TYPE Original Research
PUBLISHED 23 January 2023
DOI 10.3389/fenrg.2022.1083005

OPEN ACCESS

EDITED BY
Zijun Zhang,
City University of Hong Kong, Hong Kong SAR, China

REVIEWED BY
Tianwei Cao,
University of Chinese Academy of Sciences, China
Linfei Yin,
Guangxi University, China
Ruyun Yu,
China Electronics Technology Research Institute of Cyberspace Security CO. LTD., China

*CORRESPONDENCE
Xiong Luo,
xluo@ustb.edu.cn
Manman Yuan,
yuanman@ustb.edu.cn

SPECIALTY SECTION
This article was submitted to Smart Grids, a section of the journal Frontiers in Energy Research

RECEIVED 28 October 2022
ACCEPTED 24 November 2022
PUBLISHED 23 January 2023

CITATION
Xu Y, Luo X, Yuan M, Huang B and Malof JM (2023), Soft-masks guided faster region-based convolutional neural network for domain adaptation in wind turbine detection.
Front. Energy Res. 10:1083005.
doi: 10.3389/fenrg.2022.1083005

Soft-masks guided faster region-based convolutional neural network for domain adaptation in wind turbine detection

Yang Xu[1,2,3], Xiong Luo[1,2,3]*, Manman Yuan[1,2,3]*, Bohao Huang[4] and Jordan M. Malof[5]

[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China, [2]Shunde Innovation School, University of Science and Technology Beijing, Foshan, China, [3]Beijing Key Laboratory of Knowledge Engineering for Materials Science, University of Science and Technology Beijing, Beijing, China, [4]Department of Electrical and Computer Engineering, Duke University, Durham, NC, United States, [5]Department of Computer Science, University of Montana Missoula, Missoula, MT, United States

Wind turbine generator system plays a fundamental role in electricity generation in industry 4.0, and wind turbines are usually distributed separately and in poor locations. Unmanned Aerial Vehicles (UAV) which could overcome the above challenges are deployed to collect photographs of wind turbines, could be used for predictive maintenance of wind turbines and energy management. However, identifying meaningful information from huge amounts of photographs taken by drones is a challenging task due to various scales, different viewpoints, and tedious manual annotation. Besides, deep neural networks (DNN) are dominant in object detection, and training DNN requires large numbers of accurately labeled training data, and manual data annotation is tedious, inefficient, and error-prone. Considering these issues, we generate a synthetic UAV-taken dataset of wind turbines, which provides RGB images, target bounding boxes, and precise pixel annotations as well. But directly transferring the model trained on the synthetic dataset to the real dataset may lead to poor performance due to domain shifts (or domain gaps). The predominant approaches to alleviate the domain discrepancy are adversarial feature learning strategies, which focus on feature alignment for style (e.g., color, texture, illumination, *etc.*) gaps without considering the content (e.g., densities, backgrounds, and layout scenes) gaps. In this study, we scrutinize the real UAV-taken imagery of wind turbines and develop a synthetic generation method trying to simulate the real ones from the aspects of style and content. Besides, we propose a novel soft-masks guided faster region-based convolutional neural network (SMG Faster R-CNN) for domain adaptation in wind turbine detection, where the soft masks help to extract highly object-related features and suppress domain-specific features. We evaluate the accuracy of SMG Faster R-CNN on the wind turbine dataset and demonstrate the effectiveness of our approach compared with some prevalent object detection models and some adversarial DA models.

KEYWORDS

soft-masks, faster region-based convolutional neural network (Faster R-CNN), wind turbine detection, synthetic data, domain adaptation (DA)

Introduction

Wind turbine generator system generates renewable and affordable energy worldwide in industry 4.0, which relieves the energy shortage situation across the world. The wind turbines are the fundamental infrastructure of wind farms, and they are widely distributed in space and may be located in remote mountains or rough sea regions which are geographically challenged for human surveillance, monitoring, and maintenance (Ciang et al., 2008). Investigations (Motlagh et al., 2021) have shown that working in a harsh outdoor environment for a long time, the surface of wind turbines will inevitably be damaged, and giant turbine damages will result in catastrophes, so patrol inspection is highly necessary. Besides, it's important for the government management in power grid integration, and counting the number of wind turbines can be conducive for surveillance.

The development of Unmanned Aerial Vehicles (UAV) has presented to be exponential growth (Motlagh et al., 2021). With the exploration of UAV, abundant photographs from different viewpoints even in a harsh environment can be obtained. The detection of the wind turbines from the UAV-taken images is requested for patrol inspection and surveillance.

Object detection plays a fundamental role in computer vision, especially for industrial applications, and deep neural network (DNN) has been one of the research hotspots due to its excellent learning ability which obtains outstanding detection performance. Training DNN requires plenty of accurately labeled imagery, and manually annotating the drone-taken images is tedious, time-cost, and error-prone.

To address the data annotation problem, the synthetic method (Xu et al., 2022) is exploited to synthesize pseudo images by implanting 3 dimensional (3D) object models into real background imagery. It's cheap and easy to generate a huge amount of synthetic images, and the significant advantage of synthetic imagery is that RGB images, object bounding boxes, and accurate pixel-level annotations could be obtained simultaneously without manual labor. With the inspiration of SIMPL (Xu et al., 2022), we examine the characteristics (e.g., size, color, orientation, illumination, densities, *etc.*) of real wind turbines in the drone-taken imagery, and expand the SIMPL approach to generate synthetic UAV-taken imagery.

With large amounts of synthetic UAV-acquired wind turbine imagery, DNN models could be trained to get promising performance. Faster region-based convolutional neural network (Faster R-CNN) (Ren et al., 2016) is a predominant approach for object detection. However, directly transferring the model trained on the synthetic dataset to the real one will lead to significant performance drops due to domain shifts or dataset bias (Zhao et al., 2022). Domain adaptation (DA), one kind of transfer learning (Wilson and Cook, 2020), is explored to learn a model from the source domain (labeled datasets) and generalize well to the target domain (unlabeled datasets with different distributions). The DA has been extensively explored for classification (image-level prediction) (Saito et al., 2018; Lee et al., 2019), and semantic segmentation (pixel-level prediction) (Sankaranarayanan et al., 2018; Tsai et al., 2018). Object detection, which involves bounding-box location and category prediction, faces more challenges compared with classification and semantic segmentation in DA.

