



This is a repository copy of *Automated detection of greenhouse structures using cascade mask R-CNN*.

White Rose Research Online URL for this paper:
<https://eprints.whiterose.ac.uk/187934/>

Version: Published Version

Article:

Oh, H.Y., Khan, M.S., Jeon, S.B. et al. (1 more author) (2022) Automated detection of greenhouse structures using cascade mask R-CNN. *Applied Sciences*, 12 (11). 5553.

<https://doi.org/10.3390/app12115553>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown


If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

Article

Automated Detection of Greenhouse Structures Using Cascade Mask R-CNN

Haeng Yeol Oh¹, Muhammad Sarfraz Khan², Seung Bae Jeon¹ and Myeong-Hun Jeong^{1,*} 

¹ Department of Civil Engineering, Chosun University, Pilmun-daero 309, Gwangju 61452, Korea; 5matrix0427@hanmail.net (H.Y.O.); zeon6779@gmail.com (S.B.J.)

² Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK; muhammad.s.khan@sheffield.ac.uk

* Correspondence: mhjeong@chosun.ac.kr; Tel.: +82-62-230-7086

Abstract: Automated detection of the content of images remains a challenging problem in artificial intelligence. Hence, continuous manual monitoring of restricted development zones is critical to maintaining territorial integrity and national security. In this regard, local governments of the Republic of Korea conduct four periodic inspections per year to preserve national territories from illegal encroachments and unauthorized developments in restricted zones. The considerable expense makes responding to illegal developments difficult for local governments. To address this challenge, we propose a deep-learning-based Cascade Mask region-based convolutional neural network (R-CNN) algorithm designed to perform automated detection of greenhouses in aerial photographs for efficient and continuous monitoring of restricted development zones in the Republic of Korea. Our proposed model is regional-based because it was optimized for the Republic of Korea via transfer learning and hyperparameter tuning, which improved the efficiency of the automated detection of greenhouse facilities. The experimental results demonstrated that the mAP value of the proposed Cascade Mask R-CNN model was 83.6, which was 12.83 higher than baseline mask R-CNN, and 0.9 higher than Mask R-CNN with hyperparameter tuning and transfer learning considered. Similarly, the F1-score of the proposed Cascade Mask R-CNN model was 62.07, which outperformed those of the baseline mask R-CNN and the Mask R-CNN with hyperparameter tuning and transfer learning considered (i.e., the F1-score 52.33 and 59.13, respectively). The proposed improved Cascade Mask R-CNN model is expected to facilitate efficient and continuous monitoring of restricted development zones through routine screening procedures. Moreover, this work provides a baseline for developing an integrated management system for national-scale land-use planning and development infrastructure by synergizing geographical information systems, remote sensing, and deep learning models.

Keywords: deep learning; computer vision; object detection; instance segmentation; aerial photograph



Citation: Oh, H.Y.; Khan, M.S.; Jeon, S.B.; Jeong, M.-H. Automated Detection of Greenhouse Structures Using Cascade Mask R-CNN. *Appl. Sci.* **2022**, *12*, 5553. <https://doi.org/10.3390/app12115553>

Academic Editors: Wen-Hsiang Hsieh, Jia-Shing Sheu and Minvydas Ragulskis

Received: 6 April 2022

Accepted: 27 May 2022

Published: 30 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



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

Continuous monitoring of restricted development zones is essential for managing and preserving the environment and the sustainability of the national heritage and territories of states. Accordingly, the Republic of Korea's regulations on the "prevention and control of illegal activities in restricted development zones" state that local governments should conduct special inspections at least four times each year. However, the authorities have encountered difficulties in handling complaints regarding illegal greenhouse facilities and encroachments in designated restricted development zones. Moreover, it is impractical to periodically monitor the entire 3800 km² nationwide expanses of the restricted development areas, owing to limitations in human resources and budget constraints [1]. Although greenhouses have positive agricultural importance, illegal encroachments dedicated to such facilities in restricted development zones damage building infrastructure and forests. Hence, continuous monitoring of greenhouse structures is critical.

Alongside recent advancements in computing technology, cutting-edge research on machine learning has been actively conducted in various fields, such as computer vision and natural language processing. In the case of computer vision, object extraction methodologies such as You Only Look Once (YOLO)-based models [2–5], and deep learning-based convolutional neural networks are rapidly being developed. These methodologies have been applied to datasets of images collected from satellite, aerial, and drone-based platforms to detect objects and record their information in a geospatial database for real-time object detection and monitoring. In the case of object detection, these methodologies have been extended to restricted development zones to perform automatic detection and monitoring. They can facilitate the continuous monitoring of restricted zones and enforcement actions against illegal developments and land encroachments.

Therefore, the primary purpose of this study is to improve the initial sensing stage of object detection methods through object recognition, and to ascertain the status of the artificial building and greenhouse structures in restricted development zones. To achieve this goal, we integrated a deep learning-based Cascade Mask region-based convolutional neural network (Cascade Mask R-CNN) model [6]. This study used a total of 2054 selected aerial images obtained from three cities (Hanam, Hwaseong, and Gunpo) in the Republic of Korea. Here, the data used in this study consist primarily of information obtained from the RGB bands. Previous literature studies reveal that, in order to detect objects from images accurately, YOLO is applied to thermal images [7]. In contrast, other studies have integrated both RGB and thermal images to detect objects [8].

