Object Roughly There: CAM-based Weakly Supervised Object Detection

How to train a deep learned object detector without needing to label position information by exploiting heatmap based explainability methods?

1 INTRODUCTION

- Object detection is concerned with classifying and localizing objects in an image.
- Highly performing object detectors require large training datasets, with class and bounding box annotations.
- Weakly Supervised Object Detection (WSOD) is concerned with training object detectors from only class labels, as opposed to Fully Supervised Object Detection (FSOD), as in Figure 1.
- MIL-based [1] WSOD methods achieve good performance, but have a high computational cost.
- CAM-based methods have primarily been studied for Weakly Supervised Object Localization (WSOL) i.e. images contain a single object. Their lightweight architecture makes them faster than MIL-based methods.
- GradCAM++ [2] is a CAM-based method that can localize where a CNN-based classifier paid attention to in an image in the form of heatmaps.
- Pin Pointing [3] entails indicating the general location of an object with a point.

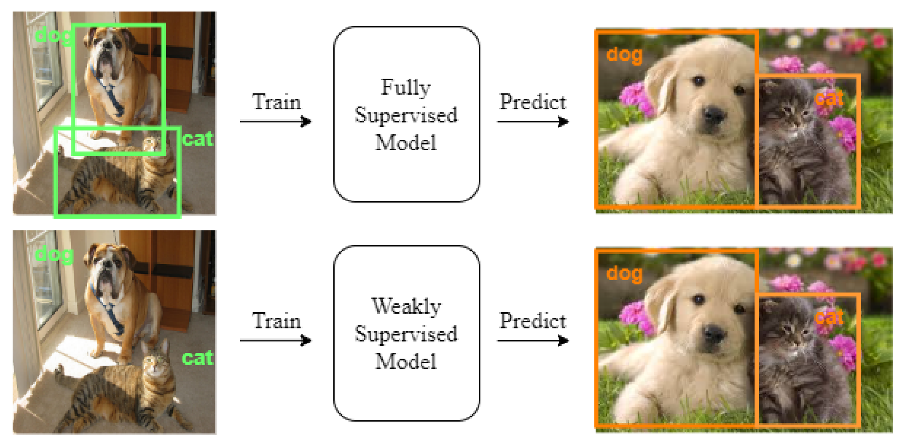

Figure 1: WSOD vs FSOD

2 METHOD - ORT

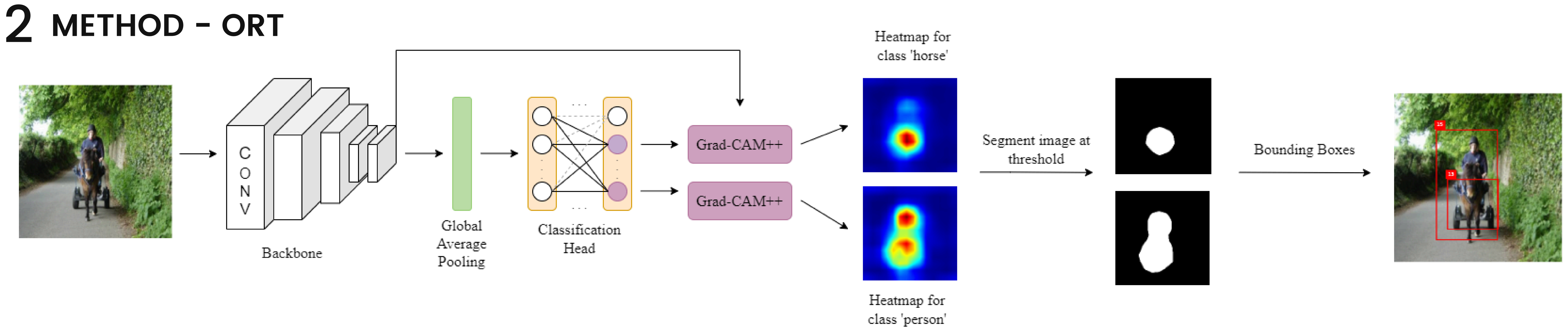

Figure 2: Proposed one-stage detection pipeline: A classifier made up of a CNN backbone, followed by a GAP layer and a classification head with multiple FC layers is used to extract feature maps from an image. GradCAM++ backpropagates the predicted class scores to the final convolutional layer for each class to obtain the CAMs represented as heatmaps. The heatmaps are segmented and contour detection is used to extract the locations of the objects.

