Distant Vehicle Detection: How Well Can Region Proposal Networks Cope With Tiny Objects at Low Resolution?

Ann-Katrin Fattal[1,2], Michelle Karg[2], Christian Scharfenberger[2], and Jürgen Adamy[1]

[1] Technische Universität Darmstadt, Institute for Control Methods and Robotics, Germany
[2] Continental AG, Germany

Abstract. High-performance faster R-CNN has been applied to many detection tasks. Detecting tiny objects at very low resolution remains a challenge, however, and a few studies addressed explicitly the detection of such objects yet. Focusing on distant object detection at very low resolution images for driver assistance systems, we introduce post-trained net surgery to 1) analyze the network activation patterns, 2) study the potential of prior information to improve localization and binary classification performance, and 3) to support the development of priors for improving the network performance.
We use post-trained net surgery to analyze the feature maps used for bounding box regression and classification for RPNs in detail, and to discuss the complexity of the network activation patterns. Using these findings, we show that incorporating prior maps into the network architecture improves the performance of bounding box regression and binary classification for small object detection in low resolution images.

Keywords: Saliency Maps, Region Proposal Network, Low Resolution, Object Detection

1 Introduction

Deep learning-based detection approaches combine three stages in one architecture for high efficiency: 1) feature extraction, 2) region proposals/localization and 3) object classification. The objective of feature extraction is to learn a meaningful set of features representing target objects. Region proposal methods use the features to identify potential object regions and to reduce the number of regions fed into an expensive classification stage. Approaches to proposing regions are of great importance due to the need of detecting regions containing distant objects very quickly for the purpose of robust object detection for autonomous driving. Hardware and cost constraints on the sensor set-up impose the need for an inexpensive yet still accurate algorithm. Classical approaches such as [1–3] make use of hand-crafted models to identify object regions. Recent work on neural networks with trained features for object detection showed better results in localization and classification, where trained filters in layers decompose input information and produce feature maps for each layer.

The Fast R-CNN [4] combines localization and classification in one architecture and shows good results on a variety of data sets for object detection. In [5], a first concept for bounding box regression using features solely from the CNN was proposed. Other architectures such as Faster R-CNN [6], YOLO [7], Overfeat [8] and SSD [9] combine trained feature extraction and region proposals in neural network architectures (RPN). The detection of small objects in low-resolution images, however, remains a challenge for all approaches presented. Li *et al.* [10] use feature maps

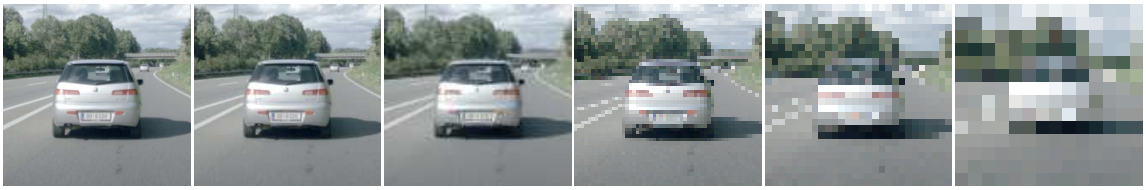


Fig. 1: The feature quality depends highly on the image resolution (from left to right: 443×421px, 222×211px, 111×106px, 55×53px, 28×27px, 14×14px). For this reason, detection of low resolution objects is a challenging task.

from different layers of the convolutional feature extractor for bounding box regression to consider more low-level features. The performance of this approach depends strongly on the feature maps chosen, and is hardly feasible for detecting small objects in low resolution images. Brazil *et al.* [11] apply a ground-truth mask on the feature maps inside an RPN during training to help the network focus on relevant object regions, resulting in less cluttered feature maps. However, a dedicated segmentation branch is needed for this approach. Huang *et al.* [12] evaluate different CNN architectures and show the excellent accuracy of Faster R-CNN [6] over a variety of convolutional object localizers. Faster R-CNN [6] makes use of feature maps from a convolutional feature extractor which may base on any fully convolutional network architecture such as ZF-Net or VGG16. The RPN uses an extra layer to extract objectiveness-specific features feeding two consecutive branches for bounding box regression and binary object classification (BCN). A set of base anchors, fitted to the data sets properties, are set on each location of the last feature map inside the RPN. The bounding box regression learns the deviation to the base anchors pre-defining position, width, and height. Finally, the binary classification branch derives a score for objectiveness, i.e., how likely an object is present in each base anchor.

