Two-Stage Object Detectors: A Comparative Evaluation

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

***Abstract*— Object detection is a fundamental task in computer vision with many applications, such as self-driving cars, security, and medical imaging. Recent advances in deep learning have led to significant improvements in the performance of object detectors. This paper presents a comparative performance analysis of generic object detectors, focusing on two-stage detectors. Two-stage detectors are a type of object detector that first generates region proposals and then classifies and refines those proposals. The paper first provides an overview of the taxonomy of two-stage object detection algorithms. It then presents a detailed performance comparison of two-stage detectors on two datasets, Microsoft COCO and PASCAL VOC 2012. The results show that DetectoRS is a state-of-the-art two-stage object detector, outperforming all other two-stage models. However, it is also more complex. The more practical of the two-stage object detectors that performed well in the comparison are Neural Architecture Search-Feature Pyramid Network (NAS-FPN), cascade R-CNN, and Mask R-CNN.**

Keywords-deep learning; object detection; computer vision; two-stage detectors; and performance analysis.

I. Introduction

Object detection is a fundamental computer vision task that identifies and localizes objects in images or videos. It has recently received a great deal of attention due to its wide range of applications. Deep convolutional neural networks (CNNs) have enabled significant advances in object detection. CNNs can learn powerful features from images, which can be used to identify and localize objects accurately. This paper provides a comparative performance analysis of two-stage object detectors, such as Fast R-CNN [5], Spatial Pyramid Pooling (SPP) [11], Region-Convolution Neural Network (R-CNN) [7], Faster R-CNN 2], Mask R-CNN [21], and DetectoRS [24]. We also discuss the trade-offs in object detection. We believe that this paper will be valuable to researchers and practitioners who are interested in object detection.

The rest of this paper is organized as follows. Section II provides background on object detection. Section III discusses two-stage detectors. Section IV presents a comparative performance analysis of two-stage detectors. Section V concludes the paper with the future direction.

II. Background

Object detection is a fundamental task in computer vision that has been revolutionized by deep learning in recent years. Deep learning models can learn powerful features from images, which can be used to identify and localize objects accurately. Object detection can be divided into four tasks:

- Object classification: Assign a class label to an object, such as "car" or "person."
- Object localization: Predict the bounding box of an object.
- Semantic segmentation: Assign each pixel in an image to a class label.
- Object instance segmentation: Predict the bounding box and class label of each object instance in an image.

Object detection is used in a wide range of applications, such as self-driving cars, video surveillance, and medical imaging.

A. *Structure of Target Detection*

Object detection can be divided into two approaches: region proposal-based detectors and single-stage detectors. Region proposal-based detectors generate region proposals (bounding boxes) and then classify each proposal into an object category [8]. Single-stage detectors directly predict the bounding boxes and class labels of objects in an image.

B. *Historical Roadmap Taxonomy of Object Detectors*

The development of object detection can be divided into two historical periods:

- Before 2012: This period is often called the traditional object detection period. During this time, object detection algorithms were primarily based on handcrafted features and shallow machine learning models.
- 2012 and after: This period is often called the deep learning-based detection period. During this time, object detection algorithms have been revolutionized by the introduction of deep learning models, such as R-CNN [7].

C. *Challenges in object detection*

One of the biggest challenges in object detection is dealing with complex scenes. Complex scenes may contain many objects, some of which may be partially occluded or

ISBN: 978-1-68558-104-6

overlapping. Additionally, the objects in a scene may vary in size and scale. Another challenge in object detection is real-time performance. For many applications, such as self-driving cars, it is important to be able to detect objects in real time. However, deep learning models can be computationally expensive, which can make it difficult to achieve real-time performance.

D. Future of object detection

The future of object detection is bright. As deep learning models continue to improve, object detection algorithms will become more accurate and efficient. This will enable object detection to be used in a wider range of applications.

Additionally, researchers are exploring new ways to improve the performance of object detection algorithms. For example, some researchers are developing new deep learning architectures that are specifically designed for object detection. Others are developing new training techniques to help deep learning models learn to detect objects more effectively.

Overall, object detection is a rapidly evolving field with a bright future. As deep learning models continue to improve, object detection algorithms will become more accurate, efficient, and versatile.

