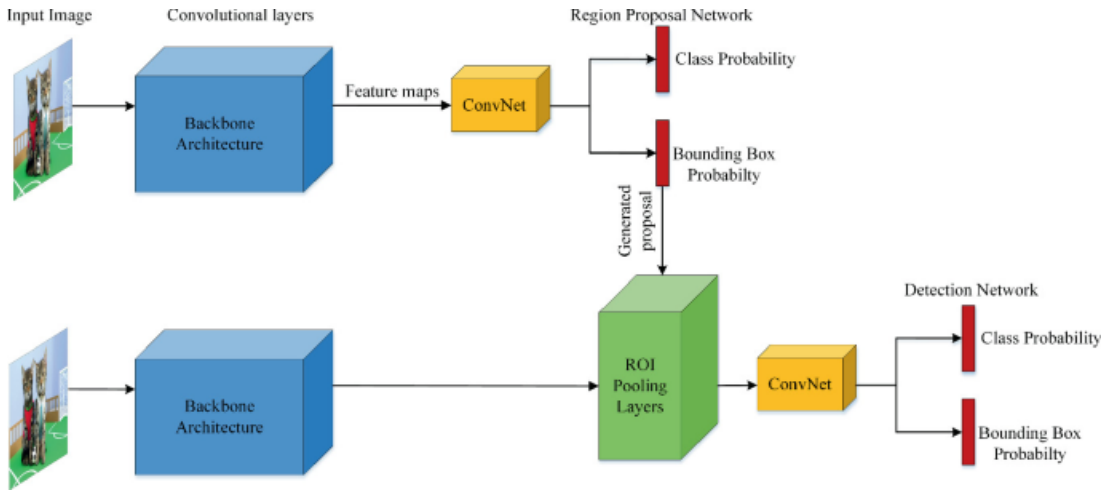
Two-Stage Object Detectors A Comparative Evaluation

IARIA INNOV 2023 Paper # 70007

Jihad Qaddour
School of Information Technology
Illinois State University
Normal, IL, USA
jqaddou@ilstu.edu



University of Tsukuba

IARIA

Outcome

- Introduction
- Background on object detection.
- Two-stage detectors.
- A comparative performance analysis of two-stage detectors.
- Conclusion

Dr. Jihad Qaddour

Dr. Qaddour received a B.S.E. degree from Damascus University in 1982, an M.E. degree in Communication and control Theory from Wichita State University in 1987, and a Ph.D. degree from Wichita State University, KS, USA, in 1990, and an M.S. degree in Applied Mathematics and statistics from Wichita State University, in 1993. He is currently an associate Professor with the Illinois State University School of IT. His research interests include artificial intelligence, computer vision, information processing, wireless security, IoT and its applications, smart city, and 5G and 6G technology.

Aims and contributions of our paper

In our paper, we aimed at:

1. Overview of the two-stage object detectors
2. Provides a comparative performance analysis of many two-stage object detectors.

Contributions of our study are threefold:

1. Presented a comparative performance analysis of two-stage object detectors
2. Evaluated the performance of different detectors on two datasets, MSCOCO and PASCAL VOC 2012.
3. We use the average precision AP0.5 and AP[0.5:95] on the above both datasets.
4. Results showed that DetectoRS outperformed all other two-stage models
5. DetectoRS achieved an AP0.5 of 53.30% and an AP[0.5:95] of 71.60% on MSCOCO.
6. DetectoRS achieved an AP0.5 of 83.00% and an AP[0.5:95] of 90.30% on PASCAL VOC 2012. However, it is also more complex.
7. Other two-stage object detectors that performed well in the comparison include:
8. NAS-FPN, Mask R-CNN, and Cascade R-CNN.
9. These models also use various techniques to improve performance, such as region proposal networks (RPNs), RoIAlign, and focal loss.

