DISSERTATION

Computer Vision for Distant Vehicle Detection: How to Find Region Proposals for Low-Resolution Objects?

Vom Fachbereich
Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt
zur Erlangung der Würde einer
Doktor-Ingenieurin (Dr.-Ing.)
genehmigte Dissertation

von

Ann-Katrin Fattal, M.Sc. Civ.Ing.
geborene Batzer am 05. November 1990 in Frankfurt am Main

Referent:	Prof. Dr. Jürgen Adamy
Korreferent:	Prof. Dr. Sören Hohmann
Tag der Einreichung:	08. Mai 2019
Tag der mündlichen Prüfung:	30. Oktober 2019

D17
Darmstadt 2019

Fattal, Ann-Katrin:
Computer Vision for Distant Vehicle Detection:
How to Find Region Proposals for Low-Resolution Objects?

Darmstadt, Technische Universität Darmstadt,
Jahr der Veröffentlichung der Dissertation auf TUprints: 2020
Tag der mündlichen Prüfung: 30.10.2019

Veröffentlicht unter CC BY-NC-ND 4.0 International
`https://creativecommons.org/licenses/`

Preface

This dissertation is the result of my work as a PhD student at the laboratory of control methods and robotics at the Technical University of Darmstadt and in the department of advanced engineering within Continental AG in Frankfurt. Here, I wish to thank my colleagues, friends and family who made this thesis possible with their advice, encouragement, feedback and support.

At first, I wish to thank Prof. Jürgen Adamy for his motivating discussions and sharing of knowledge. I am also grateful to Prof. Sören Hohmann for accepting to act as a second referee. I further owe my deepest gratitude to Dr. Michelle Karg and Dr. Christian Scharfenberger as my technical supervisors. Thank you for all the fruitful, deep and insightful discussions we had within these years.

Furthermore, I want to thank Dr. Sighard Schräbler, Dr. Stephan Kirstein and Dr. Ralph Grewe for providing the needed hardware for this project, and our great discussions within the broad field of machine learning.

Special thanks to all my past colleagues and students for making my PhD-time an unforgettable time at Continental: Philippe, Historei, Christopher, Lucas, Patrick, Daniel D., Thomas L., Kerstin, Annemarie, Matthias, Fabian, Attila, Florian, Daniel K., Dennis, Rex, Georg, Enno, Enrico, Marian, Julien, Thomas B., Jürgen, Johannes, Stefan H., Jonas, Nicolai, Rujiao, Andreas, Frank.

Last but not least, I want to thank my parents and brothers for their support in all my decisions. The same holds for Thomas, whom I deeply thank for his extensive support and understanding in all matters.

Abstract

Safety is crucial to the development and acceptance of assisted and highly automated driving functions. In 2017, 69.3% of German fatal accidents happened on roads where the speed limit was not enforced or higher than 100km/h. At this speed, to perform safe driving maneuvers, the environment perception is a key element. Detecting objects in distances up to 200m is instrumental in anticipating potential obstacles.

Due to hardware limitations, an automotive camera maps cars in e.g. 200m distance to an image of only 8px width. Hence, the absence of local details degrades the state-of-the-art detection methods designed for detecting bigger sized objects.

The scope of this thesis is to develop, extend and evaluate object region localizers to improve the detection range of cameras. A saliency inspired voting map is proposed that highlights anomalies in automotive scenes. The environment is modeled with few homogeneous regions representing the background within the image. Such global features allow detecting small object regions.

Inspired by the concept of learning features, this thesis presents machine learning methods detecting small objects. Existing labeled data sets such as the KITTI data set only have object regions which sizes are larger than 25px height. The presented methods in this thesis are performed against a newly created data set with 67% of object regions having a width of 8-30px, a range that has rarely been subject to research yet.

Convolutional Neural Network based localizers have been evaluated and extended. To maintain a low computational power, only small networks can be used. However, such networks are limited to the usage of local features. An incorporation of global generic priors to local networks is proposed, which increases the recall especially for small object regions.

The parameters to adjust Region Proposal Networks (RPNs) for the special case of small objects are further optimized and the main parameters are identified. A novel relevance based net-surgery is introduced, allowing to select the most relevant features while maintaining the recall of the RPN. It is then possible to reduce the network size to these few features.

Zusammenfassung

Sicherheit ist ein wichtiger Aspekt in der Entwicklung und Akzeptanz von assistierten und hoch automatisierten Fahrfunktionen. Im Jahr 2017 geschahen 69.3 % aller tödlichen Unfälle in Deutschland auf Landstraßen oder Autobahnen mit hohen Geschwindigkeiten über 100Km/h. Bei solchen Geschwindigkeiten ist eine robuste Umfelderfassung zu sicheren Fahrmanövern notwendig. Um in solchen Situationen sicher zu fahren, muss eine robuste Objekt Detektion bis 200m Entfernung gewährleistet sein.
Aufgrund von limitierter Hardware kann eine automotive Kamera ein Auto in z.B. 200m Entfernung nur auf ca. 8 px abbilden. Dadurch sind lokale Details nicht abgebildet und aktuelle Objekt Detektionen verlieren stark an Performanz, da sie für größere Objekte ausgelegt sind.
In dieser Thesis werden Methoden zur Objektregionen Lokalisierung in Bildern entwickelt, erweitert und ausgewertet um die Detektionsrate in der Entfernung zu erhöhen.
Dazu wird ein von Aufmerksamkeitskarten inspirierter Ansatz entwickelt, der Besonderheiten in automotiven Szenen hervorhebt. Der Hintergrund im Bild wird dabei durch wenige homogene Bereiche modelliert. Solche globalen Ansätze erlauben die Detektion für kleine Objekte. Aktuelle Datensätze, wie bspw. der KITTI Datensatz, beinhalten minimal Objekte der Höhe 25px. Zur Evaluation der hier entwickelten Methoden wurde ein anspruchsvoller Datensatz generiert, in dem 67% der Objekte 8-30px breit sind. Dies zeigt, dass dieser Bereich von kleinen Objekten noch nicht Gegenstand aktueller Forschung ist.
Faltende Netzwerke, welche die Merkmal Extraktion durch Parameteroptimierung lernen, eignen sich ebenfalls zur Lokalisierung von Objektregionen in Bildern (RPNs). Um jedoch den Rechenaufwand für automotive Anwendungen gering zu halten, eigenen sich vergleichbar kleine und dadurch lokal beschränkte Netzwerke. Daher wird die Einarbeitung von ausgewählten globalen Priors vorgeschlagen und Untersuchungen zeigen eine Verbesserung des Recalls für RPNs. In einer Parameteranalyse für faltende Lokalisierungsmethoden wird der Recall weiter optimiert und die wichtigsten Parameter identifiziert. Zur weiteren Optimierung von Merkmalen innerhalb des Netzwerks, wird eine neuartige Relevanz basierte Netzwerk-Operations-Methode entwickelt, die es ermöglicht, die markantesten Merkmale des Netzwerks zu identifizieren und das Netzwerk auf diese Merkmale zu verkleinern bei nahezu gleichbleibendem Recall.

Contents

List of Abbreviations **xi**

1 Introduction **1**

1.1 Technical Motivation 4
1.2 An Object in Computer Vision 5
    1.2.1 The Camera and Image 5
    1.2.2 Physical Limitations 7
    1.2.3 An Object 8
    1.2.4 Object Localization 10
    1.2.5 Feature Extractor 11
1.3 Limitations in Automotive Applications 13
1.4 Goals and Contributions 17
1.5 Structure of the Thesis 19

2 Fundamentals and Related Approaches **21**

2.1 Visual Attention or Saliency Maps 22
    2.1.1 Application on Small Object Regions 23
2.2 Hand-engineered Object localization methods 25
    2.2.1 Merging Methods 25
    2.2.2 Scoring Methods 25
    2.2.3 Application on Small Object Regions 26
2.3 Region Proposal Network 27

2.3.1 Background to Convolutional Neural Networks . . 28
2.3.2 Network Architecture 30
2.3.3 Application on Small Object Detection 33
2.4 Feasibility of Super Resolution 34
2.4.1 Concept of Super Resolution 34
2.4.2 Experiment 35
2.4.3 Registration Errors of Existing Algorithms 35
2.4.4 Reconstruction Errors of Existing Algorithms . . . 37
2.4.5 Discussion 38
2.5 Conclusion 39

3 Anomaly Detection / Hypothesis Generation 41
3.1 Motivation 42
3.2 Outline 43
3.3 Proposed Approach 45
3.3.1 Patch Selection in Image Zones 45
3.3.2 Generation of Voting Maps 47
3.3.3 Combination of Voting Maps 48
3.3.4 Generating Hypotheses 49
3.4 Evaluation 50
3.4.1 Data set and Labeling 50
3.4.2 Evaluation Metrics 50
3.4.3 Results 51
3.4.4 Comparison to Objectiveness Measure 57
3.5 Summary and Conclusion 59

4 Incorporation of Global Priors 61
4.1 Motivation 61
4.2 Outline 63
4.3 Incorporation of External Data 64

4.3.1 Faster R-CNN with ZF Net 65
4.3.2 Incorporation of Global Prior 66
4.3.3 Prior Maps 67
4.4 Metric and Results 69
4.4.1 Data set and Data Augmentation 69
4.4.2 Training 71
4.4.3 Evaluation Metrics 73
4.4.4 Results 76
4.4.5 Discussion 84
4.5 Summary and Conclusion 85

5 Relevance Based Net-Surgery 87
5.1 Motivation 88
5.2 Outline 89
5.3 Network Architecture 91
5.4 Net-Surgery 92
5.4.1 Relevance-based Clustering of Feature Maps 93
5.4.2 Incorporation of Prior Information 95
5.5 Metric and Results 98
5.5.1 Data set and Data Augmentation 98
5.5.2 Network Architecture and Training Details 99
5.5.3 Evaluation Metrics 100
5.5.4 Results of the Net Surgery 100
5.5.5 Influence of External Data on the Recall 102
5.6 Summary and Conclusion 108

6 Conclusion and Outlook 109
6.1 Conclusion 109
6.2 Outlook 111

A Gauss-Carpet Super Resolution Method 113

References 115

List of Abbreviations

ADAS	Advanced Driver Assistance Systems.
AuC	Area under Curve.
BB	Bounding Box.
BB-rect	Bounding Box in form of a binary map.
BCN	Binary Classification Branch.
CCD	Charged-Coupled Device.
CFT	Continuous Fourier Transformation.
CMOS	Complementary Metal-Oxide-Semiconductor.
CNN	Convolutional Neural Network.
DB-Scan	Density Based Spatial Clustering for Applications with Noise.
DFT	Discrete Fourier Transformation.
EXT	External data in form of a feature map.
FN	False Negative.
FOV	Field of View.
FP	False Positive.
FT	Fourier Transformation.
GMACs	Giga MAC (multiply-add/accumulate unit) per second.
GRBG	Certain constellation of red, green and blue sensitive photoactive regions.
GT	Ground Truth data.
GT-adjusted	Ground Truth data for regression tasks only.
HOG	Histogram of Oriented Gradient.

