IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 1, February, 2024, pp. 485-491
Journal homepage: http://iieta.org/journals/ts

Enhanced Detection of Electric Power Facilities Utilizing a Re-Parameterized Convolutional Network

Jinglin Han[1*], Zhiyong Chen[2], Ping Hu[1], Hongtao Li[1], Guangyi Li[2]

[1] State Grid Hebei Electric Power Company, Shijiazhuang 050081, China
[2] State Grid Hebei Economic Research Institute, Handan 056305, China

Corresponding Author Email: jinglinh783@gmail.com

https://doi.org/10.18280/ts.410143

Received: 26 July 2023
Revised: 18 November 2023
Accepted: 3 December 2023
Available online: 29 February 2024

Keywords:
region-based convolutional network (R-CNN), unmanned aerial vehicle (UAV), deep learning, digital twin, electric power facility detection

ABSTRACT

In electrical grid management, the integration of deep learning and digital twin technology constitutes a pivotal component of contemporary power network systems. The foundation of the intelligent digital electrical grid rests upon the meticulous collection of edge facility information, necessitating rapid and precise identification of electric power facilities for both civilian and military utilization within digital grid systems. This study introduces a novel object detection methodology tailored for a diverse array of electric power facilities, leveraging a re-parameterized Mask Region-based Convolutional Neural Network (Mask R-CNN) augmented by transfer learning techniques. A multi-scale dataset of electric facilities was developed, facilitating the training and testing of the proposed model on images featuring manually annotated electric power facilities. These facilities are categorized into two distinct groups based on target scale, encompassing utility poles, transformers, insulators, cross arms, and wire clips. To enhance the efficiency of bounding region localization, the Mean Shift (MS) algorithm was employed to adjust the size of anchors within the Region Proposal Network (RPN), thereby streamlining the detection process. Experimental outcomes reveal that, in comparison to the original model, the re-parameterized Mask R-CNN (Rep-Mask R-CNN) demonstrates a 6.17% increase in mean Average Precision (AP) and a 33% reduction in inference time. Equipped with a geolocation module, Unmanned Aerial Vehicles (UAVs) deploying this model can achieve comprehensive digital base map management, encompassing geographic and equipment information, while also supporting visual display services within the digital electrical grid. This study underscores the potential of re-parameterized convolutional networks in enhancing the accuracy and efficiency of electric power facility detection, contributing significantly to the advancement of intelligent digital grid management systems.

1. INTRODUCTION

In recent years, the management and planning of electrical grids have increasingly relied on digital twin technology, marking a significant advancement in modern power network infrastructure. The role of electric power facilities is central to the functioning of electrical grids, with the smart management of these facilities being crucial for the development of digital and intelligent grid systems. Research in computer vision has primarily focused on the real-time detection, identification, and categorization of objects in images. Utilizing UAVs equipped with geolocation modules and advanced edge computing models enables the integration of geographical and equipment data management, enhancing digital mapping and visual services within digital electricity networks. Despite this, there remains a scarcity of research focused on the detection of objects within common power facility environments. Implementing object detection algorithms can greatly benefit digital electricity networks, including intelligent detection of facility faults and digitalized management of data collection systems. Figure 1 illustrates how data from on-site measurements, digital mapping, scene management, and digital twins are integrated to form a comprehensive, modern, digitally intelligent power network system.

Conventional methods for object detection typically depend on pre-defined templates [1] or rely on geometrical shapes and prior knowledge [2, 3], while algorithms for object-based image analysis [4, 5] focus on segmenting and classifying objects. These traditional techniques often require manual data extraction, which complicates achieving precise detection results. Recent advancements in machine learning have led to the development of enhanced techniques for feature extraction from images [6, 7], including support vector machines (SVM), AdaBoost, and decision trees. These approaches are capable of identifying statistical characteristics such as the gray-level run length matrix (GLRLM) [8], histogram of oriented gradients (HOG) [9], bag of words (BOW) [10], local binary pattern [11], DCT coefficients [12], scale-invariant feature transform (SIFT) [13], and the deformable part model (DPM) [14]. The advent of deep learning algorithms, especially those based on CNN [15], has significantly advanced object detection, enabling the analysis of large datasets without the

need for pre-existing templates and facilitating the recognition of complex patterns.

Deep learning's efficacy has been underscored by the success of CNN in tasks such as image classification, detection, and recognition, leading to the proposal of several high-performance deep network models like AlexNet [16], GoogleNet [17], and VGGNet [18], known for their straightforward yet deep structures. ResNet [19] introduced a novel residual structure, addressing the learning degradation issue present in very deep networks by simplifying the model's internal framework. Subsequent developments have introduced complex neural networks featuring skip connections and multi-branch architectures, which excel at feature extraction and are widely applied in computer vision. However, these increasingly complex structures can lead to unnecessary computational demands and reduced efficiency and interpretability. RepVGG [20], a simpler model that forgoes complex connections for a streamlined approach, matches ResNet's performance with faster inference times through the Rep-Conv-Block method, reflecting a shift towards models that balance accuracy with lightweight, rapid processing.