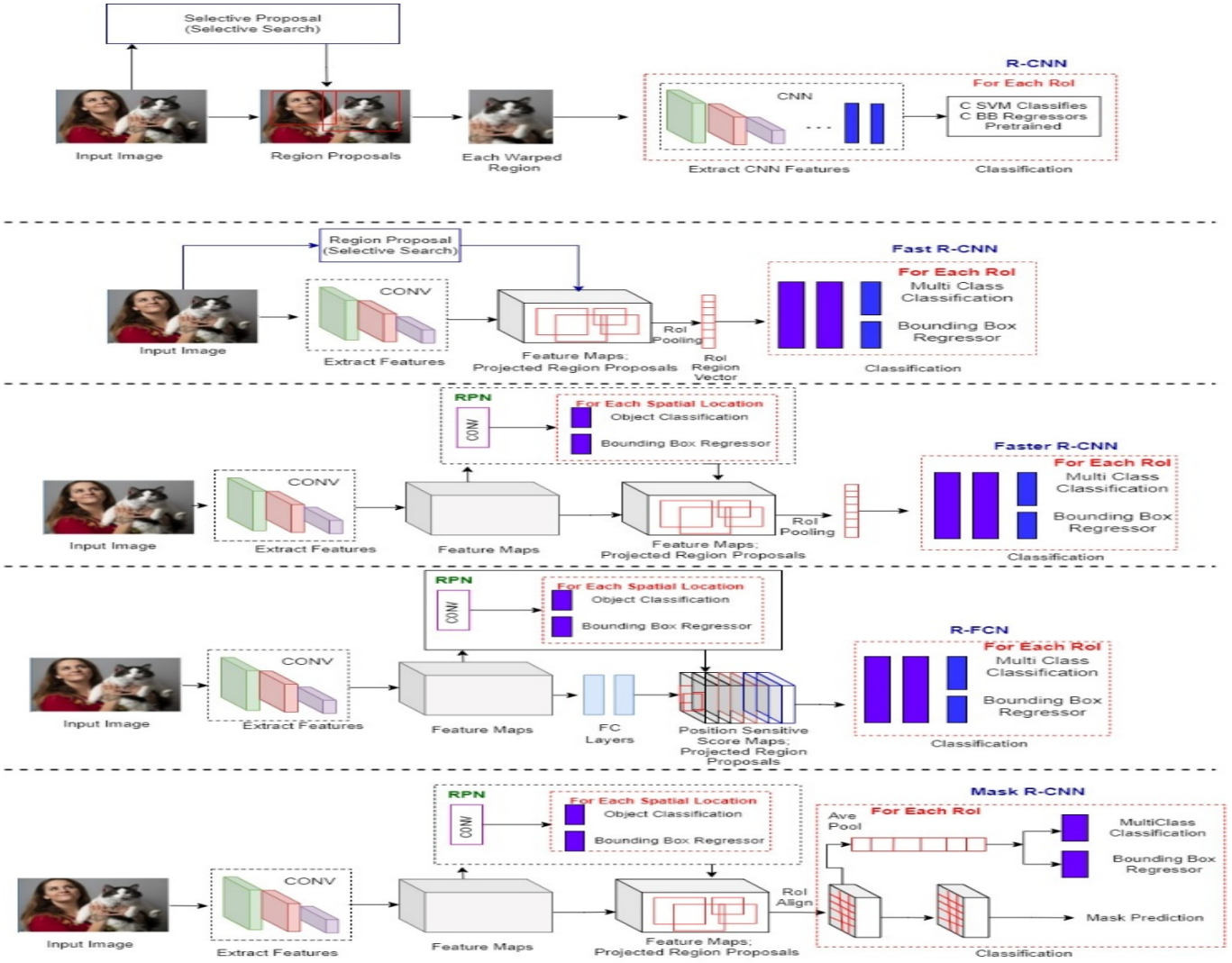
Introduction

- Deep convolutional neural networks (CNNs) have enabled significant advances in object detectors.
- Provides a comparative performance analysis of many two-stage object detectors:
  1. Region-Convolution Neural Network (R-CNN)
  2. Spatial Pyramid Pooling (SPP)
  3. Fast R-CNN
  4. Faster RCNN
  5. Feature Pyramid Network (FPN)
  6. R-FCN
  7. Mask R-CNN
  8. Cascade R-CNN
  9. DetectoRS
  10. Neural Architecture Search-Feature Pyramid Network (NAS-FPN)

Summary of How Detectors Work

Detector	Summary of how it works
R-CNN	Uses a selective search algorithm to generate candidate object regions, which are then classified and localized using a convolutional neural network (CNN).
SSP-NET	A single-stage object detector that uses a CNN to predict bounding boxes and class probabilities for each pixel in an image.
Fast R-CNN	An improved version of R-CNN that uses a shared convolution layer for all regions of interest (ROIs), which speeds up computation.
Faster R-CNN	A further improvement over R-CNN that uses a region proposal network (RPN) to generate ROIs, which further speeds up computation.
R-FCN	A single-stage object detector that uses a CNN to predict fully-convolutional networks (FCNs) for each pixel in an image. These FCNs can then be used to predict bounding boxes and class probabilities for all objects in the image.
FPN	A feature pyramid network that combines features from different layers of a CNN to improve the accuracy of object detection for small objects.
Mask R-CNN	An extension of Faster R-CNN that can also segment objects in an image.
NAS-FPN	A neural architecture search (NAS) method for finding the optimal FPN architecture for a given task.
DetectoRS	A modular framework for object detection and other computer vision tasks. DetectoRS provides a variety of different detectors, including R-CNN, Fast R-CNN, Faster R-CNN, R-FCN, and Mask R-CNN.