The analysis of convolutional object detection performed by Huang *et al.* [12] showed that the detection of small objects with dimensions smaller than 20 pixels is three times worse than the detection of medium-sized objects. Zhang *et al.* [13] emphasized that small objects are the most common source of false negatives. This has a high relevance to distant vehicle detection for assisted/autonomous driving, and implies a low miss rate for distant objects that can only be achieved when the Recall of the region proposal network is high. The Hypernet [14] and the Feature Pyramid Network [15] use feature maps at higher resolution with more low level features to detect small objects. However, in the case of a low resolution images, using different feature maps is difficult as object regions provide weak local features.

In this work, we wish to make explicit use of incorporating prior information into a Faster R-CNN [6] to improve the overall network performance with focus on very small objects in low-resolution for driver assistance systems. As shown in Figure 1, vehicles in 130m distance to the camera may occupy only 8×8 pixels on the camera image. In particular, we introduce *post-trained net surgery* to 1) cluster network activation patterns, 2) to study the potential of prior information to improve localization and classification performance, and 3) to support the development of priors for improving an overall network performance of RPNs. We further study the impact of the feature maps and priors on bounding box regression and binary object classification for very small objects. First, we evaluate the importance of the feature maps by clustering them based on their correlation to the task specific ground truth data. The clustered feature maps show that very few maps cover the most features of the small objects and contribute the most to the overall localization and

classification performance of an RPN. Second, we incorporate different external priors into the RPN chain before bounding box regression and binary classification and study their contributions to the overall performance of the RPN.

This analysis allows for several important conclusions: Post trained net surgery, with selecting the most important cluster of feature maps, helps identify the important feature maps for bounding box regression and binary classification, and understand the contribution of features to obtain decent feature maps. This allows to adapt priors or external data to the most prominent features to increase the Recall for the task of improving distant, and hence object detection in low-resolution images. Finally, evaluations demonstrate the need for incorporating priors into the network architecture to increase the Recall for small object detection significantly.

2 Network Architecture

We wish to choose a network architecture for evaluation purposes that is designed and optimized for the detection of small objects such as distant vehicles for assisted and autonomous driving. Given the limited computational resources on embedded systems for driver assistance systems, small and inexpensive neural networks are preferred over larger and more complex architectures.

Faster-RCNN with ZF Net We chose the Faster R-CNN [6] as it combines a region proposal network (RPN) and a binary classification network (BCN) in one architecture and achieves higher detection accuracies than single-shot architectures like Yolo or SSD [12].

Faster R-CNN uses a common feature representation for both the RPN and BCN, is learned in an end-to-end fashion and shares convolutional layers between the RPN and BCN to compute locally restricted feature maps for feature extraction [16]. Different core network architectures can be used for feature extraction. Here, we choose as a small network architecture the ZF-Net to meet the limited computational resources for running CNNs on embedded devices for automotive applications. The ZF-Net consists of five convolutional layers, with two pooling and two fully connected layers. The stride of the network is 16 px, however the input image is upscaled by a factor of 2.4 as suggested by Fan *et al.* [17] to improve the performance. This still allows learning of important features due to the down-sampling layers as well to use pre-trained networks of the shelf.

Region Proposal Network (RPN) An RPN proposes bounding boxes for the subsequent binary classification network (BCN). The RPN aims to decrease the false negative detection rate, resulting in a high Recall, and the BCN reduces the false positive detection rate, resulting in a high Precision [11]. Since a high Recall is prerequisite for detecting small and distant objects, we focus on improving the performance of the RPN in this work.

The RPN consists of the three main stages: (1) a set of convolutional layers for feature extraction, (2) binary classification to compute a score indicating the object probability in each anchor, and (3) bounding box regression for the center coordinates (x,y), width and height of each anchor. Overall, refining the bounding boxes using regression improves the overall classification results.

A set of N anchors predefines aspect ratio and scale and is fitted to the application. Binary classification and bounding box regression are computed for each anchor at each position in the final feature map of the RPN, including two scores for the presence of objects and four values for the bounding box. Given the fix locations of anchor areas, the deviation to the object region center

Table 1: weights: pretrained RPN weights, anch.-o.: original anchors, anch.-a.: adjusted anchors, 2.4x: upscaling by factor 2.4x, batchs.256: batchsize of 256 within the RPN, batchs.20: adjusted batchsize of 20 within the RPN.

Parameters						Mean recall in % for object sizes			
weights	anch.-o.	anch.-a.	2.4x	batchs.256	batchs.20	8-20px	20-30px	30-60px	60-100px
-	✓	-	-	✓	-	16.89	70.32	87.95	87.88
✓	✓	-	-	✓	-	18.84	74.24	87.23	85.61
✓	-	✓	-	✓	-	29.03	71.24	82.31	90.91
✓	-	✓	-	-	✓	32.93	75.00	86.23	93.18
✓	-	✓	✓	✓	-	56.74	83.39	94.71	95.13
✓	-	✓	✓	-	✓	**68.54**	**90.93**	**96.51**	**99.57**

in x and y direction and the deviation of the bounding boxes in height and width are considered in the network.

