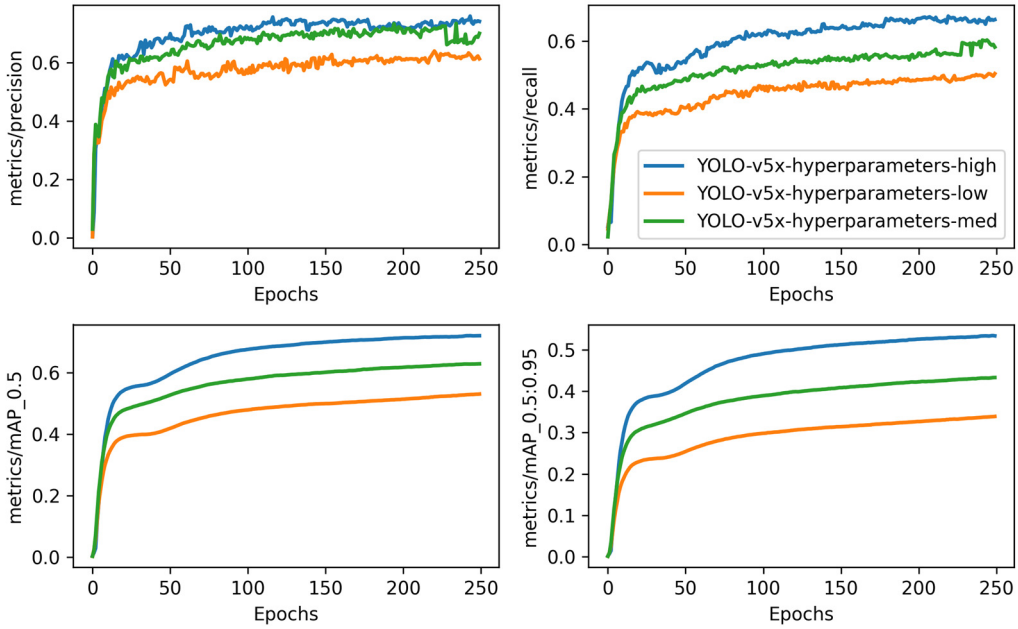


Figure 14. Evaluation prediction performance of the proposed VPP framework using three different hyperparameters and initialization strategies: hyp.scratch-low, hyp.scratch-med, and hyp.scratch-high.

The quantitative average evaluation results of the best YOLO model using three hyperparameters and initialization strategies are summarized in Table 2. As a result of varying training settings, the prediction performance in terms of mAP is increased from 53% using hyp.scratch-low to 71% with hyp.scratch-high. Indeed, the hyperparameter optimization process shows a significant improvement with 18% mAP of the prediction performance. It is important to investigate the multiple factors that could evolve the hyperparameters to boost the model's prediction performance.

Table 2. The evaluation performance of YOLOv5x with three different hyperparameters and initialization strategies.

AI Model	Precision	Recall	F1-Score	mAP_0.5
hyp.scratch-low	0.61	0.50	0.55	0.53
hyp.scratch-med	0.70	0.58	0.63	0.62
hyp.scratch-high	0.74	0.66	0.70	0.71

By using such training remedies and training setting optimization the prediction performance of the proposed VPP framework is significantly improved. Comparing the results in Tables 1 and 2, we can clearly show an improvement in performance by 15% and 5% in terms of F1-score and mAP, respectively.

4.2. Prediction Evaluation Performance during the Deep Active Learning (DAL) Strategy

After selecting the best AI model (i.e., YOLOv5x) and optimizing the model's training activation functions and hyperparameters, the DAL strategy is used to automatically annotate the reset of the unlabeled VP images in our MOMRAH private database. For the DAL query image selection strategy, we use the visual similarity approach of voxel51 brain, which can easily query and sort images to automatically find similar image examples with initial annotated ones through an app's point-and-click interface.

For the first DAL cycle, the new subset of unlabeled images is selected based on higher similarity with the previous labeled set, which is the initial annotated VP images. The new subset of selected images is then automatically labeled via the DAL strategy based on the previous fine-tuned AI model using the initial annotated images. Then, the new and initial labeled image sets are merged and used to fine-tune the deep learning model again

for the next DAL cycle. This means that the number of annotated images for the coming DAL cycle will be increased, which makes the prediction results better than those of the previous cycle. For each DAL cycle, we select 500 new images based on high similarity with the previous labeled images. As shown in Figure 15, the prediction performance of the AI model is dramatically increased with an increase in the number of labeled images of each DAL cycle. Indeed, the visual similarity approach is compared with other approaches, such as random sampling and entropy-based sampling for instance selection and finding images or objects within similar examples. For each selection approach the DAL strategy based on YOLOv5x is separately conducted, and the prediction results over all of the cycles are presented in Figure 15. This means that YOLOv5x is fine-tuned for each DAL-based query selection approach using the same deep learning structure and optimized training settings as concluded in Section 4.1. Each point in Figure 15 represents the mean of three trials utilizing different shuffled initial labeled images. In the last active learning cycle, the prediction performance of 89% mAP is achieved using the visual similarity approach, which is better than the random baseline approach by 9.88%. The entropy selection approach achieves prediction performance with mAP of 85.05%, outperforming the random baseline approach with mAP of 80.65%. Indeed, the entropy method could not capture the uncertainty of bounding box regression, which is the essential part of object detection. Thus, we decide to use the annotation results using the visual similarity selection approach to conduct our experimental results in this study. We can conclude that the query selection approach plays a crucial role in improving the final prediction performance of the proposed VPP framework.