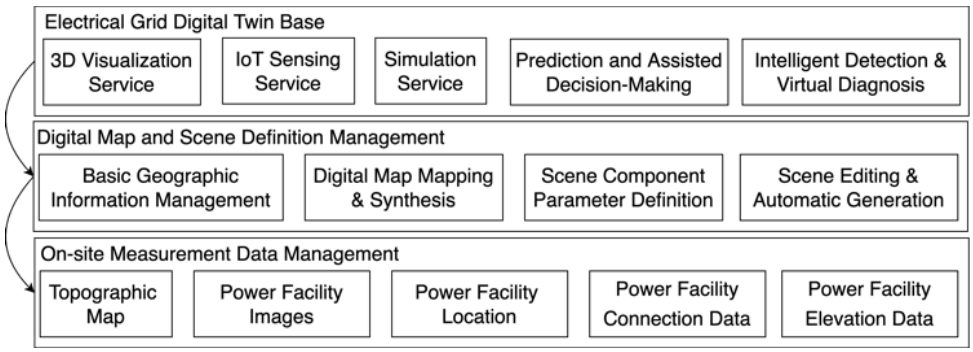

Figure 1. The relationship between on-site measurement data, digital mapping, scene management, and digital twin bases

The advancement of deep learning frameworks has significantly contributed to the evolution of methods for target detection. These algorithms are broadly categorized into one-stage and two-stage models. In one-stage models, object detection is achieved through direct regression of the target box, with default boxes pre-set based on the feature map. This category includes models like YOLO [21], SSD [22], and subsequent enhancements such as YOLOv3 [23] with its upsampling layer and layer-skipping concatenation, as well as YOLOv4 [24] and YOLOv5 [25], which incorporate various advanced techniques like expanding the receptive field, e.g., SPP-Net [26], RFB-Net [27], and attention mechanisms [28, 29], and YOLOx [30] that refine the prediction branch and candidate box approach. The two-stage models, initiated by R-CNN [31], pre-generate candidate regions through Selective Search before performing feature extraction. Building on this, Fast R-CNN [32] and Faster R-CNN [33] were developed to address the slow detection speed of candidate boxes, high training costs, and inefficiencies in the R-CNN by employing the RPN [34] and Feature Pyramid Network (FPN) [35] for enhanced feature fusion across different layers. Mask R-CNN [36] introduced a distinct binary mask branch for separate prediction, and improvements in alignment issues were made in Faster R-CNN with the adoption of RoIAlign over RoIPooling.

Mask R-CNN has demonstrated remarkable success in object detection, leading to the development of numerous enhanced versions. Innovations include the use of a non-quantization rounding pooling layer [37] to tackle issues related to the fixed count of interpolated pixels, and the creation of an improved ResNet-FPN [38] designed for preprocessing synthetically altered images. Beyond modifications to the architecture, further enhancements [39] have been made to optimize Mask R-CNN's loss function, refine the dimensional output of RoIs [40], and incorporate transfer learning strategies to improve performance on datasets with limited samples. Additionally, Mask R-CNN has been adapted to include RoIWarping [41] for better feature mapping, and a Light-Head subnet has been introduced to reduce the model's complexity. The development of Cascade R-CNN [42] employs a multi-stage cascading architecture and revises the IoU thresholding approach to enhance feature utilization.

This paper explores the identification of electric energy facilities by tailoring the Mask R-CNN model to a manually curated facility dataset for multi-scale object detection. Our contributions include creating a comprehensive dataset of electric energy facilities, segmenting this dataset by facility scale to enhance detection across different sizes, and implementing a Rep-Mask R-CNN model with innovations in model structure, such as adaptive anchors in RPN, a recursive FPN structure, and a re-parameterized backbone. Additionally, we employ transfer learning to address small dataset convergence challenges, achieving effective training outcomes by leveraging weights from larger datasets for objects of varying sizes and numbers.

2. RELATED WORKS

Traditional object detection algorithms are broadly categorized into one-stage and two-stage methods. One-stage methods directly analyze images to produce detection outcomes, while two-stage methods initially extract candidate frames from the image, followed by a refinement process to finalize the detection points based on these candidate areas. Typically, two-stage methods achieve higher precision and perform better across various public datasets, albeit at the cost of increased computational demands and resource usage. The pioneering two-stage approach, RCNN, tackles the detection challenge through four main steps:

(1) Creating candidate regions from an input image;

(2) Extracting features from these regions using a deep network;

(3) Classifying these features into categories using a feature classifier (initially binary-SVM);

(4) Adjusting the positions of candidate boxes with

regressors.

