

# Dissecting of the two-stages object detection models architecture and performance

Sara Bouraya, Abdessamad Belangour

Laboratory of Information Technology and Modeling, Faculty of Sciences Ben M'sik, Hassan II University, Casablanca, Morocco

## Article Info

### Article history:

Received Apr 7, 2023

Revised Jul 18, 2023

Accepted Sep 27, 2023

### Keywords:

Computer vision

Convolutional neural network

Deep learning

Deep neural networks

Neck models

Object detection

Two stage detectors

## ABSTRACT

Artificial intelligence (AI) is the discipline focused on enabling computers to operate autonomously without explicit programming. Within AI, computer vision is an emerging field tasked with endowing machines with the ability to interpret visual data from images and videos. Over recent decades, computer vision has found applications in diverse fields such as autonomous vehicles, information retrieval, surveillance, and understanding human behavior. Object detection, a key aspect of computer vision, employs deep neural networks to continually advance detection accuracy and speed. Its goal is to precisely identify objects within images or videos and assign them to specific classes. Object detection models typically consist of three components: a backbone network for feature extraction, a neck model for feature aggregation, and a head for prediction. The focus of this study lies on two stage detectors. This study aims to provide a comprehensive review of two stage detectors in object detection, followed by benchmarking to offer insights for researchers and scientists. By analyzing and understanding the efficacy of these models, this research seeks to guide future developments in the field of object detection within computer vision.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Sara Bouraya

Laboratory of Information Technology and Modeling, Faculty of Sciences Ben M'sik

Hassan II University

Casablanca, Morocco

Email: sarabouraya95@gmail.com

## 1. INTRODUCTION

Object detection is often called image detection, object identification, and object recognition; and all these concepts are synonymous (Figure 1). It is a computer vision method for locating instances of objects in an image or video sequence. Object detection algorithms, therefore, typically benefit from machine learning techniques or deep learning techniques to gain meaningful results. When humans look at images or videos, they could locate and recognize objects of interest easily. The goal of object detection is to mimic this intelligence using a computer. With recent advancements in deep learning-based computer vision models, object detection use cases are spreading more than ever before. A wide range of applications is implemented, for instance, self-driving cars, object tracking, anomaly detection, and video surveillance.

The paper explores two-stage detector models, focusing on their relevance and advancements within the field of object detection. In the related works section, existing research is reviewed to contextualize the study. Background details fundamental concepts, including deep neural networks and model architecture. comparison evaluates various models based on performance metrics. Results present empirical findings, while discussion interprets and discusses implications. Finally, the conclusion summarizes key findings and suggests future research directions.

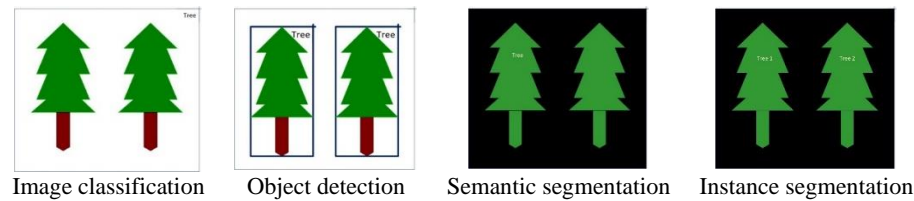


Figure 1. Comparison of visual recognition tasks in computer vision

## 2. RELATED WORKS

Several scientific works and research have been implemented to develop and evolve object detection applications and systems and depend on enormous methodologies of the deep learning era, machine learning era, and, other eras. Several researchers and scientists are expanding their implementation and research to develop and apply enormous methodologies (Figure 2). Such is the case of feature aggregation methods that are used to make a connection between low and high features for better object recognition in video sequences and images. Feature aggregation is used widely in action recognition [1]–[5], and video description [6], [7]. Most of these methods use recurrent networks (RNNs) to aggregate features from consecutive frames on the one hand. Exhaustive temporal-spatial convolution is used to extract temporal-spatial features, on the other hand. U-Net [8] was proposed to concatenate features from low-level to high-level for medical image segmentation, and it achieved great success in that field. To gain an outstanding feature for object detection, the feature pyramid networks (FPN) aggregated both the transformed feature from the bottom-up weighted pyramid and the top-down lateral convolutions through a simple sum operation. Relying on feature pyramid networks, several extensive works [9]–[12] define a new option for connectivity between scales. Attention-based models also prove their efficiency in several applications of deep learning era [13]–[18]. Self-attention models by measuring and apply. The unified architecture of two-stage detector methodologies (Figure 3) typically consists of three main components: a backbone network, a proposal generation stage, and a refinement stage. These methodologies are commonly used in object detection tasks to localize and classify objects in an image.

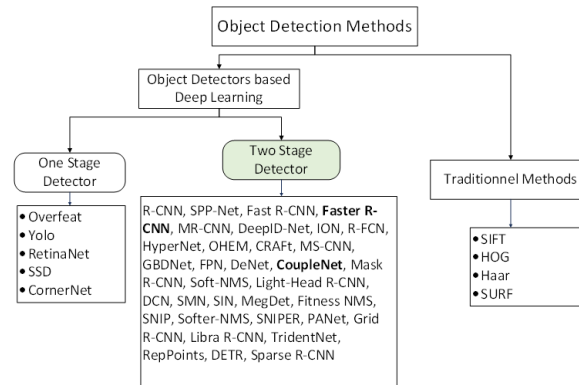


Figure 2. Taxonomy of two-stage detector models based on deep learning

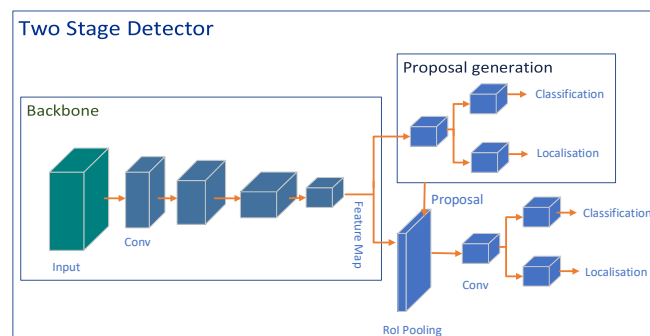


Figure 3. The unified architecture of two-stage detector methodologies

### 3. BACKGROUND

Faster region-based convolutional neural network (R-CNN) [19], as the name faster R-CNN refers is an extension of fast R-CNN, as well as the name, suggests faster R-CNN is faster than its previous fast R-CNN which emphasizes the strongness of the region proposal network (RPN). By the use of RPN which refers to a fully convolutional network that is responsible for generating proposals with different aspect ratios and various scales. In their paper, they introduce the anchor boxes concept, rather than the use of pyramids of filters. An anchor box is a specific aspect and scale ratio reference (Figure 4).