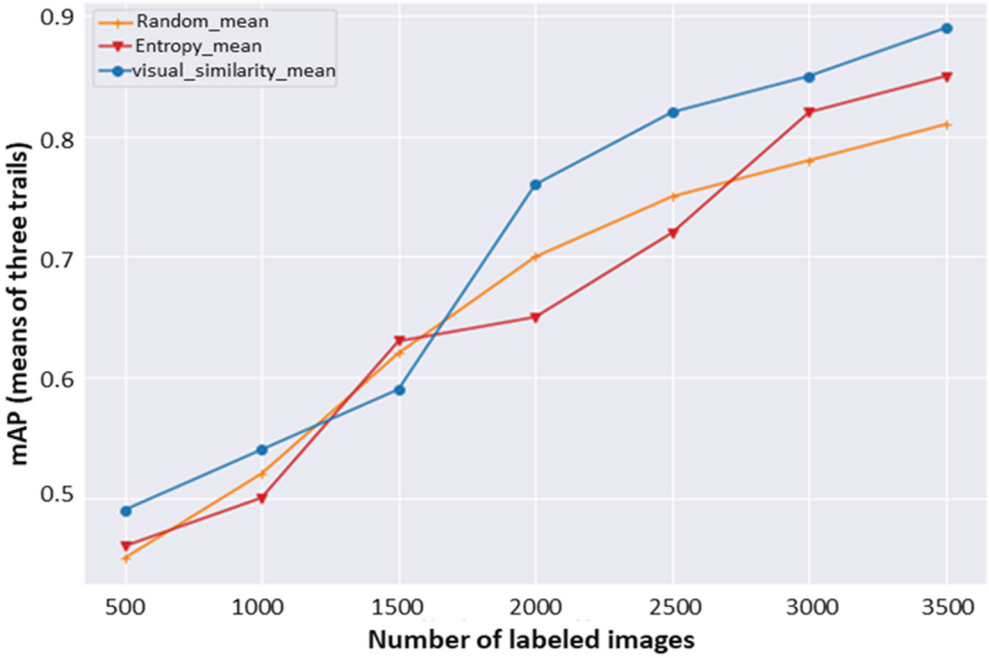


Figure 15. Active learning results of the object detection via the proposed VPP framework used to automatically annotate the VP objects in the VP images in our MOMRAH database.

4.3. Prediction Evaluation Results Using the Whole Annotated Dataset

Another study is conducted after annotating all of the VP images in our MOMRAH database. This is to investigate the capability of the proposed AI-based VPP framework using the manipulated MOMRAH big data and check the prediction performance improvements. Figures 16 and 17 illustrate the prediction behavior of the proposed AI framework using the best AI model (i.e., YOLOv5x). Up to the tenth epoch, the loss values of the box, object, and classification loss functions decrease dramatically for the validation dataset, exhibiting a rapid decline. Meanwhile, the prediction performance reached its peak in terms of evaluation metrics with 88% precision, 89% recall, and 92% mAP. Such performance is achieved as the

best training weights, which are fine-tuned at epoch number 50 by using the early stopping strategy. The prediction performance is improved in comparison with the small initial dataset by 18% precision, 27% recall, and 25% mAP. This means that the DAL annotation process of the VP images is a key to achieving such promising evaluation performance.