Most DA detection models aim to measure the feature distribution distance of different domains and then minimize the discrepancy, therefore adversarial manner is exploited between the feature extractor and domain discriminator (Chen et al., 2018; Saito et al., 2019; Li et al., 2020; Chen et al., 2021). Adversarial feature learning aims to decrease style gaps (e.g., color, texture, illumination) between domains for improving the generalization ability, however, for the challenges of content gaps such as various locations, densities, and distributions, the adversarial feature learning may lead to feature misalignment, which decreases the discriminability of the detector (Jiang et al., 2022; Yu et al., 2022).

With scrutiny of characteristics of real UAV-taken wind turbine images, we develop SIMPL to synthesize the images with carefully designed wind turbines, backgrounds, distributions, illuminations, and so on, which could decrease the style gaps and content gaps to some extent. Besides, We propose a novel SMG Faster R-CNN without traditional domain adaptation components, where soft masks could help the detector focus on highly object-related areas (both discriminative and domain-invariant). What's more, large-scale variation (various scales of the same class objects) of targets impairs the performance of detection models, and the SMG Faster R-CNN could alleviate this problem by integrating soft masks with large-scale feature maps to enhance accuracy for small targets.

This study addresses these issues: 1) Difficulties in training data annotation: synthetic dataset in this study is generated with both RGB images and corresponding annotations.

2) Performance drop in domain adaptation: with soft masks, the feature extractor could suppress some domain-specific information and obtain discriminative regions highly related to the target objects, which plays a positive role in detection. 3)Large-scale variation of targets: soft masks are resized to combine with different scale feature maps, which contributes to the detection performance on multiscale objects. The contributions of this article are as follows.

1) Synthetic UAV-acquired dataset is generated to simulate authentic UAV-taken wind turbine imagery, which makes the acquisition of training data cheap and effective. And carefully designed scenarios and 3D wind turbine models could make realistic rendering characteristics of the synthetic imagery.
2) Soft masks are designed to guide Faster R-CNN for DA detection, and soft masks ensure that more weights are assigned to foreground targets and lower weights to the background, tending to extract more discriminative features and suppress the domain-specific information in DA detection.
3) The combination of soft masks and different scale feature maps positively impact the detection of targets with large-scale variations.
4) Our proposed detection model outweighs not only the representative object detection models but also some popular adversarial DA models in accuracy.

The rest of the article is organized as follows. A series of related works are briefly described in **Section 2**. The proposed method is detailed in **Section 3**. Experiments, performance comparisons, and analysis are shown in **Section 4**. **Section 5** is the conclusion.

Related work

This section explains why wind turbine detection is necessary and describes existing object detection approaches, synthetic data for object detection, DA in object detection as well as mask contributions for object detection.

Necessary of wind turbine detection

There are various situations such as acid rain, turbulent wind, and sand storm that may lead to surface damages or defects on wind turbines (Du et al., 2020). These damage may cause a catastrophe, therefore it's of great significance to detect the defects to wind turbines in time. Machine learning has already been used for the identification of wind turbines as it's cost-effective. UAV-taken images have been used to detect damage to wind turbines (Stokkeland et al., 2015; Wang and Zhang, 2017), where traditional image processing methods (e.g., Hough transform (Dalal and Triggs, 2005) and Haar-like features (Lienhart and Maydt, 2002)) are exploited to obtain meaningful features. In Moreno et al. (2018), a deep learning vision-based approach for defect detections was proposed, and convolution neural network (CNN) was exploited for feature extraction. However, drone-taken images contain not only wind turbines but also various backgrounds, the first step is to locate the wind turbines and then apply defect detection algorithms. In this study, we explore a deep learning model (DNN) to identify wind turbines from UAV-captured imagery.

Deep neural networks for object detection

Recently, object detection has drawn dramatically increasing amounts of attention, due to its outstanding performance on a wide range of academic and commercial applications, such as remote monitoring, security surveillance, autonomous driving, and so on. Besides, the impressive GPU computing ability makes a positive contribution to the great performance of object detection. In general, object detection models broadly fit into two categories: one-stage object detector [e.g., SSD (Liu et al., 2016), RetinaNet (Lin et al., 2020)], and two-stage object detector [e.g., Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2016)]. One-stage detector considers every grid of the feature map as a potential proposal, while two-stage detectors obtain potential proposals by region proposal network (RPN), then the proposals are used for further predictions. Thus, the two-stage detectors are more accurate than one-stage detectors at the cost of speed. In this study, we prefer a two-stage detector.

As wind turbine imagery is taken at eye level, wind turbines vary in scale due to different distances to the camera. The closer the distance, the larger the wind turbines (shown in **Figure 1**). Multi-layer detection can effectively handle various scale problems (Luo et al., 2020). Similarly, the feature Pyramid Network (FPN) (Lin et al., 2017), built upon feature pyramids, has shown an impressive ability dealing objects with different scales. With the addition of literal connections, FPN-based Faster R-CNN could make full use of both lower-level and higher-level features and obtain outstanding detection performance. Our work is based on the FPN-based Faster R-CNN.