R-CNN paved the way for its successors, Fast R-CNN and Faster R-CNN, which improved upon the original by incorporating advancements in region-based networks. Fast R-CNN was introduced by Ross Girshick in 2015, enhancing over R-CNN, and Faster R-CNN further refined this by using anchors of specific scales and proportions, along with more sophisticated feature extraction methods like the FPN.

Mask R-CNN, an evolution of Faster R-CNN, significantly boosts performance by optimizing two-stage operations. It distinguishes itself through a refined multi-scale detection approach and RoI processing technique. Unlike Faster R-CNN, which primarily outputs the final layer, Mask R-CNN utilizes FPN to generate multi-scale feature maps for RPN RoI extraction. It introduces RoIAlign, replacing RoI pooling, to standardize RoIs' sizes, which are then processed for bounding box regression and classification predictions. An additional convolution layer post-RoIAlign facilitates mask prediction. Essentially, Mask R-CNN extends Faster R-CNN by incorporating the ResNet-FPN module and a segmentation branch for masks. The advent of multi-scale pooling layers in mainstream detection methods aims to extract both semantic and spatial details more effectively. The FPN's architecture, combining bottom-up, lateral, and top-down connections, integrates features across all levels to handle images containing objects of varying sizes simultaneously.

In Faster R-CNN, the RoIPooling step could lead to misalignment issues, where candidate boxes might shift from their intended positions. Mask R-CNN addresses this with the RoIAlign layer, which maintains fractional components through bilinear interpolation, avoiding the rounding off that occurs in RoIPooling. RoIAlign meticulously interpolates within subdivided regions, using the center points for floating-point coordinate sampling, thus ensuring more precise candidate box alignment. RoIWarp, an alternative approach, pre-empts feature map warping before pooling, opting for quantization followed by nonlinear interpolation, differing from the dual processes in RoIPooling and RoIAlign. By minimizing the sample points traversed, both RoIAlign and RoIWarp offer improved outcomes compared to RoIPooling, prompting further innovations in pooling operations like PSRoIPooling and PSRoIAlign [43].

3. METHODOLOGY

In this study, we delve into the unique challenges of detecting electric energy facilities, characterized by their strong class homogeneity, significant variations across scenes, and size disparities among targets. To address these challenges, we enhance and tailor our dataset model accordingly. To ensure robust generalization capabilities, we compile a comprehensive dataset of electrical power facilities using web scraping and on-site photography. These images are meticulously labeled by hand, with each being classified and annotated with accurate bounding boxes and masks according to scale. This meticulous preparation enhances the model's ability to detect facilities of varying sizes with greater precision, especially when identifying smaller objects within large-scale images. To better match the anchor sizes to the actual dimensions of the targets, instead of defaulting to the conventional anchor ratios of 1:1, 1:2, and 2:1 for bounding box regression, we adapt the MS clustering algorithm. The process of our modified MS clustering algorithm is as follows: (1) Take the ratio of length to width of all ground truth boxes as the basis for clustering to initialize the sample space; (2) Initially determine a central point, and calculate vectors from all points to this central point within a specified step distance; (3) Compute the mean of all vectors within the space to find a mean offset value; (4) Shift the central point to this mean offset location; (5) Continue shifting the central point until it meets specific criteria.

The above process can be summarized in a mathematical form as follows:

$$M(p) = \frac{1}{K} \sum_{p_i \in S_h} (p_i - p), \tag{1}$$

$$p^{t+1} = M(p^t) + p^t, \tag{2}$$

$$p_i^* = arg_p \min M(p_i), \tag{3}$$

where, p_i represents all points in the sample space, M represents the mean error, after several iterations, the result p^* is the clustering result, which is the final choice of anchor ratio in RPN that best fits our current dataset.

A key enhancement within Mask R-CNN is the integration of the FPN. Opting for ResNet as its foundational architecture, FPN incorporates residual elements and skip connections, leveraging the strengths of ResNet's design. However, given the inherently complex nature of this two-stage model, we opted for a re-parameterized approach to simplify the architecture. This strategy avoids the need for the intricate skip connections and multi-branch configurations found in models like ResNet or the Inception modules, despite their proven efficacy in boosting model performance.