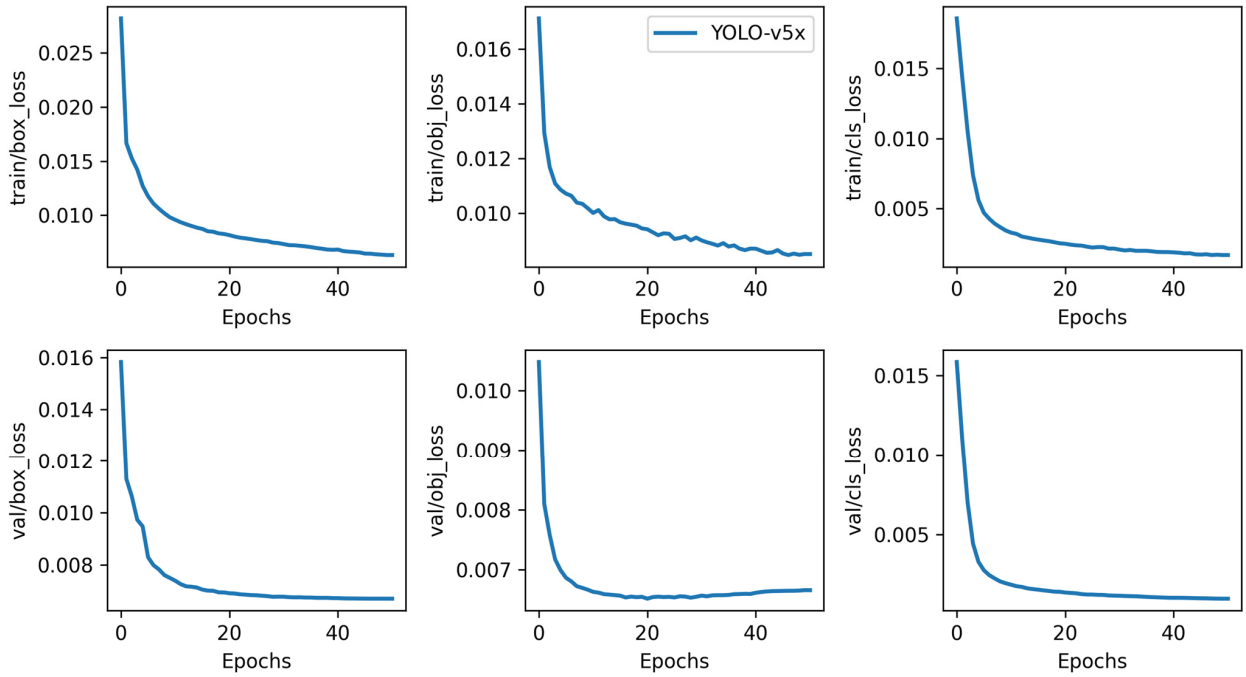


Figure 16. Training and validation loss functions of the proposed VPP framework based on YOLOv5x over 50 epochs using whole DAL-annotated VP images.

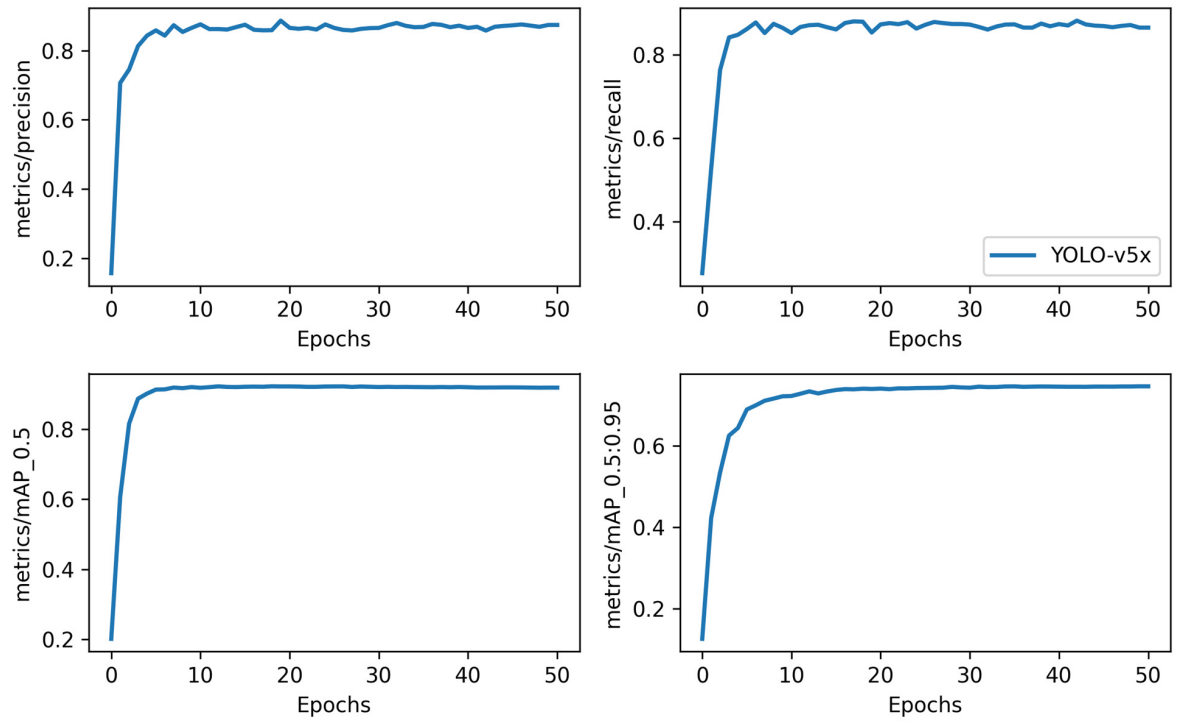


Figure 17. Evaluation prediction performance of the proposed VPP framework based on YOLOv5x using whole DAL-annotated VP images.