Synthetic data for object detection

Synthetic data has been increasingly exploited in computer vision for various problems (Hattori et al., 2015; Peng et al., 2015; Ros et al., 2016; Kong et al., 2020; Liu et al., 2022; Xu et al., 2022). Specifically, in Hattori et al. (2015), a synthetic pedestrian dataset was designed by using geometric scenes and a customizable database of virtual pedestrian motion

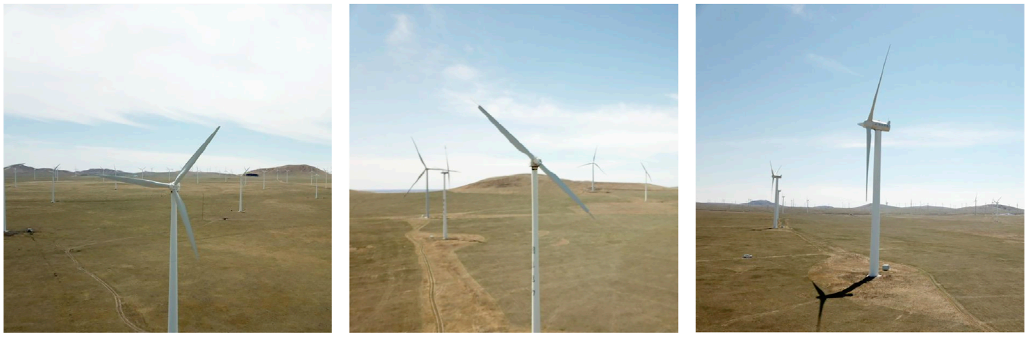


FIGURE 1
Examples of real wind turbine images.

simulations for scene-specific pedestrian detection. In the study by Peng et al. (2015), a synthetic 2D dataset was generated by exploiting publicly available synthetic 3D computer-aided design (CAD) models, textures, and category-related scene images, to train object detectors. In Liu et al. (2022) and Xu et al. (2022), synthetic aircraft datasets were generated by implanting 3D airplane models onto real background images for remote sensing detection. In Ros et al. (2016), a large synthetic dataset was created for segmentation in urban scenes. In Kong et al. (2020) synthetic overhead imagery was exploited for building segmentation.

Inspired by the work of Xu et al. (2022), we design specific-geometric scenes similar to real ones, and randomly place 3D wind turbine models in the scene, then set the virtual camera at a certain height to take photos, both RGB images and pixel-wise ground truth annotations will be obtained easily.

Domain adaptation in object detection

Domain adaptation (DA) is one type of transfer learning, and it learns a model from a well-annotated source domain and generalizes well to an unlabeled target domain (Li et al., 2022). The prevalent idea in addressing the DA problem of object detection is based on an adversarial learning manner to align feature distributions across domains, which helps the detector to produce domain-invariant features. For example, the adversarial manner with gradient reversal layers (GRL) (Ganin and Lempitsky, 2015) was exploited for both image-level and instance-level feature alignments (Chen et al., 2018), where image-level alignment included not only object categories but also scene layouts, backgrounds, *etc.* In the study by Saito et al. (2019), they argued that image-level alignment worked well for small domain shifts but may hurt performance for large domain shifts, to address this issue, they proposed Strong-Weak DA (SWDA) model which utilized strong local alignment to match colors or textures across domains and used weak global alignment to match layouts or backgrounds. In Zhu et al. (2019), the authors argued that conventional DA methods focus on bridging the whole-image gaps while neglecting local characteristics of object detection, they, therefore, proposed a region-level alignment framework, which focused on the region proposals pertinent to object detection. In Chen et al. (2021), the authors observed that object scales are crucial challenges for DA object detection, they proposed a scale-aware DA Faster R-CNN (SA-DA-Faster) model to incorporate the object scales into the adversarial learning for better feature alignment. Although the domain-invariant features obtained by feature alignments are favorable to the transferability, they may also impair the discriminability of detectors, which harms the detection performance. Our method leverages soft masks instead of adversarial learning to extract domain-invariant and discriminative features.

Mask contributions for object detection

As mentioned above, objects in UAV-acquired imagery vary in scale, and large-scale objects coexist with small-scale objects (shown in **Figure 1**), which is a challenging problem. To address these issues, a pixel attention-aided detection model was proposed by Yang et al. (2019), where the pixel attention branch was exploited to suppress the noise and highlight the target features, and the pixel-level saliency masks were obtained by fulfilling object-bounding boxes. In Pang et al. (2019), mask-guided attention (MGA) network was integrated into a standard pedestrian detection model, where MGA was used to generate pixel masks supervised by object bounding boxes. In Sharma and Mir (2022), a saliency-guided Faster R-CNN was proposed for camouflaged object detection, where the saliency map was obtained by a convolutional neural network to identify important regions of the input images. In the study by Zhou et al. (2019), they introduced a semantic attention CNN for pedestrian detection, which considered the segmentation results as self-attention cues to identify target regions and suppress backgrounds. In Yang et al. (2019), fulfilling the bounding box to obtain pixel-level masks was inaccurate and unsuitable for large objects. In Pang et al. (2019); Sharma and Mir (2022), the approaches to get pixel masks would introduce additional computing costs and be inaccurate. In Zhou et al. (2019), segmentation branches were required, and the whole architecture was complex.

Inspired by all the above, we explore DA for wind turbine detection with a synthetic UAV-acquired dataset, and a novel SMG Faster R-CNN is developed, where soft masks are exploited to extract regions highly related to target objects, which is consistent with the DA approaches trying to find the domain-invariant features. Besides, we consider the object scales by incorporating the soft masks with multi-scale feature maps to extract discriminative features.

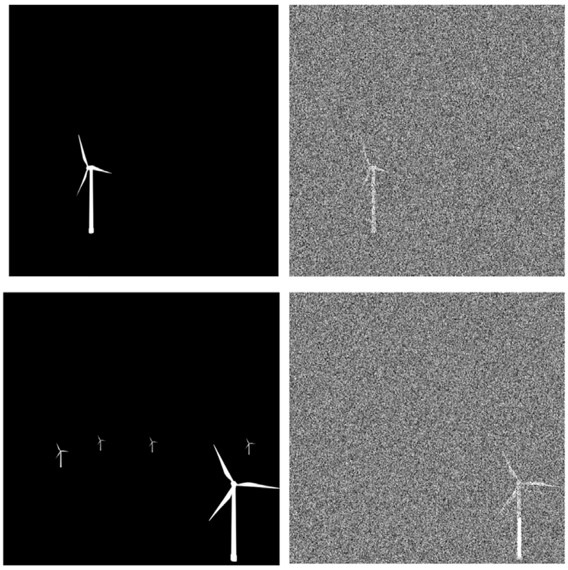


FIGURE 3
Examples of original masks and soft masks. The left column is original masks, and the right column is soft masks.