III. Two-stage Object Detectors

Region-based object detection is inspired by the human visual system, which scans images and focuses on regions of interest. R-CNN [7] was the first region-based detector to show that CNNs are better than handcrafted features, such as HOG, for object detection. In this paper, we review many two-stage detectors that have been proposed since R-CNN.

A. R-CNN object detection

The R-CNN object detection model [7] is a region-based approach that was the first to demonstrate the superiority of convolutional neural networks (CNNs) over handcrafted features, such as HOG.

R-CNN works as follows:

1. Region proposal generation: R-CNN uses selective search [15] to generate 2000 region proposals from an image.
2. Feature extraction: R-CNN extracts 4096-dimensional features from each region proposal using a pre-trained CNN.
3. Classification and localization: R-CNN uses a linear SVM to classify each region proposal and predict its bounding box.

R-CNN has several limitations, including:

- Slow testing speed: R-CNN has to recalculate the CNN for each region proposal, which adds to the testing time.
- Time-consuming training: R-CNN has to fine-tune the CNN on a dataset of region proposals.
- High memory usage: R-CNN has to store the features extracted from each region proposal.
- Prone to overfitting: R-CNN is prone to overfitting, as the region proposals generated by selective search are not always accurate.
- Object localization errors: R-CNN uses bounding boxes to localize objects, which can lead to errors, as the boxes may not be perfectly aligned with the objects.

Researchers have proposed several solutions to these limitations, such as:

- The MCG system [7]: This system uses a variety of techniques to generate region proposals, which helps to reduce the risk of overfitting.
- The GOP system [27]: This system uses a geodesic-based segmentation technique to split the voters, which helps to improve object localization.
- Edge box techniques: These techniques return objects with fewer outlines crossing their bounds, which helps to reduce object localization errors [28].
- Pre-extracted reranking: This method removes duplicate region proposals from the recommendation lists, which helps to improve the accuracy of object detection.
- Semantic segmentation [29]: This technique can be used to improve object localization by providing more accurate information about the objects in an image.

The R-CNN object detection model is a landmark paper in the field of computer vision. It was the first to show that CNNs could be used to achieve state-of-the-art results in object detection. However, R-CNN has several limitations, such as slow testing speed and high memory usage. Researchers have proposed several solutions to these limitations, which have led to the development of more efficient and accurate object detection models.

B. SPP-Net object detection

He et al. [11] proposed Spatial Pyramid Pooling (SPP)-Net to address the limitations of R-CNN, such as the loss of object content and geometric deformation caused by cropping and wrapping.

SPP-Net uses spatial pyramid pooling to create a new CNN design that allows the SPP layer to be reused for different region proposals, regardless of their size. This makes SPP-Net more efficient and scalable than R-CNN.

SPP-Net has been shown to achieve better results than R-CNN, especially when the corresponding scale of different region proposals is precisely determined. However, SPP-Net can be slower than R-CNN at test time due to the pooling computation expenses.

Overall, SPP-Net is a significant improvement over R-CNN, and it has laid the foundation for many modern object detection algorithms.

ISBN: 978-1-68558-104-6

C. *Fast R-CNN object detection*

Fast R-CNN [5] addresses the limitations of R-CNN and SPP-Net, such as the need to train different systems individually and the high storage capacity requirements. Fast R-CNN works as follows:

1. Feature extraction: Fast R-CNN extracts a single feature map from the entire image using a CNN.
2. Region proposal generation: Fast R-CNN uses a region proposal network (RPN) to generate region proposals from the feature map.

Classification and localization: Fast R-CNN uses a single linear Support Vector Machine (SVM) to classify each region proposal and predict its bounding box.

Fast R-CNN is more efficient and accurate than R-CNN and SPP-Net, and it has become the basis for many modern object detection algorithms.

D. *Faster R-CNN object detection*

Faster R-CNN [4] addresses the limitations of previous object detection algorithms, such as the need for external region proposal generation methods and the slow speed of Fast R-CNN. Faster R-CNN introduces a region proposal network (RPN) that is fully integrated into the CNN architecture. The RPN generates region proposals directly from the CNN feature maps, which eliminates the need for external region proposal generation methods. Faster R-CNN also uses a single-stage training procedure, which further improves speed. In a single pass, the RPN generates region proposals, and the CNN classifies and localizes the objects in the region proposals.