4.4. Evaluation Comparison Results

This section presents an evaluation comparison of the proposed VPP framework using various state-of-the-art AI-based object detectors: MobileNetSSDv2 [25], EfficientDet [26], Faster R-CNN [27], Detectron2 [28], and YOLO [6,12,14,29,30]. All of these AI detectors are trained and evaluated using our annotated MOMRAH dataset in a multi-class prediction scenario. Meanwhile, the same training settings are used to fine-tune these deep learning detectors. Such target detection methods are selected to find the optimal prediction performance of the proposed VVP framework applicable for real-time VP applications. Table 3 shows the evaluation comparison results of the proposed VPP framework based on five different AI detectors. It is clearly shown that the optimal prediction performance is achieved using YOLOv5x, with 88% precision, 89% recall, and 92% mAP. Comparing the detection capabilities of several object detection methods, the proposed method achieved the best balance between detection performance and detection speed, while also being hardware-friendly and hence more practical. After optimization, the proposed VPP framework could recognize 319 frames per second (FPS), which is better than other predictors. Recently, YOLOv7 was released after we finalize our methodology and experimental studies; however, we also evaluated it as the most current version of YOLO [7], which provided good performance in terms of mAP and FPS. In the future, YOLOv7 will be considered as the backbone of the suggested framework for more prediction improvements. Such impressive results provide us with evidence that the proposed VPP framework based on the YOLO predictor is the best solution, since it shows an encouraged capability to be applicable for real-time applications.

Table 3. Direct evaluation prediction performance of the proposed VPP framework using our annotated MOMRAH dataset. Different state-of-the-art deep learning object detectors are used for this study: MobileNetSSDv2, EfficientDet, Faster RCNN, Detectron2, and YOLO.

AI Predictor	Precision	Recall	F1-Score	mAP@0.5	Inferencing Time (Msec)	FPS
MobileNetSSDv2	0.70	0.58	0.63	0.62	600	13.2
EfficientDet	0.74	0.66	0.70	0.72	583.1	8.32
Faster R-CNN	0.84	0.77	0.80	0.80	540.2	98.2
Detectron2	0.87	0.86	0.86	0.89	342.0	120.2
YOLOv5x	**0.88**	**0.89**	**0.88**	**0.92**	**22.7**	**319**
YOLOv7	**0.89**	**0.88**	**0.89**	**0.93**	**18.5**	**325**

Moreover, some qualitative evaluation results are demonstrated in Figure 14 to show the performance of the proposed VPP-based framework using different AI predictors. The final model predictor could correctly identify all types of visual pollution. Therefore, such a sophisticated detective system might be used in real-time monitoring applications. As shown in Figure 18 and Table 3, the proposed VPP framework has the best prediction performance using the YOLOv5x perdition model. The lowest evaluation performance is recorded using MobileNetSSDv2, since an average of 62% mAP is achieved. Meanwhile, the predictors YOLOv5x and Detectron2 have almost similar prediction behaviors, with slightly better performance in the case of YOLO by 1% precision, 3% recall, and 3% mAP. As is presented in the last row of Figure 18, the proposed VPP framework has the capability to predict multiple objects in a simultaneous manner regardless of the class type. Both potholes and barriers are perfectly predicted via YOLOv5x with very high confidence scores of 93%, while EfficientDet fails to detect pothole objects. In cases where the input frame has no objects (i.e., not polluted), the VPP framework will still work and tell us that there is no pollution on this image. Therefore, no object bounding box or confidence score will be generated. This is a general aspect of any machine-based learning system (robotics, CAD systems, VP frameworks, and so on).