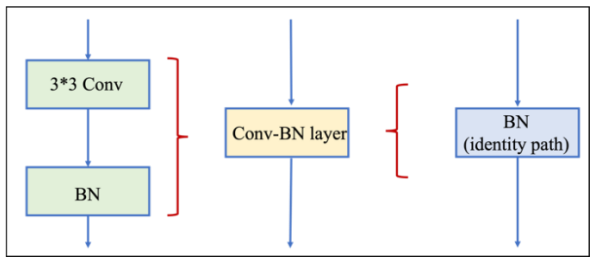

Figure 2. The method for combining a 3×3 convolution layer with a BN layer and calculating a tensor that performs like an identity operation to achieve similar fusion

In our deployment, we employ a model reconstruction strategy using re-parameterized layers. This approach differs from RepVGG's architecture, which typically includes a parallel arrangement of a 3×3 convolution layer and a 1×1 convolution layer, combined with the original data through summation. Instead, our design incorporates two parallel 3×3 convolution layers along with an identity path that includes a BN layer. These three paths are then summed together, ultimately integrating them into a single block for streamlined processing. We achieved this design via three steps, as follows:

First, we decouple the triple frame group and combine them as one block. We fuse the 3x3 convolution layer and BN layer as follows:

$$rep(x) = \gamma\big(W(x) + (x - \mu)\big)\sigma^{-1} + bias, \tag{4}$$

where, tensor x is the input, and the fusion result is calculated with convolution weight W and other convolution layer parameters. In Figure 2, the methodology for integrating the

3x3 convolution layer with the Batch Normalization (BN) layer is depicted, along with the approach for calculating a tensor that performs an equivalent identity operation to facilitate a similar fusion.

Subsequent to the vertical and horizontal merging processes, zero-padding is utilized to adjust the 1x1 convolution layer and the identity layer to a 3×3 format. These adjusted layers are then amalgamated with our redesigned 3×3 layer, resulting in two distinct module sequences: the 3×3-1×1 combined block and the 3×3-identity combined block.

$$M^{(3,1,0)} = rep\left(M^{(i)} * \left(W^{(3)} + W^{(1)} + 1\right), \mu, \sigma, \gamma, \beta\right), \quad (5)$$

And our re-parameterized convolution block replaces the 1×1 layer with 3×3 layers as follows:

$$M^{(3,3,0)} = rep\left(M^{(i)} * \left(W^{(3)_1} + W^{(3)_2} + 1\right), \mu, \sigma, \gamma, \beta\right), \quad (6)$$

where, $\mu, \sigma, \gamma, \beta$ represent the parameters in convolution layers, $M^{(i)}$ denotes the input tensor, $M^{(3,3,0)}$ and $M^{(3,1,0)}$ denote the outputs of the Rep Convolution block layer and the original Rep block, the weights of 3×3 layers, $W^{(3)_1}$ is differ from $W^{(3)_2}$ to get different results from the 3×3 convolution operation. Following the layering of modules with analogous structures, an adaptive average pooling layer is introduced to adjust the output size based on the input parameters. Consequently, the output from our layered feature mapping operation, facilitated by Rep-RPANet, serves as the input for the subsequent module.

Ultimately, Figure 3 illustrates the entire re-parameterization procedure of the Rep-Convolution-block, transitioning from a three-branch architecture to a Conv-BN setup, and thereafter converting to a unified layer structure, culminating in the model's reconstruction.

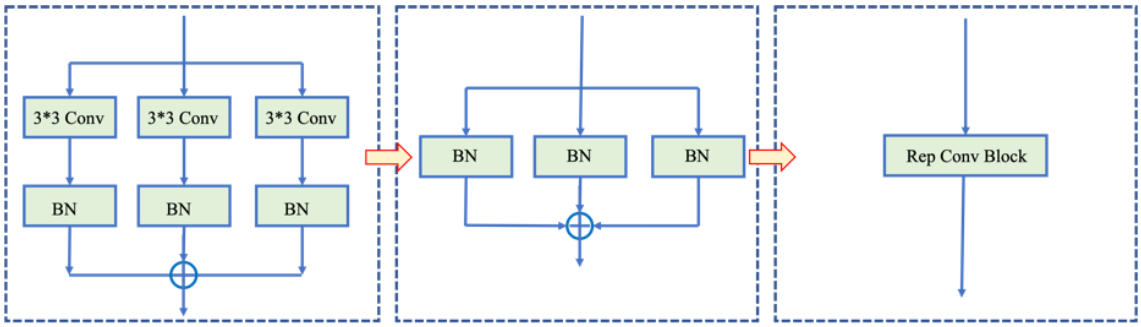

Figure 3. The re-parameterization of the Rep-Conv block involves transitioning from a three-branch architecture to a Conv-BN setup, eventually consolidating into a single-layer structure