For the special use case of small object detection at low resolution, the anchor sizes and scales are fitted to the data set. Furthermore, the input image size is upscaled. As the number of positions with sufficiently overlapping anchors is smaller for small objects, the batch size during training is reduced drastically to obtain a better balanced RPN training set. The minimal allowed scaled bounding boxes are chosen to be larger than the stride of the feature map within the RPN because the stride defines the smallest possible detected object. Table 1 summarizes the adapted parameters for small object detection.

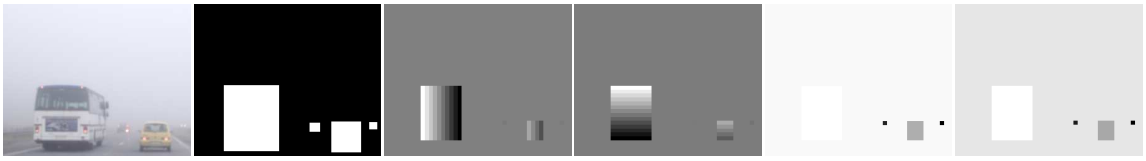
The output size of the 1×1 convolutional layer is $2 \times N$ for the classification, and $4 \times N$ for bounding box regression. The input size is the number of feature maps of the last convolutional layer of the RPN. To investigate the capacity of this 1×1 convolutional layer for binary classification and bounding box regression, we alter the input of the last convolutional layer in the RPN by both removing feature maps using net surgery and by providing prior information.

3 Net Surgery

To analyze the feature representation prior to binary classification and bounding box regression in detail, we propose two extensions to net surgery. The first extension addresses the clustering of feature maps of similar relevance and the second extension addresses the search for the optimal representation of prior information.

3.1 Relevance-based Clustering of Feature Maps

Understanding the underlying functionality of CNNs has attracted the interest of the research community. This interest is driven by the fact that neural network architectures are generally understood as black-box technologies. Understanding of this functionality is especially relevant for safety-critical applications such as autonomous driving. Zeiler *et al.* [18] introduced the use of back-propagation to generate the optimal input image for a convolutional network given the desired output. Using this method, it is possible to understand which kernel filters are responding maximal for different object classes. We are interested in how many feature maps include class-relevant information, in the redundancy of the class-relevant information, and whether or not the



(a) (b) (c) (d) (e) (f)

Fig. 2: Visualization of the ground truths for the RPN: (a) original image, (b) ground truth for the classification and pre-processed ground truth for the bounding box regression: (c, d) deviation to the center in x and y direction, (e,f) deviation of width and height

(a) Input (b) BB-rect (c) BB-rect.-gauss (d) Spectral Residual (e) Voting Map

Fig. 3: (a) shows the original image, (b) the ground truth in BB-rect, (c) ground truth BB-rect smoothed by Gaussians, (d) spectral residual saliency map [19] and (e) the voting map [3].

information content is depended on object attributes such as size. For this reason, we introduce post-trained, relevance-based clustering of feature maps, where only a subset of feature maps is used during inference.

Clustering of RPN Feature Maps Given that several feature maps evolve similar features [18], we wish to make use of clustering to merge feature maps with similar content. Clustering can be based on several characteristics such as homogeneity or correlation to external data. In this work, we chose correlation to ground truth as our cluster criterion, such as shown in Figure 2.

Clustering is conducted for each trained model using the data of the test set. For each test-image$_i$, the absolute correlation is calculated for each feature map fm$_{i,j}$ and ground truth BB-rect$_i$ as shown in Figure 3b. The feature maps are then sub-divided into Q equally large groups q where the feature maps with the highest correlation score is assigned to the group with the highest z_Q with $z_q \in [1, .., q, .., Q]$. This is done for all images in the test set. The final cluster is then computed by finding the highest occurrence of z_q for each fm$_j$.