However, due to the prevailing agro-meteorological conditions of the Republic of Korea, the greenhouse's temperature is similar to the outdoor environment, thereby making it challenging to use thermal bands in our study area. In particular, greenhouses installed in restricted development zones are in the form of unauthorized temporary buildings. These greenhouses do not heat up in the summer season, which means that the thermal bands will not add a big difference, and will only make our methodology more complex and data-intensive. It is not consistent with the purpose of this study to use the thermal band in the summer season. In addition, some greenhouses only have frames without coating, and it is not easy to detect these objects with thermal bands. Therefore, aerial images consisting of only RGB bands were used in this study.

The experiment results in the current manuscript demonstrated that our optimized model yields better prediction accuracy compared with existing models. We further investigated which model and parameter combination can yield a better prediction of the greenhouse by exploiting transfer learning and hyperparameter tuning, which were undertaken to build an optimized model for greenhouse detection in the restricted development zone of the Republic of Korea. We have experimentally investigated the potential performance of our optimized model in the automated detection of greenhouse structures.

The remainder of this study is organized as follows. Section 2 provides a review of the relevant literature. The experimental design, datasets, and research methodologies are presented in Section 3, followed by the results in Section 4, along with some additional discussion. Finally, Section 5 concludes the work and suggests possible future research avenues.

2. Background

In computer vision, image recognition using machine learning and deep learning algorithms is generally performed to identify the key objects present in images for active surveillance and monitoring purposes. A CNN algorithm was first introduced to perform this function in 1998 as a step towards the effective recognition and processing of images [9]. Recently, a modified and optimized CNN model was published in the literature; Khan et al. [10] conducted a study on gap-filling in flux tower observations by transforming the data into time-dependent images while training a two-dimensional CNN algorithm to predict missing values more effectively. The inception of CNN models led to subsequent progress in deep learning algorithms, and to the development of extended one-

and two-stage object detection models. Single-stage models are mathematical structures in which region proposal and classification are performed simultaneously; the two functions are performed separately in two-stage models. Owing to the simultaneous calculation, single-stage structures exhibit a faster calculation speed compared to two-stage models.

Based on the faster calculation speed of one-stage object detection models, they have been widely used in various applications that require real-time detection at a relatively faster speed. For example, YOLO [2], a one-stage object detection model, enables real-time object detection by delivering input images simultaneously to an efficient backbone network. YOLO has since developed into a more efficient and accurate model, YOLOv4 [5], and is being used in various fields such as guest recognition [11]. Despite its faster computation speed, YOLO has a significant drawback: its input object size must be smaller than a 7×7 grid. This limitation was subsequently addressed by the Single Shot MultiBox Detector (SSD) algorithm [12], which can handle objects of more diverse sizes by integrating feature maps of various sizes. Despite their greater computational efficiency, these one-stage models are limited in that they cannot provide higher accuracy than two-stage models. For example, Lin et al. [13] developed the RetinaNet model as an optimized single-stage model. It introduced a focal loss function to maintain good accuracy and better computational efficiency, solving the severe class imbalance problem between foreground and background classes, and eventually obtaining accuracy similar to that of two-stage models. This means that the RetinaNet model provided a better trade-off between accuracy and computational efficiency, which enabled these one-stage models to be applied in a relatively wide range of application areas.

Similarly, the single-stage CornerNet model [14] also demonstrated good performance compared with the existing two-stage models by detecting a bounding box of objects with a pair of key points without utilizing the anchor-box concept introduced in the Fast R-CNN [15] model. Moreover, the single-stage RefineDet model [16] improved class imbalance by filtering negative anchors with an Anchor Refinement Module (ARM), enabling an Object Detection Module (ODM) to generate more accurate predictions. It is noteworthy that all these single-stage methodologies are actively used in fields where real-time detection is essential, such as autonomous driving [2,17–19]. However, when static objects, such as buildings and greenhouses, need to be detected, applying single-stage models is inappropriate because such applications require higher accuracy than computational speed.

Of note, two-stage object detection models have the disadvantage of a slower computational speed; however, they have the advantage of providing a relatively higher computational accuracy than single-stage models. For example, the two-stage model R-CNN [20] demonstrated higher computational accuracy than conventional single-stage models by combining a Region Proposal Network (RPN) and a CNN. Similarly, the two-stage SPP-Net model solved the problem of distortion and slower computational speed by adding a spatial pyramid pooling layer to solve the image information loss caused by crop and warp operations in R-CNN. Moreover, the two-stage Fast R-CNN model [15] introduced regional feature extraction ideas to solve the problem of multi-stage pipelines, which was a bottleneck of existing R-CNN and CNN methods. The Faster R-CNN model [21] showed an unprecedented speed improvement by introducing RPN to the Fast R-CNN model. Since then, numerous studies based on faster R-CNN have been conducted to solve various problems in the models to address the current limitations in their computational speed. The R-FCN model [22] was proposed along these lines, which effectively solved the translation invariance dilemma through a position-sensitivity score and pooling by encoding the location information of objects by class. In contrast, recently, the Mask R-CNN model [23] replaced Region of Interest pooling (RoI pooling) with Region of Interest Alignment (RoI-Align) to estimate the exact location of the object on the Faster R-CNN for instance segmentation, and maintained a better trade-off between computational speed and accuracy.

To simultaneously improve computational speed and accuracy, researchers have recently proposed multi-stage algorithms, such as Cascade object detection methods,

in which the algorithms mentioned above are modeled in multi-stage settings. For example, the MR-CNN model [24] was proposed as an iterative region adaptation module to create a more accurate bounding box in a multi-stage structure. Similarly, the Attention-net model [25] was proposed as an alternative to sequential localization of bounding boxes of detected objects, which repeatedly used the same regression to locate them accurately. Moreover, Li et al. [26] proposed moving the face area of the ground truth to the closest location by adjusting the face candidate area through a CNN of the Cascade structure for accurate face detection.