The widespread adoption of the FPN framework is attributed to its superior capability in extracting features from objects across multiple scales [44]. Enhancements to the original FPN, such as cascade FPN, up-sample FPN, and recursive FPN [45], have been developed to further refine its performance. Despite the ability of a traditional FPN to extract features at multiple levels, it typically relies on the output from the k-th layer alone for RoI computations. The k is obtained as follows:

$$k = \left\lfloor k_0 + \log_2\left(\sqrt{wh}\right) - \log_2 H \right\rfloor, \quad (7)$$

where, w and h denote the width and height of the input image, respectively, with H typically set to 224. This is because large-scale RoIs are better extracted from low-resolution feature maps to enhance the detection of larger targets, whereas small-scale RoIs benefit from high-resolution feature maps for improved detection of smaller targets. Addressing the challenge of potentially losing low-level feature information with this distribution approach, we integrate low-level features with high-level features. Specifically, we introduce the RPANet by incorporating side connections within the FPN framework. This enhances the integration of multi-scale features at various levels. Furthermore, a two-level recursive structure is added to facilitate a second cycle of feature fusion. In this model, let n represent the number of recursion epochs, $C_i^{(n)}$ the output of the i-th stage of the model backbone (e.g., our Rep-blocks), $P_i^{(n)}$ the output of each layer in the top-bottom feature fusion path employing down-sampling to match the feature map sizes, and $N_i^{(n)}$ the output of each layer in the bottom-top feature fusion path using up-sampling to equalize the feature map sizes. The sum of these outputs serves as the input for the next recursion epoch or as the final outputs of RPANet. Should the n-th epoch be the final one, the complete model structure of RPANet is illustrated in Figure 4.

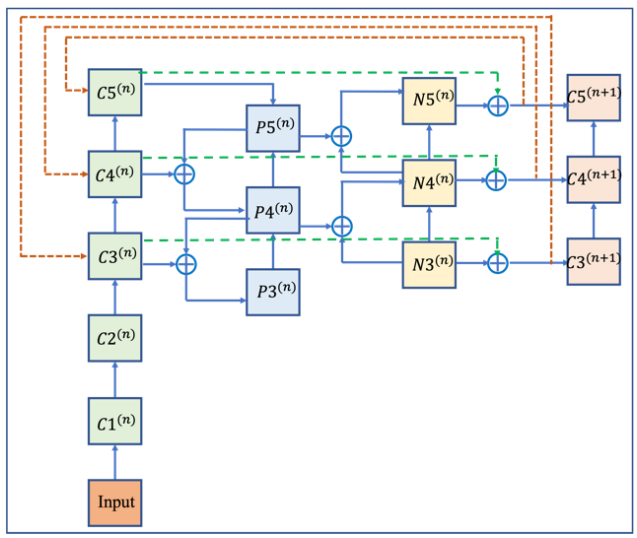

Figure 4. The structure of RPANet

4. EXPERIMENTS AND DISCUSSION

We compiled a dataset of electric power facilities through manual efforts, sourcing images from the internet and direct photography. The images are either originally 1024x728 or resized to this dimension, and subsequently zero-padded to

achieve a uniform size of 1024x1024. The dataset is categorized based on the scale of the power facilities; it includes larger objects like "telegraph poles" and "transformers," and smaller objects such as "insulators," "wire clips," and "cross arms," totaling 150 images for larger objects and 100 for smaller objects, culminating in a dataset of 250 images. Table 1 outlines the detailed distribution of the dataset, and Figure 5 presents a selection of these images, each marked with ground truth bounding boxes. For this project, our model's training and testing were conducted on Tensorflow 1.13 and Keras 2.2.4, utilizing an Ubuntu 20.04 system equipped with an NVIDIA GeForce RTX 2080 Ti graphics card and an Intel i9-9900K processor. The training process involved 25 epochs for the model's head layers and another 25 epochs for the entire model, with learning rates set at 1e-3 for the head layers and 1e-2 for the full model layers.

Tabel 1. Each target was classified as either "Large" or "Small" based on size

Label Name	Number of Objects	Group	Mean H/W
UtilityPole	240	Large	4.1
Transformer	40	Large	1.2
Insulator	940	Small	0.8
WireClip	177	Small	0.8
CrossArm	207	Small	0.5

Several tests were conducted to assess the effectiveness of our suggested model. Initially, we juxtaposed our model against the original Mask R-CNN, Mask R-CNN integrated with Rep-blocks, SSD, YOLOv3-SPP, and YOLOv5x, aiming to gauge its performance on the dataset of power energy facilities, as depicted in Figure 6.