Cluster-based Inference With the different feature map clusters, it is now possible to determine the influence of each cluster to the performance of the RPN to find the most important features. With net surgery, only feature maps in cluster q are contributing to following layers inside the network. A filter kernel in a convolutional neural network is described by its size K, with its depth the number of input data channels (here feature maps). X kernel use then all feature maps fm$_m$ of layer output m, and after a convolution it produces X new feature maps fm$_{m+1}$. During net surgery, the input data of the kernel is modified so that only feature maps within certain clusters

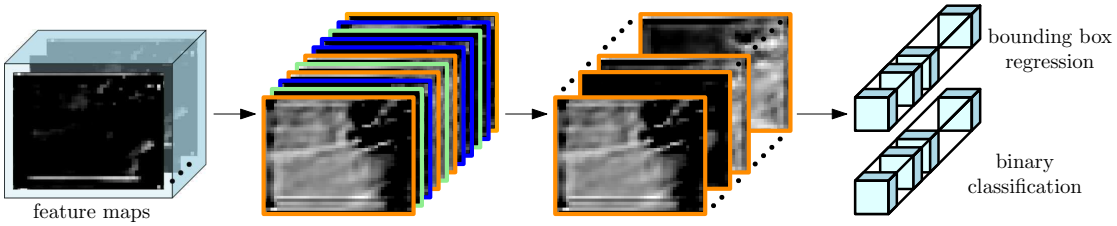


Fig. 4: The feature maps of the last layer of the RPN are clustered using the similarity between the feature map activation and ground truth as cluster criterion. We analyzed the features maps within each cluster regarding: their contribution to the bounding box regression and binary classification.

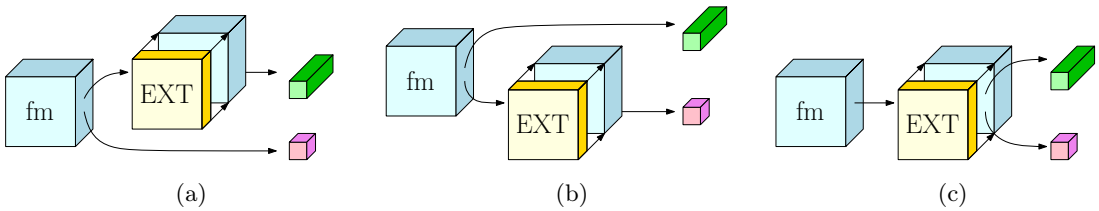
are processed. To perform net surgery with cluster q, the weights inside the kernel are set to zero except for weights that correspond to all feature maps fm_j with $\mathrm{cl}_j == z_q$. Hence, only feature maps fm_m within cluster q transport information to fm_{m+1}. Then, it is possible to evaluate the performance of the network based on different feature clusters. In this work, net surgery is performed on the bounding box regression kernel and/or binary classification kernel. Figure 4 visualizes the net surgery.

3.2 Incorporation of Prior Information

Using the idea of net surgery we wish to study the positive impact of external data on the performance of an RPN. Figure 5 shows three different ways of incorporating of prior information during feed-forward and training phase into the last layer of the region proposal network.

Motivated by improving the performance of the RPN by incorporating prior feature maps, as in [20], we study the potential performance gain that can be obtained when incorporating perfect feature maps. To estimate such an upper bound, prior information is computed based on ground truth data. The effect of feature representation is studied for the 1×1 convolutions for binary classification and bounding box regression. Hence, the optimal feature maps are incorporated using net surgery prior to bounding box regression (Figure 5a), pior to classification (Figure 5b), and prior to both classification and bounding box regression (Figure 5c). In doing so, we analyze how efficient the information from additional feature maps can be learned by the 1×1 convolutions for binary classification and bounding box regression using stochastic gradient descent. Furthermore, this approach enables studying the optimal representation for prior information. In the following, a set of different representations for priors are summarized, both for theoretically studying the optimal performance gain using the ground truth as prior and for application-relevant priors, such as saliency maps.

Adapted Ground Truth Data The ground truth data for binary classification and bounding box regression are different due to the different nature of the underlying task. The ground truth for only binary classification architectures as shown in Figure 5b are created by setting all pixel values inside an object region to one and zero otherwise (see Figure 3b). This form of ground truth is called *BB-rectangular*. The BB-rectangular data is designed in the same way as the label data for the binary classification branch during the training phase of the RPN. The ground truth data for the bounding box regression branch as shown in architectures of Fig 5a follows another pattern



(a) (b) (c)

Fig. 5: Prior information is added as external data (EXT) to the output of the last convolutional layer (fm) of the RPN for either (a) bounding box regression, (b) binary classification, or (c) both bounding box regression and binary classification. The blue boxes represent data from the feature map, the yellow boxes external data and green and pink boxes the 1×1 kernels for regression (green) or classification (pink).

which is based on the underlying anchor size. It contains the deviation of the anchor center (x_a, y_a) to the object center (x, y), and the deviation of the anchor size (h_a, w_a) in height and width to fit the objects size (h, w) as shown in Figure 2 (c)-(f) [21]. The ground truth data $t_{x,y,w,h}$ for the regression is then given as following

$$t_x = (x - x_a)/w_a \qquad t_w = log(w/w_a)$$
$$t_y = (y - y_a)/h_a \qquad t_h = log(h/h_a)$$

"BB-rectangular" refers to the ground truth suitable for classification, "GT-adjusted" refers to the transformed ground truth for bounding box regression.