The proposed method

Generation of synthetic datasets of wind turbines

Although Xu et al. (2022) provided insight for generating synthetic datasets and released an implementation of the generation process, the work was for remote sensing overhead imagery and didn't consider scene terrains. In our study, synthetic images are captured by a visual camera at eye level rather than overhead view. To make the scenarios more realistic, we carefully examine the real UAV-taken imagery and download similar 3D wind turbine models from open-source websites, find

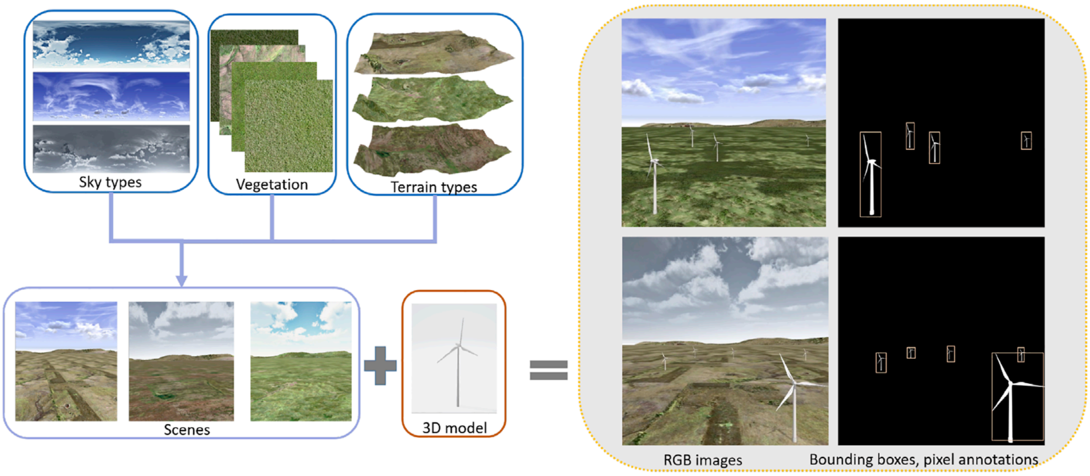


FIGURE 2
The process of generating synthetic wind turbine datasets. The three blue boxes in the upper left are the basic materials to simulate real scenes of wind farms. Randomly place the public available 3D wind turbine models in the synthetic wind farms, and set height and moving step of the virtual camera, synthetic RGB images are captured. For pixel annotations, ignore the environment scene, then repeat the steps for capturing RGB images.

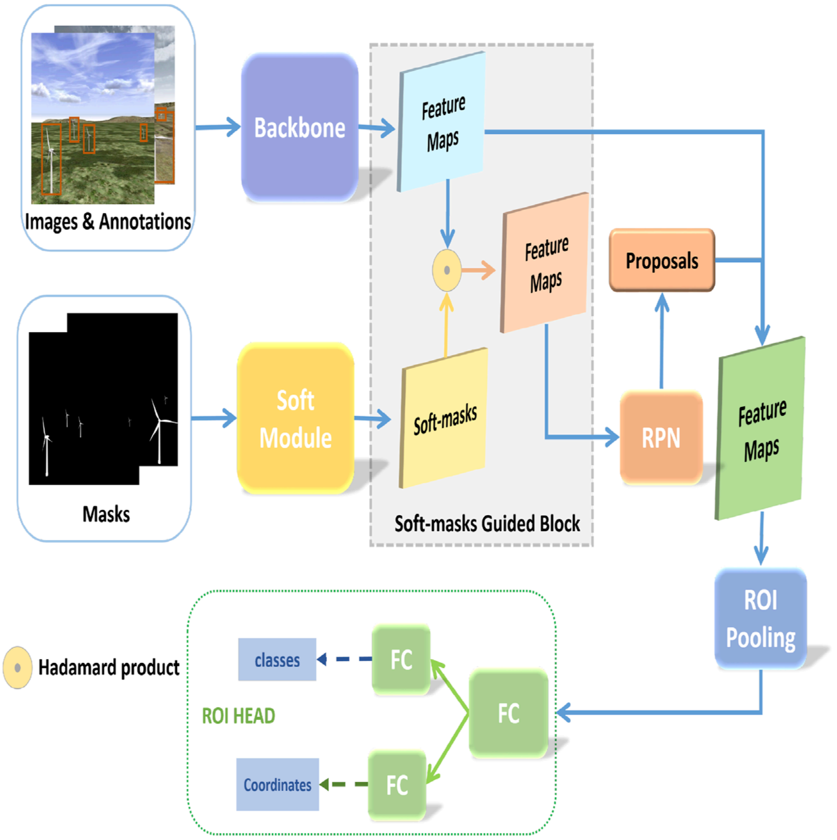


FIGURE 4
The architecture of SMG Faster R-CNN.

similar sky images, and design terrains with similar vegetation. To increase the sample diversity, we designed 3D wind turbine models with various scales, orientations, illumination, dense distributions, *etc.*

Figure 2 presents a pictorial illustration of synthetic dataset generation. Materials of sky types, vegetation, terrain types, and 3D wind turbines are basic components to create virtual reality scenes of wind farms, and all these components could be easily obtained from open-source websites. Randomly distribute 3D wind turbines in the scene, and set the virtual camera height close to that of wind turbines, then move the virtual camera by fixed step to capture images. Remove all colorful backgrounds, and paint wind turbines white, the corresponding annotations are obtained similarly. The bounding boxes are obtained by enclosing the connected white pixels with rectangles.

It needs about 0.8*s* for rendering and capturing one synthetic RGB image (608 × 608) and less for an annotation image on a Windows 10 operating system with an Intel(R) Core(TM) i9-7920X CPU@2.90 GHz.

Soft-masks guided faster region-based convolutional neural network

As the sizes of wind turbines in UAV-taken imagery vary (shown in Figure 1), the large variety of sizes hampers the detection of wind turbines. Faster R-CNN is the most representative model of two-stage detectors, which balances accuracy and speed very well. However, the original Faster R-CNN handles features from one scale, which makes it inferior to multiscale detectors (Lin et al., 2017).