HR	High Resolution.
ImageNet	A large visual database organized according to the WordNet hierarchy and designed for use in visual object recognition software research.
IoU	Intersection over Unit.
KITTI	A data set provided by [40].
LR	Low Resolution.
mA	mean Area.
mD	mean normalized Distance.
nm	Nanometer = 10^{-9} meter.
Pascal VOC 2007	A visual object classes challenge 2007 [31].
PDF	Probability Density Function.
POCS	Projection onto Convex Sets Approach.
PSF	Point Spread Function.
px	Pixel.
RGB	Certain constellation of red, green and blue sensitive photoactive regions.
ROI	Region of Interest in an image.
RPN	Region Proposal Network.
SGD	Stochastic Gradient Descent.
SIFT	Scale Invariant Feature Transform.
SNR	Signal to Noise Ratio.
SSD	Single Shot Detector.
SVM	Support Vector Machine.
THz	Frequency in terra hertz.
TP	True Positive.
VGG16-Net	A network model proposed by K. Simonyan and A. Zisserman in 2016.
Yolo	You Only Look Once.
ZF-Net	Network model proposed by Zeiler and Fergus.

Chapter 1

Introduction

In this introduction, the statistics of traffic accidents is analyzed to emphasize the motivation of this work. The following technical motivation introduces possible sensing methods for assisted and highly automated vehicles, succeeded by an explanation of common terms and main concepts within the thesis. This Chapter concludes with the limitations in an automotive environment and contributions presented within the thesis.

Since the invention and usage of cars, accidents contributed to the development for better safety in vehicles. In this Section, the statistics of accident occurrence in Germany since 1991 is examined to understand the need for further improvements on the safety for drivers and their passengers. In Figure 1.1, the relative change of motorized vehicles, injured persons and deaths due to accidents over the years since 1991 to 2017 in Germany is depicted. It shows a decrease of overall accidents even though the number of registered vehicles increases during this time period. One factor contribution to this decrease might have been the increasing usage and development of active and passive driver assistant and safety systems such as seat belts, collision warnings, lane-departure warnings, adaptive cruise control, traffic jam assist, and park assist.

To understand the accident occurrence with fatal injuries in more detail, a closer look on the year 2017 in Germany is taken. During this period, 3180 accidents with fatal injuries happened. An amount of 69.3% of these deaths occurred on countryside highways and motorways, while only 31.4%[1] of all accidents including injured persons and deaths happened in the same region. This statistic shows that accidents outside of towns and

---

[1] Data taken from [26].

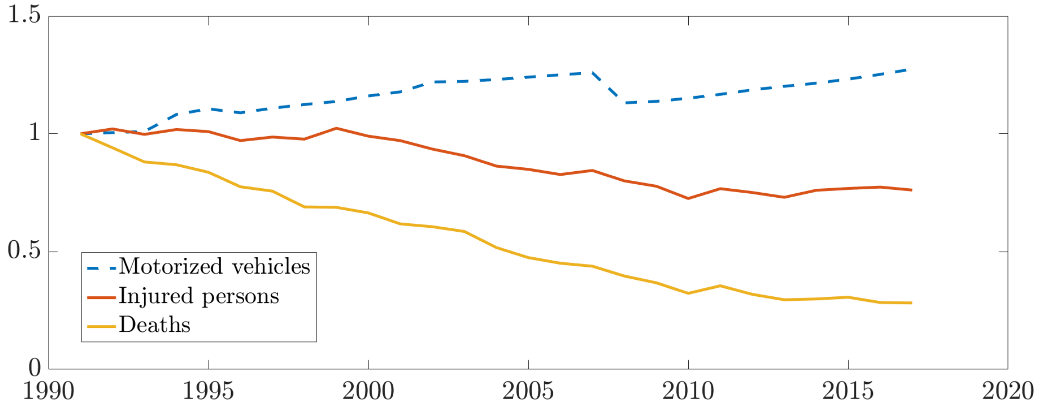


Figure 1.1: Relative change of number of motorized vehicles, injured persons and deaths in the years 1991-2017, following [26, 72].

cities comprise a higher fatality than inside cities. One factor might be the speed during the accident. While the speed limit inside towns is set to 50 km/h, the speed limit on countryside highways is set to 100 km/h and there exists no speed limit on motorways in Germany.

At high speeds, the breaking distance is much larger than with lower speeds. Also, the distance driven while obstacles or other unexpected road occurrences are realized by the driver (the time of realization and reaction) is much larger. In Figure 1.2, the velocity and distance to start-breaking point is shown at different breaking powers in an ideal world[2]. Commonly maximal up to 3 m$/s^2$ of breaking power is seen as comfortable, while a breaking power of 6 m$/s^2$ is categorized as emergency braking. Hence, to avoid any accidents on countryside highways and motorways, obstacles or incidents on the road should be recognized at least 64.3 m prior at a velocity of 100 km/h to initiate an emergency braking. At a higher velocity such as 140 km/h this distance increases to 126 m. To drive more comfortable with a breaking power of 3 m$/s^2$, this distance increases even more to 128.6 m for a velocity of 100 km/h and 252 m for 140 km/h respectively.

These statistics show the need to further increase the safety of passengers especially on regions where vehicles can drive in high speeds such as countryside highways and motorways. As of today, there exists many driver assistance systems, which warn the driver in the event of danger such as e.g. Forward Collision Warning, Adaptive Cruise Control and

---

[2] An ideal world means no external influences such as e.g. coefficient of friction of the street and tire are taken into account.

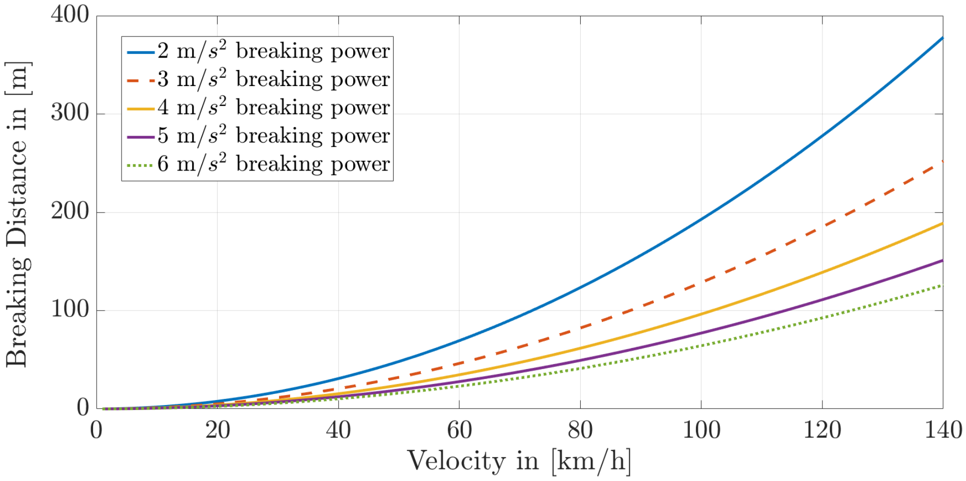


Figure 1.2: Breaking distance in m at different velocities and breaking power in an ideal world scenario.

Minimal Risk Maneuver and increase therefore the safety of the vehicle. Such safety functions make use of an environment perception, which is designed to describe the world in front or around the vehicle so that all important information is captured and safe maneuvers can be triggered. As those functions are only assistance functions, the driver remains as a supervisor of all driving functions and is able to react at any moment. Nowadays, there is an increasing demand for advanced safety [33, 46] and comfort functionality [109, 122] in today's vehicles due to the development of highly and fully autonomous vehicles. In such vehicles, the driver does not supervise the functions any more at all time. Hence, the driver supervision redundancy is not given anymore and in case of system or sensor failures, the remaining redundant systems and sensors need be capable to provoke safe maneuvers until the vehicle reaches a secure state. Such highly or fully autonomous driving offers the possibility to increase security as the system runs constantly without any inattention. High range sensing and redundancies within the whole vehicle system are necessary. Despite the high requirement on safety and reliability, the acceptance of such highly or fully automated driving is directly influenced by the comfort of the driving maneuvers. Too rapid acceleration or braking and cornering sharply reduce the acceptance of autonomous functions and undermine possibilities for increased safety on roads.

This thesis proposes improvements for the environment perception in large distances of up to 200 m distance to the car. For optimization, the camera-output is chosen as it has the theoretical possibility to sense the whole environment such as distance to obstacles, depth, and object classification. The scope of this thesis is to detect any distant obstacles in front of the vehicle with a front-mounted camera so that eventually following maneuvers can be initiated well ahead to increase safety and remain comfort.

1.1 Technical Motivation

This Section motivates the use of a camera sensor for developing improvements for far range sensing by a short comparison of the different possible sensor systems.