Gauss-degradation to Blobs To examine the influence of degradation of the quality on the ground truth data as external priors, the ground truth bounding boxes are transformed to Gaussian blobs which follow the elongation of the object based on ground truth labels. The center of each object obtains the highest overall pixel value within the data map, and all remaining data is reduced following the multivariate normal distribution:

$$y = \frac{1}{\sqrt{|\Sigma|(2\pi)^d}} \cdot exp\left(-0.5(x-\mu)\Sigma^{-1}(x-\mu)'\right)$$

where y is the pixel value on the data map, and Σ a $d \times d$ symmetric positive definite matrix. For each object region at an object center position $d = 2$, $\mu = 0$ and Σ

$$\Sigma = \begin{bmatrix} w & 0 \\ 0 & h \end{bmatrix}$$

For all external data, that is degraded by such distribution the flag "-gauss" is added.

Saliency-inspired Maps All previous presented external data is based on the knowledge of the ground truth and can therefore only be used hypothetically for analysis. Saliency-inspired maps can be calculated without any ground truth knowledge as the fundamental idea of saliency maps is to focus on the essential information within an image. Hence, saliency maps can be computed directly on the image and can be considered as external data. Potential priors that can be computed in

Table 2: Details of the used data set. The table illustrates the number of images used for training, test and validation, and the distribution of object sizes.

Data set	Train	Val	Test	**Object width (px)**	8-20	20-30	30-60	60-100	>100
# **Samples**	826	104	104	# **Samples**	2792	2770	1767	631	332

real-time are Spectral Residual [19] and a Voting Scheme [3]. The Spectral Residual map analyses anomalies in the frequency spectrum of an image. Since regions containing small objects exhibit high spatial frequencies when compared to background, this map is suitable for detecting small objects. The Voting Map is adaptively modeling the background within an image with few homogeneous areas to distinguish between background and possible foreground. This approach is tailored to the task of guiding detection towards small object regions in motorway scenarios. Both the Spectral Residual and Voting maps are chosen as saliency-inspired external data in this study.

4 Experiments

The proposed approach is evaluated on a motorway and highway traffic data set that includes many distant cars and trucks. The evaluation includes the results for the model without any external data or net surgery as well as different architectures including net surgery or external data.

4.1 Data set and Data Augmentation

A data set for distant vehicle detection has been recorded and includes 1034 different motor- and highway scenes. The data set contains over 5.000 distant cars and trucks with object sizes smaller than a width of 30 px, and is subdivided into training, validation and test set as illustrated in Table 2. Objects that are occluded by less than 50% are included in the data set. The image size is 1024×640 px. Table 2 shows the occurrence of object widths within the data set. In addition, data set augmentation is used for both training and testing to avoid over-fitting during training. Therefore, 10 crops of size 300×250 px are taken from each image using a random jittering inside the image, ensuring that each crop includes at least one object and no objects are truncated. Additionally each image/crop is flipped vertically.

4.2 Network Architecture and Training Details

A ZF-net architecture with weights pre-trained on the ImageNet and Kitti data set is used as the core network for the Faster R-CNN. The input size of the Faster R-CNN using the ZF-net is 600×720 px. The image crops are upscaled by a factor of 2.4 as suggested by Fan *et al.* [17] to improve the performance especially for the task of small object detection. For our data set the best ratios are [0.5, 1, 2] as it contains trucks as well as cars from the side. The anchor sizes are set to $[10, 20, 40]$ px to suit especially small object regions, and to fit to the object occurrence. Forward feed allowed 8 px wide boxes. For the following experiments the faster RCNN was trained as proposed in [6] with original parameter set-up. The trained weights of this base model are then frozen except for the binary classification and bounding box regression branch in the RPN. During all following experiments only the RPN is refined in one stage. To fit to the data set with many small objects, the

Table 3: Details of the training parameters to train the RPN and classification head of the faster RCNN for a data set with small objects

Parameter	Value
Iterations within the RPN	9.000
Upscaling factor	2.4
Max value of external data	10
RPN batchsize	20
anchor sizes	10, 20, 40 px
anchor ratios	[0.5, 1, 2]
Minimum bounding box	8 px

batch size for the RPN is reduced drastically to 20 to generate a balanced foreground/background set of possible anchors during training (Table 3).