However, in the case of existing object detection methods, useful information in the bounding box with an Intersections over Union (IoU) value of less than 0.5 was not utilized because the IoU was generally set to a value of 0.5 or higher than 0.5. The Cascade R-CNN identified this problem. Object detection models with different IoU criteria were merged into Cascade structures to utilize the information from the low IoU bounding box, which verified performance improvements for Cascade structures with different IoU criteria [27]. In contrast, the HTC model [28] modified the object detection and object segmentation layers connected in parallel to the Cascade R-CNN into a cross-array structure. This approach improved the information flow connecting the mask features extracted during each stage. Similarly, ref. [29] demonstrated the high performance of their proposed multi-stage models by combining bi-directional pyramid networks (BPN) and three IoU learning steps with the SSD structure.

Notably, all of these studies used the Cascade structure to achieve high-quality detection. Therefore, considering the potential of the Cascade-based structure, in this study, we aimed to detect houses and greenhouses in various environmental settings prevalent in aerial images by investigating the proposed optimized Cascade Mask R-CNN model. In this study, hyperparameter tuning and transfer learning were performed to optimize the Cascade Mask R-CNN model to detect and optimize greenhouses using 2054 aerial images. The results obtained from the Cascade Mask R-CNN were further compared with the baseline Mask R-CNN and Mask R-CNN to scrutinize the accuracy of our optimized Cascade Mask R-CNN model.

In addition, the recent advancements in sensor and data processing technologies result in better depth maps in conjunction with RGB information to facilitate improved object detection [30–32]. However, object detection using RGB bands and depth information may not be appropriate under all environmental conditions, specifically for images taken at a time or place with little or no light at all. On the other hand, thermal images are less sensitive to light, which means they can be contaminated with haze and cloudy atmospheric conditions. Therefore, few researchers have reported that the thermal images may complement the RGB images for higher-accuracy object detection, which needs to be further investigated [33,34]. However, considering the prevailing environmental conditions of our study area, this study utilizes RGB images from aerial photographs to detect greenhouses accurately. Thus, this study focuses on deep learning approaches such as the Cascade Mask R-CNN and the Mask R-CNN.

3. Materials and Methods

3.1. Area of Study and Data

This study utilized 2054 aerial images of restricted development zones in the Republic of Korea. The Ministry of Land, Infrastructure, and Transport (MOLIT), which is a government agency of the Republic of Korea, provides data on restricted development zones designated throughout the country in Shapefile (.SHP) format [35]. We can construct 2054 aerial images based on the retrieved dataset, including restricted development zones.

The dataset comprised a set of aerial photographs obtained from three big cities (e.g., Hanam, Hwaseong, and Gunpo) of the Republic of Korea, including two types of object information, such as greenhouse facilities and residences. The cameras for capturing all three cities have the same resolution of 25 cm. All aerial images are 410×410 pixels in size, including various environments, such as forests, urban areas, roads, and agricultural

landscapes. In the Republic of Korea, restricted development zones are designated to preserve the green environment necessary for the health of citizens, so the environment is not very different in different cities. Therefore, there was no significant difference between the three cities. Accordingly, out of 2054 images containing data from three cities, 1643 randomly selected aerial images (80%) were used to train and validate the proposed model, and 411 images were used for model evaluation. The ground-truth values required for model training and evaluation were created using the VGG Image Annotator (VIA) tool [36] to create a mask of houses and greenhouses appearing in aerial images (Figure 1).

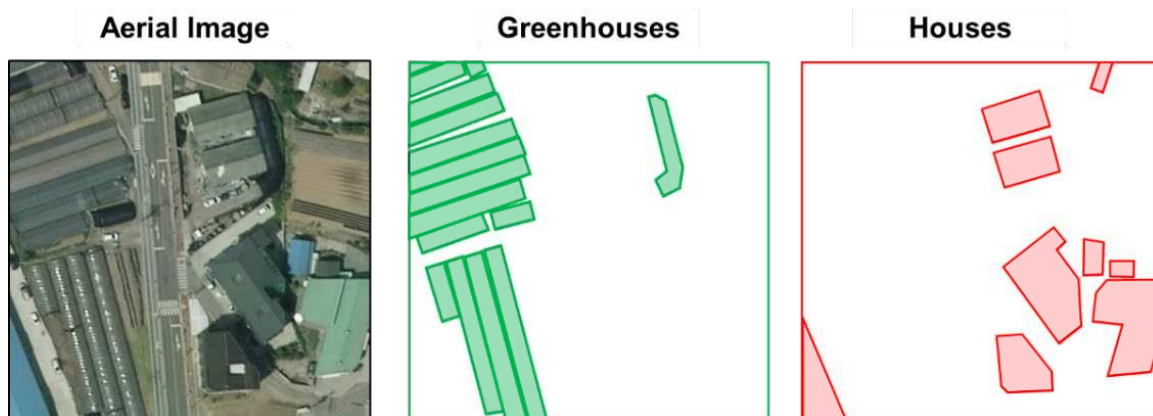


Figure 1. Object information extracted from an aerial photograph, including greenhouse structure and residence annotations.