Faster R-CNN has achieved state-of-the-art results on many object detection benchmarks, and it has become the basis for many modern object detection algorithms.

E. *Feature Pyramid Network (FPN)*

FPN is a deep convolutional network that can generate high-level semantic features of varying sizes. It is a flexible and powerful tool for computer vision tasks. It can be used in various applications, including object detection, instance segmentation, key point detection, image classification, and semantic segmentation. FPN uses a top-down approach to combine features from higher and lower levels of the network. This allows FPN to preserve high-level semantic information while also providing fine-grained details. FPN can be used with any CNN architecture and has been shown to improve performance on various computer vision tasks.

One of the main advantages of FPN is that it can achieve state-of-the-art performance on object detection tasks. This is because FPN can generate features at multiple scales, which allows it to detect objects of different sizes. FPN is also not tied to a specific CNN architecture, which makes it more flexible and adaptable. This means that FPN can be used with various CNN architectures. FPN has been shown to be effective in various computer vision applications. For example, FPN has been used to improve the performance of object detectors on the Microsoft COCO and Pascal VOC datasets. FPN has also been used to improve the performance of instance segmentation algorithms on the Cityscapes dataset.

Overall, FPN is a powerful and versatile tool for computer vision tasks. It is easy to implement and can be used with any CNN architecture.

F. *R-FCN object detector*

Region-based Fully Convolutional Network (R-FCN) [26] uses fully connected layers that share almost all processing across the entire image instead of convolutional layers for object detection. This makes R-FCN faster and more efficient than previous region-based detectors, such as Faster R-CNN.

R-FCN addresses the translation invariance problem by using position-sensitive score maps. These score maps are generated for each object category, and they indicate the likelihood of an object of that category being present at a particular location in the image. The position-sensitive score maps are combined with region proposals to generate bounding box predictions.

R-FCN can be easily adapted to fully convolutional image classifier backbones, such as Residual Networks (ResNets). This makes it easy to train and deploy R-FCN models. R-FCN has been shown to achieve competitive performance on the PASCAL VOC datasets. For example, R-FCN with ResNet-101 achieves 83.6% mAP on the 2007 set.

Finally, R-FCN consists of four convolutional networks:

1. The input image is passed through a CNN to obtain feature maps.
2. The feature maps are then passed to a region proposal network (RPN) to identify potential object locations.
3. The potential object locations are then passed to the R-FCN network, which generates position-sensitive score maps.
4. The position-sensitive score maps are then used to classify and regress the bounding boxes of the objects.

G. *Mask R-CNN object detector*

Mask R-CNN [14] is a deep learning algorithm that can perform both object detection and instance segmentation. It is an extension of Faster R-CNN and adds a branch for each region of interest (ROI) to predict segmentation masks. The mask branch is a small, fully convolutional network (FCN) that is added to each RoI and predicts a pixel-by-pixel segmentation mask. The FCN is trained to predict a binary mask for each pixel, indicating whether the pixel belongs to the object or not. Mask R-CNN uses a two-stage approach to instance segmentation:

1. The first stage uses the Region Proposal Network (RPN) to generate a set of candidate RoIs.
2. The second stage uses the mask branch to predict segmentation masks for each RoI.

Mask R-CNN also introduces a new RoI pooling layer called RoIAlign, which is designed to improve the alignment of RoIs with the original image regions. RoIAlign uses bilinear interpolation to sample the feature map at the RoI's center and four corners, which results in a more accurate alignment than the traditional RoI pooling layer because it preserves the spatial information of the RoI.

Mask R-CNN has been shown to be very effective, for instance, segmentation, achieving state-of-the-art results on several benchmarks. It is a simple and efficient algorithm that can be easily extended to other object detection and instance segmentation tasks, such as pedestrian detection, car detection, and instance segmentation of medical images.

Advantages of Mask R-CNN:

- Accurate and efficient instance segmentation.
- Can be easily extended to other object detection and instance segmentation tasks.
- Simple to implement.