Two-Stage Object Detectors

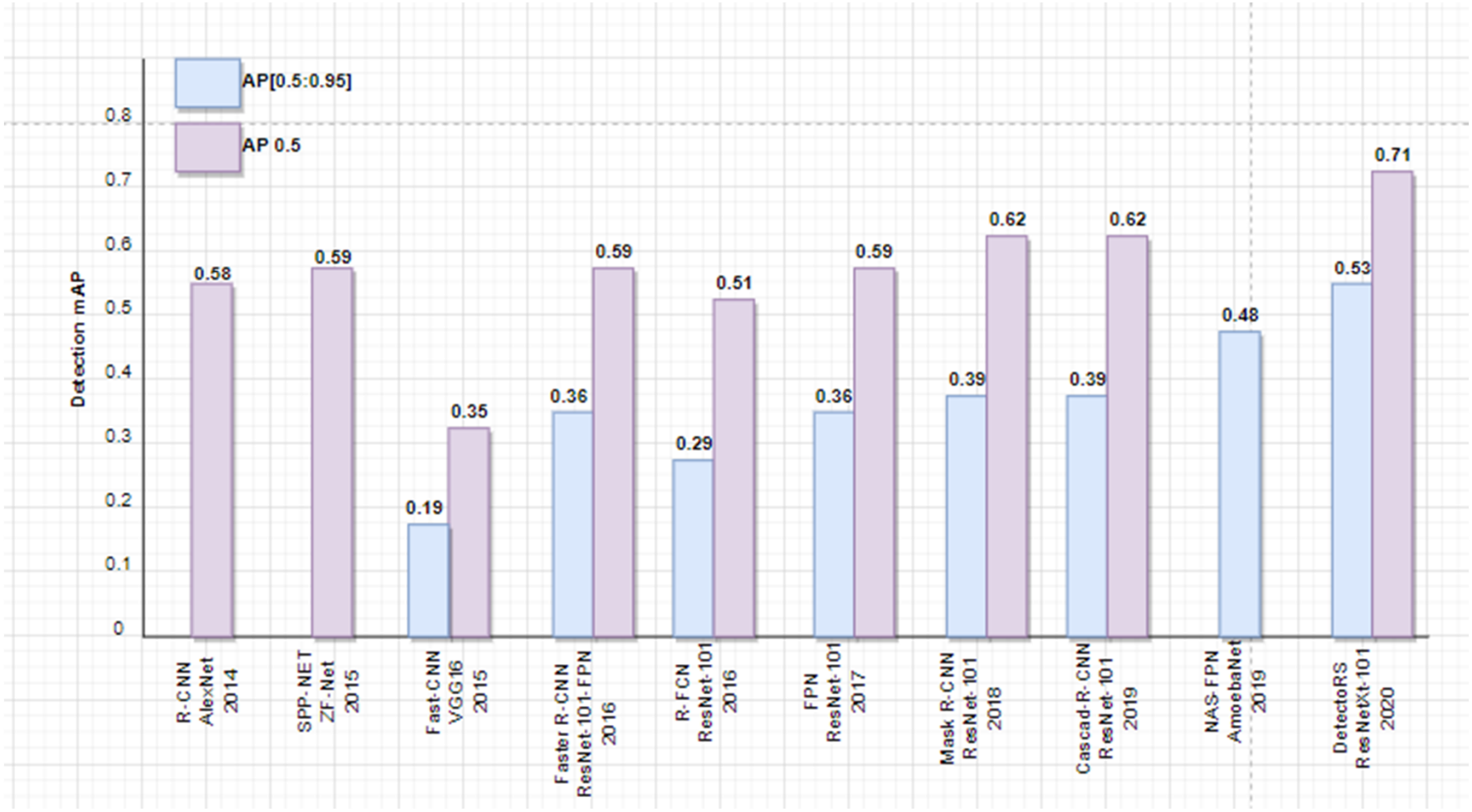


COMPARATIVE PERFORMANCE ANALYSIS OF TWO-STAGE OBJECT DETECTORS

TABLE I. TWO-STAGE OBJECT DETECTORS PERFORMANCE COMPARISON ON MS COCO AND PASCAL VOC 2012 DATASETS AT SIMILAR INPUT IMAGE SIZES FOR THE TWO-STAGE OBJECT DETECTORS.