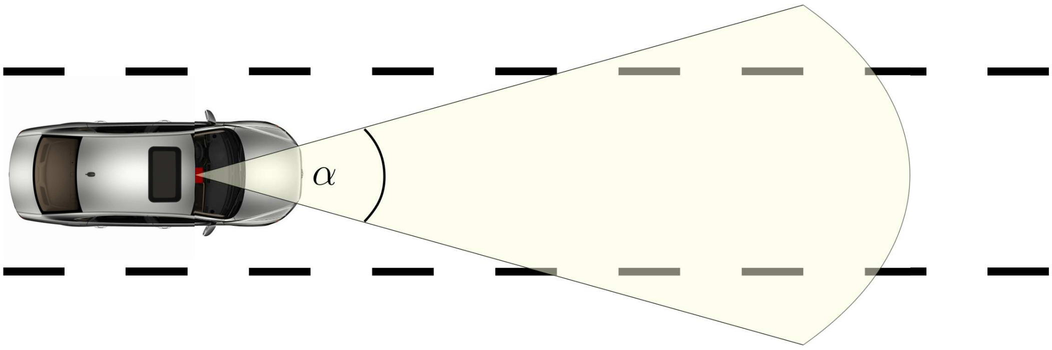


Figure 1.3: Scene perception of a front mounted sensor on a car with an aperture angle α.

A redundant region of coverage is one fundamental requirement on sensory systems to avoid a single reliance on one sensor. Being able to rely on a set of sensors allows for higher security and reliability of the whole environment perception. For vehicles in autonomous mode, a variety of sensors play a role to detect an object in automotive scenarios: radar, camera, ultrasound and lidar scanner[3]. Ultrasound sensors are only feasible for parking maneuvers due to its short sensing range, while radar, camera as well as lidar scanner can deliver sensing data in larger ranges to the sensor. For autonomous cars, the radar[4] is a reliable and necessary sensor due to its precise distance measurements with ranges up to 250 m [25]. However, this data does not allow to draw a detailed conclusion on the type or criticality of the detected object.

---

[3] In a survey paper of F. Ponte [25] the ranges of sensors for automotive applications have been evaluated.

[4] radio detection and ranging

The lidar[5], on the contrary, allows to obtain 3D data points of an object in distances up to 200 m [25]. Unlike radar sensors, the lidar sensors are expensive due to the complexity of mechanical components and are therefore less interesting for automotive applications. The lidar is a sensor which is not yet under development and is not yet set as default for newly built autonomous cars. In the future however, this sensor may show robust and good capabilities for far range sensing.

The camera is a sensor which allows to capture 2D-information of a large section of a scene and its containing objects, due to the digitalized representation of the overall scene. This sensor can be used in close proximity to the vehicle as well as in greater distance. However, the detection range can be estimated for nowadays existing camera systems up to 100 m. The camera is a comparable cheap sensor and is built in all higher class cars nowadays. Hence, this sensor can be considered as a candidate for a redundant sensor to the radar and/or lidar for far ranges. In Figure 1.3, a front mounted sensor and its hypothetical perception area are shown. The apertures angle α of the camera and resolution of the imager define the pixel density of a mapped object in the image. In the following Section, common terms and camera specific limitations are presented.

1.2 An Object in Computer Vision

In this thesis, the performance of different object detection steps is analyzed. To achieve this, it is eminent to understand what exactly an object in computer vision and especially in this thesis is. In this Chapter, the generation of a digital image and its mathematical limitation are explained. The concept of an object in a digital image and the basic ideas on how to detect an object in computer vision are presented.

1.2.1 The Camera and Image

A camera captures discretely the electromagnetic radiation that is either reflected or emitted from the surrounding or scene in front of the camera. This electric radiation is hereby in the visible frequency range of 400-789 THz with wavelengths between 390-700 nm. This radiation is then

---

[5] light detection and ranging

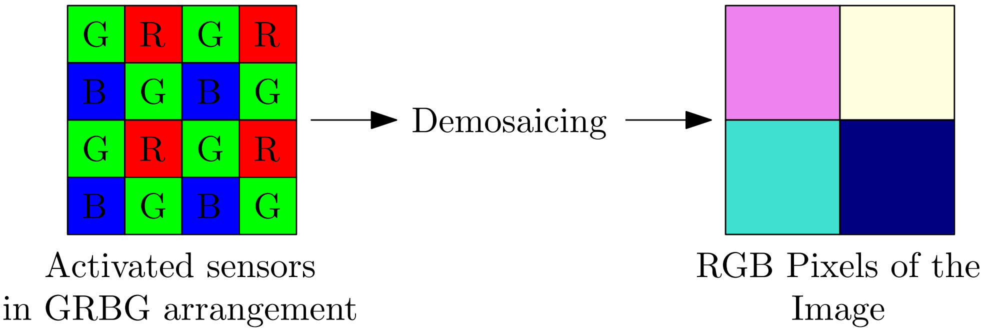


Figure 1.4: A Sensor with GRBG arrangement with a Bayer filter captures light and this light is transformed to image pixels.

focused through a lens to form an image on a CCD- or CMOS-detector (Charge-Coupled Device or Complementary Metal-Oxide-Semiconductor). The detector, often also called imager, is composed of an array of photoactive regions. When visible light hits such a region, an electric charge proportional to the light intensity at that location is accumulated and transferred to digital values. The digital output is a matrix where each entry of the matrix corresponds to one photoactive region from the sensor. This matrix is called an 'image' in computer vision and each entry of the matrix is a pixel in the image. Since each region is only sensitive to a specific wavelength range such as green, red or blue, a Bayer filter [6] is applied on the sensor to be able to capture different color information of the scene. Four different sensitive regions are arranged in a certain pattern such as e.g. GRBG as shown in Figure 1.4. The signals from those four regions are then demosaiced to form one pixel with three values for red, green and blue. A whole camera detector is consisting of many of such sensor arrangements to transform as much different light emissions form the scene it is pointing to. The number of such single sensors in width and height is referred to the image resolution. The image is then a discrete representation of the captured scene in RGB values per pixel. In this thesis, the camera is front-mounted on the windshield and the images are captured with a Bayer filter in GRBG arrangement with a rate of 16 frames per second. The resolution is changing depending on the camera, however, in this thesis, only one camera set up is used with a resolution of 1024×640 px and an aperture angle of 53°. In such a camera system, one degree is described by 20 px. A car in 100 m distance to the camera is then described by 18 px in width and in 200 m distance by 8 px on the imager.

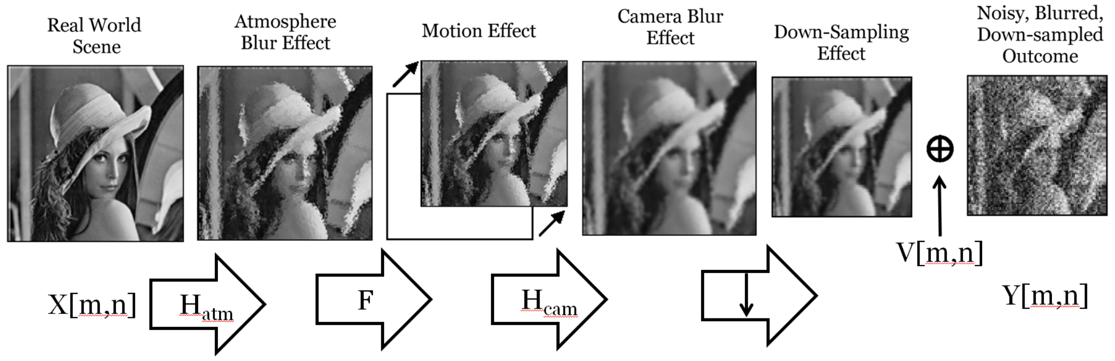


Figure 1.5: Model of capturing an image with an optical camera. Small pictures are taken from [34].

1.2.2 Physical Limitations

Theoretically, it is possible to capture the whole scene in every detail in front of such a camera. An ideal picture would contain all information: color and brightness of each location and/or all frequencies in the pictures. A digital image does not provide all this information since all information is captured discretized as described in the previous Chapter. As a result, it is not possible to obtain the whole information of the scenery observed by the camera.

Image observation model The camera lens can introduce additional optical distortion during image capturing besides certain loss of information due to aliasing. optical distortion can be modeled with a matrix H_{cam} convolving the captured image. Motion during shutter time leads to warping captured by the matrix F and the atmosphere H_{atm} adds blur which impacts the quality of the captured image. Furthermore for an image of the size $m \times n$, statistical noise $V\left[m,n\right]$ in the photo detectors decreases the signal to noise ratio SNR in each pixel. Combining all these aspects together, it is possible to sketch the whole image capturing process as described in the following passage.

Starting with a real scene world $X\left(x,y\right)$ the signal finally obtained is a noisy, blurred and down-sampled representation $Y\left[m,n\right]$ of the real scene.

As any signal can be written as a sum of periodic functions, it is possible to see the limitations best in the discrete Fourier domain. An image with N pixels width has N discrete Fourier coefficients in x-direction and is therefore band limited [130]. In the situation where the scene contains