- Bounding Boxes are created with a low segmentation threshold of 20%.
- Pin Points are created at the highest activation within the contours obtained with a high segmentation threshold of 50%.
- Backbone classifier is trained with Binary Cross-Entropy Loss, which allows for one image to contain multiple class labels.
- Two backbone classifier architectures are experimented with: VGG16 and a novel FPN-based classifier.

Incorporating features from shallow layers

- To improve on the performance of the method, features from shallow and deep layers of the CNN-based backbone can both be incorporated.
- While other methods aggregate CAMs computed from different layers, I propose leveraging the architecture of the backbone classifiers directly, by using Feature Pyramidal Networks (FPNs) [4].

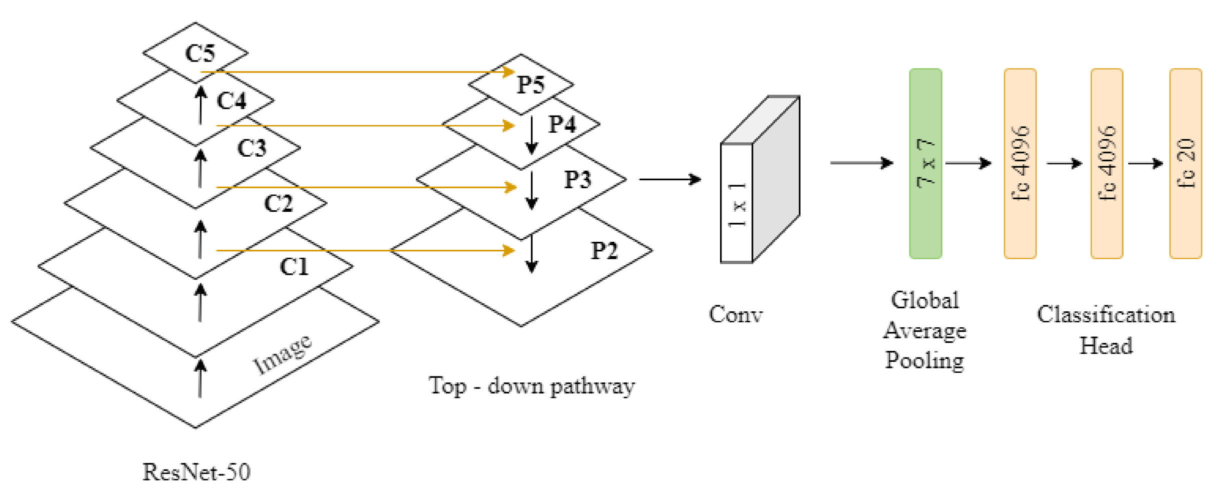

Figure 3: Proposed architecture for FPN Classifier: The ResNet50 backbone serves as a bottom-up pathway encoding the image into feature maps across five modules. The top down pathway upsamples the resulted low resolution feature maps and aggregates them with the corresponding bottom-up pathway maps via lateral connections. One of the resulting feature maps containing spatial and semantic information is passed through a 1 x 1 convolution, followed by a 7 x 7 GAP layer and 3 FC layers used for classification.

Two-stage method: WSOD to FSOD

- MIL-based methods have sucesfully been using the Weakly Supervised to Fully Supervised (WSOD to FSOD) method to increase their performance.
- It entails using a weakly supervised detector to generate labels for the training set and then training a fully supervised detector on these pseudo-labels.
- I use the SOTA fully supervised Faster-RCNN [5] for its robust architecture.

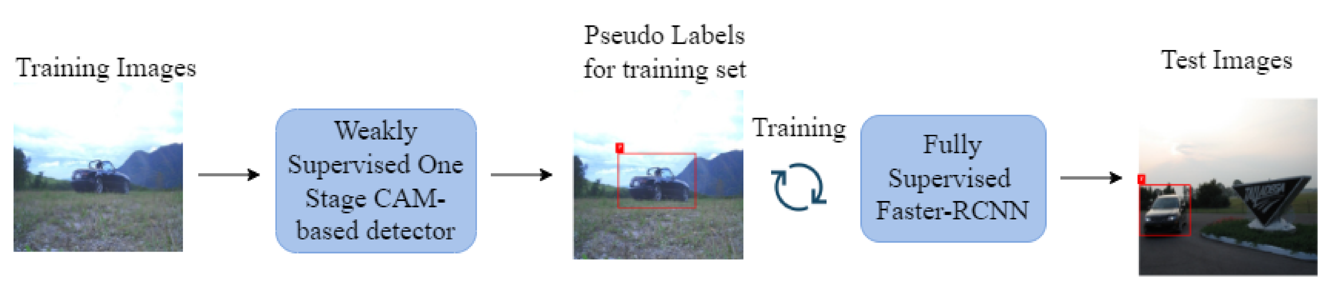

Figure 4: Proposed two stage detection pipeline: The one stage weakly supervised object detector is used to generate pseudo bounding boxes for the training set. A fully supervised Faster-RCNN is then trained and used to make predictions.