Detector & year	Backbone	Image Size	AP[0.5:0.95]	AP0.5
R-CNN, 2014	AlexNet	224	-	58.50%
SSP-NET, 2015	ZFNet	Variable	-	59.20%
Fast-R-CNN, 2015	AlexNet, VGG16	Variable	-	65.70%
Faster-R-CNN, 2016	ZFNet, VGG	600	-	67.00%
R-FCN, 2016	ResNet101	600	31.50%	53.20%
FPN, 2017	ResNet-101	800	36.20%	59.10%
Mask-R-CNN, 2018	ResNetXt101, FPN	800	39.80%	62.30%
NAS-FPN, 2019	ResNet-50	1280	48.3	-
DetectoRS, 2020	ResNeXt-101	1333	53.30%	71.60%

A COMPARATIVE ANALYSIS OF THE PERFORMANCE OF TWO-STAGE OBJECT DETECTORS.



Detectors, Pros, Cans, and Application

Detectors	Pros	Cons	Applications
R-CNN	High accuracy	Slow	Object detection and instance segmentation
Fast R-CNN	Faster than R-CNN	Less accurate than R-CNN	Object detection and instance segmentation
Faster R-CNN	Fast and accurate	Requires a lot of training data	Object detection and instance segmentation
R-FCN	Faster than Faster R-CNN	Less accurate than Faster R-CNN	Object detection and instance segmentation
FPN	Improves accuracy for small objects	More complex than other detectors	Object detection and instance segmentation
Mask R-CNN	Can detect and segment objects in a single stage	Less accurate than Cascade R-CNN	Object detection and instance segmentation
Cascade R-CNN	Very accurate	Slow	Object detection and instance segmentation
NAS-FPN	Searches for the optimal FPN architecture for a given task	Can be time-consuming to train	Object detection and instance segmentation
DetectoRS	A modular framework for object detection and other computer vision tasks	Can be complex to use	Object detection, instance segmentation, keypoint detection, and more

Merit and Limitations of Detectors

Detector & year	Merit and Limitations
R-CNN 2014	**Merit: H**as improved performance on the **PASCAL VOC** datasets better than **HOG**-based methods. **Limitation**: It is slow and expensive to train because of its sequentially trained multistage pipeline.
SSP-NET 2015	**Merit**: accelerates R-CNN without sacrificing performance. **Limitation:** SPP-Net inherits the disadvantages of R-CNN
Fast-R-CNN 2015	**Merit**: Enhances performance over SPPNet by designing RoI pooling layer and eliminating disc storage for features. **Limitation**: External RP computation becomes a bottleneck, and real-time applications are sluggish.
Faster-R-CNN 2016	**Merit: I**ntroduces multi-scale regression anchor boxes, making it faster than Fast RCNN without sacrificing performance. **Limitation:** Real-time detection is slow, and training is hard due to the sequential training process.
R-FCN 2016	**Merit It** is a fully convolutional detector network that is faster than Faster R-CNN. **Limitation**: is still too slow for real-time use, and the training process is not streamlined.
FPN 2017	**Merit:** FPN is significantly faster and improved over several competition winners using **densely sampled image pyramids**. **Limitation**: FPN is computationally expensive due to the use of **densely sampled image pyramids**.
Mask-R-CNN 2018	**Merit**: IT is a refined version of the Faster R-CNN framework that can **perform instance segmentation** with an additional branch for mask detection in parallel with the BB prediction branch. **Limitation**: Falls short of real-time applications due to its computational complexity.
NAS-FPN 2019	**Merit: It** exceeds Mask R-CNN with **less computation time** and achieves **2mAP accuracy in mobile detection**, because of a combination of top-down and bottom-up connections. **Limitation**: NAS-FPN is **still slow for real-time applications.**
DetectoRS 2020	**Merit**: Makes a significant difference in **efficiency and effectiveness** by achieving state-of-the-art accuracy for **object identification and instance segmentation**. **Limitation:** DetectoRS is still unsuitable for real-time detections due to its computational complexity.

Conclusion and Future Work

Conclusion

- Presented a comparative performance analysis of two-stage object detectors
- Evaluated the performance of different detectors on two datasets, MSCOCO and PASCAL VOC 2012.
- Using the average precision AP0.5 and AP[0.5:95] on both datasets.
- The results showed that DetectoRS outperformed all other two-stage models
- DetectoRS achieved an AP0.5 of 53.30% and an AP[0.5:95] of 71.60% on MSCOCO.
- DetectoRS achieved an AP0.5 of 83.00% and an AP[0.5:95] of 90.30% on PASCAL VOC 2012. However, it is also more complex.
- Other two-stage object detectors that performed well in the comparison include:
- NAS-FPN, Mask R-CNN, and Cascade R-CNN.
- These models also use various techniques to improve performance, such as region proposal networks (RPNs), RoIAlign, and focal loss.