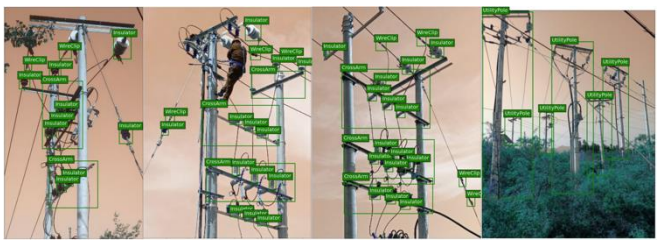

Figure 5. Images in dataset manually annotated with ground truth bounding boxes

In Figure 6, our model achieves the highest average mAP at 88.62%, surpassing YOLOv5, a notable and efficacious single-stage model, and significantly outperforming traditional models. Detecting small targets remains a challenge within object detection tasks. While YOLOv5x demonstrates commendable performance in identifying larger targets, its accuracy diminishes with small and overlapping targets, leading to a 10.71% decrease in mAP. The escalation in model complexity could further decelerate inference speeds. Nonetheless, in comparison to more complex models, our model maintains a quicker operational speed, striking an optimal balance between detection efficacy and processing velocity.

Secondly, by conducting experiments with ResNet50, ResNet101, RepVGG, and Rep-Conv-Blocks (our model) as backbones, we assess their impact on mAP, illustrated in Figure 7. This includes an evaluation of the benefits derived from utilizing re-parameterized operations. Among these, the model equipped with Rep-Conv-Blocks achieves the highest mAP at an IoU threshold of 0.5, registering 6.05% higher than ResNet50 and 4.29% higher than ResNet101. Moreover, it exhibits a 33% improvement in speed over ResNet50 and a 35% advantage over ResNet101. This demonstrates that Mask R-CNN, when employing Rep-Conv-Blocks as its backbone, outperforms traditional Mask R-CNN configurations using ResNet in both inference speed and recognition capabilities. Additionally, RepVGG, when used as a backbone, achieves commendable inference speeds post-deployment but suffers a 9.5% loss in mAP. This performance discrepancy suggests that the feature extraction capacity of the 1×1 convolution layer falls short compared to the 3×3 convolution layer. To address this, we substitute the 1×1 convolution layer in RepVGG with a newly devised 3×3 convolution layer to enhance feature extraction, culminating in the design of Rep-Conv-Blocks.

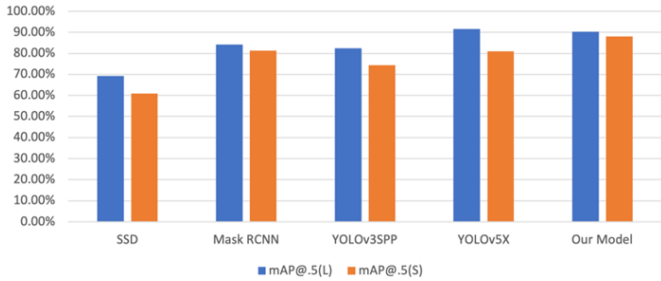

Figure 6. The experimental results including SSD, Mask RCNN, YOLOv3-SPP, YOLOv5x, and our model

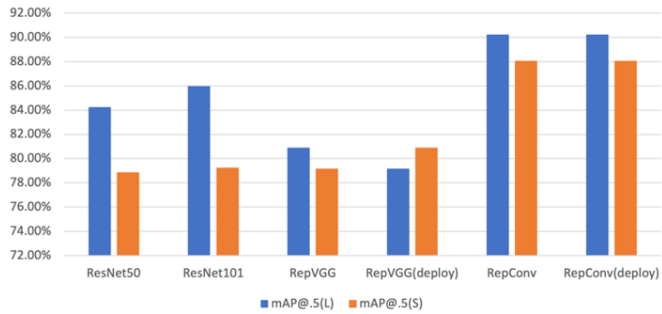

Figure 7. The outcomes from testing various model backbones in our experiments

Lastly, we evaluate the influence of the RPANet and various pooling layers on the model's efficacy. The findings indicate that in the absence of RPANet, while a simpler FPN may achieve quicker inference speeds, it experiences a 3.6% reduction in mAP due to its inferior capability for feature fusion relative to RPANet. This substantiates RPANet's contribution to enhancing the model's recognition performance through improved feature extraction capabilities.

5. CONCLUSION

The primary focus of this study is on the intelligent detection and identification of specific electric power facilities, crucial for on-site data measurement and management. In our research, we analyze the pros and cons of various architectural models for detection and introduce a Rep-Mask RCNN model equipped with RPANet, aimed at identifying electrical power facilities. In our proposed model, we utilize Rep-Conv blocks as the backbone with varied convolutional layer structures, replace the traditional FPN with RPANet, and strike a balance

between computational efficiency and detection precision. Additionally, we constructed a dataset of electric power facilities from scratch, conducted anchor clustering, segmented the data according to the scale of target boxes, and applied transfer learning techniques to optimize detection outcomes across different target sizes. The experimental outcomes demonstrate that our Rep-Mask RCNN model not only achieves a high inference speed but also maintains robust capabilities in box detection and classification. Looking ahead, our future endeavors will focus on two main areas: firstly, expanding our dataset with more images captured under various conditions such as different lighting, weather, and shooting angles to enhance the model's generalizability; and secondly, addressing challenges associated with the Rep-Conv blocks backbone, such as the initial training issue of substantial loss, despite its effective performance enhancements. Our forthcoming work will explore refining the model structure, potentially incorporating attention mechanisms and transformer architectures, to further elevate detection performance and accelerate inference speeds.