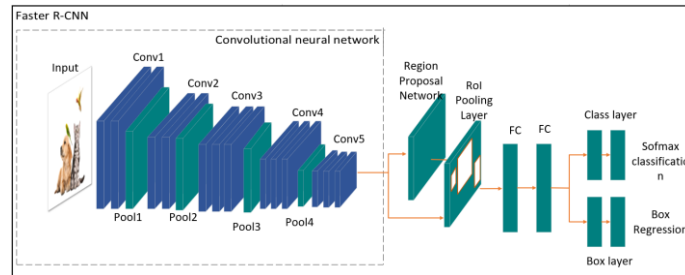


Figure 4. Faster R-CNN architecture

CoupleNet [20] is an object detection that gathers the object proposals gathered by RPN and then fed them into the coupling module that combines two branches. The first branch captures the local part feature of an object using position-sensitive RoI (PSRoI), and the other branch for encoding the context and global features using RoI pooling. The ResNet-101 is used as a backbone for removing the FC layer, and average pooling. Then each proposal is fed into two branches global fully convolutional network (FCN) and local FCN. Then, finally, both local and global FCNs are combined to produce the final result (Figure 5).

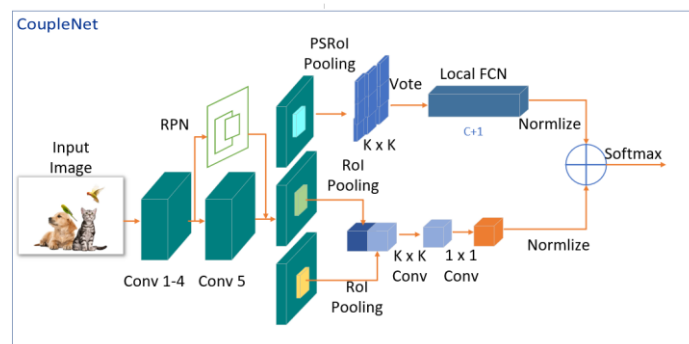


Figure 5. CoupleNet architecture

Fast R-CNN [21], as the name refers fast R-CNN is an extension of R-CNN, and it overcomes several of its issues. As the name refers to the fast R-CNN is faster than R-CNN. Fast R-CNN proposed a layer called region of interest or ROI pooling which tries to extract feature vectors from proposals. Compared to the R-CNN model, which covers multiple stages starting from region proposal generation then feature extraction, and finally classification using support vector machine (SVM), faster R-CNN uses just one neural network that has only just one stage. Faster R-CNN spread convolutional layer calculations across all proposals. By making use of ROI pooling layer that makes fast R-CNN faster and more accurate than R-CNN. The fast R-CNN model does not cache extracted features which decrease the use of disk storage compared to R-CNN (Figure 6).

SPP-Net [1], is one of the convolutional neural network (CNN) models that utilize spatial pyramid pooling which removes the fixed size of the neural network. On top of the last layer, a SPP is added, which pools the features as well as generates a fixed length of outputs that will be used in a fully connected layer. For avoiding the need for wrapping and cropping, they perform information aggregation at a deeper stage of the neural network (Figure 7).

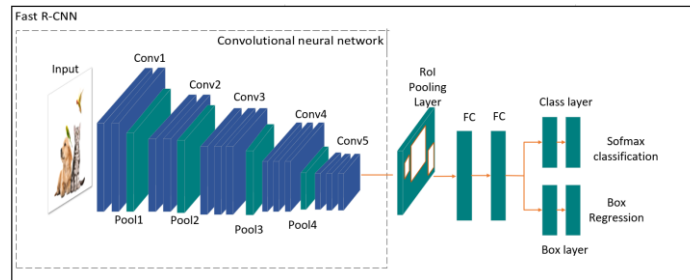


Figure 6. Fast R-CNN architecture

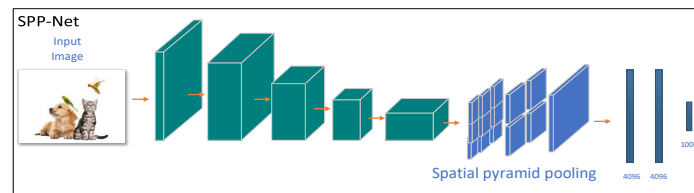


Figure 7. SPP-Net architecture

RepPoints detector (RPDet) [2] is one of the two-stage detectors, which are based on deformed CNNs and it is an anchor-free model. RepPoints is used as the sample and basic object representation inside the object detection system. The starting RepPoints are acquired based on regressing offsets over center points. The learning process of these gained RepPoints is driven by two goals: the object recognition loss of the stages as well as the bottom right and top left points distance loss among the ground truth and the induced pseudo box (Figure 8).

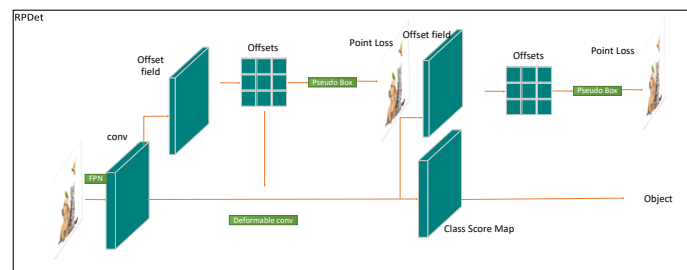


Figure 8. RepPoints architecture

Libra R-CNN [3] is an object detector that tries to reach a balanced training process, not like the past detectors that have suffered from an imbalanced training process, which in general combines three different levels sample level, feature level as well as an objective level. Libra R-CNN covers three other different levels: IoU-balanced sampling, balanced feature pyramid, and finally balanced L1 loss to reduce the imbalance at the feature stage, sample stage, and objective stage (Figure 9).