3 EXPERIMENTS

How do different backbone classifiers affect the detection capabilities of ORT?

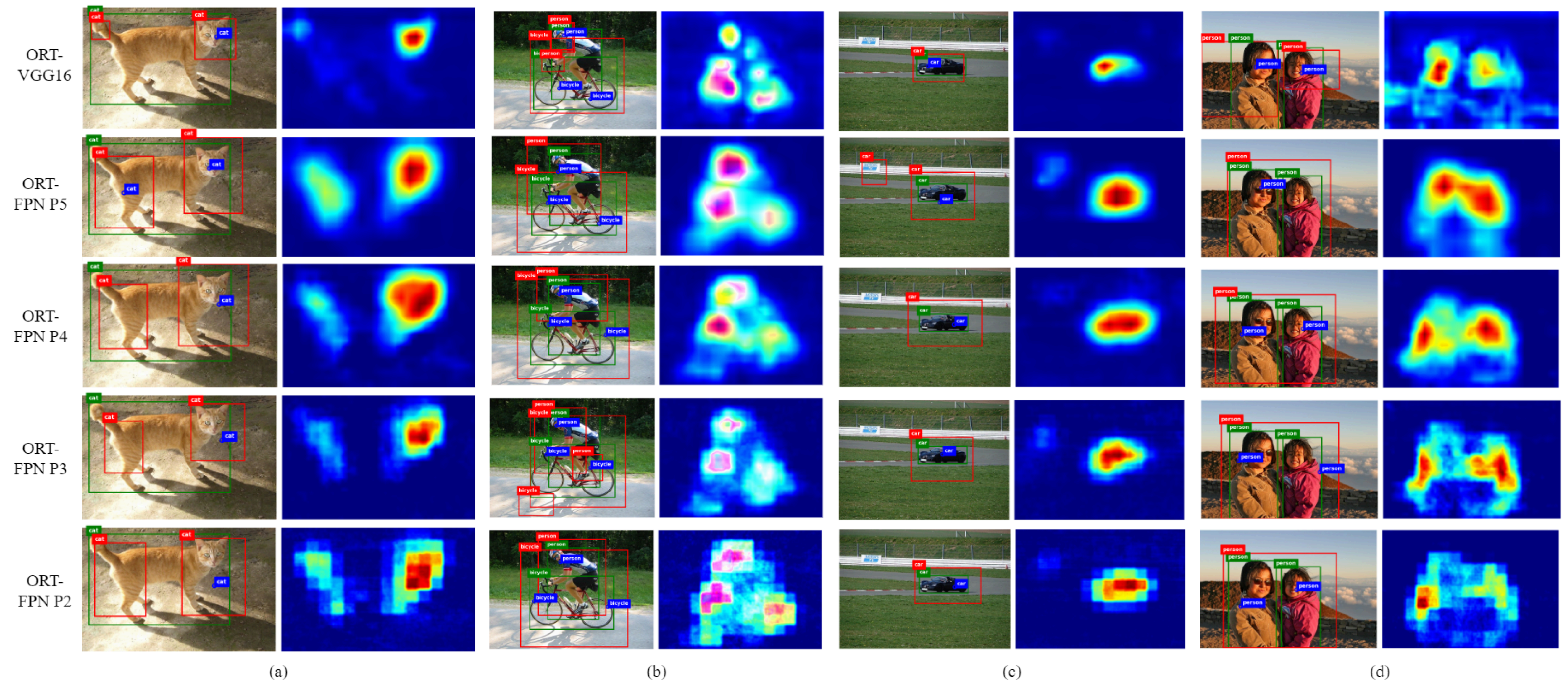

Figure 5: Comparison between the detection and pin pointing performance between the different backbone classifiers used for ORT on images of the VOC 2007 test set. Each column contains the original image with the ground truth bounding boxes in green, the predicted bounding boxes in red, the predicted pin points in blue and the heatmap generated with GradCAM++. **The FPN-based models manage to detect more of the objects, compared to the VGG16 that mainly looks at the most discriminative parts. The deeper feature maps in the FPN detect more fine-grained features and have a less uniform aspect.**

How does ORT perform across different backbone architectures compared to other object detectors?