Applications of Mask R-CNN:

- Object detection.
- Instance segmentation.
- Pedestrian detection.
- Car detection.
- Instance segmentation of medical images.

Finally, Mask R-CNN is a powerful and versatile instance segmentation algorithm that can be used for various tasks. It is easy to implement and can be extended to other object detection and instance segmentation tasks.

H. Cascade R-CNN

Cascade R-CNN [10] is an object detection model that uses a cascade of detectors to gradually increase the quality of hypotheses while ensuring that all detectors have access to a positive training set of similar size. This technique eliminates the quality mismatch between hypotheses and detectors during inference, which can lead to overfitting and reduced inference speed.

Cascade R-CNN consists of a series of detectors that are trained with increasing intersection of union (IoU) thresholds. This means that the first detector is trained to only detect objects with high IoU scores, while the second detector is trained to detect objects with lower IoU scores, and so on. This allows the Cascade R-CNN to gradually increase the quality of hypotheses while ensuring that all detectors have access to a positive training set of similar size.

Cascade R-CNN has been shown to be effective in reducing and eliminating overfitting. Overfitting occurs when a model learns the training data too well and is unable to generalize to new data. The Cascade R-CNN addresses this problem by training the detectors in a sequential manner. This means that the first detector is only trained on the most difficult examples, while the later detectors are trained on easier examples. This helps to prevent the model from overfitting to the training data. In addition to reducing overfitting, the Cascade R-CNN also improves the speed of inference. This is because the later detectors only need to process the examples that were not detected by the earlier detectors. This can significantly reduce the amount of time it takes to process an image.

Overall, Cascade R-CNN is a promising approach to object detection. It has been shown to be effective in reducing overfitting and improving the speed of inference. As a result, it is a promising technique for achieving high-quality object detection. Some of the applications of Cascade R-CNN are object detection, self-driving cars, robotics, and surveillance.

Finally, Cascade R-CNN is a powerful and versatile object detection model that can be used for various tasks. It is easy to implement and can be extended to other object detection tasks.

I. DetectoRS

Detection and Retrieval System (DetectoRS) [24] is a new object detection and retrieval system that combines the Recursive Feature Pyramid (RFP) and Switchable Atrous Convolution (SAC) techniques. It achieves state-of-the-art accuracy for object detection and instance segmentation on the COCO test-dev platform, with 55.7% box accuracy, 48.5% mask accuracy, and 50.0% PQ for panoptic segmentation. Key Features of DetectoRS include:

- Recursive Feature Pyramid (RFP): RFP is a new feature pyramid network that provides additional feedback connections from the Feature Pyramid Networks (FPN) into the bottom-up backbone layers. This allows for more efficient processing and allows the network to learn to use different atrous rates for different objects.
- Switchable Atrous Convolution (SAC): SAC is a new type of convolution that allows the network to learn to use different atrous rates for different parts of an object. This is useful for detecting objects of different sizes and shapes.

Advantages of DetectoRS include:

- State-of-the-art accuracy: DetectoRS achieves state-of-the-art accuracy on the MSCOCO test-dev platform for object detection, instance segmentation, and panoptic segmentation.
- Efficiency: DetectoRS is an efficient object detection system due to the use of RFP and SAC.
- Flexibility: DetectoRS is built on top of the Faster R-CNN framework and uses the ResNet-50 backbone network. This makes it flexible and adaptable to different needs.

Applications of DetectoRS include:

- instance segmentation: DetectoRS can be used, for instance, for segmentation tasks, such as medical image segmentation and scene parsing.
- Panoptic segmentation: DetectoRS can be used for panoptic segmentation tasks, which involve simultaneously detecting and segmenting all objects in an image.

Finally, DetectoRS is a powerful and versatile object detection system that can be used for a variety of tasks. It achieves state-of-the-art accuracy and efficiency and is built on top of a flexible and adaptable framework.

J. NAS-FPN

Neural Architecture Search-Feature Pyramid Network (NAS-FPN) [22] is a modified version of neural architecture search (NAS) that allows for feature fusion at different scales through top-down and bottom-up connections. It achieves state-of-the-art accuracy on object detection tasks while using less computation time than other methods.