ACKNOWLEDGMENTS

This work was supported by Science and Techology Proiect of Hebei Electric Power Company "Research on Key Technologies for Planning of Digital Active Distribution Network in Xiong' an New Area" (Grant No.: SGHEJY00GHJS2000103).

REFERENCES

[1] Chaudhuri, D., Kushwaha, N.K., Samal, A. (2012). Semi-automated road detection from high resolution satellite images by directional morphological enhancement and segmentation techniques. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5(5): 1538-1544. https://doi.org/10.1109/JSTARS.2012.2199085

[2] Ok, A.O. (2013). Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. ISPRS Journal of Photogrammetry and Remote Sensing, 86: 21-40. https://doi.org/10.1016/j.isprsjprs.2013.09.004

[3] Zhang L. M., Cong Y., Meng F. Z., Wang Z. Q., Zhang P., Gao S. (2021). Energy evolution analysis and failure criteria for rock under different stress paths. Acta Geotechnica, 16(2): 569-580. https://doi.org/10.1007/s11440-020-01028-1

[4] Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Tiede, D. (2014). Geographic object-based image analysis–towards a new paradigm. ISPRS Journal of Photogrammetry and Remote Sensing, 87: 180-191. https://doi.org/10.1016/j.isprsjprs.2013.09.014

[5] Li, Y., Wang, S., Tian, Q., Ding, X. (2015). Feature representation for statistical-learning-based object detection: A review. Pattern Recognition, 48(11): 3542-3559. https://doi.org/10.1016/j.patcog.2015.04.018

[6] Dong, C., Liu, J., Xu, F. (2018). Ship detection in optical remote sensing images based on saliency and a rotation-invariant descriptor. Remote Sensing, 10(3): 400. https://doi.org/10.3390/rs10030400

[7] Zhang L. M., Chao W.W., Liu Z.Y., Cong Y., Wang, Z.Q. (2022). Crack propagation characteristics during progressive failure of circular tunnels and the early warning thereof based on multi-sensor data fusion. Geomechanics and Geophysics for Geo-Energy and Geo-Resources, 8: 172. https://doi.org/10.1007/s40948-022-00482-3

[8] Mushtaq, S., Mir, A.H. (2014). Novel method for image splicing detection. In 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Delhi, India, pp. 2398-2403. https://doi.org/10.1109/ICACCI.2014.6968386

[9] Zhang, D., Han, J., Cheng, G., Liu, Z., Bu, S., Guo, L. (2014). Weakly supervised learning for target detection in remote sensing images. IEEE Geoscience and Remote Sensing Letters, 12(4): 701-705. https://doi.org/10.1109/LGRS.2014.2358994

[10] Yokoya, N., Iwasaki, A. (2015). Object detection based on sparse representation and Hough voting for optical remote sensing imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(5): 2053-2062. https://doi.org/10.1109/JSTARS.2015.2404578

[11] Rahim, M.A., Hossain, M.N., Wahid, T., Azam, M.S. (2013). Face recognition using local binary patterns (LBP). Global Journal of Computer Science and Technology, 13(4): 1-8.

[12] Shi, Y.Q., Chen, C., Chen, W. (2007). A natural image model approach to splicing detection. In Proceedings of the 9th Workshop on Multimedia & Security, United States, pp. 51-62. https://doi.org/10.1145/1288869.1288878

[13] Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60: 91-110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

[14] Felzenszwalb, P., McAllester, D., Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, pp. 1-8. https://doi.org/10.1109/CVPR.2008.4587597

[15] Bharati, A., Singh, R., Vatsa, M., Bowyer, K.W. (2016). Detecting facial retouching using supervised deep learning. IEEE Transactions on Information Forensics and Security, 11(9): 1903-1913. https://doi.org/10.1109/TIFS.2016.2561898

[16] Lv, H., Zhao, D., Chi, X. (2017). Deep learning for early diagnosis of Alzheimer's disease based on intensive AlexNet. Computer Science, 44(6): 50-60.