Feature pyramids (Lin et al., 2017) building hierarchical feature maps of multi scales, contain plentiful semantic information. The key components of feature pyramids (bottom-up route, top-down route, and lateral connections) help to capture coarse information of low-level and strong semantic information of high-level at multi scales. To handle the large-size variation problem of wind turbines, we prefer a variant of Faster R-CNN with a backbone of ResNet50 and feature pyramids as the baseline, which is considered the default architecture of Faster R-CNN in this article.

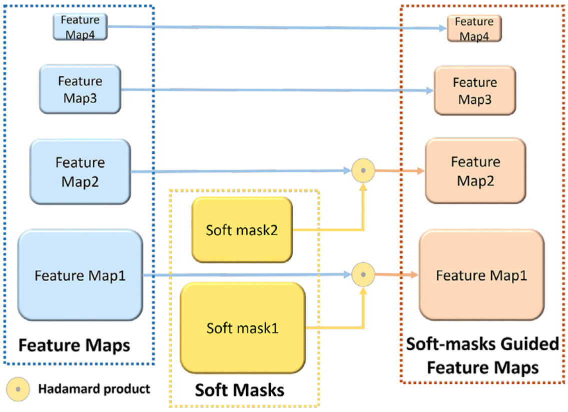


FIGURE 5
Soft-masks guided block.

Figure 2 shows that the synthetic dataset contains not only bounding boxes but also pixel-level masks. The pixel masks are grayscale images that contain only black and white colors, where black represents the background while white represents target objects (shown in the left column of Figure 3). We assign random values from $[0, 255)$ to background pixels, called soft masks, shown in the right column of Figure 3.

To further improve the model detection performance in the target domain, we combine soft masks of the source domain with feature maps at different scale levels to guide the detector to focus on areas that highly are related to target objects and suppress the domain-specific object-less features, which contributes to extracting discriminative regions.

Moreover, as feature pyramids output feature maps at each layer block with different scales, the larger the output scale, the smaller the size of detected objects. To alleviate large-scale variation problems, the soft masks will be multiplied to larger outputs of the last two blocks of feature pyramids (illustrated in Figure 4), which means highlighting the targets while suppressing the background. Note that, it's suppressing rather than ignoring the background.

Figure 4 depicts the whole architecture of our proposed approach. The backbone, composed of ResNet50 and feature pyramids, is used to extract feature maps from multi scales. Input the RGB images and annotations into the backbone (no details here, we refer readers to (He et al., 2016) for the details of ResNet50), and output different scale feature maps. Before feeding to the region proposal network (RPN) module, we input the feature maps into the soft-masks guided block (shown in Figure 5), where soft-masks are multiplied to the two larger-scale feature maps of backbone separately with Hadmard product (Aguiar and Mahajan, 2020), which is profit to focus on discriminative regions. Then the proposals are flattened and concatenated for the region of interest (ROI) pooling, and the results are fed to the ROI head for final prediction (shown in Figure 4). The loss of SMG Faster R-CNN, $\mathcal{L}^{det}$, is the same as that of baseline Faster R-CNN:

$$\mathcal{L}^{det} = \mathcal{L}^{rpn}_{cls} + \mathcal{L}^{rpn}_{reg} + \mathcal{L}^{roi}_{reg} + \mathcal{L}^{roi}_{reg}, \tag{1}$$

where $\mathcal{L}^{rpn}_{cls}$ and $\mathcal{L}^{rpn}_{reg}$ indicate the RPN classification loss and regression loss respectively, $\mathcal{L}^{roi}_{reg}$ and $\mathcal{L}^{roi}_{reg}$ mean the classification loss and regression loss of the ROI head.

Experiments and analysis

Datasets

Real UAV-taken dataset: We use one UAV of DJI MAVIC PRO of Platinum version, and take images in Xilin Gol League, Inner Mongolia, China, named Xilin dataset. We manually labeled 96 images (608×608), which contain about 992 instances of wind turbines.

Note that, the adversarial DA methods in the following experiments, need both unlabeled target-domain data and labeled source-domain data for training, we randomly split the real dataset into a training set (42 images, 512 instances) and a test set (54 images, 480 instances), where the training set is for the adversarial DA methods, and all models in following experiments evaluate on the test set for comparison.

Synthetic dataset: We use CityEngine to simulate realistic scenes and take photographs by a virtual camera, and set camera height close to the height of wind turbines in the virtual scenes, then take photos at eye-level. The total number of synthetic images is 12087, 70% for training, and 30% for validation.

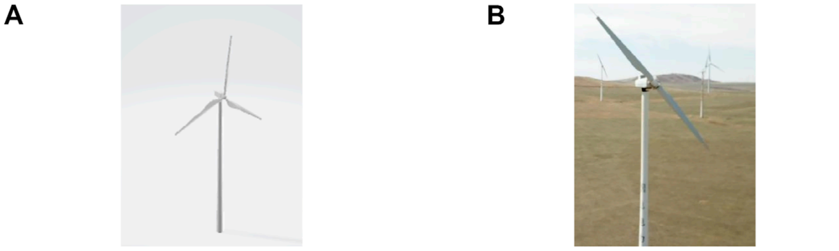
Training environment and settings

For all experiments, we run on the Linux platform with NVIDIA GPU of GeForce RTX 2080 Ti. We compare our proposed approach with some dominant DNN-based object detection models [e.g., SSD (Liu et al., 2016), RetinaNet

TABLE 1 Results comparison on real xilin dataset.

Model	*AP@IoU*0.5 on Xilin
SSD Liu et al. (2016)	0.281
RetinaNet Lin et al. (2020)	0.222
Faster R-CNN Ren et al. (2016)	0.361
DA Faster R-CNN Chen et al. (2018)	0.131
SWDA Saito et al. (2019)	0.126
SA-DA-Faster Chen et al. (2021)	**0.420**
SMG Faster R-CNN	0.409

The bold value means the best performance in terms of *AP@IoU0.5*.