This study used Amazon Web Service (AWS) to implement our proposed model, which provided high-performance computations. The overall system architecture of our study is presented in Figure 2. First, aerial image data were uploaded to an AWS S3 bucket, and then the data were preprocessed using the Python programming language. The Cascade Mask R-CNN used in this study was implemented using TensorFlow and the Keras framework. Model learning and performance evaluation were conducted in environments of p2.xlarge (61 GB memory and a graphics processing unit (GPU) with 11,441 MB) and CUDA version 10.0.

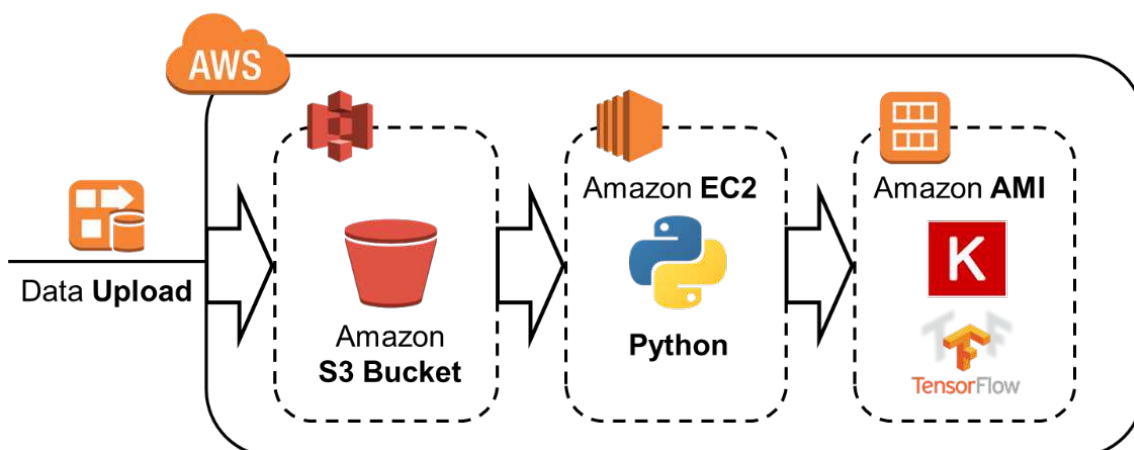


Figure 2. The system architecture implemented in the current study.

3.2. Mask R-CNN

The structure of Mask R-CNN models includes mask branches and RoI-Align for “instance segmentation”, and a Feature Pyramid Network (FPN) is added to reduce operations in Faster R-CNN. A mask branch is a form in which a small Full Convolutional Network (FCN) is added to each RoI. To perform segmentation tasks more effectively, an RoI-Align layer that preserves spatial information of objects was added to the Mask R-CNN.

The process of the Mask R-CNN is as follows. First, the Mask R-CNN adjusts the image through bilinear interpolation, and then adjusts the image size to 1024×1024 with zero padding to fit the input size of the convolutional backbone. Next, feature maps (C1, C2, C3, C4, C5) are created at each layer through ResNet-50 or ResNet-101, and feature maps P2, P3, P4, P5, and P6 are created at the FPN. Each of the final generated feature maps are classified by applying an RPN, and the bounding box regression output value is derived. The bounding box regression value obtained through this method is projected onto the original image to generate an anchor box. Then, all the anchor boxes, except those with the highest score, are removed through Non-Maximum Suppression (NMS) anchor boxes of different sizes, and are aligned with RoI-Align. Finally, the anchor box is passed through the head of the Mask R-CNN structure. Of note, the Mask R-CNN performs localization and instance segmentation of objects through this process.

3.3. Cascade Mask R-CNN

Single-classifiers involve two main problems in general object detection methods. First, if IoU is too high, the number of positive samples decreases, resulting in overfitting. Second, the IoU differs between “training” and “inference” procedures. To address this limitation, Cai and Vasconcelos [6] proposed a Mask R-CNN-based Cascade structure. The model comprises 4-stage structures. First, based on the bounding box generated through the RPN, the region proposal is generated using the lowest detector with an IoU threshold of 0.5. The generated region proposal is used to train a detector with an IoU of 0.6. A detector with a value of 0.7 IoU is trained from the corresponding detector to derive a final output value. Here, the cascade structure is not only applied to the training stage, but also used for inference. Therefore, the Cascade Mask R-CNN model implemented in the current study used the Cascade structure proposed by Cai and Vasconcelos [6]. We apply transfer learning and hyperparameter tuning to the aforementioned Mask R-CNN and Cascade Mask R-CNN to verify which model and parameter combinations can find the greenhouse correctly.

4. Results and Discussion

This section is divided into the following three sub-sections to investigate the performance of our proposed model.

1. We explain the results obtained from our experimental design, in which the transfer learning, hyperparameter tuning, and the values of the final hyperparameters are explained.
2. We scrutinize the performance of our proposed model, and compare it with two existing models (baseline Mask R-CNN and Mask R-CNN).
3. We visualize the object-detection results obtained from the Cascade Mask R-CNN.

To optimally detect objects in aerial images, model optimization and tuning are paramount, specifically the estimation of hyperparameters that are useful for obtaining the most out of our datasets. Various combinations of hyperparameters are possible in the proposed model. However, to achieve the best accuracy, it is crucial to search for the optimal parameters for the input datasets.

Their performance was then investigated for each combination of hyperparameters against each experiment based on a statistical indicator value, such as Average Precision (AP) and mean Average Precision (mAP) [19]. AP is an indicator of precision and recall, which is derived by calculating the area below the precision–recall curve. After calculating the AP for each image, the value obtained by averaging all values corresponds to the mAP.