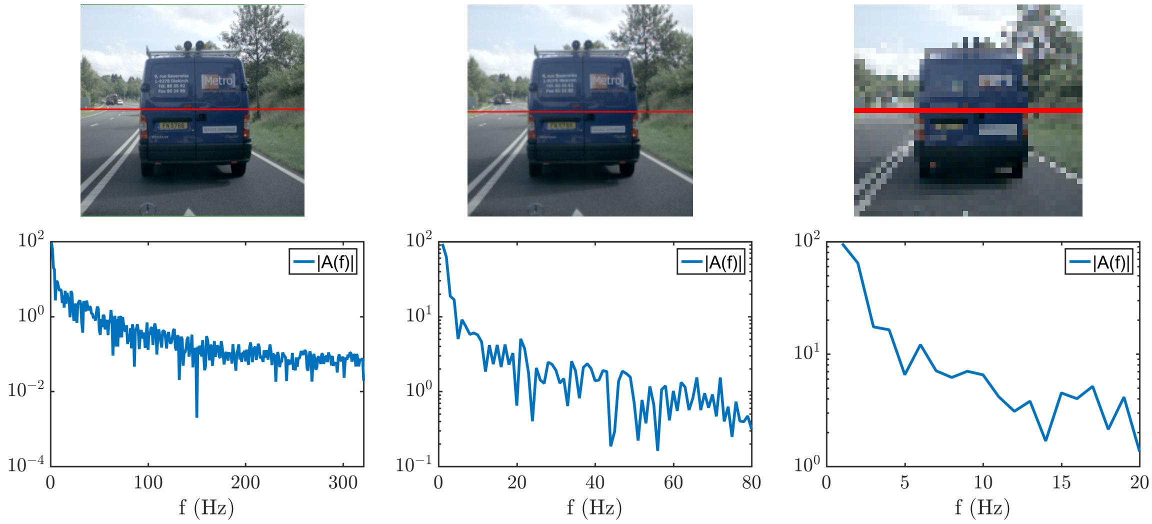
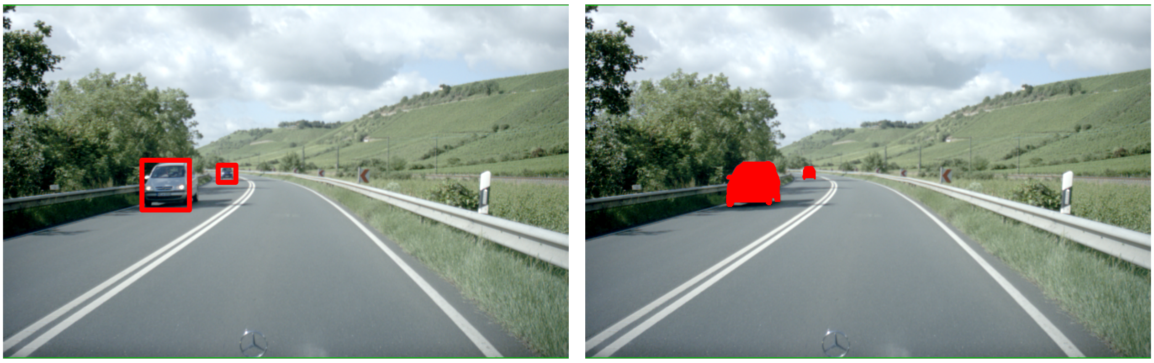


Figure 1.6: In this Figure, the impact on the amplitude of a discrete Fourier transformation with decreasing resolution is shown. For a clearer visualization, only one horizontal column of the image is used resulting in a 1D-transformation. The red line within the Figures indicates the chosen column. The resolution of the images from (a)-(c) is decreased subsequently by four. In the second row, the corresponding Fourier transformations are shown. Higher frequencies vanish with decreasing resolution.

only low frequencies, a lower number of Fourier coefficients is needed to describe the whole scene. For additional higher frequencies such as well defined edges or small objects, more Fourier Coefficients and therefore more pixels per image are needed to describe the scene [124]. This is a direct result of the Nyquist–Shannon sampling theorem [98], which states that 'If a function x(t) contains no frequencies higher than B hertz, it is completely determined by giving its ordinates at a series of points spaced 1/(2B) seconds apart'. As a result of a limited pixel number, sharp edges can not be represented perfectly and edges as well as small objects, which need to be described by high frequencies, seem to be blurred or are in extreme cases not visible anymore. Information on high frequencies which got lost this way can not be recovered without further input (see Figure 1.6).

1.2.3 An Object

In contrast to the real physical world, an object in computer vision is described by its pixel values in the image. Depending on the application, an object can have different properties. Unlike in physical objects, in



(a) (b)

Figure 1.7: An image captured by a front mounted car camera with a resolution of 1024×640 px with two front facing cars as objects. In 1.7a, the objects are described by bounding boxes (red rectangles) around the front appearance of the cars. In 1.7b, the objects are described by all pixels belonging to each object.

computer vision all objects are represented by 2-D data and hence, for each type of object, it has to be predefined which parts of the object in 2-D representation belong to the object. In the automotive environment a car in an image is often defined only by its front or back appearance in the image. All pixels belonging to one object contain information that was captured by the camera sensors of a physical object of the scene in front of the camera. To describe an object within an image, its location, size and type of object is of interest. In this work, Chapter 4 and 5 make use of bounding boxes which are rectangular boxes large enough to enclose all pixels that belong to one object. The corners of such a bounding box are given in image coordinates x and y as $[x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}}]$ (see Figure 1.7). Within this thesis, only objects that are captured by a front-mounted car camera are of interest. For each algorithm and evaluation presented in this thesis, it is stated which object and which appearance is of interest (Chapters 3, 4, 5).

To detect an object, two tasks need to be fulfilled:

- localization of the object within the image in x and y coordinates
- classification of the object type or class such as e.g. car or truck.

Humans can fulfill both tasks easily. In computer vision, these tasks can be automatically learned by machine and/or deep learning. Thereby, extracting relevant information from the pixel values is referred to as feature extraction. For localization, less class specific features are needed since it is only of interest, where an object exists within an image. Often localization is a background and foreground detector, which determines which regions within the image belong to an object. Those regions without any object specific classification are called Regions of Interest (ROI) [12]. Class specific features are then determining through a classification method which type of object is present within the ROI.

1.2.4 Object Localization

A naive method to localize an object region within an image is to use a sliding window approach. A predefined window is slid pixel by pixel over the entire image. For each window, a trained classifier such as a SVM or CNN decides whether any and which object is present in the window. Using this method comprises an exhaustive search for objects over the whole image, implying that all possible locations within the image are considered. Additionally to this exhaustive search, several aspect ratios and scales need to be considered for all object sizes as object classifiers are trained for certain scales and aspect ratios. Hence, the usually expensive object classification method has to be performed on the maximum possible regions to detect certain objects. A typical region of interest detection requires 10^6 classifications per image of size $\sim 500 \times 300$ px [55]. Hence, a localizer is desired, which proposes a subset of regions which contains an object with a higher probability than other regions.

Object region proposal methods intend to distinguish between background and foreground with object specific cues. The output of most region proposal methods is given as a bounding box around an object region within the image. The main goal of object region proposal methods is to find a limited number of object regions for a following object classification. Different aspect ratios as well as sizes impose challenges to the used methods. The main objective of a region proposal method is to gather all possible object regions since the subsequent classification step can not retrieve any missing object regions. Such missing object regions impact severely the detection rate. Slightly too many proposed regions can be adjusted during the classification. Too many incorrectly proposed regions might impose a higher chance for a false classification. Especially for small objects where only little information is available, these false classifications need to be avoided by the localizer.

1.2.5 Feature Extractor

To detect an object, an abstract representation of the image is needed to transform pixel values into object specific features. A localization and classification method is then trained with features to determine its position and type. Feature extraction is a field of constant research as any method works better when the quality of features is better adapted to the application. Until approximately 2012, only hand-engineered feature extractors were used which were developed by scientists and engineers. After 2012, the feature extraction process is learned by representation learning and the feature extractor becomes more and more class or application specific.

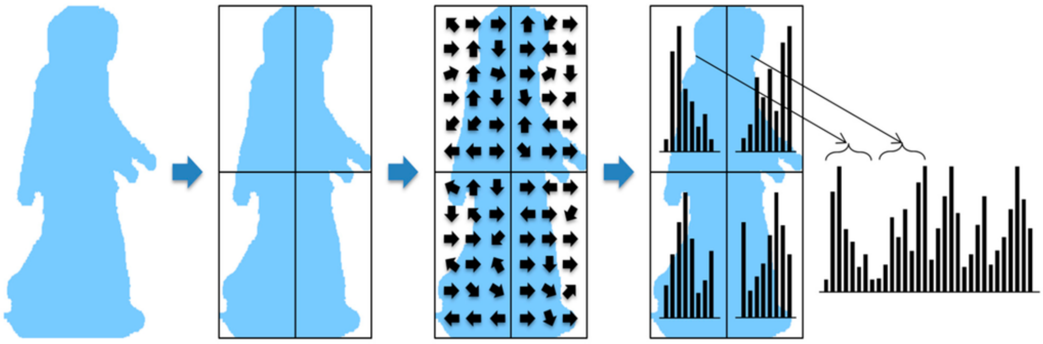
Hand-engineered Features

Although the following hand-engineered feature extractors are not used within this work, it is useful to understand the influence on state-of-the-art techniques. Limitations on hand-engineered features are presented shortly.

Edge based Edge based feature extractors consist of kernels or filters which are matrices with predefined entries. Those filters are convolved with the image to generate a feature map. The Soebel-filter [120] uses the following 3×3 matrices $\mathbf{S}_x$ and $\mathbf{S}_y$ as edge detectors in x- and y-direction.

$$\mathbf{S}_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} , \mathbf{S}_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

Figure 1.2.5 shows feature maps based on the Sobel-filter. An extension to the edge detection in x- and y-direction are the Gabor filters [24], which can cover any direction of edges as well as the possibility to detect different frequencies within the image than only explicit edges, which is often referred to wavelets [121]. Rather than detecting edges, the Harris Corner Detector [48] detects corners as features inside the image. The basic idea behind the edge detector is to compare the change of edges in different directions as a corner implies a large change of edge direction in at least two dimensions. Corners are often referred to keypoints within an image as they are significant and distinctive features.



MHI image Cell division Gradient orientation Histogram Vectorization

Figure 1.8: How to obtain HOG features of an image. Taken from [30].

Gradient/histogram-based Another idea of feature extractors is based on the gradients in different directions. The SIFT feature extractor (**S**cale **I**nvariant **F**eature **T**ransform) finds at first keypoints within an image and analyses then the surrounding of these keypoints [85]. To obtain the feature of the keypoints, an oriented histogram, meaning a histogram in different directions to the keypoint, is calculated. For rotation invariance, a certain main direction is defined and all directions are displayed relatively to this main direction. The angle of directions are discretized and the histogram contains the information to each region (f.ex. 0-10 degrees, 11-20 degrees and so on). The formed histogram is then transformed to a description vector and this vector is then normalized. The final vector is called SIFT feature of one keypoint [125].

A HOG feature (**H**istogram of **O**riented **G**radients) incorporates a similar idea, however, the difference resides in how the histogram is calculated namely over uniformly distributed blocks comparable to a sliding window approach [22, 39]. Each block is divided in cells and for each cell, an oriented histogram is calculated. The vector which contains the histograms of all cells in one block is the HOG feature for the block (see Figure 1.8).