Synthetic 3D wind turbine with 3 blades

Real wind turbine with 2 blades

FIGURE 6
The mismatch of real wind turbine and synthetic 3D model. **(A)** Synthetic 3D wind turbine with 3 blades **(B)** Real wind turbine with 2 blades.

(Lin et al., 2020), Faster R-CNN (Ren et al., 2016)], and some adversarial DA object detection models [e.g., DA Faster R-CNN (Chen et al., 2018), SWDA (Saito et al., 2019) and SA-DA-Faster (Chen et al., 2021)]. All DNN-based models are trained on a synthetic dataset for 20 epochs and the batch size of each model is 8, the optimizer of each model is stochastic gradient descent (SGD) solver. The adversarial DA models are trained on a synthetic dataset and an unlabeled real dataset for 70000 iterations with batch size one. The result is an average of three trials for each model.

Evaluation metric and results analysis

The Intersection over Union (IoU) ratios of model predictions and ground truth bounding boxes determine whether ground truth labels should be assigned to the predictions. The standard metric is average precision (AP) with 50% overlaps, $AP@IoU0.5$, for single-class object detection.

Testing on the real wind turbine dataset, Xilin, the detection results are shown in **Table 1**, and the optimal result is in bold. None of these results is over 0.5, for one reason is that the domain gap between the synthetic dataset and the real dataset, for another reason is the shape mismatch of 3D models and real wind turbines, specifically, most 3D wind turbine models have 3 blades (shown in **Figure 6A**), but the real ones may have two blades (shown in **Figure 6B**). Compared to the traditional DNN-based object detection models (SSD, RetinaNet, and Faster R-CNN), our approach outperforms all of them, which demonstrates the advantage of our method in extracting discriminative features for detection. Further, compared with two predominant adversarial DA models (DA Faster R-CNN, SWDA), our approach wins by a large margin, and the two adversarial DA methods show poor performances, which may be caused by the large-scale variation. SA-DA Faster takes not only image-level and instance-level feature alignment but also object size alignment, which dramatically improves the DA object detection performance. Similarly, our method takes object scale into consideration, although our method is mildly lower than the SA-DA Faster in accuracy, the difference is small, which demonstrates the comparable ability in DA object detection and the advantage in handling large-scale variation detection problems. Besides, the training process of our method doesn't need adversarial components and complex computation, which makes it more efficient.

Figure 7 shows the predicted bounding boxes of Faster R-CNN and SMG Faster R-CNN, and the confidence scores are attached. Most large-scale wind turbines are detected by both of them, and our approach could detect more small wind turbines than Faster R-CNN, which corroborates the ability of SMG Faster R-CNN in handling large-scale variation problems.

Effect of soft values

To probe the effect of soft mask values, we design two schemas to generate soft values. The purpose is to help the object detector focus on the target area and suppress the background. One schema is filling the background with fixed values. The easiest way is to use the original masks, which means the background is black with pixel values of 0, and another way is to set the background area with fixed non-zero values, such as 125. The other schema is to use dynamic values. We sample soft values from $[0, 125)$ and $[0, 255)$ separately. **Table 2** shows the performances of SMG Faster R-CNN with different soft values, the first two rows are the results of fixed value schema, which indicates the larger the value, the better the performance. The last two rows indicate the performance of dynamic value schemas, and the larger the range, the better the results. The soft values control the weights of the corresponding area in feature maps by multiplying soft masks with feature maps. The larger the soft values, the higher weights will be assigned to the corresponding feature area. The results of this experiment demonstrate that background information is also important for target detection.

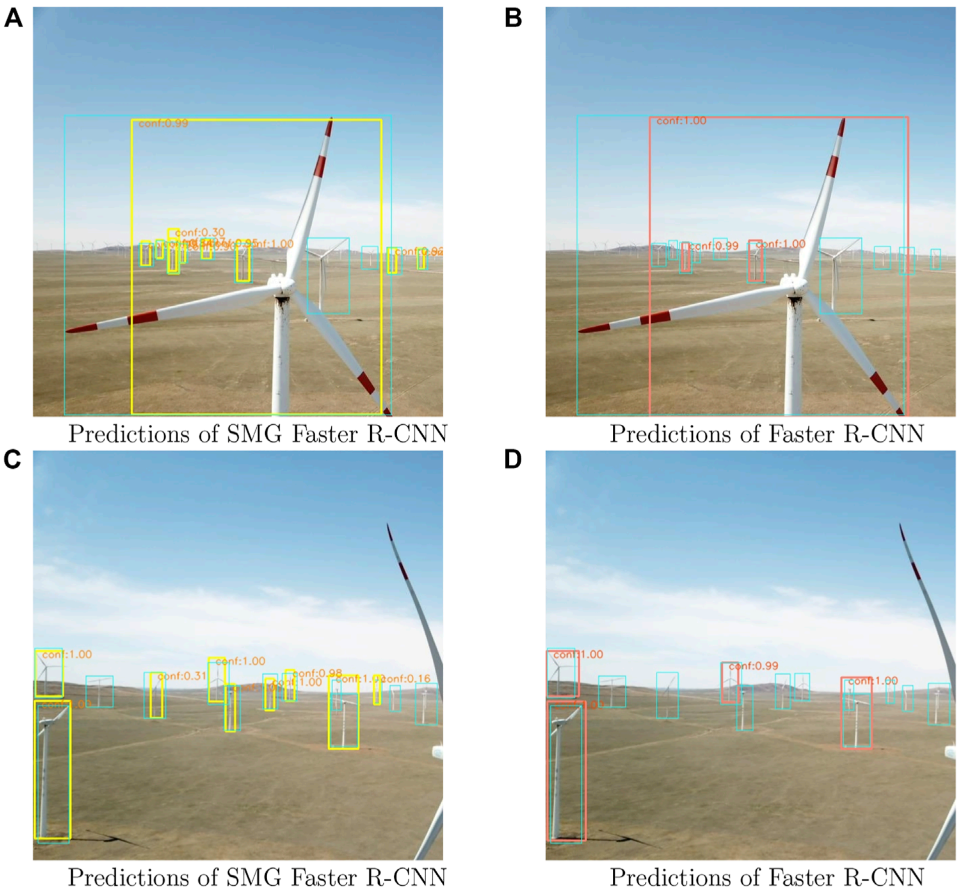


FIGURE 7
Predictions of SMG Faster R-CNN and Faster R-CNN. The cyan boxes are ground truth, yellow boxes mean predictions of SMG Faster R-CNN, and salmon boxes are predictions of Faster R-CNN. The "conf" in orange means the confidence score of the detected instance. **(A)** Predictions of SMG Faster R-CNN **(B)** Predictions of Faster R-CNN **(C)** Predictions of SMG Faster R-CNN **(D)** Predictions of Faster R-CNN.