Table 1 shows the experimental values for each of the four parameters that significantly affect the model performance among the parameters of the Cascade Mask R-CNN and Mask R-CNN. In the case of transfer learning, the layer to be learned is determined according to the ResNet backbone stage based on a model pre-trained with the Microsoft Common Object in the Context (MSCO) dataset. Notably, based on our results, transfer learning in the Cascade Mask R-CNN showed high performance when set to 3+, whereas

all other parameters, such as learning rate, were set to 0.0003, the backbone to ResNet-101, and image sizes to 768×768 and $1,024 \times 1,024$, respectively. The Mask R-CNN showed high performance when transfer learning was set to 3+, learning rate to 0.001, the backbone to ResNet-101, and image size to 768×768 and $1,024 \times 1,024$, respectively. These results are presented in Table 1.

Table 1. Results of the experiment by major hyperparameters.

Hyperparameter	Input	Cascade Mask R-CNN mAP	Mask R-CNN mAP
Transfer Learning	Heads	49.29	48.28
	5+	77.11	55.08
	4+	78.95	74.24
	3+	79.79	74.52
	all	79.96	72.64
Learning Rate	0.005	67.00	59.63
	0.001	75.47	66.72
	0.0001	76.23	63.18
	0.0003	78.86	61.13
Backbone	Resnet-50	77.32	67.94
	Resnet-101	78.95	70.57
Image Size	256×256	60.62	48.68
	512×512	65.63	67.59
	768×768	76.38	70.45
	1024×1024	76.85	70.56

Tables 2 and 3 show the hyperparameter values finally determined for the Cascade Mask R-CNN and Mask R-CNN. Here, 3+ was set as the layer to perform transfer learning, and the image size was set to 768×768 . It was determined that there was no significant difference in performance compared to all sizes and 1024×1024 , which showed the highest performance, as shown in Table 1. For example, the learning Cascade Mask R-CNN at 768×768 took 585 min, whereas learning models at 1024×1024 took 773 min, an increase of about 32.1% in training time. The Mask R-CNN model also took 553 min when learning at 768×768 , but learning at 1024×1024 took 716 min, an increase of about 29.5%. Therefore, we decided to slightly lower the training time by selecting a one-step lower value. In addition, to prevent too many objects from being detected with lower accuracy, the minimum detection confidence was set to 0.9.

Table 2. Values of *Cascade Mask R-CNN* hyperparameters.

Hyperparameters	Input
NUM classes	1 + 2
Step per epoch	200
Detection min confidence	0.9
Backbone	Resnet-101
Image resize mode	square
Image max dim	768
Image channel count	3(RGB)
Max GT instances	70
Train ROIS per image	256
RP anchor scales	(12, 32, 64, 128, 256)
Detection max instances	50
Learning rate	0.0003
Weight decay	0.00003
RPN train anchors per image	320
Layers	3+

Table 3. Values of *Mask R-CNN* hyperparameters.

Hyperparameters	Input
NUM classes	1 + 2
Step per epoch	200
Detection min confidence	0.9
Backbone	Resnet-101
Image resize mode	square
Image max dim	768
Image channel count	3(RGB)
Max GT instances	70
Train ROIS per image	256
RP anchor scales	(12, 32, 64, 128, 256)
Detection max instances	50
Learning rate	0.001
Weight decay	0.00003
RPN train anchors per image	320
Layers	3+

Based on the hyperparameter values, the performances of the three selected models in the current study were evaluated using all available input datasets. The results presented in Table 4 demonstrate that the proposed Cascade Mask R-CNN model performed better than the baseline Mask R-CNN and the Mask R-CNN models. The performance evaluation results show that our proposed Cascade Mask R-CNN model exhibited better performance compared with the two selected models, with an mAP value of 83.60, compared with 70.77 for the baseline Mask R-CNN, and 81.70 for the Mask R-CNN models. These results show that the proposed Cascade Mask R-CNN model demonstrated superior performance, with an mAP value of 1.9, larger than that of the Mask R-CNN, when we considered hyperparameter tuning and transfer learning, and an mAP value of 12.83, larger than that of the baseline Mask R-CNN model.

Table 4. Performance of proposed model compared with two other models (baseline Mask R-CNN, and Mask R-CNN).

Model	mAP	F1-Score
baseline Mask R-CNN	70.77	52.33
Mask R-CNN (hyperparameter tuning and transfer learning)	81.70	59.13
Cascade Mask R-CNN (hyperparameter tuning and transfer learning)	83.60	62.07

In addition, this study utilized an alternative metric (i.e., the F1-score) to prove the superiority of the proposed model. The F1-score summarizes both the precision and recall values as the harmonic mean [37]. The proposed Cascade Mask R-CNN model outperformed (i.e., achieved higher F1-scores, by 2.94) the Mask R-CNN with hyperparameter tuning and transfer learning. In comparison with the baseline Mask R-CNN model, the proposed Cascade Mask R-CNN model also outperformed (i.e., a 9.74 higher value).

Although the calculation time of the Mask R-CNN is better than the Cascade Mask R-CNN, the Cascade Mask R-CNN for greenhouse detection performed better than the Mask R-CNN in terms of the accuracy.

In this subsection, we demonstrate the prediction accuracies of the proposed Cascade Mask R-CNN model in predicting objects using two experiments, and present the results in Figure 3. For the first experiment using the input dataset (Figure 3a), the results presented in Figure 3b demonstrate that two small greenhouses located inside the forests were accurately predicted, which shows that the model exhibited good performance. To assess the performance of our proposed model in a complex landscape, we conducted the second

experiment in a more complex environmental setting in which a large number of mixed objects were present in the input imagery (Figure 3c). Figure 3c shows a total of 35 objects, including a house and a greenhouse house, in a single image. It is noteworthy that the proposed Cascade Mask R-CNN model accurately detected ~94% of the ground-truth objects, as presented in Figure 3d, in which 33 out of 35 objects were successfully detected; two objects could not be detected. The better performance of our proposed Cascade Mask R-CNN model is attributed to the adopted localization scheme, instrument segmentation, and the optimization of the model for the selected study area.