Machine Learned Features

The previous Section 1.2.5 described hand-engineered filters or feature extractors. Recently, a new type of feature extractor is used, which are machine learned arrays of filters on a labeled data set to generalize the feature

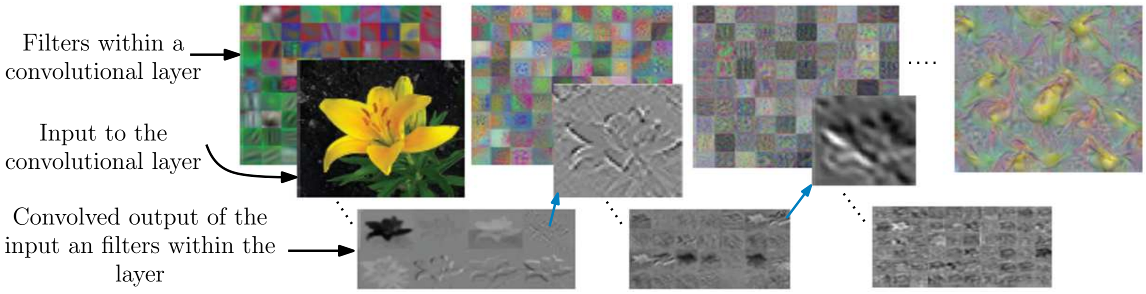

Filters within a convolutional layer
Input to the convolutional layer
Convolved output of the input an filters within the layer
....

Figure 1.9: Different features through learned filters in different layers within an convolutional neural network. The output of one array of filter is the input to the next learned filter array.

extraction. The main difference is that the filter entries are learned and adjusted to the underlying problem or application. It can be seen that machine learned filters often exhibit similar appearance to hand-engineered edge or corner filters, while other filters contain very class specific features (see Figure 1.9). In several cases, the learned features comprise a better representation of the object than hand-engineered features. In convolutional neural networks, machine learned features are stacked behind each other via convolutions. Such stacking gives the possibility to evolve features based on different field of views on the input image. In Section 2.3, such convolutional networks are explained in detail.

1.3 Limitations in Automotive Applications

In the automotive environment and especially autonomous environment, several characteristics are given such as demand of high availability, low cost and close to real time evaluation of the environment as shown in Figure 1.10. Therefore, an automotive camera is used which comprises a certain maximal resolution, certain size and lens quality as well as an imager. Scenes in front of the camera are hereby represented discretized on the imager. Following such representation high frequencies, which in general describe details within a scene, are not captured. However, such discretized representation is necessary to decrease the data rate and flow per image while still capturing the main information of the scene. In Figure 1.11 and 1.12, an image captured by a typical front mounted camera is shown. The resolution of the image was decreased as indicated and edge based features as well as histogram based features are displayed for each

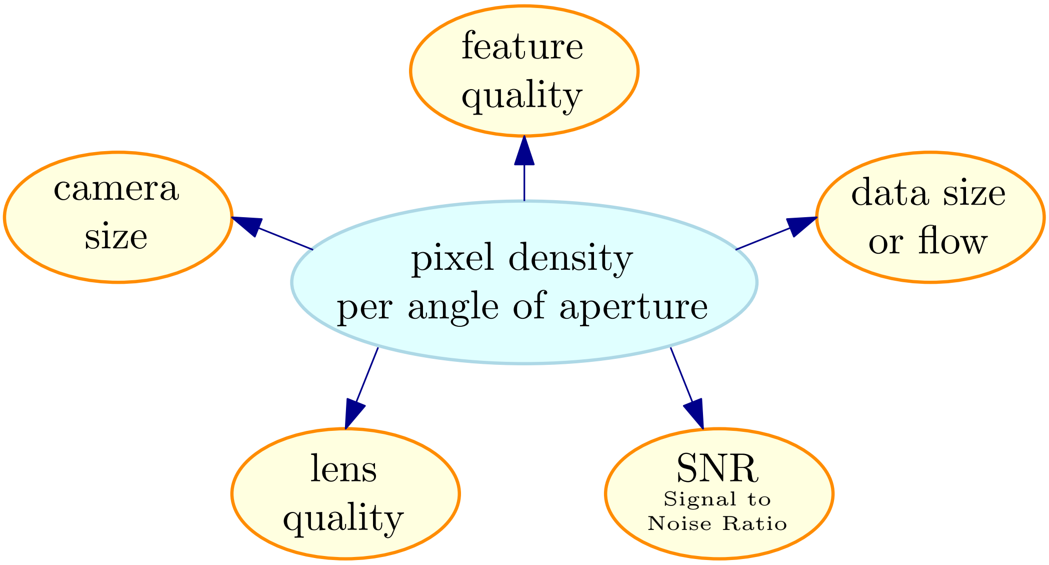


Figure 1.10: The parameter "pixel density" relates to the shown parameters.

resolution. Here, the loss in the feature characteristic becomes visible and it shows - without any additional classification algorithm using the features - the significant loss of information. An object which is located close to the camera is mapped in higher detail than the same object in larger distance to the camera due to the optical path of the reflected light of the objects.

While an vehicle in just ~ 20m distance to the camera is occupying roughly ~ 90×90px, a car in ~ 100m distance is only described with ~ 18×18px on the imager with a resolution of 1024×640 px and an aperture angle of 53°. In this thesis, small objects are defined as objects which are covered by less than 30px in width. A vehicle in ~ 60m distance to the above mentioned camera system is mapped to only ~ 30px in width on the imager. Features such as license plates, rear lights and tires show less distinctiveness or are partly not visible anymore. Hence, the underlying problem to detect small objects with comparable low number of pixels is affected by the absence of strong and distinctive features.

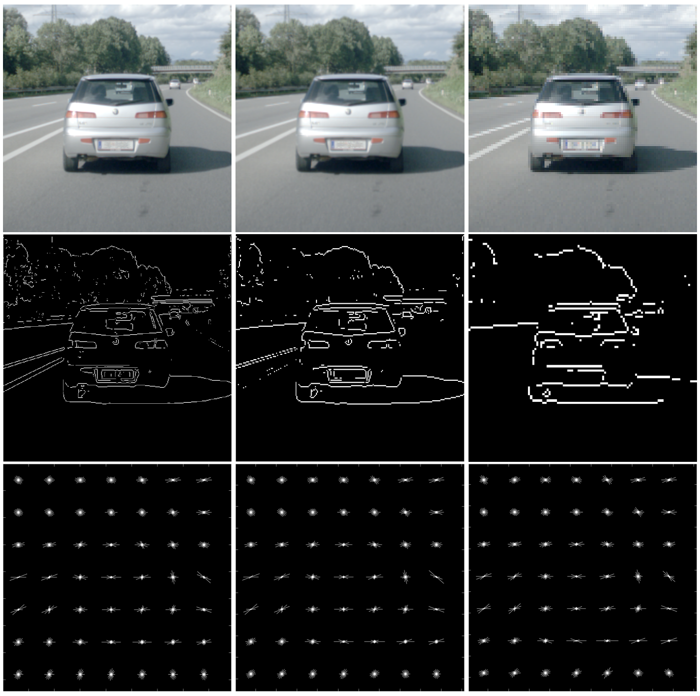
One possibility to overcome this shortcoming would be to increase the number of pixel per angle unit resulting in a more detailed image of the environment. Such higher pixel density can be used to increase the resolution and information of the whole scene in front of the camera. However, the sensor arrangement needs electronics around each photoactive region to transport, transmit and amplify the signal. Such electronics is placed around each photoactive region and hence, the density of light-receiving regions is limited. Additionally, a higher amount of active electronics increases the chance of camera failures due to excess heat evolution.

More photoactive regions on the detector increase the size of the camera as well as the size or number of needed lenses to form an image. An increase in camera size requires also an increase installation space. Both implies higher costs for the camera and is therefore not desired especially in automotive applications. Another possibility is to use smaller but more photoactive region placed side by side. In such an arrangement, the Signal to Noise Ratio (short SNR), which is a measure to compare the level of a desired signal to background noise, gets worse since less light hits each photoactive region. In consequence, the whole image becomes noisier and imposes a worse night performance.

The discussion of camera hardware limitations shows that in principle it is possible to increase the resolution which results directly in stronger features. However, the strategy for any camera constructor will be to keep the resolution as low as possible for the here described reasons. As discussed in Section 1.1, cameras have nowadays a range of approximately 80 m. In the above mentioned camera system with a resolution of 1024×640px and apertures angle of 53°, a car would be covered by 23 px in width. A car in 160 m distance and a doubled resolution in width and height, would then also be mapped on 23 px in width.

Given the high speed and comfort driving sensation for high acceptance (see Figure 1.2), even such a resolution would yet not be enough. This implies the need of work on the algorithmic basis even in the far future of camera usage. For the special case of distant objects, which are located close or on the optical axis of a camera system, only a low optical flow is present and can hardly be used for distant object detection on e.g. motorways.

Additionally, larger images impose a higher computational cost for data extraction as well as for data flows between several units. Increased computational costs demand a larger computational unit within the camera, which is also limited for automotive applications. Hence, it is desired to keep the computational costs of the whole object detection method as low as possible. For the localizer, this implies tt only a minimum number of possible object regions should be proposed to keep the whole object detection computationally as low as possible. The region proposal method itself should operate with minimal required computations.



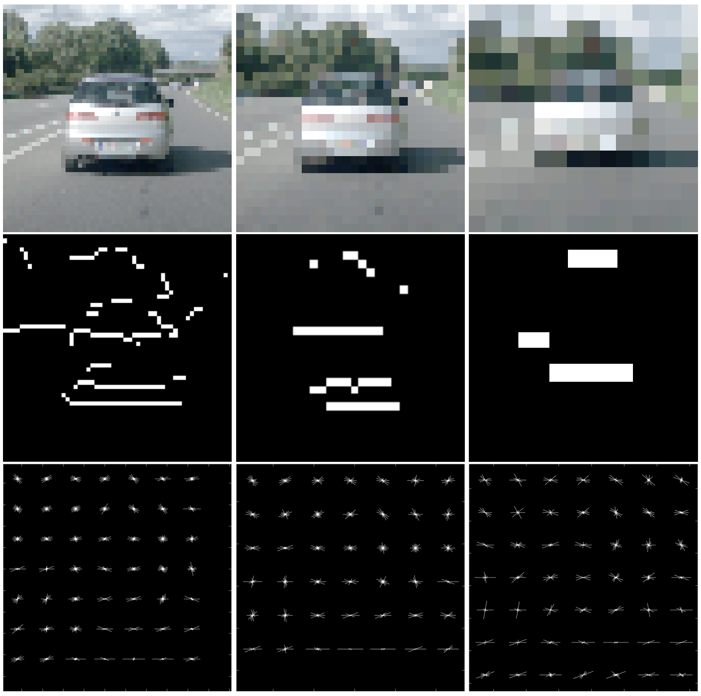
(a) 443×421px (b) 222×211px (c) 111×106px

Figure 1.11: In (a)-(c), the same scene is shown with different image resolutions[6]. The first row shows the color camera image. The second row shows the edge image computed with a edge based Soebel filter. The third row shows gradient based HOG features. To obtain the HOG features, the block size was chosen so that 7 HOG features are always calculated in x-direction.