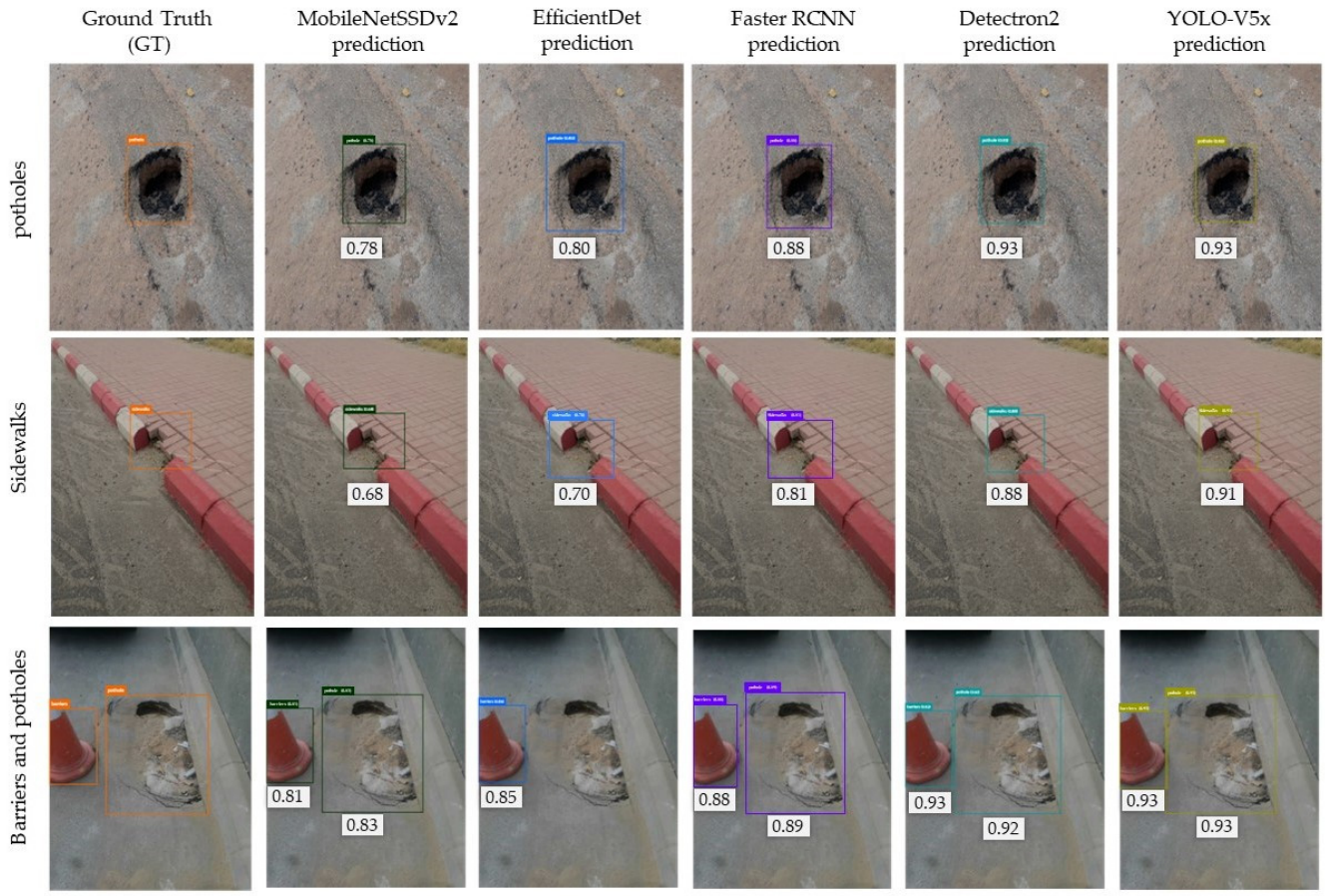


Figure 18. Qualitative evaluation results of the proposed VPP framework for VP detection and classification using AI predictors: MobileNetSSDv2, EfficientDet, Faster RCNN, Detectron2, and YOLO. The prediction object surrounding the box with its confidence score is superimposed on the original image for each AI predictor. The confidence score or classification probability is highlighted inside a small white box besides a detected object.

For indirect comparison with the existing research findings, we summarize in Table 4 some relative studies that have been conducted for VP prediction. Major research studies were conducted to identify solely potholes from road images. For our study, we propose a comprehensive AI-based framework to predict multiple objects simultaneously, such as excavation barriers, potholes, and dilapidated sidewalks. Additionally, we show our performance using precision, recall, and F1-score alongside the impressive mAP evaluation index, which is important for providing us with an impression about model prediction reliability and feasibility. Such an indirect compression always lacks a fair work comparison since the datasets, execution environments, parameter settings, and AI models are totally different. However, our study is compared with recent AI studies to understand the objective of the research area and investigate the work limitations as well as future work.

Table 4. Comparison evaluation results of the proposed VPP framework against the latest works available in the literature.

Reference	Dataset	Target Classes	Methodology	Evaluation Performance (mAP) (%)			
				Precision	Recall	F1-Score	mAP
Aparna et al. (2019) [43]	Road thermal images	Pothole	Classification via CNN-based ResNet	81.15	-	-	-
M. H. Yousaf et al. (2018) [44]	Private dataset: 120 pavement images	Pothole	Classification via SVM	71.59	-	-	-
Ji-Won Baek et al. (2020) [45]	Private road damage images	Pothole	YOLO-based algorithm	83.45	-	-	-
Pham et al. (2020) [28]	2020 IEEE Global Road Damage Cup Challenge	Longitudinal crack, transverse crack, alligator crack, and pothole	Faster-RCNN	-	-	51.40	-
Proposed *	**Private MOMRAH Dataset**	**Excavation barriers, potholes, and dilapidated sidewalks**	**Simultaneous detection and classification via AI-based VPP framework**	**89.0**	**88.0**	**89.0**	**93.0**

* The evaluation result of the proposed VPP is recorded using YOLOv7.

4.5. Work Limitation and Future Work

The scarcity of annotated VP images in a multi-class manner for both detection and classification tasks is always a challenge for supervised AI models. The deep active learning strategy is used for the automatic labeling process but still needs a lot of labor attention, concentration, and effort, since experts must be involved with the machine to correct the automatic labels. Including more classes in our dataset is another challenge, since the individuals that collect the VP images always have different mobile phones with different camera settings, which leads to diversity in image settings.