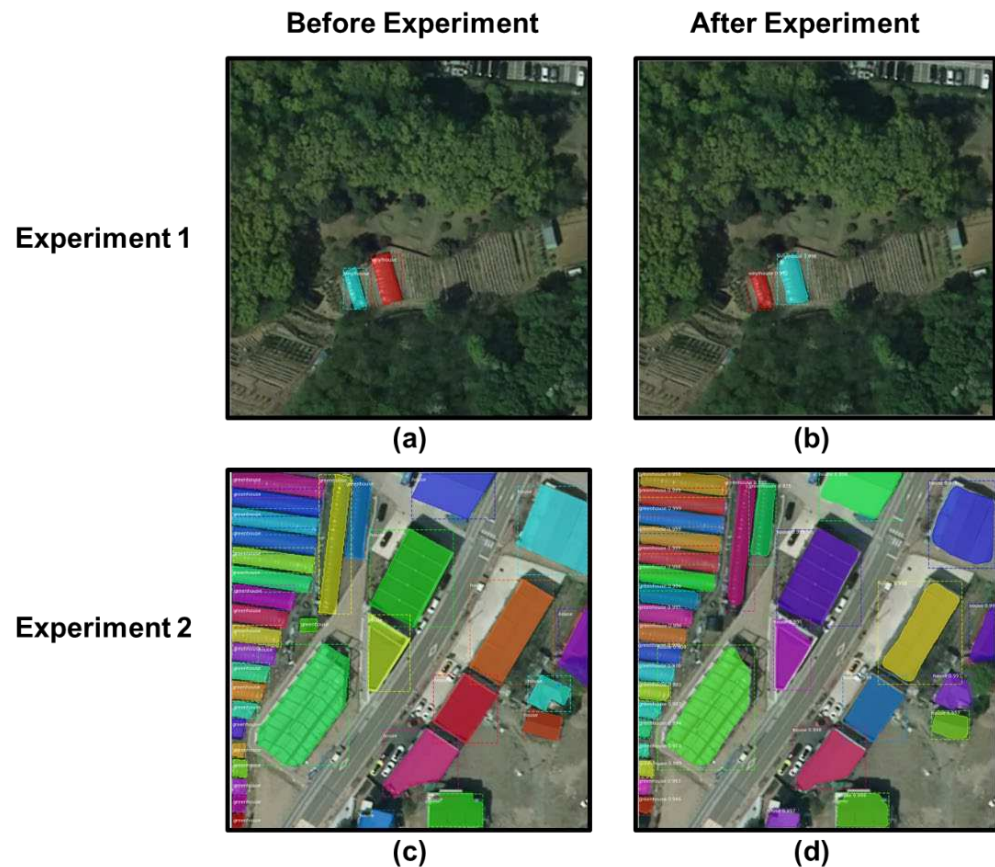


Figure 3. The model results before and after the experiment. (a) Annotations in original image before experiment 1. (b) Mask and bounding box of detected objects after applying the Cascade Mask R-CNN model for Experiment 1. (c) Annotations in the original image before experiment 2. (d) Mask and bounding box of detected objects after applying the Cascade Mask R-CNN model for Experiment 2.

The visual comparison of the object detection for the proposed Cascade Mask R-CNN and Mask R-CNN is shown in Figure 4. Figure 4a presents the input dataset with annotations. In Figure 4b, the Mask R-CNN with hyperparameter tuning and transfer learning shows that it could not detect one small greenhouse in the center of the forest. However, the proposed cascade Mask R-CNN accurately detected the greenhouses, and it is presented in Figure 4c.

In this subsection, we detail the limitations of the proposed model. Our results revealed that the proposed Cascade Mask R-CNN model optimized in the current study detected objects of the same class poorly when they were placed at some inclined angle. This implies that the performance of the optimized Cascade Mask R-CNN model deteriorated slightly based on the complex angle information of the objects. For example, the results presented in Figure 5 demonstrate that when greenhouses were closely located in a complex setting with inclined objects, all objects were successfully detected, except for four greenhouses placed at an angle. Hence, if the angle is inclined in the case of objects with a large aspect

ratio, the overlapping area with the bounding box of the surrounding objects is relatively increased. This results in a foreground instance class ambiguity problem [38,39], which was responsible for the deterioration of the performance of our proposed model. Therefore, in future studies, a robust experimental design should be developed by incorporating a rotated bounding box during the training and model optimization process, which might increase the prediction accuracies of the models under complex environmental settings.