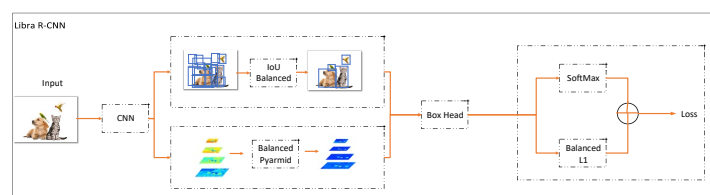


Figure 9. Libra R-CNN architecture

Multi-region CNN (MR-CNN) [4] is an object representation utilizing several regions to gather various aspects of an object. The first stage consists of passing an image through an activation map module and getting an activation map. The different bounding box candidates or region proposals are generated utilizing the selective search. Additionally, VGG-16 is used as a backbone and the last max pooling is removed (Figure 10).

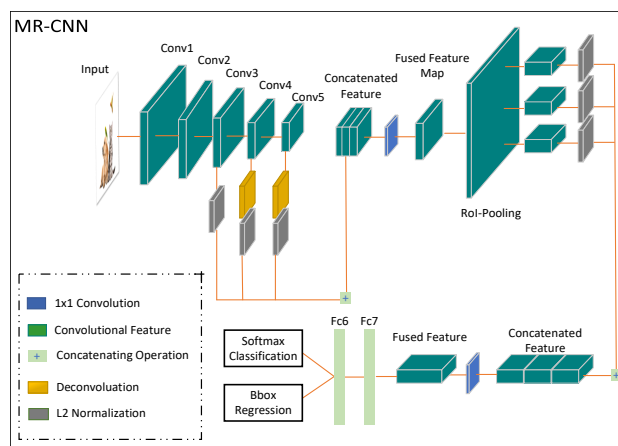


Figure 10. MR-CNN architecture

DeepID-Net [5] is an object detection belonging to multi-stage models and deformable CBBs, that have multiple innovations in several aspects, in the new proposed architecture, a new structure pooling layer is proposed. The integration of multiple classifiers optimized the path samples at several stages and levels. As well as defining a new training approach to learn the deep feature representation for reaching an important generalization capability and a good object detection result. This model improves the modeling by relying on several techniques including changing the neural network structure, as well as the training strategies by adding some stages and removing others inside the detection pipeline. Which gives us a crucial diversity of performant models (Figure 11).

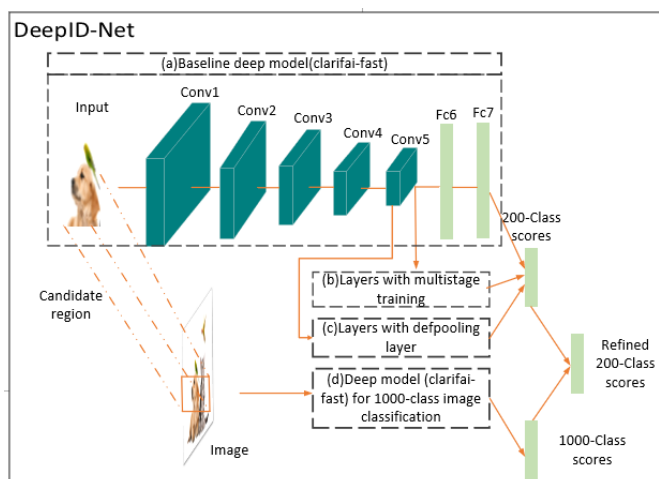


Figure 11. DeepID-Net architecture

Region based fully convolutional network (R-FCN) [6] is an object detection region based. This model is a fully convolutional network that shares the computation on the entire image, in comparison to the other ones such as fast/faster R-CNN which are region-based object detection that applies a subnetwork many times. To reach this R-FCN uses position-sensitive score maps to address an issue between translation variance in object detection and translation invariance in image classification (Figure 12).

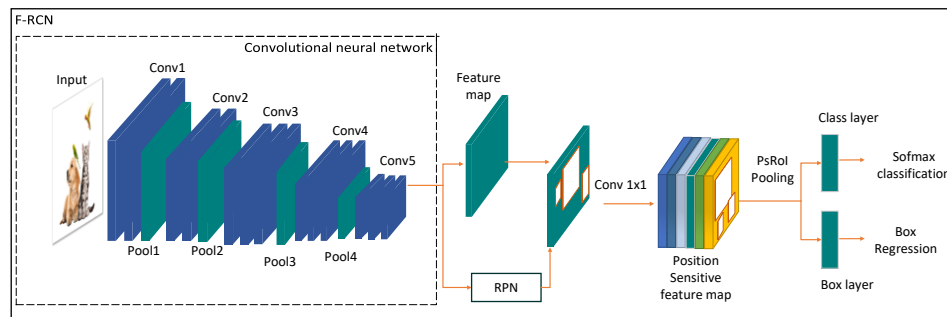


Figure 12. R-FCN architecture

Grid R-CNN [7], is an object detection model, where a grid point-guided localization is made in the traditional regression formulation place. The model divides the bounding box region into grids and then implements the FCN. Grid R-CNN gathers the grid point location and the explicit special information due to fully CNN architecture and they can be obtained in pixel levels. Based on the grid points, this model can detect performant bounding boxes. This grid vision can make it better than the regression methods (Figure 13).

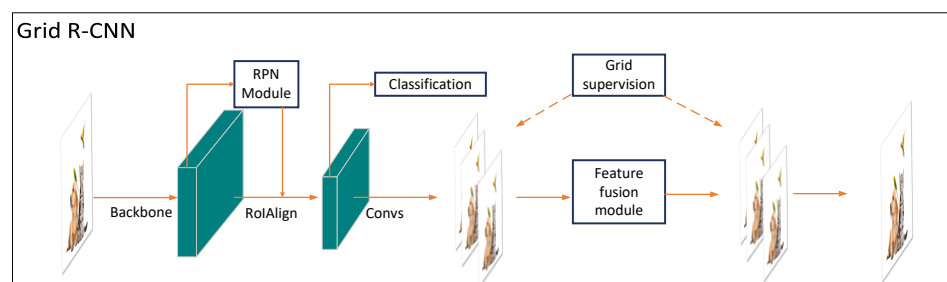


Figure 13. Grid R-CNN architecture

TridentNet [8], is an object detection model that is used to tackle scale variation issues. Both categories of object detection one or two-stage detectors don't handle scale variation. Indeed, there are different ways to solve that issue but the problem it increases the time inference (Figure 14).