Consequently, the motivation for this thesis is to increase the safety through better far range detection with limited camera hardware. Additionally, distant object detection with the camera is necessary to enable and/or increase sensor redundancy for far range sensing. For the acceptance of autonomous driving, long range sensing is necessary to support comfortable driving through smooth dynamical steering or acceleration. The focus of this work is to develop new far range sensing methods for an already existing camera set up while keeping the computational costs limited for objects far away from the camera.

---

[6] For the resolution in (a) and (b) the license plate of the shown vehicle is blurred due to personal data protection



(a) 55×53px (b) 28×27px (c) 14×14px

Figure 1.12: In (a)-(c), the same scene as in Figure 1.11 is shown with different image resolutions as indicated. Due to resizing compression in (c) less HOG features in y-direction are computed.

1.4 Goals and Contributions

In this thesis, the camera sensor is used to detect distant objects and the use of this sensor imposes several goals on the method to sense objects in great distance to the vehicle.

- The sensing method needs to be able to deal with real world complexity such as different street scenes and changing object visibility due to distance to the camera. Furthermore, the method needs to adapt to different weather and light conditions such as rain, fog, sun, and driving through a tunnel, or under a bridge with fast changing light conditions.
- Localizing objects within the image which are at distances of up to 200 m to the camera. The camera used within vehicles is bound to hardware limitations such as number of pixels, quality of lens and size. Hence, the method must be able to cope with a limited image quality such as blurring and low pixel density per object.

- To keep the camera system feasible for the use in the automotive environment, the computational power needs to be kept as low as possible while still delivering robust detection performance. High computational power leads to higher system costs and larger power consumption, which makes it less feasible for autonomous cars.

Following these goals and requirements on the method for a camera sensor, in this thesis, the main contributions can be summarized as following:

- Investigation of the feasibility of Super Resolution on automotive images captured by a front mounted camera, Section 2.4.
- Introduction of a novel object localization method based on a saliency-inspired voting scheme for the special use case of small object detection on motorway scenes. It comprises a high adaptability to contrast changes within the image due to different weather conditions or driving under bridges, Chapter 3
- Comparison to the state-of-the-art objectness measure of Alexe *et al.* [3], Chapter 3.
- Adjustments of Region Proposal Networks (RPNs) to detect small objects at low resolution, Chapter 4 and Chapter 5.
- Incorporation of global priors based on saliency maps to combine global and local features for locally restricted region proposal networks, Chapter 4.
- Detailed metric for validating the performance for the special use case of detecting small objects between 8 to 100 px width, Chapter 4.
- Development of a novel investigation method called net-surgery to obtain most relevant features within a network to understand the need of feature characteristics inside RPNs, Chapter 5.
- Feature-relevance induced pruning to reduce the size of the used networks which results in a reduction of computational cost, Chapter 5.
- Creation of data sets for small objects down to object region widths of 8 px to train and evaluate methods. A car in 200 m distance is mapped to˜8 px in width with the used camera setup. Such low pixel-density has rarely been subject of research since most data sets do

not contain such small objects. The e.g. KITTI data set [40] provides only labeled objects of minimal 25 px height. Different weather conditions, day times, tunnel, and bridge scenes are present within the data set, Chapter 3, 4 and Chapter 5.

Many of the contributions have been previously published in conferences and are indicated with references in each Chapter.

As motivated in the previous Sections, the classification step uses most of the computational power due to its extensive class specific feature extraction. Hence, in this thesis, the main focus lies on the detection of object regions to decrease the number of the expensive classification step.

1.5 Structure of the Thesis

The thesis is structured as shown in Figure 1.13. This graphic will be shown at the beginning of each Chapter to lead the reader through the thesis. Topics highlighted in yellow are covered within the Chapter.

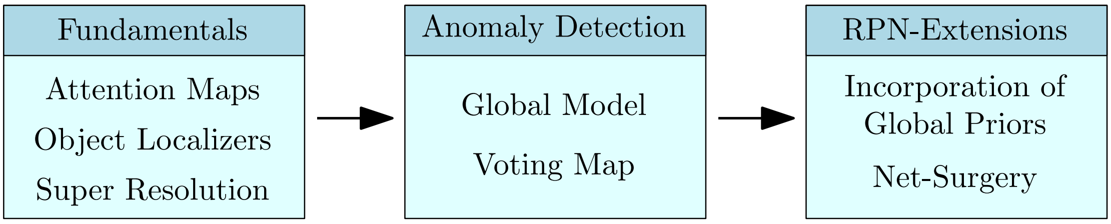


Figure 1.13: Main structure of the thesis. The presented thesis is roughly divided in three main parts. At first related approaches are discussed and related to the requirements of the detection of distant objects. In the second part a novel approach for anomaly detection is presented and discussed for the use in automotive environments. The last part focuses on the limitations and adaptions of convolutional neural networks to cope with the detection of small object regions.

Within Chapter 1, the main motivation, challenges and a summary of contributions of this thesis are described. In Chapter 2, fundamentals and related approaches to cope with the detection of small objects in the automotive environment are first explained and categorized. Each explanation of an already existing approach is followed by a short discussion on its feasibility for distant object detection. Following the fundamentals in Chapter 3, a novel method for small object region detection based on a saliency inspired voting scheme is introduced and compared to a state-of-the-art region proposal method. In Chapter 4, a region proposal method using machine learned features within a convolutional network is extended by the incorporation of a global prior to increase the recall for small objects. Chapter 5 introduces the novel relevance-based net-surgery on region proposal networks. Net-surgery allows to determine the most relevant features within a region proposal network and to reduce the network complexity. All contributions and findings are then summarized and concluded in the last Chapter 6.

Chapter 2

Fundamentals and Related Approaches

In this Chapter, the fundamentals and already existing related approaches are presented, see Figure 2.1. At first, the concept of attention maps is described, followed by hand engineered object localizers and a feasibility test of super resolution.

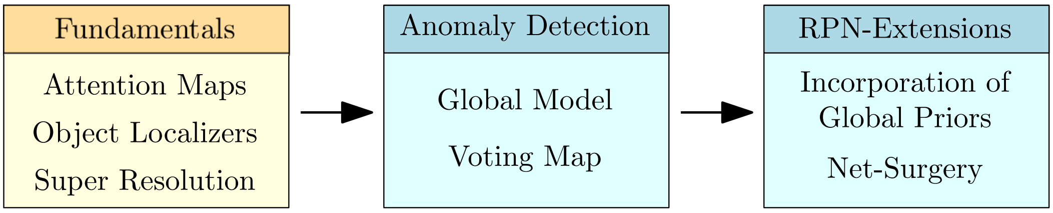


Figure 2.1

To localize object regions within images, several methods have been proposed. In this Chapter, the background to region proposal methods is elaborated and the possible suitability for the detection of small object regions is discussed. Typically, object localization is based on weak or generic i.e. not object specific features. The subsequent classification step uses distinct object specific features. In Figure 2.2, a typical object localization method with a subsequent classification of the object region is sketched. Different hypothesis generation methods can be characterized by the different approaches of feature extraction. In this Chapter, at first, the concept of visual attention maps, often referred to saliency maps, is described. An attention map highlights anomalies inside an image based on either local or global distinctiveness, which can be a base for further object region extraction. Hand engineered hypothesis generators or object

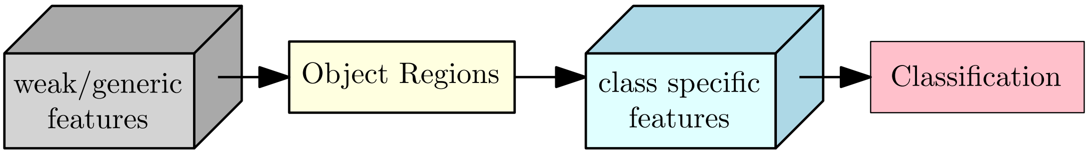


Figure 2.2: Typical flow for object localization based on weak/ generic features with a subsequent classification of the object regions based on class specific features.

region localizers make us of designed generic features such as super pixel straddling or edge density evaluation. The main concepts of such methods are described in this Chapter. Contrary to hand engineered methods, convolutional neural networks make extensive use of machine learned features. With region proposal networks, such features are used to determine object regions. In this Chapter, the concept of region proposal network is presented. The Chapter concludes with a feasibility study on Super Resolution for automotive scenes, followed by a conclusion of current object region generation methods.

2.1 Visual Attention or Saliency Maps

The main idea of attention or saliency is defined by the distinctiveness of a region compared to the remaining regions within an image. Mathematically, salient information is defined as a subset of all available information for further processing. In the case of computer vision, saliency maps highlight e.g. object regions for further classification purposes.

The effort to find the salient regions within the image need to be kept low so high cost computations such as expensive feature extraction is only conducted on the interesting regions.

Based on defined light features, an attention map is created which highlights areas of interest within the image. The difference of salient methods lies in the definition of the different light features. The maps itself have ideally the same size as the input image in which each pixels brightness corresponds to the likelihood of being salient, i.e. of belonging to an object. Hence, a binarization of the attention map with a following segmentation can be used to determine object regions. Object regions are therefore rather pixel-based output than instances like bounding boxes or windows. Such behavior is one of the differences to object localizers.