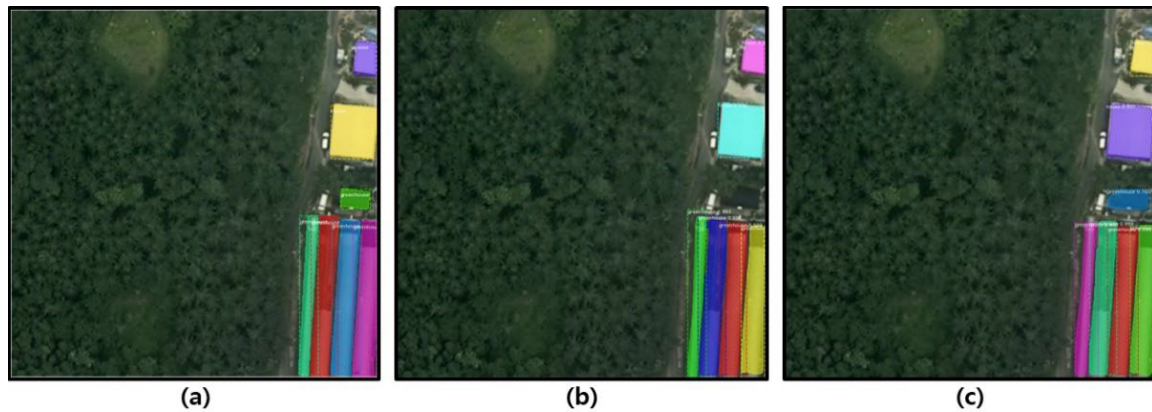


Figure 4. The model results before and after the experiment. (a) Annotations in the original image before the experiment. (b) Mask and bounding box of detected objects after applying the mask R-CNN model with hyperparameter tuning and transfer learning. (c) Mask and bounding box of detected objects after applying the Cascade Mask R-CNN model.

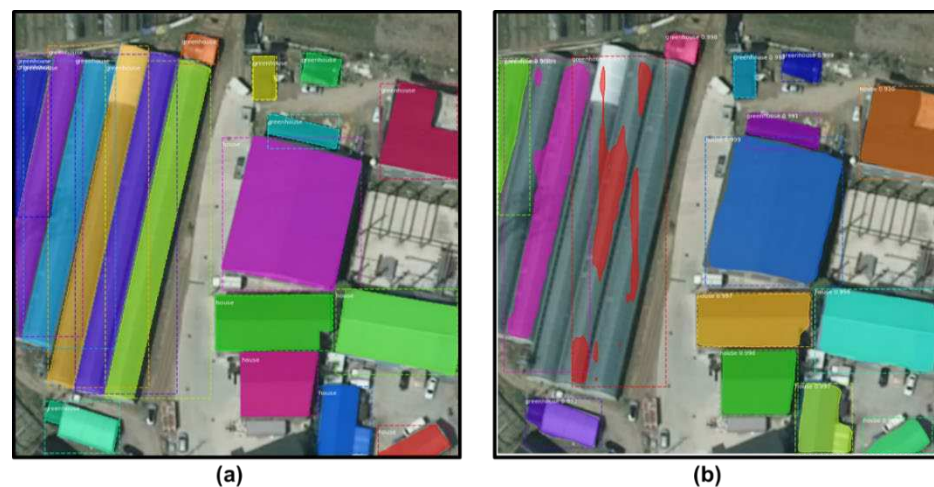


Figure 5. The model results before and after the experiment. (a) Annotations in the original image before the experiment. (b) Mask and bounding box of detected objects after applying the Cascade Mask R-CNN model.

5. Conclusions and Future Prospects

For the protection, restoration, and advancement of the sustainable use of land and ecosystems, the efficient management of restricted development zones is essential, and is among the Sustainable Development Goals established by the United Nations. This study has provided an optimized and improved deep-learning-based Cascade Mask R-CNN model designed to detect greenhouse facilities automatically. For this purpose, the present work has investigated 2054 aerial images of three large cities (Hanam, Hwaseong, and Gunpo) in the Republic of Korea. Our Cascade Mask R-CNN model was optimized by integrating transfer learning and hyperparameter tuning based on a set of randomly selected aerial images comprising 80% of the original dataset.

The results obtained from our proposed model have been evaluated compared with those of two existing models (baseline Mask R-CNN and Mask R-CNN) using statistical indicators such as mean Average Precision (mAP). The results show that our proposed model obtained an mAP value of 83.60, thus outperforming the baseline Mask R-CNN model, which had a lower mAP value of 70.77, and the Mask R-CNN, with an mAP value of 81.70. In terms of F1-score, our proposed model obtains a value of 62.07, outperforming the baseline Mask R-CNN model, with a lower F1-score value of 52.33, and the Mask R-CNN model, with an F1-score value of 59.13.

Although the operation time of the Mask R-CNN is slightly better than the Cascade Mask R-CNN, the accuracy of the Cascade Mask R-CNN is better than the Mask R-CNN, which we have reported in the current study. This study aims to detect illegal encroachments, such as greenhouses, in the restricted development zones. Therefore, it is crucial to detect greenhouses with better prediction accuracy through routine screening procedures regardless of the increased operation time.

The superior performance of our proposed Cascade Mask R-CNN model is attributed to the localization scheme adopted, as well as to instrument segmentation and the optimization of the model for the selected study area. We noticed a slight deterioration in the performance of our proposed model, which is attributed to the instance class ambiguity problem, which is related to the horizontally-oriented bounding box structure used during the training stage. Future works should focus on incorporating rotated bounding boxes into the structure of the object detection pipeline to address this limitation and further improve the performance of the proposed Cascade Mask R-CNN model.

Further, recently, new approaches for object detection have been proposed. For example, training a model by separating a training sample by considering the different sensitivities of classification and regression [40] or segmenting instances and inferring relative importance rankings [41] have been proposed. Further investigation should focus on the design of a robust algorithm for detecting greenhouses accurately.

The optimized Cascade Mask R-CNN model proposed in this study is expected to facilitate the automatic and accurate detection of greenhouses and similar structures, as well as the development of plans to control illegal encroachments, which will provide policymakers with a baseline for the sustainable development of restricted zones in the Republic of Korea, as well as similar regions across the globe.