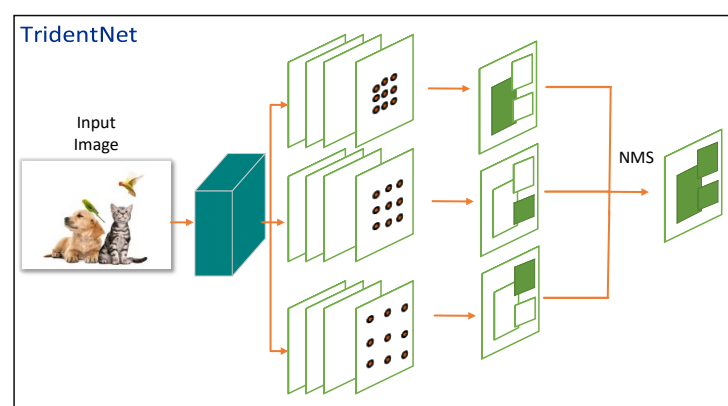


Figure 14. TridentNet architecture

ION [9], is an object detection model that tries to utilize information both outside the region of interest and inside the region of interest. Regarding the information outdoors the region of interest is integrated utilizing a special recurrent neural network (RNN). Indoor, skip pooling is used to extract and gather features at multiple levels and scales of abstraction (Figure 15).

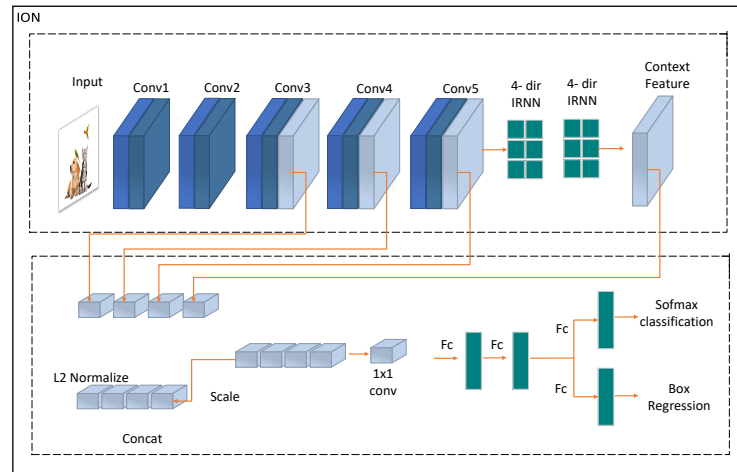


Figure 15. ION architecture

Gated bi-directional CNN (GBD-Net) [10], is an object detection model that is implemented under fast R-CNN, which concentrates on bi-directional CNN named GBD-Net to transfer features between different regions inside two stages of feature extraction and feature learning. Those features transferred can be implemented based on convolution among neighbored regions and transferred in two directions among several layers. Thus, contextual and local patterns can emphasize the existence of their self by gathering the nonlinear relationships as well as their complex interactions. This model affirms that message passing and transferring is not always helpful. The messages transmitted are controlled by the use of gated functions. Model handed used a set of backbones for instance ResNet and inceptions as feature extraction models (Figure 16).

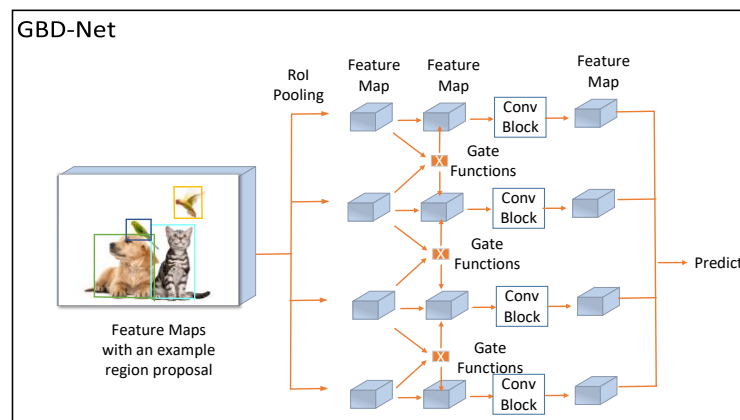


Figure 16. GBD-Net architecture

Mask R-CNN [11], is an object detection model that is used for object detection as well as instance segmentation. It is used for generating segmentation masks for each instance. It is an extension of faster R-CNN, which add a branch used for predicting objects' masks. This makes Mask R-CNN easy to be implemented by using just an overhead (Figure 17).

Light-head R-CNN [12], the heavy-head design of the two-stage object detection model makes them slow in comparison with two-stage detectors. Light-head R-CNN is trying to tackle this shortcoming of two-stage detectors, by making and designing the head as light as possible, by utilizing a thin feature map as well as a cheap R-CNN combining fully connected layers and pooling. ResNet-101 is used as a backbone (Figure 18).

Structure inference network (SIN) [13], combined faster R-CNN with a graphical model that tries to infer object state. SIN model makes vital the act of taking into consideration not only visual appearance but also making use of object interaction as well as scene information. This model transforms the object detection issue into a graph structure inference. The objects are seen as nodes among a graph and the relation among them is seen as edges (Figure 19).