4.3 Evaluation Metrics

The performance is evaluated using the Recall metric based on the Intersection over Unit (IoU = 0.5). The Recall measures how many of the relevant objects are successfully detected:

$$\text{Recall} = \frac{TP}{TP + FN}$$

TP is the number of relevant matches retrieved, and FN the number of relevant matches missed. As the RPN is the localizer of the Faster R-CNN, the Recall representing the number of object regions detected initially is of interest only. We compute the Recall for each object size over the number of selected bounding boxes, for an IoU equal to 0.5 and 600 proposed bounding boxes.

4.4 Results of the Net Surgery

To understand the diversity and spread of information of the RPN within the feature maps, the absolute correlation of the feature maps and ground truth as BB-rect data (see Figure 3b) is used. The clustering uses the absolute correlation values, and 3, 5 and 10 clusters are formed. In Figure 6, the Recall of the different active feature map clusters for 5 clusters is shown. The Recall is decreased in all object size classes for just some of the clusters activated.

It can be seen that the feature map cluster with highest absolute correlation value (yellow bars) contains the most valuable features for the RPN among the object sizes 8-60 px. For larger object sizes the feature maps with the 40-60% highest absolute correlation value shows the highest Recall among the model which were modified by net surgery. Here, it is shown that different feature maps evolve object size specific feature. Hence, the relevance of a feature map cluster depends on the size of the object detected, where feature maps including fine-grained information support the detection of small objects, and feature maps omitting detailed information the detection of larger objects. For small objects a high similarity to the BB-rectangular data is most favorable. This can be used when e.g. a large network is trained on several object classes/sizes but only certain sizes/ classes are of interest. Then all feature maps not useful for the interesting object-size/class can be removed or added as needed.

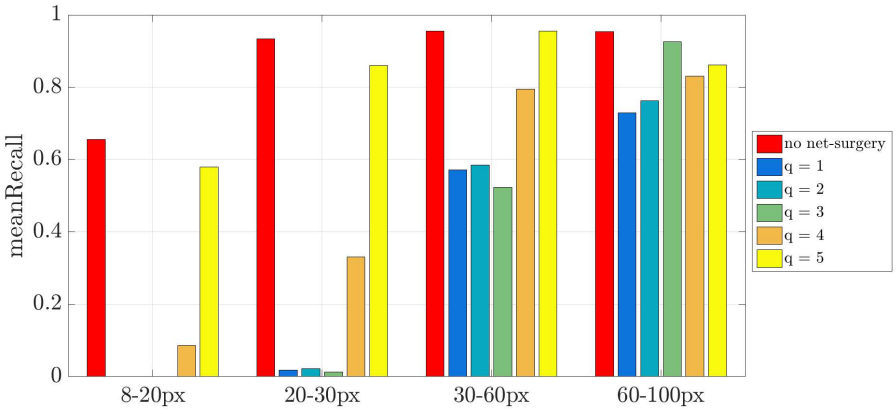


Fig. 6: The Recall for cluster-based net surgery for 600 boxes and 5 different clusters is shown. For each cluster q, the feature maps with highest range of absolute correlation to BB-rectangular are activated. The red bar corresponds to an architecture without net-surgery, while each bar to the left represents a different activated cluster with increasing correlation. We can see that different features impact the performance for detecting differently sized objects.

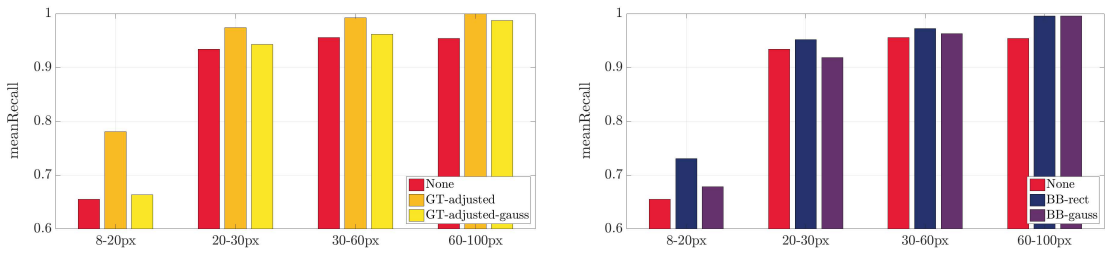
Table 4: The table shows the potential of branch size reduction Δbranch-size and reduction of recall $\Delta q_{\max}$ for the different object sizes and different number of clusters Q. The baseline is the model without any net surgery.