We have a future plan to continue improving prediction performance using advanced AI approaches such as explainable AI (XAI) to also provide explainable results besides label predictions. Meanwhile, the latest emerging AI techniques, such as transformer-based and knowledge distillation, could be good candidates for more prediction improvement once they are integrated with YOLO in a hybrid scenario, as in our preliminary study [46]. Another plan is to increase the number of classes of visual pollution (VP) to improve the proposed VPP framework to be able to predict several objects in different environments.

4.6. Ablation Study

Our private dataset is publicly published with three classes of excavation barriers, potholes, and dilapidated sidewalks. Unfortunately, we could not find similarly categorized public datasets from different sources with multiple classes to perform an ablation study using multiple classes. However, to conduct an ablation study using unseen VP images from different resources, we found a public dataset called "Pothole detection dataset" [47] but with a single pothole class with 1482 VP images. The proposed VPP framework is re-tested and verified using all VP pothole images. We achieved 81% precision, 75% recall, and 70% mAP, which is more reasonable and acceptable performance, as shown in Figure 19. Moreover, transfer learning is a recent emerging strategy that is expected to assist in producing better predication evaluation results than those received by training from scratch. As mentioned above, transfer learning is a great technique for rapidly retraining a model on new data while retraining the whole network. The proposed model is initialized with weights from a pretrained COCO model (YOLOv5X), where the backbone layers serve as feature extractors by passing the freeze argument while training. Therefore,

the domain adaptations are automatically archived when evaluating the trained model on different datasets.

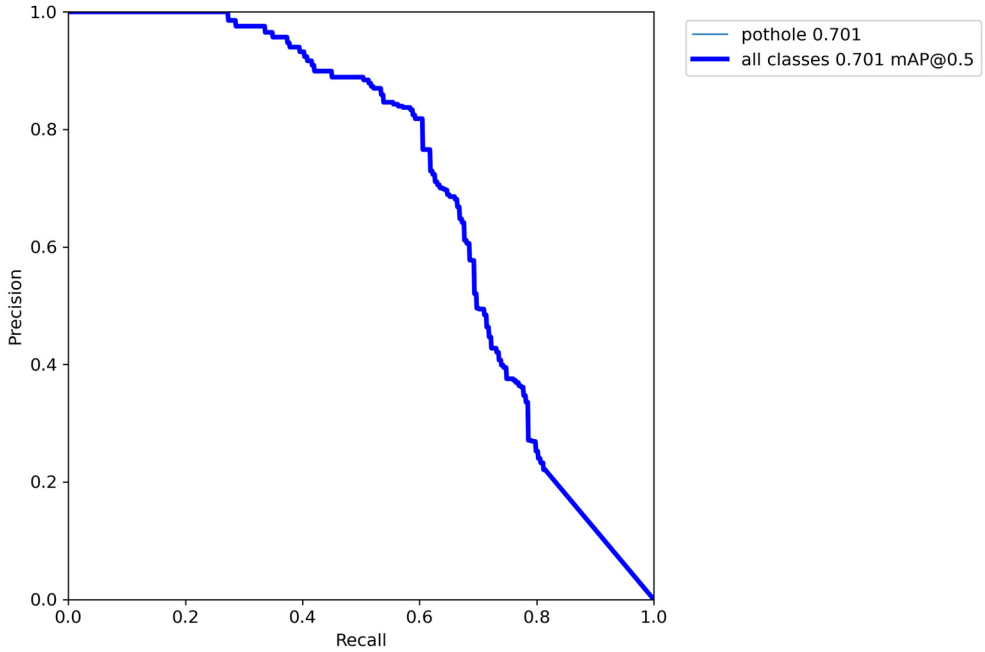


Figure 19. Precision–recall curve on a public dataset which only has images of potholes.

5. Conclusions

This paper proposes an AI-based VPP framework to detect and classify different VP objects in a multi-class simultaneous and classification scenario. To train and evaluate the proposed VPP model, the deep active learning (DAL) approach plays a crucial role in annotating our MOMRAH dataset's VP images. The DAL strategy is applied via three different query image selections: random, entropy-based, and visual similarity, achieving mAP performances of 80.65%, 85.05%, and 89%, respectively. Using annotated big data via DAL, the prediction performance of the proposed VPP framework is improved by 18% precision, 27% recall, and 25% mAP. Indeed, the VPP framework is constructed based on five state-of-the-art AI predictors: MobileNetSSDv2, EfficientDet, Faster RCNN, Detectron2, and YOLO. A comprehensive experimental evaluation study is conducted to select the best AI predictor. The derived evaluation results show that VPP-based YOLO outperforms other predictors, achieving mAP of 92% compared with the figures of 62%, 72%, 80%, and 92% for MobileNetSSDv2, EfficientDet, Faster RCNN, and Detectron2, respectively. Based on the recognition objects, the hyperparameters of the best detector are determined via a comprehensive optimization strategy where transfer learning is used to improve prediction performance. This study compared the backbone of YOLOv5 networks with various widths and depths, and the results demonstrated that, under identical setting conditions, YOLOv5x had superior usability in terms of detection performance, model weight size, and detection speed. This method achieved an optimal balance between detection performance and detection speed while being hardware-friendly, making it more applicable. Over public roads, the optimized YOLOV5x achieved 92 percent mAP in detecting barriers, potholes, and sidewalks.