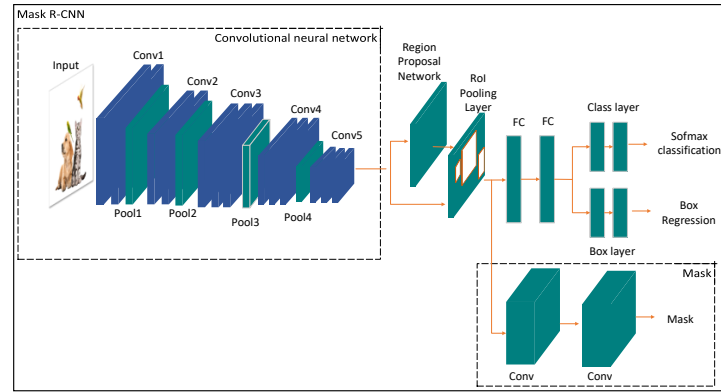


Figure 17. Mask R-CNN architecture

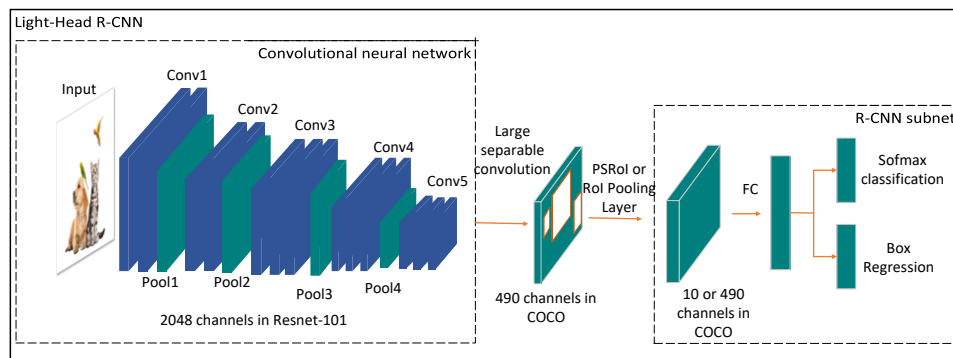


Figure 18. Light-head R-CNN architecture

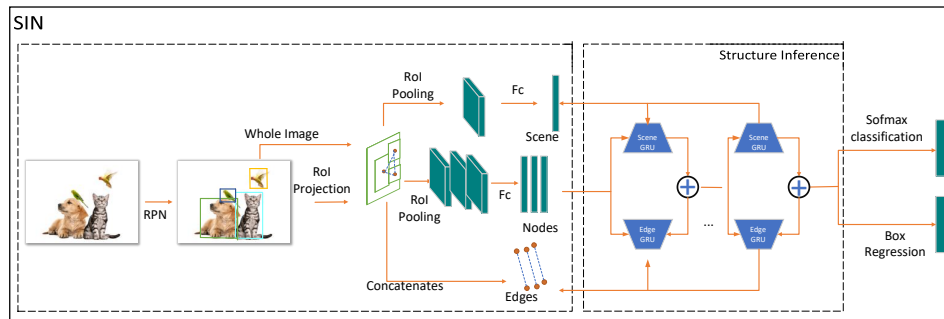


Figure 19. SIN architecture

Detection transformer (DETR) [14], tries to remove the need for some stages such as non-maximum suppression as well as anchor generation that encode the prior knowledge. The new model stages are a set-based global loss that forces unique prediction, as well as a transformer encoder-decoder model. This model, DETR, reasons about the relation between the context which is the image, and the object to output a set of predictions. DETR is conceptually easy and can be simply investigated to perform panoptic segmentation (Figure 20).

HyperNet [15], is an object detection model that employs region proposals to control an object's instance search. To get high recall among regions proposal methods need enormous proposals, which hurts the object detection efficiency. As well as these models are still struggling with small-size objects. HyperNet aims to handle object detection jointly and region proposal generation. HyperNet tries to aggregate feature maps and then compresses them into a unique uniform space (Figure 21).

Multi-scale CNN (MS-CNN) [16], is a unified deep neural network, which is proposed to tackle fast and multi-scale detection. MS-CNN proposed two sub-networks a detection one and a proposal one. In the first sub-network, proposal one the detection is reached at multiple output layers, so that receptive match targets of several scales. These scale detectors produce a multi-scale object detection model. Feature upsampling is reached by deconvolution to reduce computation costs and memory (Figure 22).



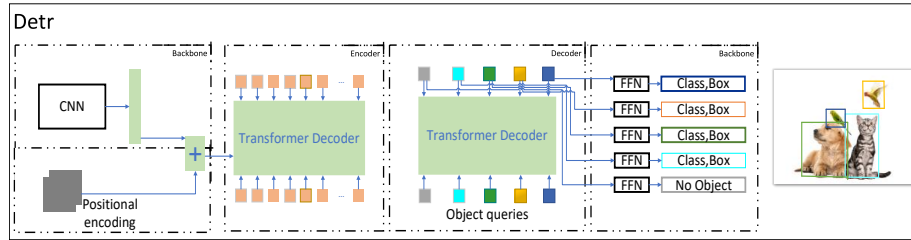


Figure 20. DETR architecture

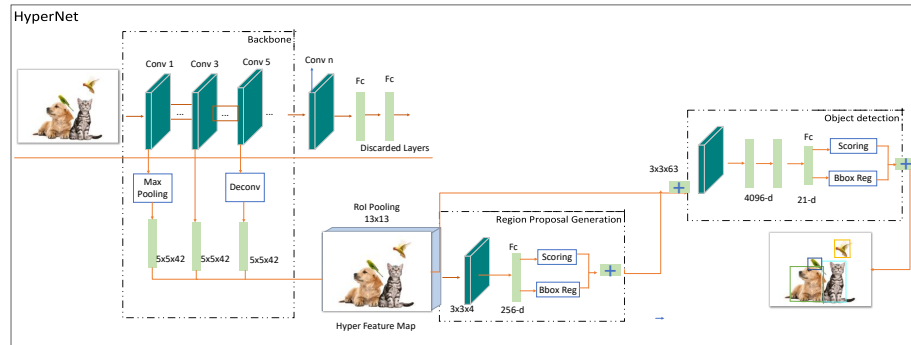


Figure 21. HyperNet architecture

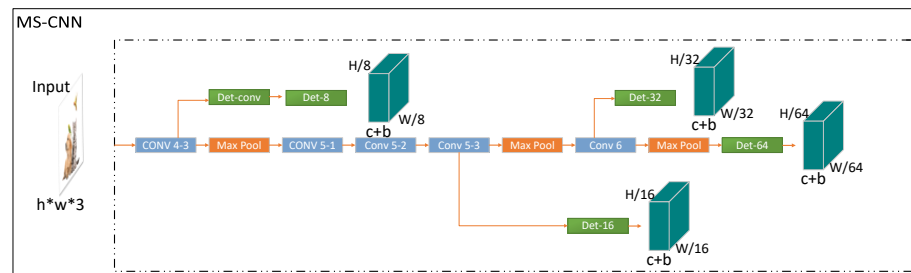


Figure 22. MS-CNN architecture

#### 4. COMPARISON

The table handed contains famous one-stage object detection models that are trained on the common object in context (COCO) dataset result the different metrics (Figure 23). Table 1 combines the different models with their backbone models which are used for feature extraction as well as feature fusion models that are used for feature fusion. Additionally, to evaluate metrics at the same time papers' names and their published years.