Q	$\Delta q_{\max}$[tiny, small, medium, large objects]	Δbranch-size
3	-7.2%, -5%, -5.1%, -3.6%	-66%
5	-12.8%, -13.5%, -6%, -2.7%	-80%
10	-36%, -31%, -19%, -13.8%	-90%

Execution Time Using cluster-based network surgery it is possible to reduce the size of the network architecture while only loosing comparable low Recall. This reduces the computational cost or execution time of the branches. In Table 4 it is shown that the branch size can be reduced by e.g. 66% during net surgery with 3 clusters while the Recall decreases only in average 5.2%. Hence, the execution time/computational cost can be decreased also by 66%. Especially in the automotive environment the reduction of network size due to limited computational power while keeping the performance high is desired [22].

4.5 Influence of External Data on the Recall

In two experiments the bounding box regression or binary classification branch uses ground truth external data as additional information to understand the capability of each branch and to determine a upper bound of performance. In the last experiment both RPN branches were trained and executed using several different external data maps. In Figure 5 the different architectures are shown.



(a) (b)

Fig. 7: (a) Mean Recall with ground truth external data only for the bounding box regression branch and (b) with BB-rect external data only for the binary classification branch of the RPN.

Comparison of only Regression or Classification Branch Figure 7a shows the Recall for different external data composed with ground truth knowledge in the regression branch only (architecture as in Figure 5a). It can be seen that especially for small objects with 8-20 px width the external data increases the overall Recall by more than 10%. For objects larger than 20 px, the increase is 5% in Recall. For Gaussian degraded ground truth the performance degrades and shows that Gaussian blobs are less suitable features for the regression branch.

When only the binary classification branch is trained with external ground truth data, Figure 7b, performance only increases for the BB-rectangle format of the ground truth. This shows that the binary classification branch can only transform external data with a high similarity to the BB-rectangle format and even gauss-degraded ground truth data is not improving recall values.

Comparing both architectures with each corresponding ground truth respectively, it shows that the ground truth adjusted for the regression branch improves the overall recall more, than the ground truth BB-rectangular for the binary classification. This finding shows that the features for the regression seems to be less evolved inside the feature maps of the region proposal network than the features for binary classification. This applies especially for objects smaller than 20 px but is also present for larger objects. Neither of the branches exhibits a recall of 100 % which showed as well, that the improvement of only one branch is not sufficient to reach high Recall values.

Bounding Box Regression and Classification Branch In Figure 8 the Recall of the RPN is shown where both branches of the RPN use external data during training and testing. It shows that the RPN has in general less Recall when the objects get smaller. For object sizes larger than 20 px in width the average Recall is more than 93% without any external data, while it is only 65.5% for objects smaller than 20px in width. Here it shows that the adjusted ground truth (GT & BB-rect) optimal for both branches reaches in all objects sizes best results as expected. However, even for objects smaller than 20 px only 86.6% of Recall is reached and is an indicator, that small objects are even with best possible data difficult to detect for the RPN. To understand the gap of performance even with best possible data it is important to discuss the evaluation metric. E.g. a bounding box proposal of correct object size with a 3 px displaced center has an IoU of 0.65 for an object of size 20 px, while the IoU is only 0.45 for an object of 8 px size. Hence, any dislocation or error in box size to the object is more severe for small objects than for larger ones. The external saliency-like maps increase the Recall for small objects down to 8 px by 2.6%, while it only slightly

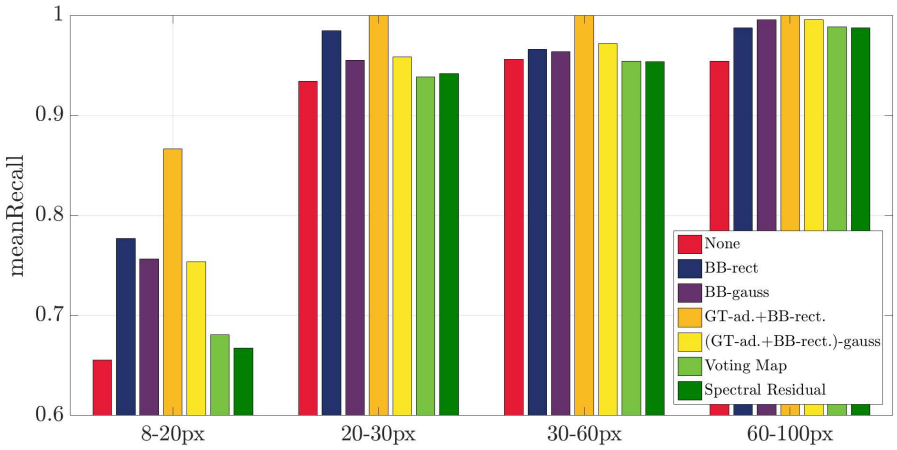


Fig. 8: The Recall for low resolution objects is significantly lower than for objects with a minimum size of 20 px. Adding prior maps as input to bounding box regression and binary classification improves the recall and the localization of low resolution objects dramatically, where ground truth prior maps estimate upper bounds.