[17] Tang, P., Wang, H., Kwong, S. (2017). G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition. Neurocomputing, 225: 188-197. https://doi.org/10.1016/j.neucom.2016.11.023

[18] Qi, Y., Zhao, Z., Du, L., Qiao, H., Wang, L. (2018). A classification method of aerial targets based on VGGNet and label distribution learning. Dianli Jianshe/Electric Power Construction, 39: 109-115. https://doi.org/10.3969/j.iisn.1000-7229.2018.02.014

[19] He, K., Zhang, X., Ren, S., 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, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[20] Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J. (2021). RepVGG: Making VGG-style convnets great

again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, pp. 13733-13742. https://doi.org/10.1109/CVPR46437.2021.01352

[21] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). 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, pp. 779-788. https://doi.org/10.1109/CVPR.2016.91

[22] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016). SSD: Single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference, Proceedings, Part I 14, Amsterdam, The Netherlands, pp. 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[23] Redmon, J., Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767. https://doi.org/10.48550/arXiv.1804.02767

[24] Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934. https://doi.org/10.48550/arXiv.2004.10934

[25] Kuznetsova, A., Maleva, T., Soloviev, V. (2020). Detecting apples in orchards using YOLOv3 and YOLOv5 in general and close-up images. In Advances in Neural Networks–ISNN 2020: 17th International Symposium on Neural Networks, ISNN 2020, Cairo, Egypt, pp. 233-243. https://doi.org/10.1007/978-3-030-64221-1_20

[26] Huang, Z., Wang, J., Fu, X., Yu, T., Guo, Y., Wang, R. (2020). DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection. Information Sciences, 522: 241-258. https://doi.org/10.1016/j.ins.2020.02.067

[27] Liu, S., Huang, D. (2018). Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, pp. 385-400. https://doi.org/10.1007/978-3-030-01252-6_24

[28] Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[29] Woo, S., Park, J., Lee, J.Y., Kweon, I.S. (2018). CBAM: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, pp. 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[30] Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J. (2021). YOLOx: Exceeding YOLO series in 2021. arXiv preprint arXiv:2107.08430. https://doi.org/10.48550/arXiv.2107.08430

[31] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014). 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, pp. 580-587. https://doi.org/10.1109/CVPR.2014.81

[32] Girshick, R. (2015). Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, pp. 1440-1448. https://doi.org/10.1109/ICCV.2015.169

[33] Shi, J., Zhou, Y., Zhang, Q. (2019). Item recognition based on faster R-CNN in service robot. Application Research of Computers, 36(10): 3152-3156.

[34] Ren, S., He, K., Girshick, R., Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, Canada, pp. 91-99. https://doi.org/10.5555/2969239.2969250

[35] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 2117-2125. https://doi.org/10.1109/CVPR.2017.106

[36] He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, Canada, pp. 2961-2969. https://doi.org/10.48550/arXiv.1703.06870

[37] Zuo, L., He, P., Zhang, C., Zhang, Z. (2020). A robust approach to reading recognition of pointer meters based on improved mask-RCNN. Neurocomputing, 388: 90-101. https://doi.org/10.1016/j.neucom.2020.01.032

[38] Ahmed, B., Gulliver, T.A., Alzahir, S. (2020). Image splicing detection using mask-RCNN. Signal, Image and Video Processing, 14: 1035-1042. https://doi.org/10.1007/s11760-020-01636-0

[39] Mahmoud, A., Mohamed, S., El-Khoribi, R., AbdelSalam, H. (2020). Object detection using adaptive mask RCNN in optical remote sensing images. International Journal of Intelligent Engineering Systems, 13(1): 65-76.

[40] Zhang, Q., Chang, X., Bian, S.B. (2020). Vehicle-damage-detection segmentation algorithm based on improved mask RCNN. IEEE Access, 8: 6997-7004. https://doi.org/10.1109/ACCESS.2020.2964055

[41] Shi, J., Zhou, Y., Zhang, W.X.Q. (2019). Target detection based on improved mask R-CNN in service robot. In 2019 Chinese Control Conference (CCC), Guangzhou, China, pp. 8519-8524. https://doi.org/10.23919/ChiCC.2019.8866278

[42] Cai, Z., Vasconcelos, N. (2018). 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, pp. 6154-6162. https://doi.org/10.1109/CVPR.2018.00644

[43] Yao, S., Chen, Y., Tian, X., Jiang, R. (2020). GeminiNet: combine fully convolution network with structure of receptive fields for object detection. IEEE Access, 8: 60305-60313. https://doi.org/10.1109/ACCESS.2020.2982939

[44] Liu, S., Qi, L., Qin, H., Shi, J., Jia, J. (2018). Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 8759-8768. https://doi.org/10.1109/CVPR.2018.00913

[45] Qiao, S., Chen, L.C., Yuille, A. (2021). Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, pp. 10213-10224. https://doi.org/10.1109/CVPR46437.2021.01008