##### Average Precision (AP):

AP % AP at IoU=.50:.05:.95 (primary challenge metric)  
 AP<sub>IoU=.50</sub> % AP at IoU=.50 (PASCAL VOC metric)  
 AP<sub>IoU=.75</sub> % AP at IoU=.75 (strict metric)

##### AP Across Scales:

AP<sub>small</sub> % AP for small objects: area < 322  
 AP<sub>medium</sub> % AP for medium objects: 322 < area < 962  
 AP<sub>large</sub> % AP for large objects: area > 962

Source Information : <https://cocodataset.org/#detection-eval>

Figure 23. Evaluation metrics evaluation

Table 1. Different two stage detector models relied on different evaluation metrics

Model	Backbone+Neck model	AP	AP50	AP75	APS	APM	APL	Paper	Year
Deformable DETR	ResNeXt-101+DCN	52.3	71.9	58.1	34.4	54.4	65.6	Deformable DETR: Deformable Transformers for End-to-End Object Detection [17]	2021
RepPoints v2	ResNeXt-101, DCN, multi-scale	52.1	70.1	57.5	34.5	54.6	63.6	RepPoints V2: Verification Meets Regression for Object Detection [19]	2020
Trident Net	ResNet-101-Deformable, image pyramid	48.4	69.7	53.5	31.8	51.3	60.3	Scale-Aware Trident Networks for Object Detection [8]	2019
PANet	ResNeXt-101, multi-scale	47.4	67.2	51.8	30.1	51.7	60.0	Path Aggregation Network for Instance Segmentation [19]	2018
SNIPER	ResNet-101	46.1	67.0	51.6	29.6	48.9	58.1	SNIPER: Efficient Multi-Scale Training [20]	2018
Mask R-CNN	HRNetV2p-W48+cascade	46.1	64.0	50.3	27.1	48.6	58.3	Deep High-Resolution Representation Learning for Visual Recognition [21]	2021
Faster R-CNN	LIP-ResNet-101-MD w FPN	43.9	65.7	48.1	25.4	46.7	56.3	LIP: Local Importance-based Pooling [22]	2019
SNIPER	ResNet-50	43.5	65.0	48.6	26.1	46.3	56.0	SNIPER: Efficient Multi-Scale Training [20]	2018
D-RFCN+SNIP	ResNet-101, multi-scale	43.4	65.5	48.4	27.2	46.5	54.9	An Analysis of Scale Invariance in Object Detection-SNIP [23]	2018
Grid R-CNN	ResNeXt-101-FPN	43.2	63.0	46.6	25.1	46.5	55.2	Grid R-CNN [7]	2019
Libra R-CNN	ResNeXt-101-FPN	43.0	64	47	25.3	45.6	54.6	Libra R-CNN: Towards Balanced Learning for Object Detection [3]	2019
Trident Net	ResNet-101	42.7	63.6	46.5	23.9	46.6	56.6	Scale-Aware Trident Networks for Object Detection [8]	2019
Faster R-CNN	HRNetV2p-W48	42.4	63.6	46.4	24.9	44.6	53.0	Deep High-Resolution Representation Learning for Visual Recognition [21]	2021
RPDet	ResNet-101	41	62.9	44.3	23.6	44.1	51.7	RepPoints: Point Set Representation for Object Detection [2]	2019
Mask R-CNN	ResNet-101-FPN, CBN	40.1	60.5	44.1	35.8	57.3	38.5	Cross-Iteration Batch Normalization [24]	2021
Fast R-CNN	Cascade RPN	40.1	59.4	43.8	22.1	42.4	51.6	Cascade RPN: Delving into High-Quality Region Proposal Network with Adaptive Convolution [25]	2019
ION	ResNeXt-101+DCN	33.1	55.7	34.6	14.5	35.2	47.2	Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks [9]	2016

## 5. RESULTS

After plotting the main table that contains the different models Table 1, we get the handed plot (Figure 24) which emphasize the strongness of the newer model, such as deformable detr based on ResNeXt-101-DCN as well as RepPoints-V2 which is relied on ResNeXt-DCN also.

### 5.1. Based on average precision

Box average precision (AP): AP % AP at IoU=50:05:95 (primary challenge metric). In terms of AP, deformable DETR has reached the highest score of 52.3 as mentioned in the figure. Deformable DETR which combines ResNeXt-101 and DCN as the backbone in addition to RepPoints v2 which occupied the second position with a difference of 0.02 which relied on ResNeXt-101, DCN, multi-scale as a backbone (Figure 25).

AP50: AP IoU=50% AP at IoU=50 (PASCAL VOC metric). As for the AP metric, Deformable DETR occupied the first position in terms of AP50 based on the ResNeXt-101 and DCN as the backbone. As well as RepPoints v2 is reaching the second position with a difference of 1.8 (Figure 25).

AP75: % AP at IoU=75 (strict metric). Regarding AP75, deformable DETR with the same backbone, as mentioned in AP and AP50, occupied the first position with 58.1. The second position is taken by RepPoints v2 with a difference of 0.6 (Figure 25).