However, visual attention maps and object localization is closely related as both wish to distinguish between background and foreground. In many cases, a visual attention map is first used followed by object specific cues to determine the extent of the object region. The first work on "salient" regions within an image was inspired by the visual search process of humans [62–64]. In Itti et. al [64], the idea to use similar techniques for object detection as humans use, is first introduced. Based on neural findings of the human visual system, Itti et. al proposed to filter the image in colors, intensities and orientations. To cope different scales, the filters are applied based on a Gaussian pyramid and combined through linear combination in a bottom-up fashion to form one saliency map. In the last years many saliency maps methods are proposed which are based on different underlying bottom-up concepts which are presented shortly in the surveys of Borji *et al.* [9,10].

2.1.1 Application on Small Object Regions

Due to the current existing data sets like Pascal VOC 2007 [31], many saliency map methods are evaluated and designed for a data set with well centered objects. Most objects within such data sets are large with respect to the total image size. Additionally, the objects are well distinguishable such as a red ball on green grass or one fish in sea water in the Pascal VOC 2007 data set [31]. In Figure 2.3, several different visual attention maps are displayed for the same input image, which contains several distant vehicles. The Figure shows, that only few maps exhibit consistent features for the detection of small object regions. Two examples of good suitability are shown in Figure 2.3d and 2.3g, which are frequency based saliency maps.

Frequency Based concepts As the basic idea of saliency is to determine salient regions with low computational costs, several methods based on the usage of the frequency spaces have been proposed [2,8,44,45,60]. To generate the saliency of each pixel within an image the frequency space is analyzed and discontinuities are used as a characteristic for saliency. Small discontinuities can therefore be recognized fast and such frequency concepts show a higher capability to detect small object regions. Among others, the frequency based approaches to generate saliency maps are the fastest to compute, and hardware acceleration is possible. Hence, frequency based methods are the most interesting methods for automotive

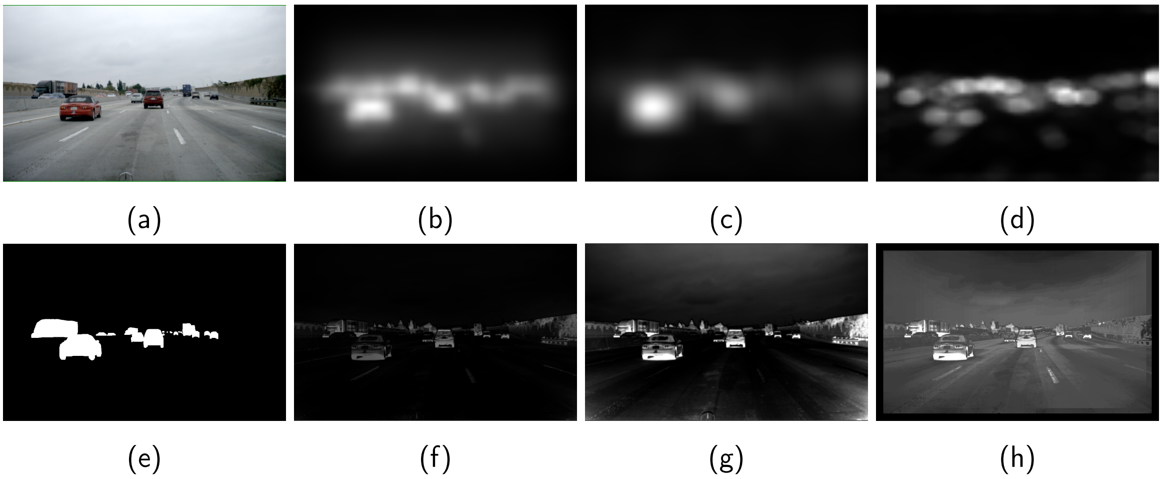


Figure 2.3: Different Saliency Maps. 2.3a Original image, 2.3b GBSV-map [47], 2.3c Signature-map [56], 2.3d Spectral Residual map [57], 2.3f Segmentation map [1], 2.3g Frequency-tuned map [2], 2.3h ESA-map [104] and 2.3e the ground truth.

applicability, due to its close to real time execution.

Hou and Zhang [57] used statistical singularities in the frequency spectrum to determine saliency regions within the image space. For an input image $I(x)$, amplitude $A(f)$ and phase $P(f)$, the log spectrum $L(f)$ of the down-sampled image is computed. The spectral Residual $R(f)$ is then determined by multiplying $L(f)$ with $h_n(f)$ which is an $n \times n$ local average filter and subtracting its results from the original $L(f)$. The saliency map is then transformed back to spatial space through inverse Fourier transform followed by a smoothing in the spatial space.

In contrast to the direct frequency space, Achanta *et al.* [2] uses the difference between the gaussian-blurred image I_{wc} and the arithmetic mean image feature vector I_μ. The saliency map S is computed only on the low-level features of color and luminosity as:

$$S(x, y) = \|I_\mu - I_{wc}\|.$$

Discussion As saliency maps are the intuitive first method to determine object regions without high computational costs within an image, several applications of saliency have been proposed in the automotive environment [69, 93, 117]. The challenge for saliency maps in real applications is mostly the applicability on real images as most saliency methods are tested on considerably easy data sets like e.g a red ball on green grass or a plane in the sky. In real images, the contrast of fore- and background pixels is not as distinct. For the detection of object regions, such saliency maps are not designed to detect small scale objects. Frequency based concepts however, may be possible candidates for the detection of distant and small objects.

2.2 Hand-engineered Object localization methods

In this Section, the most common object proposal methods based on hand-engineered feature extractors are presented. Object proposal methods can be divided in two different categories: (1) based on region merging or (2) based on scoring of windows [54]. For both categories, the most popular concepts are presented and discussed.

2.2.1 Merging Methods

The basic idea behind window merging methods is to segment the image first in small units with high similarities within each unit. Afterwards, the units are merged depending on a chosen metric until the whole image is composed of only one unit. The different merged units correspond then to the proposed object regions. To reach a sufficient amount of windows, a low-level segmentation technique such as the in [37] proposed superpixels can be used at the base [88, 105, 127, 128]. The idea behind Selective Search [127] is to use segmentation in form of superpixels, which are merged in a greedy iterative algorithm with the next most similar super pixel. The similarity measure is a combination of color, texture similarity, and penalty terms that encourage smaller regions to merge first without leaving gaps within the merged regions. After each iteration of merging, bounding boxes around each segment are added to the list of possible regions, until the whole image forms one bounding box, see an example in Figure 2.4.

2.2.2 Scoring Methods

Methods using the scoring concept generate a score for a window. This score implies how likely the window contains an object based on different object cues [3, 17, 38, 66, 103, 129, 141, 142]. One of the first methods using

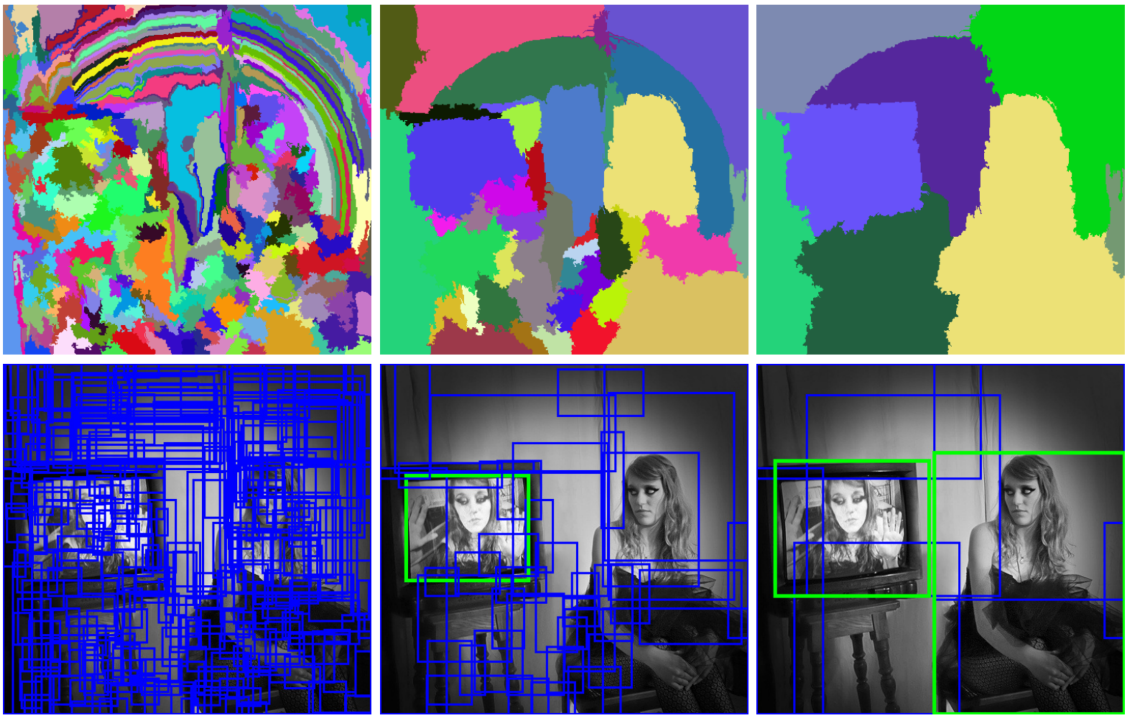


Figure 2.4: The Selective Search algorithm. The top row shows the superpixels merged in different iterations from left to right. The bottom row shows the resulting proposed object windows. Image taken from [127].

objectness cues is proposed by Alexe *et al.* [3]. The idea behind Alexe's Objectness measure is to determine object windows through a combination of different cues such as multi-scale saliency, color contrast, edge density, super pixels straddling as well as location and size. For each possible window, the cues are computed. For the multi-scale saliency cue, the map from Hou et. al [57] is used and to generate the super pixels the method proposed in [37] is used. In Chapter 3, the method of Alexe *et al.* is evaluated for the detection of small object region detection. Another scoring method is the lately proposed region proposal networks based on machine learned filters on a convolutional basis. Such techniques are described in Section 2.3.