TABLE 2 Performances of our approach with different soft values.

soft pixel value	$AP@IoU0.5$ on Xilin
0	0.303
125	0.332
$[0,125)$	0.345
$[0,255)$	0.409

Conclusion

With exponentially increasing amounts of data in industry 4.0 applications, deep-learning-based methods have been applied for data analysis. In this article, we proposed SMG Faster R-CNN for DA in wind turbine detection. With a low-cost, easy-to-get, and accurately labeled synthetic dataset, we propose SMG Faster R-CNN for wind turbine detection, which is a fundamental step for further surveillance and management in industrial applications. Besides, to handle the DA detection problem, we make full use of synthetic annotations (pixel masks) to guide Faster R-CNN to extract discriminative regions and suppress the domain-specific features. Furthermore, the masks multiplied to different scale feature maps alleviate the large-scale variation problem of wind turbines in the UAV-acquired imagery. We evaluate the effectiveness of our approach by comparing it with other representative models on the real dataset.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

Funding

This work was supported by the Beijing Natural Science Foundation under Grant L211020, in part by the Beijing Natural Science Foundation under Grant 19L2029, in part by the National Natural Science Foundation of China under Grant U1836106, in part by the Scientific and Technological Innovation Foundation of Foshan under Grants BK20BF010 and BK21BF001, and in part by the National Natural Science Foundation of China under Grants 62106020, and in part by the Postdoctoral Research Foundation of Shunde Graduate School of University of Science and Technology Beijing under Grant 2021BH008.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aguiar, M., and Mahajan, S. (2020). Hadamard product. *Encycl. Math. its Appl.* 2020, 335–383. doi:10.1017/9781108863117.012

Chen, Y., Li, W., Sakaridis, C., Dai, D., and Van Gool, L. (2018). "Domain adaptive faster r-cnn for object detection in the wild," in *Proceedings of the IEEE/CVF international conference on computer vision and pattern recognition* (Salt Lake City, UT: IEEE), 3339–3348. (Accessed June 18–22, 2018). doi:10.1109/CVPR.2018.00352

Chen, Y., Wang, H., Li, W., Sakaridis, C., Dai, D., and Van Gool, L. (2021). Scale-aware domain adaptive faster r-cnn. *Int. J. Comput. Vis.* 129, 2223–2243. doi:10.1007/s11263-021-01447-x

Ciang, C. C., Lee, J. R., and Bang, H. J. (2008). Structural health monitoring for a wind turbine system: A review of damage detection methods. *Meas. Sci. Technol.* 19, 122001. doi:10.1088/0957-0233/19/12/122001

Dalal, N., and Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.* 1, 886–893. doi:10.1109/CVPR.2005.177

Du, Y., Zhou, S., Jing, X., Peng, Y., Wu, H., and Kwok, N. (2020). Damage detection techniques for wind turbine blades: A review. *Mech. Syst. Signal Process.* 141, 106445. doi:10.1016/j.ymssp.2019.106445

Ganin, Y., and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. *Proc. Int. Conf. Mach. Learn.* 37, 1180–1189. doi:10.5555/3045118.3045244

Girshick, R. (2015). "Fast r-cnn," in Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 07-13 December 2015 (IEEE), 1440–1448. doi:10.1109/ICCV.2015.169

Hattori, H., Naresh Boddeti, V., Kitani, K. M., and Kanade, T. (2015). "Learning scene-specific pedestrian detectors without real data," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 07-12 June 2015 (IEEE), 3819–3827. doi:10.1109/CVPR.2015.7299006

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27-30 June 2016 (IEEE), 770–778. doi:10.1109/CVPR.2016.90

Jiang, J., Chen, B., Wang, J., and Long, M. (2022). *Decoupled adaptation for cross-domain object detection. arXiv.* doi:10.48550/arXiv.2110.02578

Kong, F., Huang, B., Bradbury, K., and Malof, J. (2020). "The synthinel-1 dataset: A collection of high resolution synthetic overhead imagery for building segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (Snowmass, CO: IEEE), 1814–1823. (Accessed March 1–5, 2020). doi:10.1109/WACV45572.2020.9093339

Lee, C.-Y., Batra, T., Baig, M. H., and Ulbricht, D. (2019). "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF international conference on computer vision and pattern recognition* (Long Beach, CA: IEEE), 10277–10287. (Accessed June 16–20, 2019). doi:10.1109/CVPR.2019.01053

Li, C., Du, D., Zhang, L., Wen, L., Luo, T., Wu, Y., et al. (2020). "Spatial attention pyramid network for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF European conference on computer vision* (Springer, Cham: IEEE Computer Society). Online, 481–497. (Accessed August 23–28, 2020). doi:10.1007/978-3-030-58601-0_29

Li, Y. J., Dai, X., Ma, C. Y., Liu, Y. C., Chen, K., Wu, B., et al. (2022). "Cross-domain adaptive teacher for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision and pattern recognition* (New Orleans, LA: IEEE), 7571–7580. (Accessed June 21–24, 2022). doi:10.1109/CVPR52688.2022.00743

Lienhart, R., and Maydt, J. (2002). "An extended set of haar-like features for rapid object detection," in Proceedings of the International Conference on Image Processing, Rochester, NY, USA, 22-25 September 2002 (IEEE). doi:10.1109/ICIP.2002.1038171

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). "Feature pyramid networks for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15-20 June 2017 (IEEE), 2117–2125. doi:10.1109/CVPR.2017.106