Key Features of NAS-FPN

- Feature fusion at different scales: NAS-FPN uses a combination of top-down and bottom-up connections to fuse features from different scales. This allows the network to learn a more comprehensive representation of the input image, which leads to improved accuracy.
- Neural architecture search: NAS-FPN uses NAS to automatically search for the optimal network architecture. This allows the network to be tailored to the specific task at hand, which can lead to further improvements in accuracy and efficiency.

Advantages of NAS-FPN

- State-of-the-art accuracy: NAS-FPN achieves state-of-the-art accuracy on object detection tasks, outperforming other methods such as SSD and Mask R-CNN.
- Efficiency: NAS-FPN is more efficient than other methods, such as SSD and Mask R-CNN, while still achieving state-of-the-art accuracy.
- Flexibility: NAS-FPN can be used with various backbone networks, such as ResNet-50 and AmoebaNet. This makes it flexible and adaptable to different needs.

Applications of NAS-FPN

- Object detection: NAS-FPN can be used for a variety of object detection tasks, such as self-driving cars, robotics, and surveillance.
- Instance segmentation: NAS-FPN can be used for instance segmentation tasks, such as medical image segmentation and scene parsing.

Finally, NAS-FPN is a powerful and versatile object detection system that achieves state-of-the-art accuracy and efficiency. It is built on top of a flexible and adaptable framework, making it a good choice for a variety of tasks.

IV. Comparative Performance Analysis of Two-Stage Object Detectors

Two-stage object detectors are a type of object detector that uses two stages to detect objects in an image. The first stage typically involves generating a set of region proposals, and the second stage involves classifying and refining the bounding boxes of the proposed regions. Two-stage object detectors have been shown to achieve state-of-the-art performance on object detection tasks. However, there are a variety of different two-stage object detectors available, and it can be difficult to choose the best one for a particular task.

This paper presents a comparative performance analysis of several popular two-stage object detectors. It used the MSCOCO and PASCAL VOC 2012 datasets to evaluate the performance of the detectors. It also used the following metrics to evaluate the performance of the detectors:

- Average precision (AP): AP is a measure of the accuracy of an object detector. It is calculated by averaging the precision of the detector at different recall levels.
- AP0.5: AP0.5 is the AP when the predicted bounding box Intersection over Union (IoU) is greater than 0.5, and the ground truth.
- AP[0.5:0.95]: AP[0.5:0.95] is the average AP for IoU values from 0.5 to 0.95 in steps of 0.5.

The results of the comparative performance analysis are shown in Table 1 and Figure 1. The comparative performance analysis shows that DetectoRS is a state-of-the-art two-stage object detector. It achieves high accuracy on both the COCO and PASCAL VOC 2012 datasets, and it can handle a variety of object sizes and scales.

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 a variety of techniques to improve their performance, such as region proposal networks (RPNs), RoIAlign, and focal loss.

Overall, the results show that two-stage object detectors can achieve high accuracy on a variety of datasets and tasks. However, they can also be computationally expensive. As a result, it is important to choose the right model for the specific task at hand.

The following Table 1 and Figure 1 illustrate the comparative parameter values for different detectors using MSCOCO and Pascal VOC 2012 datasets using the average precision (AP) metric. The results showed that DetectoRS outperformed all other two-stage models in both AP0.5 and AP[0.5:95] on both datasets.