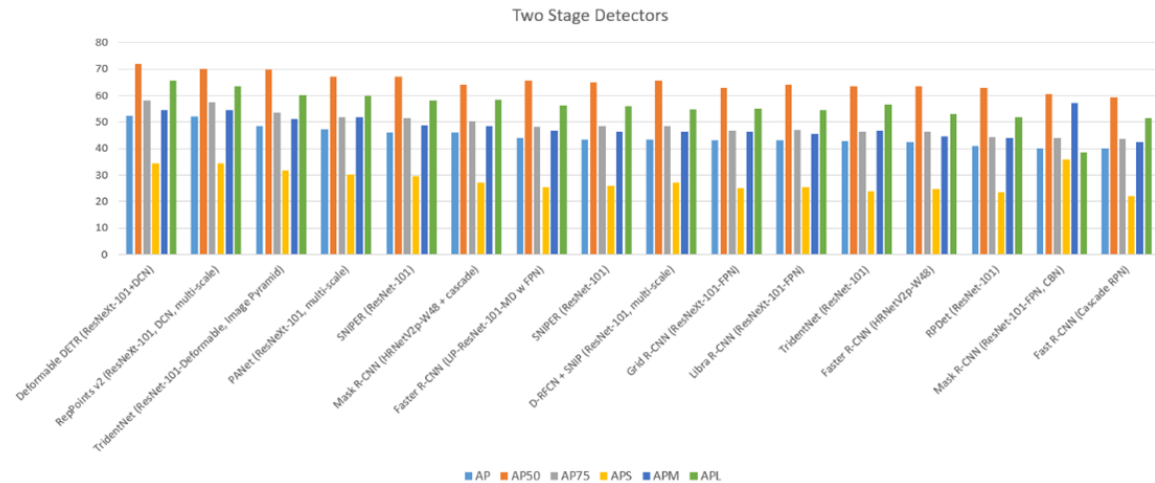


Figure 24. Two-stage detectors comparison plot based on different evaluation metrics

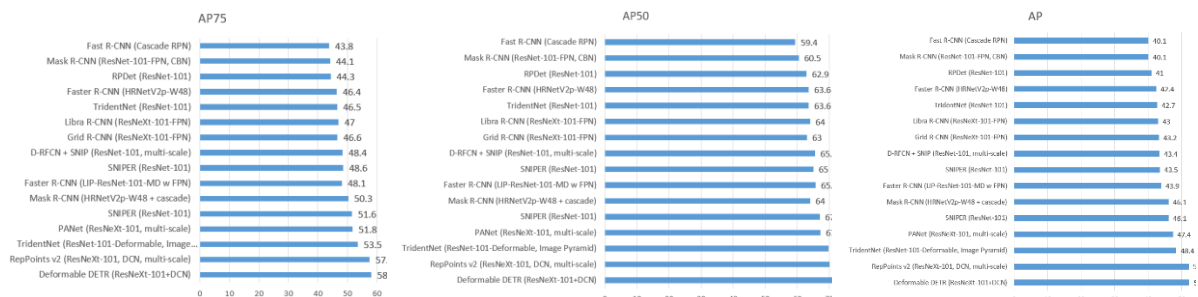


Figure 25. Two-stage detectors comparison based on AP, AP50, and AP75

## 5.2. Based on AP across scales

Average precision small (APS): % AP for small objects: area <322. This is the first time that we recognize that another model other than Deformable DETR reached the high score in terms of APS which is Mask R-CNN based on (ResNet-101-FPN, CBN) as a backbone and feature aggregation model. RepPoints V2 is saving its place with a 0.1 difference (Figure 26). Average precision medium (APM): % AP for medium objects: 322 < area <962. Regarding the APM evaluation metric Mask R-CNN with the same combined models occupied the first position, in addition to RepPoints v2 which save its second place this time also with a difference of 2.7. Average precision large (APL): % AP for large objects: area >962 see Figure 26. Deformable DETR with ResNeXt-101 as a backbone reached 65.6 in its first position additionally RepPoints v2 with ResNeXt-101 as a backbone reached 63.6 with a difference of 2 see Figure 26.

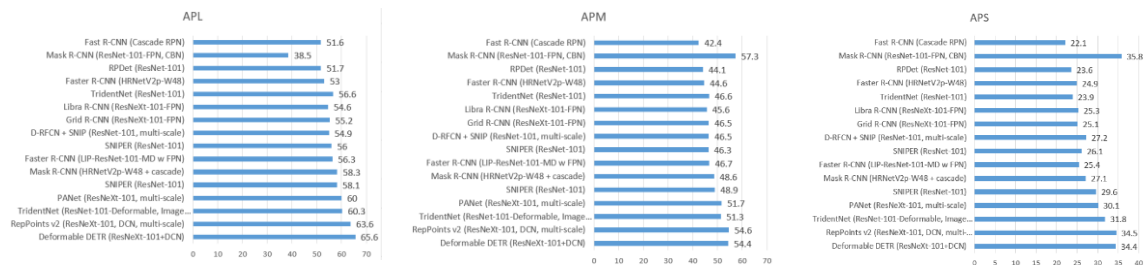


Figure 26. Two-stage detectors comparison based on APL, APM, APS

## 6. DISCUSSION

We present the widespread two-stage object detection methods. This paper covers several interesting models, we started by listing the different major branches of computer vision, image classification, object detection, semantic segmentation, and instance segmentation. After clarifying the main differences among the cited branches, we presented the part of the related work which presented the same works but after deep research, for gathering the related articles there are no studies that focus on just one category of object detection models as we do here just concentrate on two-stage object detection models.

In our third stage, we analyze each two-object detector model separately by offering the main architecture as well as a detailed description of the main components utilized starting from input then the backbone after that the neck and the head model. These components are the most common components combined. At the end of these stages and after discussing and analyzing these models' architecture, we dive into a benchmark table that combined the enormous models additionally to their evaluation metrics' score based on AP such as box AP, AP50, AP75, and those relied on across scale such as APS, APM, APL, all the cited models are implemented on COCO dataset.

After gathering a detailed benchmarking, we visualize the results carefully based on AP and across scale metrics each one separately. The results are analyzed relying on the handed plots. After visualizing and discussing the best results we can conclude that deformable DETR relied on ResNext-101 and DCN as a backbone as well as RepPoints V2 which based on the same backbone are heading the listed models in terms of different metrics. Except that Mask-RCNN based on ResNet-101 for feature extraction and FPN as a model for feature fusion deals better in terms of APS, and APM. From the noticed results we emphasize that using ResNeXt-101 with DCN are constructing a great team in term of AP and terms of across scale metrics Mask R-CNN is dealing better which clarify the strongness of ResNet-101-FPN-CNB as an architecture. The handed results going to help us in constructing other models that are inspired carefully by the best of the cited models to gain better performance. Such as the case of using ResNext-101.