Author Contributions: Conceptualization, M.A., M.A.A.-a. and J.H.; data curation, M.A., M.A.A.-a. and J.H.; funding acquisition, H.F.A.; investigation, M.A. and J.H.; methodology, M.A., M.A.A.-a. and J.H.; project administration, H.F.A.; software, M.A., M.A.A.-a. and J.H.; supervision, H.F.A.; validation, M.A.A.-a., H.F.A., B.A. and J.H.; visualization, J.H.; writing—original draft, M.A.A.-a.; writing—review and editing, M.A., M.A.A.-a., H.F.A., A.A., B.A. and J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported through the Annual Funding track by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia. (Project no. AN000673.)

Data Availability Statement: The private MOMRAH visual pollution (VP) dataset is used under permission number 4400003688 from the Ministry of Municipal and Rural Affairs and Housing, Kingdom of Saudi Arabia (KSA). The MOMRAH dataset is publicly published at the following link: https://data.mendeley.com/datasets/bb7b8vtwry (accessed on 28 October 2022).

Acknowledgments: We would like to acknowledge the project sponsors of the Annual Funding track by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia. (Project no. AN000673.) We are thankful to the Ministry of Municipal and Rural Affairs and Housing (MOMRAH), Kingdom of Saudi Arabia (KSA), without whose permission for acquiring, using, and publishing the data this project could not have been completed. (Permission letters no. 4300454067 and 4400003688).

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

References

1. Campaign to Improve Saudi Arabia's Urban Landscape. Available online: https://www.arabnews.com/node/1910761/saudi-arabia (accessed on 26 April 2022).
2. Aqeel, A.B. Quality of Life. Available online: https://www.vision2030.gov.sa/v2030/vrps/qol/ (accessed on 26 April 2022).
3. Models of Drivers of Biodiversity and Ecosystem Change. Available online: https://ipbes.net/models-drivers-biodiversity-ecosystem-change (accessed on 10 December 2022).
4. Visual Pollution, Pollution A to Z. Available online: https://www.encyclopedia.com/environment/educational-magazines/visual-pollution (accessed on 25 April 2022).
5. Ahmed, N.; Islam, M.N.; Tuba, A.S.; Mahdy, M.; Sujauddin, M. Solving visual pollution with deep learning: A new nexus in environmental management. *J. Environ. Manag.* **2019**, *248*, 109253. [CrossRef]
6. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
7. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
8. Szczepańska, M.; Wilkaniec, A.; Škamlová, L. Visual pollution in natural and landscape protected areas: Case studies from Poland and Slovakia. *Quaest. Geogr.* **2019**, *38*, 133–149. [CrossRef]
9. Chmielewski, S. Chaos in motion: Measuring visual pollution with tangential view landscape metrics. *Land* **2020**, *9*, 515. [CrossRef]
10. Liu, H.; Lei, F.; Tong, C.; Cui, C.; Wu, L. Visual smoke detection based on ensemble deep cnns. *Displays* **2021**, *69*, 102020. [CrossRef]
11. Al-Masni, M.A.; Al-Antari, M.A.; Choi, M.T.; Han, S.M.; Kim, T.S. Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Comput. Methods Programs Biomed.* **2018**, *162*, 221–231. [CrossRef]
12. Al-Antari, M.A.; Al-Masni, M.A.; Choi, M.T.; Han, S.M.; Kim, T.S. A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. *Int. J. Med. Inform.* **2018**, *117*, 44–54. [CrossRef] [PubMed]
13. Al-antari, M.A.; Hua, C.-H.; Bang, J.; Lee, S. Fast deep learning computer-aided diagnosis of COVID-19 based on digital chest x-ray images. *Appl. Intell.* **2020**, *51*, 2890–2907. [CrossRef] [PubMed]
14. Al-Antari, M.A.; Kim, T.-S. Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital x-ray mammograms. *Comput. Methods Programs Biomed.* **2020**, *196*, 105584. [CrossRef]
15. Salman, A.G.; Kanigoro, B.; Heryadi, Y. Weather forecasting using deep learning techniques. In Proceedings of the 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 10–11 October 2015; pp. 281–285.
16. Wang, X.; Chen, J.; Quan, S.; Wang, Y.-X.; He, H. Hierarchical model predictive control via deep learning vehicle speed predictions for oxygen stoichiometry regulation of fuel cells. *Appl. Energy* **2020**, *276*, 115460. [CrossRef]
17. Gunning, D.; Aha, D. DARPA's explainable artificial intelligence (XAI) program. *AI Mag.* **2019**, *40*, 44–58.
18. Al-antari, M.A.; Hua, C.-H.; Bang, J.; Choi, D.-J.; Kang, S.M.; Lee, S. A rapid deep learning computer-aided diagnosis to simultaneously detect and classify the novel COVID-19 pandemic. In Proceedings of the 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), Langkawi Island, Malaysia, 1–3 March 2021; pp. 585–588.
19. Wu, X.; Sahoo, D.; Hoi, S.C. Recent advances in deep learning for object detection. *Neurocomputing* **2020**, *396*, 39–64. [CrossRef]
20. Koch, C.; Brilakis, I. Pothole detection in asphalt pavement images. *Adv. Eng. Inform.* **2011**, *25*, 507–515. [CrossRef]
21. Shu, Z.; Yan, Z.; Xu, X. Pavement crack detection method of street view images based on deep learning. *Journal of Physics: Conference Series* **2021**, *1952*, 022043. [CrossRef]
22. Yang, F.; Zhang, L.; Yu, S.; Prokhorov, D.; Mei, X.; Ling, H. Feature pyramid and hierarchical boosting network for pavement crack detection. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 1525–1535. [CrossRef]