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2020). Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 318–327. doi:10.1109/TPAMI.2018.2858826

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., et al. (2016). Ssd: Single shot multibox detector. *Proc. IEEE Eur. Conf. Comput. Vis.* 2016, 21–37. doi:10.1007/978-3-319-46448-0__2

Liu, W., Luo, B., and Liu, J. (2022). Synthetic data augmentation using multiscale attention cyclegan for aircraft detection in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5. doi:10.1109/LGRS.2021.3052017

Luo, X., Wong, F., and Hu, H. (2020). Fin: Feature integrated network for object detection. *ACM Trans. Multimed. Comput. Commun. Appl.* 16, 1–18. doi:10.1145/3381086

Moreno, S., Peña, M., Toledo, A., Treviño, R., and Ponce, H. (2018). "A new vision-based method using deep learning for damage inspection in wind turbine blades," in Proceedings of the IEEE International Conference on Electrical Engineering, Computing Science and Automatic Control, Mexico City, Mexico, 05-07 September 2018 (IEEE), 1–5. doi:10.1109/ICEEE.2018.8533924

Motlagh, M. M., Bahar, A., and Bahar, O. (2021). Damage detection in a 3d wind turbine tower by using extensive multilevel 2d wavelet decomposition and heat map, including soil-structure interaction. *Structures* 31, 842–861. doi:10.1016/j.istruc.2021.01.018

Pang, Y., Xie, J., Khan, M. H., Anwer, R. M., Khan, F. S., and Shao, L. (2019). "Mask-guided attention network for occluded pedestrian detection," in *Proceedings of the IEEE/CVF international conference on computer vision* (Seoul, South Korea: IEEE), 4967–4975. (Accessed October 27, 2019–November 3, 2019). doi:10.1109/ICCV.2019.00507

Peng, X., Sun, B., Ali, K., and Saenko, K. (2015). "Learning deep object detectors from 3d models," in *Proceedings of the IEEE international conference on computer vision* (Santiago, Chile: IEEE), 1278–1286. (Accessed December 7–13, 2015). doi:10.1109/ICCV.2015.151

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi:10.1109/TPAMI.2016.2577031

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. (2016). "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27-30 June 2016 (IEEE), 3234–3243. doi:10.1109/CVPR.2016.352

Saito, K., Ushiku, Y., Harada, T., and Saenko, K. (2019). "Strong-weak distribution alignment for adaptive object detection," in *Proceedings of the IEEE/CVF international conference on computer vision and pattern recognition* (Long Beach, CA: IEEE), 6949–6958. (Accessed June 16–20, 2019). doi:10.1109/cvpr.2019.00712

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. (2018). "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF international conference on computer vision and pattern recognition* (Salt Lake City, UT: IEEE), 3723–3732. (Accessed June 18–22, 2018). doi:10.1109/cvpr.2018.00392

Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S. N., and Chellappa, R. (2018). "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision and pattern recognition* (Salt Lake City, UT: IEEE), 3752–3761. (Accessed June 18–22, 2018). doi:10.1109/CVPR.2018.00395

Sharma, V., and Mir, R. N. (2022). Saliency guided faster-rcnn (sgfr-rcnn) model for object detection and recognition. *J. King Saud Univ. - Comput. Inf. Sci.* 34, 1687–1699. doi:10.1016/j.jksuci.2019.09.012

Stokkeland, M., Klausen, K., and Johansen, T. A. (2015). "Autonomous visual navigation of unmanned aerial vehicle for wind turbine inspection," in Proceedings of the IEEE International Conference on Unmanned Aircraft Systems, Denver, CO, USA, 09-12 June 2015 (IEEE), 998–1007. doi:10.1109/ICUAS.2015.7152389

Tsai, Y. H., Hung, W. C., Schulter, S., Sohn, K., Yang, M. H., and Chandraker, M. (2018). "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision and pattern recognition* (Salt Lake City, UT: IEEE), 7472–7481. (Accessed June 18–22, 2018). doi:10.1109/CVPR.2018.00780

Wang, L., and Zhang, Z. (2017). Automatic detection of wind turbine blade surface cracks based on uav-taken images. *IEEE Trans. Ind. Electron.* 64, 7293–7303. doi:10.1109/TIE.2017.2682037

Wilson, G., and Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.* 11, 1–46. doi:10.1145/3400066

Xu, Y., Huang, B., Luo, X., Bradbury, K., and Malof, J. M. (2022). Simpl: Generating synthetic overhead imagery to address custom zero-shot and few-shot detection problems. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 15, 4386–4396. doi:10.1109/JSTARS.2022.3172243

Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., et al. (2019). "Scrdet: Towards more robust detection for small, cluttered and rotated objects," in *Proceedings of the IEEE/CVF international conference on computer vision* (Seoul, South Korea: IEEE), 8232–8241. (Accessed October 27, 2019–November 3, 2019). doi:10.1109/ICCV.2019.00832

Yu, F., Wang, D., Chen, Y., Karianakis, N., Shen, T., Yu, P., et al. (2022). "Sc-uda: Style and content gaps aware unsupervised domain adaptation for object detection," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 03-08 January 2022 (IEEE), 1061. doi:10.1109/WACV51458.2022.00113

Zhao, S., Yue, X., Zhang, S., Li, B., Zhao, H., Wu, B., et al. (2022). A review of single-source deep unsupervised visual domain adaptation. *IEEE Trans. Neural Netw. Learn. Syst.* 33, 473–493. doi:10.1109/TNNLS.2020.3028503

Zhou, C., Wu, M., and Lam, S. K. (2019). Ssa-cnn: Semantic self-attention cnn for pedestrian detection. *arXiv*. doi:10.48550/arXiv.1902.09080

Zhu, X., Pang, J., Yang, C., Shi, J., and Lin, D. (2019). "Adapting object detectors via selective cross-domain alignment," in *Proceedings of the IEEE/CVF international conference on computer vision and pattern recognition* (Silver Spring, MD, USA: IEEE Computer Society), 687–696. doi:10.1109/CVPR.2019.00078