## 7. CONCLUSION

In conclusion, this paper has presented an extensive overview of two-stage object detection methods, covering various models within the realm of computer vision. Beginning with an exploration of major branches such as image classification, object detection, semantic segmentation, and instance segmentation, we proceeded to delve into a comprehensive review of related works, focusing specifically on two-stage object detection models. Through meticulous analysis, each two-stage detector model was dissected, elucidating their architectures and key components, including input, backbone, neck, and head models. These components represent the fundamental building blocks shared across most models.

Furthermore, we conducted a detailed benchmarking, evaluating the performance of these models on the COCO dataset using metrics such as AP and across scale APS, APM, APL. Visualization of the results facilitated a nuanced discussion, revealing standout performers such as deformable DETR and RepPoints V2, both leveraging ResNext-101 with DCN as a backbone. notably, Mask-RCNN, utilizing ResNet-101 and FPN, demonstrated superior performance in terms of APS and APM. This underscores the efficacy of the ResNet-101-FPN-CNB architecture for across-scale tasks. These findings serve as a valuable resource for informing the development of future models, emphasizing the potential benefits of architectures such as ResNext-101 with DCN. By leveraging insights from top-performing models, we aim to enhance the performance of subsequent iterations and push the boundaries of object detection capabilities.




## REFERENCES

- [1] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *European Conference on Computer Vision ECCV 2014: Computer Vision – ECCV 2014*, 2014, vol. 8691, pp. 346–361, doi: 10.1007/978-3-319-10578-9\_23.
- [2] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point Set Representation for Object Detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9656–9665, Oct. 2019, doi: 10.1109/ICCV.2019.00975.
- [3] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards Balanced Learning for Object Detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 821–830, Jun. 2019, doi: 10.1109/CVPR.2019.00091.
- [4] Z. Liu, J. Du, F. Tian, and J. Wen, "MR-CNN: A Multi-Scale Region-Based Convolutional Neural Network for Small Traffic Sign Recognition," *IEEE Access*, vol. 7, pp. 57120–57128, 2019, doi: 10.1109/ACCESS.2019.2913882.
- [5] W. Ouyang *et al.*, "DeepID-Net: Deformable deep convolutional neural networks for object detection," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2403–2412, Jun. 2015, doi: 10.1109/CVPR.2015.7298854.
- [6] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," *Advances in Neural Information Processing Systems*, pp. 379–387, 2016.
- [7] X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan, "Grid R-CNN," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7355–7364, Jun. 2019, doi: 10.1109/CVPR.2019.00754.




- [8] Y. Li, Y. Chen, N. Wang, and Z. X. Zhang, "Scale-aware trident networks for object detection," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, pp. 6053–6062, 2019, doi: 10.1109/ICCV.2019.00615.
- [9] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016, pp. 2874–2883, 2016, doi: 10.1109/CVPR.2016.314.
- [10] X. Zeng *et al.*, "Crafting GBD-Net for Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 9, pp. 2109–2123, Sep. 2018, doi: 10.1109/TPAMI.2017.2745563.
- [11] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, Feb. 2020, doi: 10.1109/TPAMI.2018.2844175.
- [12] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Light-Head R-CNN: In Defense of Two-Stage Object Detector," *arXiv*, pp. 1–9, 2017, doi: 10.48550/arXiv.1711.07264.
- [13] Y. Liu, R. Wang, S. Shan, and X. Chen, "Structure Inference Net: Object Detection Using Scene-Level Context and Instance-Level Relationships," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6985–6994, Jun. 2018, doi: 10.1109/CVPR.2018.00730.
- [14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," *European Conference on Computer Vision ECCV 2020: Computer Vision – ECCV 2020*, 2020, vol. 12346, pp. 213–229, doi: 10.1007/978-3-030-58452-8\_13.
- [15] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 845–853, Jun. 2016, doi: 10.1109/CVPR.2016.98.
- [16] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection," *European Conference on Computer Vision ECCV 2016: Computer Vision – ECCV 2016*, 2016, vol. 9908, pp. 354–370, doi: 10.1007/978-3-319-46493-0\_22.
- [17] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable Detr: Deformable Transformers for End-To-End Object Detection," *ICLR 2021 - 9th International Conference on Learning Representations*, pp. 1–16, 2021.
- [18] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8759–8768, Jun. 2018, doi: 10.1109/CVPR.2018.00913.
- [19] Y. Chen, Z. Zhang, Y. Cao, L. Wang, S. Lin, and H. Hu, "RepPoints v2: Verification meets regression for object detection," *Advances in Neural Information Processing Systems*, pp. 1–4, 2020.
- [20] B. Singh, M. Najibi, and L. S. Davis, "Sniper: Efficient multi-scale training," *Advances in Neural Information Processing Systems*, pp. 1–11, 2018.
- [21] J. Wang *et al.*, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, Oct. 2021, doi: 10.1109/TPAMI.2020.2983686.
- [22] Z. Gao, L. Wang, and G. Wu, "LIP: Local importance-based pooling," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3354–3363, 2019, doi: 10.1109/ICCV.2019.00345.
- [23] B. Singh and L. S. Davis, "An Analysis of Scale Invariance in Object Detection - SNIP," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3578–3587, Jun. 2018, doi: 10.1109/CVPR.2018.00377.
- [24] Z. Yao, Y. Cao, S. Zheng, G. Huang, and S. Lin, "Cross-Iteration Batch Normalization," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12326–12335, Jun. 2021, doi: 10.1109/CVPR46437.2021.01215.
- [25] T. Vu, H. Jang, T. X. Pham, and C. D. Yoo, "Cascade RPN: Delving into high-quality region proposal network with adaptive convolution," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

## BIOGRAPHIES OF AUTHORS



**Sara Bouraya**    Ph.D. student at the Faculty of Science, Ben Msik, at the University Hassan 2. She specializes in the field of computer vision, which involves developing algorithms and systems that enable computers to perceive and understand visual information. She can be contacted at email: sarabouraya95@gmail.com.



**Abdessamad Belangour**    academic researcher faculty member at the University Hassan 2 Faculty of Science in Ben Msik, Morocco. He specializes in the fields of data science, big data engineering, and meta-modeling. He can be contacted at email: belangour@gmail.com.