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	Merit and Limitations
R-CNN [7] 2014	AlexNet	224	-	58.50%	**Merit:** Faster R-CNN has improved performance on the PASCAL VOC datasets than HOG-based methods. **Limitation:** Faster R-CNN is slow to train because of its sequentially trained multistage pipeline, and training is expensive in terms of storage and time.
SSP-NET [11] 2015	ZFNet	Variable	-	59.20%	**Merit:** SPP-Net accelerates R-CNN without sacrificing performance. **Limitation**: SPP-Net inherits the disadvantages of R-CNN and only provides a small improvement in results.
Fast-R-CNN [5] 2015	AlexNet, VGGm, VGG16	Variable	-	65.70%	**Merit: Faster R-CNN enhances performance over SPPNet by designing RoI pooling layer and eliminating disc storage for features.** **Limitation: External RP computation becomes a bottleneck, making real-time applications sluggish.**
Faster-R-CNN [3] 2016	ZFNet, VGG	600	-	67.00%	**Merit**: Faster R-CNN proposes RPN and introduces 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 [12] 2016	ResNet101	600	31.50%	53.20%	**Merit:** Mask R-CNN is a fully convolutional detector network that is faster than Faster R-CNN. **Limitation**: Mask R-CNN is still too slow for real-time use, and the training process is not streamlined.
FPN [13] 2017	ResNet-101	800	36.20%	59.10%	**Merit:** FPN is significantly faster and improved over several competition winners by using densely sampled image pyramids. **Limitation**: FPN is computationally expensive due to the use of densely sampled image pyramids.
Mask-R-CNN [14] 2018	ResNetXt101, ResNet101, FPN	800	39.80%	62.30%	**Merit:** Mask R-CNN 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**: Mask R-CNN falls short of real time applications due to its computational complexity.
NAS-FPN [22] 2019	ResNet-50	1280	48.3	-	**Merit**: NAS-FPN exceeds Mask R-CNN with less computation time and achieves 2mAP accuracy in mobile detection, thanks to its combination of top-down and bottom-up connections. **Limitation**: NAS-FPN is still slow for real-time applications.
DetectoRS [24] 2020	ResNeXt-101	1333	53.30%	71.60%	**Merit**: DetectoRS makes a significant difference in terms of 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.

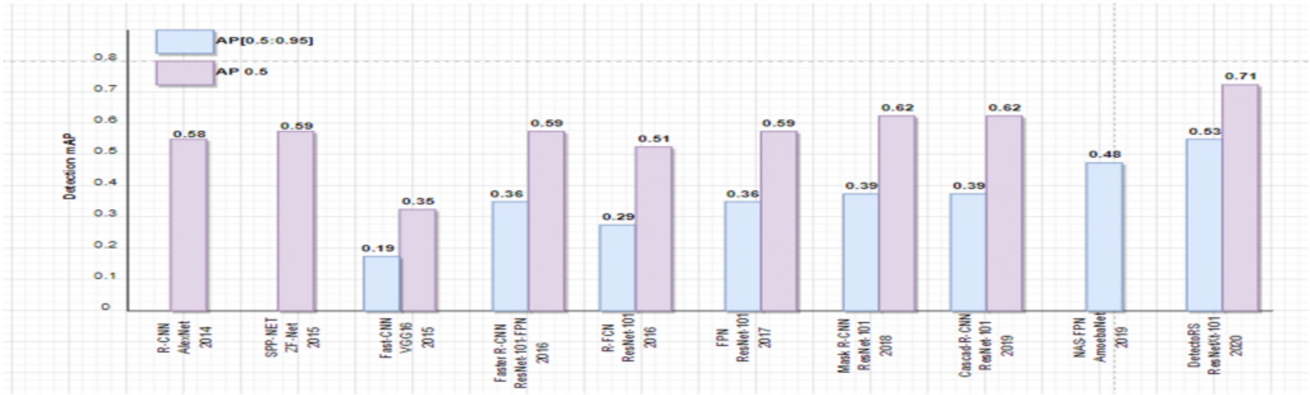

Figure 1. On the MSCOCO and PASCAL VOC2012 datasets, A comparative analysis bar graph of the performance of various two-stage object detectors.

ISBN: 978-1-68558-104-6

V. CONCLUSION AND FUTURE WORK

This paper presented a comparative performance analysis of two-stage object detectors, which are state-of-the-art in object detection accuracy. The paper evaluated the performance of different detectors on two different datasets, MSCOCO and PASCAL VOC 2012, using the average precision (AP) metric. The results showed that DetectoRS outperformed all other two-stage models in both AP0.5 and AP[0.5:95] on both datasets. DetectoRS achieved an AP0.5 of 53.30% and an AP[0.5:95] of 71.60% on MSCOCO, and 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 a variety of techniques to improve their performance, such as region proposal networks (RPNs), RoIAlign, and focal loss.

Future research in object detection and recognition should focus on 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.