Method	Full dataset	Localization			Multi Instance	Multi Class
		small	large	both		
ORT-VGG16	6.0	27.2	10.9	12.2	4.2	5.5
+Faster-RCNN	21.1	49.4	28.5	32.6	15.3	21.4
ORT-FPN P5	12.4	21.5	39.8	33.0	6.2	11.4
+Faster-RCNN	20.8	23.2	54.9	43.1	12.0	19.7
ORT-FPN P4	10.4	14.2	33.6	25.0	4.9	10.3
+Faster-RCNN	22.2	27.3	53.2	43.4	14.5	20.1
ORT-FPN P3	6.7	21.2	17.6	15.8	3.7	6.9
+Faster-RCNN	20.5	38.1	38.2	36.5	14.3	20.3
ORT-FPN P2	9.2	19.2	25.3	20.7	6.0	8.6
+Faster-RCNN	22.0	33.7	49.4	43.8	14.8	20.3
PCL*	48.8	-	-	-	-	-
Faster-RCNN	74.2	87.4	82.5	85.1	67.4	71.0

Table 1: Object Detection results with mAP@50 on VOC 2007. **Weakly supervised models are outperformed by the fully supervised detector, struggling most with multi instance and multi class images. The two stage method boosts the performance of the one stage models.**

Method	Full dataset	Localization			Multi Instance	Multi Class
		small	large	both		
ORT-VGG16	80.0	75.6	98.8	96.5	56.1	71.3
+Faster-RCNN	92.5	83.6	99.4	98.6	79.4	87.0
ORT-FPN P5	79.7	73.8	99.3	96.9	53.7	72.1
+Faster-RCNN	88.2	88.1	99.0	98.3	71.2	81.9
ORT-FPN P4	85.6	75.3	99.0	96.6	61.0	78.1
+Faster-RCNN	89.7	89.5	99.1	98.1	73.8	84.3
ORT-FPN P3	83.8	72.7	99.3	96.9	59.7	75.3
+Faster-RCNN	92.6	85.1	99.0	97.9	77.8	87.2
ORT-FPN P2	85.0	71.5	98.6	96.5	61.4	77.0
+Faster-RCNN	91.2	87.5	99.1	97.6	75.8	85.8
Faster-RCNN	95.7	90.0	99.0	99.3	85.1	92.2

Table 2: Pin Pointing results with mAP on VOC 2007, where a point is considered correct if it falls in the ground-truth bounding box. **The proposed models can successfully pin point the general location of objects, without knowing the objects' locations during training, their performance being close to the fully supervised detector.**

Can ORT achieve low inference time?

Method	ORT-VGG16	ORT-FPN P5	PCL	YOLOv3	Faster-RCNN
Inference time (seconds)	2.38	1.31	36.35	0.59	1.81

Table 3: Comparison between the inference time (in seconds) of the proposed weakly supervised models, SOTA weakly supervised MIL-based model (PCL) and two stage (Faster RCNN) and one stage (YOLOv3) fully supervised detectors. **ORT achieves near real-time inference, being faster than Faster-RCNN.**

4 CONCLUSION & LIMITATIONS

- ORT has reduced capabilities at detecting the full extent of objects with bounding boxes, but it can achieve good pin pointing performance: 85.6% and 92.6% mAP@50 on the VOC 2007 dataset with the one- and two-stage method respectively
- ORT's near real-time inference speed shows potential for real world applications (e.g. robotics, autonomous driving).
- ORT is limited by the segmentation thresholds that were the same for every backbone classifier, and would benefit from different thresholds, especially across the FPN layers.
- Improvements in the training strategy and the choice of loss function for the backbone classifier could help it learn more robust features, with better separation of the classes.
- Future research could improve on the bounding box generation with strategies used in FSOD such as Non-Matrix Factorization (NMS)
- The proposed method should be analyzed when using a transformer architecture instead of the simple classifier. In object detection, transformer based models such as DETR and DINOv2 have reached state-of-the-art, even in self-supervised settings.

TU Delft Delft University of Technology

Author
Petra Postelnicu
email: P.Postelnicu@student.tudelft.nl

Supervisors
Dr. Jan van Gemert
Dr. Osman S. Kayhan

Affiliations
EEMCS, Delft University of Technology

References

[1] Tang, Peng, et al. "Pcl: Proposal cluster learning for weakly supervised object detection." IEEE transactions on pattern analysis and machine intelligence 42.1 (2018): 176-191.
[2] Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018.
[3] Oquab, Maxime, et al. "Is object localization for free?-weakly-supervised learning with convolutional neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[4] Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[5] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015).