Future research should focus:

- Improving the speed of two-stage detectors without sacrificing accuracy
- Developing anchor-free detectors that are as accurate as anchor-based detectors but more computationally efficient.

REFERENCES

1. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes (voc) challenge,” International Journal of Computer Vision. vol. 88, pp. 303–338, Jun 2010.
2. S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, pp. 1137–1149, June 2017.
3. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788, June 2016.
4. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks in Proc. Adv. Neural Inf.Process. Syst., 2015, pp. 91_99.
5. R. Girshick, “Fast R-CNN,” ICCV in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440_1448.
6. J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, “Selective search for object recognition,” Int.J. Comput. Vis., vol. 104, no. 2, pp. 154_171, Sep. 2013.
7. R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” In 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587, June 2014.
8. A. Krizhevsky, I. Sutskever, and G. E. Hinton, Image net classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process Syst., 2012, pp. 1097_1105.
9. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2 Dpose estimation using part af_nity_elds,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 7291_7299.
10. Z. Cai, and N. Vasconcelos, “Cascade R-CNN: High-Quality Object Detection and Instance Segmentation, arXiv:1906.09756 [cs.CV], June 2019.
11. K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition,” IEEE Trans. Pattern Anal. Mach. Intel., vol. 37, no. 9, pp. 1904_1916, Sep. 2015.
12. F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and<0.5MB model size,” arXiv:1602.07360v4, 2016.
13. T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf.Comput. Vis. Pattern Recognit.,” Jul. 2017, pp. 2117_2125.
14. K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc.IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 2961_2969.
15. J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, “Selective search for object recognition,” Int.J. Comput. Vis., vol. 104, no. 2, pp. 154_171, Sep. 2013.
16. M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. “The Pascal Visual Object Classes Challenge 2012 (voc2012) Results (2012),” http://www.pascalnetwork.org/challenges/VOC/voc2011/workshop/index.html.
17. T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in Proc. 13th Eur. Conf. Comput. Vis. (ECCV). Zürich, Switzerland: Springer, Sep. 2014, pp. 740_755.
18. J. Dai, K. He, and J. Sun, “Instance-aware semantic segmentation via multi-task network cascades,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 3150_3158.
19. J. Nan and L. Bo, “Infrared object image instance segmentation based on improved mask-RCNN,” Proc. SPIE, vol. 11187, Nov. 2019, Art. no. 111871E.
20. A. O. Vuola, S. U. Akram, and J. Kannala, “Mask-RCNN and U-Netensembled for nuclei segmentation,” in Proc. IEEE 16th Int. Symp.Biomed. Imag. (ISBI), Apr. 2019, pp. 208_212.
21. J. Li, X. Liang, J. Li, Y. Wei, T. Xu, J. Feng, and S. Yan, “Multistage object detection with group recursive learning,” IEEE Trans. Multimedia, Vol. 20, no. 7, pp. 1645_1655, Jul. 2018.
22. G. Ghiasi, T. Y. Lin, and Q. V. Le, “NAS-FPN: Learning scalable feature pyramid architecture for object detection,” in Proc. IEEE Conf. ComputVis. Pattern Recognit., Jun. 2019, pp. 7036_7045.
23. L. Aziz, M. S. B. Haji Salam, U. U. Sheikh, and S. Ayub, “Exploring Deep Learning-Based Architecture, Strategies, Applications and Current Trends in Generic Object Detection: A Comprehensive Review, in IEEE Access, vol. 8, pp. 170461-170495, 2020, doi: 10.1109/ACCESS.2020.3021508.
24. S. Qiao, L.-C. Chen, and A. Yuille, “DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution,” http://arxiv.org/abs/2006.02334, 2021.
25. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deep Lab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” available: http://arxiv.org/abs/1606.00915.
26. I. Krylov, S. Nosov, and V. Sovrasov, “Open Images V5 Text Annotation and Yet Another Mask Text Spotter, https://doi.org/10.48550/arXiv.2106.12326.
27. B. Hariharan, R. Girshick, K. He, and P. Dollar, “Scalable, high-quality object detection using deep learning,” In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 418–426).
28. C. L. Zitnick, P. Dollár, “Edge boxes: Efficiently detecting salient object edges,” “In European conference on computer vision. Springer.
29. L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, Mask R-CNN. arXiv preprint arXiv:1703.06870.