23. Wakil, K.; Naeem, M.A.; Anjum, G.A.; Waheed, A.; Thaheem, M.J.; Hussnain, M.Q.u.; Nawaz, R. A hybrid tool for visual pollution Assessment in urban environments. *Sustainability* **2019**, *11*, 2211. [CrossRef]
24. Wakil, K.; Tahir, A.; Hussnain, M.Q.u.; Waheed, A.; Nawaz, R. Mitigating urban visual pollution through a multistakeholder spatial decision support system to optimize locational potential of billboards. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 60. [CrossRef]
25. Chiu, Y.-C.; Tsai, C.-Y.; Ruan, M.-D.; Shen, G.-Y.; Lee, T.-T. Mobilenet-SSDv2: An improved object detection model for embedded systems. In Proceedings of the 2020 International Conference on System Science and Engineering (ICSSE), Kagawa, Japan, 31 August–3 September 2020; pp. 1–5.
26. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
27. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
28. Pham, V.; Pham, C.; Dang, T. Road damage detection and classification with detectron2 and faster r-cnn. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 5592–5601.
29. Dima, T.F.; Ahmed, M.E. Using YOLOV5 algorithm to detect and recognize american sign language. In Proceedings of the 2021 International Conference on Information Technology (ICIT), Amman, Jordan, 14–15 July 2021; pp. 603–607.
30. Al-Masni, M.A.; Al-Antari, M.A.; Park, J.-M.; Gi, G.; Kim, T.-Y.; Rivera, P.; Valarezo, E.; Choi, M.-T.; Han, S.-M.; Kim, T.-S. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Comput. Methods Programs Biomed.* **2018**, *157*, 85–94. [CrossRef]
31. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. Huang, G.; Liu, Z.; Maaten, L. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
34. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
36. Mohammad, A.; Hafiz, A.; Jamil, H.; Mugahed, A.; Bader, A.; Areeba, A. *Saudi Arabia Public Roads Visual Pollution Dataset*; King Faisal University: Hufof, Saudi Arabia, 2022. [CrossRef]
37. Tzutalin, L. LabelImg. Available online: https://github.com/tzutalin/labelImg (accessed on 25 April 2022).
38. Kim, J.-H.; Kim, N.; Park, Y.W.; Won, C.S. Object detection and classification based on YOLO-v5 with improved maritime dataset. *J. Mar. Sci. Eng.* **2022**, *10*, 377. [CrossRef]
39. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
40. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
41. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June 2016–1 July 2016; pp. 779–788.
42. Li, Z.; Tian, X.; Liu, X.; Liu, Y.; Shi, X. A two-stage industrial defect detection framework based on improved-yolov5 and optimized-inception-resnetv2 models. *Appl. Sci.* **2022**, *12*, 834. [CrossRef]
43. Bhatia, Y.; Rai, R.; Gupta, V.; Aggarwal, N.; Akula, A. Convolutional neural networks based potholes detection using thermal imaging. *J. King Saud Univ. -Comput. Inf. Sci.* **2022**, *34*, 578–588.
44. Yousaf, M.H.; Azhar, K.; Murtaza, F.; Hussain, F. Visual analysis of asphalt pavement for detection and localization of potholes. *Adv. Eng. Inform.* **2018**, *38*, 527–537. [CrossRef]
45. Baek, J.-W.; Chung, K. Pothole classification model using edge detection in road image. *Appl. Sci.* **2020**, *10*, 6662. [CrossRef]
46. Al-Tam, R.M.; Al-Hejri, A.M.; Narangale, S.M.; Samee, N.A.; Mahmoud, N.F.; Al-Masni, M.A.; Al-Antari, M.A. Ahybrid workflow of residual convolutional transformer encoder for breast cancer classification using digital x-ray mammograms. *Biomedicines* **2022**, *10*, 2971. [CrossRef]
47. Universe, R. Pothole Detection Dataset. Available online: https://universe.roboflow.com/aegis/pothole-detection-i00zy (accessed on 17 December 2022).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.