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

Funding: This paper was supported by “Ministry of the Interior and Safety” R&D program (20017423).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy and copyright.

Conflicts of Interest: The authors declare no conflict of interest regarding the publication of this manuscript.

References

1. Park, J.H. A Study on Policy Changes the Green Belt by Analyzing of Official Gazette. *Geogr. J. Korea* **2021**, *55*, 57–72.
2. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
3. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.

4. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
5. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
6. Cai, Z.; Vasconcelos, N. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1483–1498. [[CrossRef](#)] [[PubMed](#)]
7. Krišto, M.; Ivacic-Kos, M.; Pobar, M. Thermal object detection in difficult weather conditions using YOLO. *IEEE Access* **2020**, *8*, 125459–125476. [[CrossRef](#)]
8. Devaguptapu, C.; Akolekar, N.; Sharma, M.M.; Balasubramanian, V.N. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 1029–1038.
9. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [[CrossRef](#)]
10. Khan, M.S.; Jeon, S.B.; Jeong, M.-H. Gap-Filling Eddy Covariance Latent Heat Flux: Inter-Comparison of Four Machine Learning Model Predictions and Uncertainties in Forest Ecosystem. *Remote Sens.* **2021**, *13*, 4976. [[CrossRef](#)]
11. Shang-Liang, C.; Li-Wu, H. Using Deep Learning Technology to Realize the Automatic Control Program of Robot Arm Based on Hand Gesture Recognition. *Int. J. Eng. Technol. Innov.* **2021**, *11*, 241.
12. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
13. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
14. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
15. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
16. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4203–4212.
17. Wu, B.; Iandola, F.; Jin, P.H.; Keutzer, K. Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 129–137.
18. Tao, A.; Barker, J.; Sarathy, S. Detectnet: Deep Neural Network for Object Detection in Digits. *Parallel Forall* **2016**. Available online: <https://devblogs.nvidia.com/detectnet-deep-neural-network-object-detection-digits> (accessed on 22 December 2021).
19. Lee, T.-Y.; Jeong, M.-H.; Peter, A. Object Detection of Road Facilities Using YOLOv3 for High-definition Map Updates. *Sens. Mater.* **2022**, *34*, 251–260. [[CrossRef](#)]
20. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
21. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [[CrossRef](#)] [[PubMed](#)]
22. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 379–387.
23. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
24. Gidaris, S.; Komodakis, N. Object detection via a multi-region and semantic segmentation-aware cnn model. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1134–1142.
25. Yoo, D.; Park, S.; Lee, J.-Y.; Paek, A.S.; So Kweon, I. Attentionnet: Aggregating weak directions for accurate object detection. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2659–2667.
26. Li, H.; Lin, Z.; Shen, X.; Brandt, J.; Hua, G. A convolutional neural network cascade for face detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5325–5334.
27. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
28. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W. Hybrid task cascade for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4974–4983.
29. Wu, X.; Sahoo, D.; Zhang, D.; Zhu, J.; Hoi, S.C. Single-shot bidirectional pyramid networks for high-quality object detection. *Neurocomputing* **2020**, *401*, 1–9. [[CrossRef](#)]
30. Zhu, C.; Cai, X.; Huang, K.; Li, T.H.; Li, G. PDNet: Prior-model guided depth-enhanced network for salient object detection. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 199–204.
31. Chen, Z.; Cong, R.; Xu, Q.; Huang, Q. DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection. *IEEE Trans. Image Process.* **2020**, *30*, 7012–7024. [[CrossRef](#)] [[PubMed](#)]

32. Chen, L.; Sun, J.; Xie, Y.; Zhang, S.; Shuai, Q.; Jiang, Q.; Zhang, G.; Bao, H.; Zhou, X. Shape Prior Guided Instance Disparity Estimation for 3D Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [[CrossRef](#)] [[PubMed](#)]
33. Ma, Y.; Sun, D.; Meng, Q.; Ding, Z.; Li, C. Learning multiscale deep features and SVM regressors for adaptive RGB-T saliency detection. In Proceedings of the 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 9–10 December 2017; pp. 389–392.
34. Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; Harada, T. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 5108–5115.
35. MOLIT. Restricted Development Zone Data in Republic of Korea. 2022. Available online: <http://data.nsd.go.kr/dataset/15147> (accessed on 22 December 2021).
36. Dutta, A.; Gupta, A.; Zissermann, A. VGG Image Annotator (VIA). 2016. Available online: <http://www.robots.ox.ac.uk/~{v}gg/software/via> (accessed on 22 December 2021).
37. Jeong, M.H.; Sullivan, C.J.; Gao, Y.; Wang, S. Robust abnormality detection methods for spatial search of radioactive materials. *Trans. GIS* **2019**, *23*, 860–877. [[CrossRef](#)]
38. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2cnn: Rotational region cnn for orientation robust scene text detection. *arXiv* **2017**, arXiv:1706.09579.
39. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sens.* **2018**, *10*, 132. [[CrossRef](#)]
40. Wang, D.; Shang, K.; Wu, H.; Wang, C. Decoupled R-CNN: Sensitivity-Specific Detector for Higher Accurate Localization. *IEEE Trans. Circuits Syst. Video Technol.* **2022**. [[CrossRef](#)]
41. Liu, N.; Li, L.; Zhao, W.; Han, J.; Shao, L. Instance-Level Relative Saliency Ranking with Graph Reasoning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [[CrossRef](#)] [[PubMed](#)]