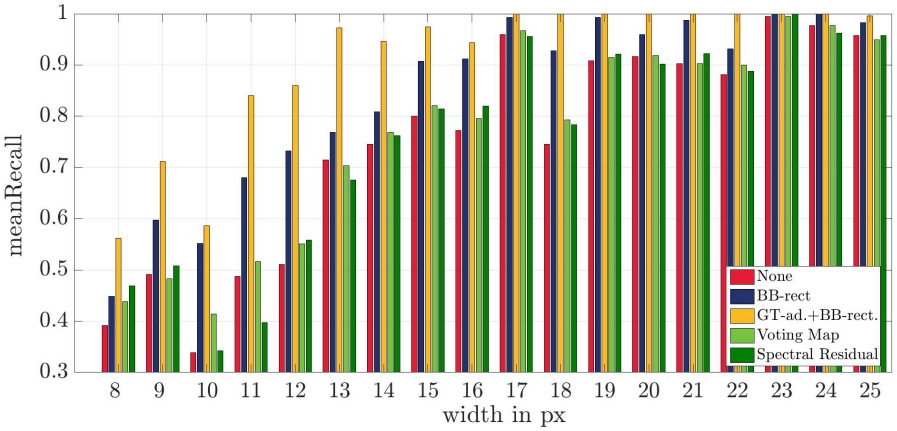


Fig. 9: Detailed visualization of the mean Recall for small objects of sizes 8-25 px width.

increase the Recall for larger objects up to 30 px. Figure 9 visualizes the Recall in detail for objects of a width between 8-25 px. Performance of the region proposal network decreases substantially for objects smaller than 13 px showing a technical boundary for the detection of small objects.

Especially in this challenging region, saliency maps as external data improve the Recall. The reasons is that both saliency-like maps are well suited for small object detection. In addition, the maps require few computations during testing time and are, hence, suitable for real-time applications.

5 Conclusions

This work studies the final feature maps of an RPN before bounding box regression and binary classification is applied to improve the overall detection performance, with the following key findings:

1) Feature map clustering and net surgery exhibit key feature maps that contribute individually to different object sizes.

2) Post-trained net surgery is used to cluster maps with similar activation patterns. For the task of detecting single objects, the information from a larger group of final feature maps is relevant. This group often includes redundant information and allows to reduce the network size by considering the key feature maps only.

3) Additional feature maps/priors improve the detection performance for very small objects. We studied a variety of prior maps to gain further understanding on how to efficiently incorporate additional prior information into the RPN. It is shown that the incorporation of additional prior information resulted in a higher performance gain for smaller than larger objects.

Finally, bounding box regression and binary classification require a different feature representation strategy, whose more detailed elaboration will be part of future work.

References

1. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. IEEE transactions on pattern analysis and machine intelligence **34**(11) (2012) 2189–2202
2. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision **104**(2) (2013) 154–171
3. Batzer, A.K., Scharfenberger, C., Karg, M., Lueke, S., Adamy, J.: Generic hypothesis generation for small and distant objects. In: Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on, IEEE (2016) 2171–2178
4. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1440–1448
5. Lenc, K., Vedaldi, A.: R-cnn minus r. BMVC (2015)
6. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. (2015) 91–99
7. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 779–788
8. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
9. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision, Springer (2016) 21–37
10. Li, W., Breier, M., Merhof, D.: Recycle deep features for better object detection. arXiv preprint arXiv:1607.05066 (2016)
11. Brazil, G., Yin, X., Liu, X.: Illuminating pedestrians via simultaneous detection & segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 4950–4959
12. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al.: Speed/accuracy trade-offs for modern convolutional object detectors. In: IEEE CVPR. (2017)
13. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1259–1267

14. Kong, T., Yao, A., Chen, Y., Sun, F.: Hypernet: Towards accurate region proposal generation and joint object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 845–853
15. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. Volume 1. (2017) 4
16. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: Advances in Neural Information Processing Systems. (2016) 4898–4906
17. Fan, Q., Brown, L., Smith, J.: A closer look at faster r-cnn for vehicle detection. In: Intelligent Vehicles Symposium (IV), 2016 IEEE, IEEE (2016) 124–129
18. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision, Springer (2014) 818–833
19. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE (2007) 1–8
20. Fattal, A.K., Karg, M., Scharfenberger, C., Adamy, J.: Saliency-guided region proposal network for cnn based object detection. In: Intelligent Transportation Systems (ITSC), 2017 IEEE 20th International Conference on, IEEE (2017) 1–8
21. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 580–587
22. Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J.: Pruning convolutional neural networks for resource efficient transfer learning. ICLR 2016 (2016)