2.2.3 Application on Small Object Regions

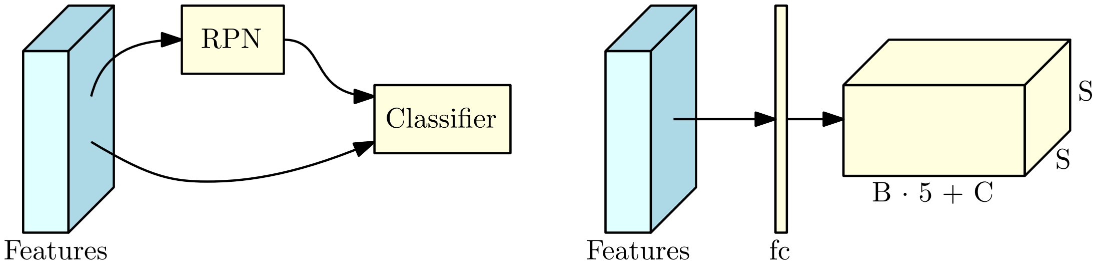
The described scoring and merging methods are based on hand engineered generic features such as color contrast, edge density, super pixels or multi-scales to determine object regions. For the special case of small object regions, multi-scale approaches lead to fast vanishing characteristics of distinct small regions. Cues such as super pixel straddling require a high capability of the used super pixel method to enclose small object region which are globally distinct. As shown in Section 1.3, edges are weak features due to few pixels per object region. Hence, the pixel density cue is not showing strong distinctiveness.

If the creation of super pixels is based only on local features, many super pixels are generated, because local features can not map global distinctiveness. Hence, many super pixels containing background only are created. As the main idea of super pixels is to merge locally related pixels together, this cue is weak for the detection of small object regions. Following this idea, merging algorithms show less capabilities to cope with the detection of small object regions and are therefore not a desired method. Using, however, a scoring method leaves the developer the chance to include global and local weak features which are designed specifically for the detection of small object regions. The concept of scoring methods is therefore promising for the detection of object hypothesis.

2.3 Region Proposal Network

To detect object regions in an image, a naïve method is to extract class specific image features and sort the extracted features to the right classes with. Since 2010, hand engineered features were extracted and sorted with a trained support vector machine (SVM).

Recently, the feature extraction process can be optimized through machine learning. Therefore, parameters in an application specific network are optimized during a training phase which uses labeled ground truth data. In the following Chapter, the theory behind neural networks and their usage and benefit for this thesis is explained. Two main developments made neural networks one of the most powerful methods in computer vision (1) the development of the backpropagation calculation and (2) hardware development in graphic cards to perform a large number of simple calculations in parallel. To perform object localization, two different architectures have been developed and adjusted to different applications. In Figure 2.5, those two architectures are sketched and explained in detail in the following Sections.



(a) (b)

Figure 2.5: In 2.5a a *two-stage object detector* with an external Region Proposal Network (RPN) is shown and 2.5b shows a *single-shot detector* with a fully-connected (fc) joint object localization and classification method. B denotes bounding box proposals per grid-cell S and C object classes.

2.3.1 Background to Convolutional Neural Networks

The concept of a convolutional neural network (CNN) is to transform input image pixels through several feature extractions to the main features of the shown image scene. In other words, it transforms an e.g. image of a car through computations to a final vector, where the entry for the class 'car' obtains the highest score. A neural network consists of neurons arranged in layers, where each layer contains feature maps which are created through the neurons within the layer. In Figure 2.6 a neural network architecture is shown.

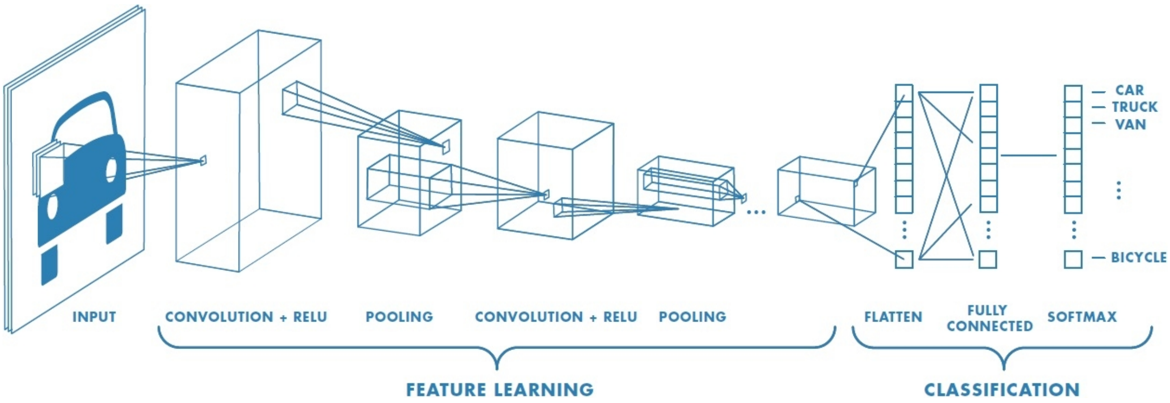


Figure 2.6: A convolutional neural network architecture is built of several layers with filters. During feed-forward / execution of the network, an image is fed from the left to the right through the layers. Each layer transforms the input of the previous layer through convolution kernels until the information is transformed to a 1 dimensional vector. Each entry of the vector belongs to a certain object class. Image taken from [1].

---

[1] https://de.mathworks.com/solutions/deep-learning/convolutional-neural-network.html

A convolution neural network is using convolving volumes to transform the input image data to a 1-dimensional classification vector. Each layer within the network consists of tensors of size $F \times K \times D$, where F is the spatial extent of the kernel, K the number of kernels and D the depth of the input data to the layer e.g. for a RGB-image[2] $D = 3$. The output of each layer are K feature maps. Several of such layers can be stacked behind each other and the hyper parameters F and K are chosen beforehand to form a specific network architecture. To fit all image sizes, zero-padding P can be added. To reduce the image information, a stride S or pooling of the feature maps can be performed. During max-pooling, a feature map is resized by a determined stride S using a kernel size F, see Figure 2.7. For each kernel position, only the maximal value is taken and due to the chosen stride, the size of the feature map decreases. After a convolution layer, a nonlinear activation layer is used to introduce non-linearity into the network and to allow to represent complex transformations needed for the object detection. Among other functions, the rectified linear unit $f(x) = \max(0, x)$ is, due to its simple calculation, the most used nonlinear activation function [95].

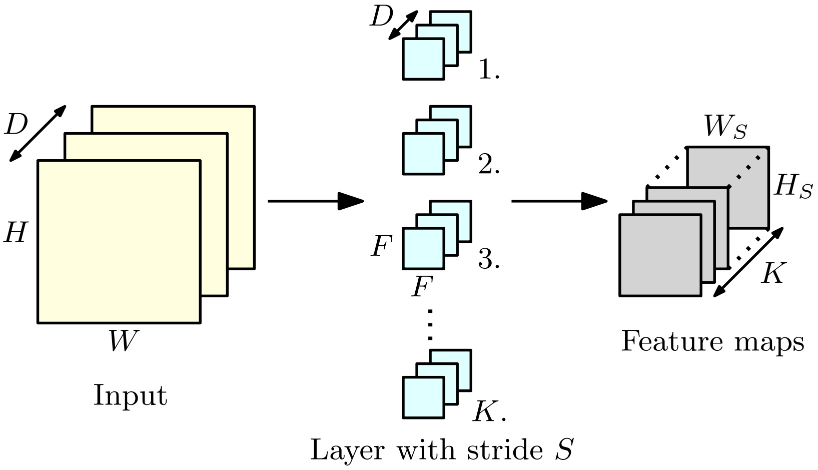


Figure 2.7: W_S and H_S are the width and height of the resulting K feature maps with a reduction of the sizes compared to the input maps of height H and width W. The size reduction is due to the used stride S during the convolutions of the K filters of size $F \times F$.

During the training-phase, the kernels are learned through backpropagation [108] with a stochastic gradient descent method [68]. The concept of the gradient descent method is that the parameters are updated following the gradient of the mean squared error at the last classification

---

[2] RGB stands for Red, Green, and Blue image i.a. the image has three color channels.

layer. As the used networks can contain many layers, a simple gradient method would vanish during updating. Hence, the deep belief backpropagation method of Hinton et. al [52] has been used since, to learn layers effectively [75].

When deciding the architecture and hyperparameters, the application and the available data set need to be considered. Solving a complex problem such as classifying an object within an image demands a higher complexity network than e.g. finding correlations between few measuring values. However, overfitting needs to be avoided, where the parameters within the network are not forming a generic solution. Overfitting can occur, when the data set to train the network is too small to provide enough different scenes or when the network is too large - and hence too complex - to represent the underlying problem. A network architecture, as shown in Figure 2.6, can classify a whole image or an image patch. Such a network is not suitable to localize and classify object regions at the same time within an image.

2.3.2 Network Architecture

To localize object regions within an image, several network architectures have been proposed from the year 2015 until now [21, 42, 43, 49, 71, 82, 83, 101, 106, 107, 114] and still ongoing.

The first neural network based object detector, which includes a localization of objects, is Overfeat [114], a sliding window CNN. No intelligent region proposal method is used, which imposes high computational costs as for every position and size of the sliding window approach, a CNN has to be computed. Using a region proposal method in combination with a CNN was first introduced by Girshick et. al [43]. They proposed to use any region proposal methods such as Selective Search [127] and crop the image to the proposed boxes. Each crop is then fed through a CNN for classification. Later on, feature extractors are used which combine localization and classification based on a shared feature extractor.

One can distinguish the proposed methods in *two-stage detectors* with a specific region proposal and classification head and so called *single-shot detectors*, where the region proposals and classification are evaluated jointly. Both detector strategies use the idea of anchors and shared features.

A *two-stage detector* is based on a network which is used as a feature extractor such as e.g. the ZF-Net [136] or ResNet [50]. The main idea is proposed by Ren et. al [107] in 2016. After the feature extraction has taken place, a region proposal network (RPN) uses the outcome of the feature extractor as input to localize possible objects. The region proposal network is based on an box-wise scoring method with an additional box regression.

To perform the scoring method for each position of a 1×1 kernel, several so called anchors with different sizes and ratios, are used as reference to perform binary fore- and background classification. Additionally, each anchor is regressed to fit the object region better, see Figure 2.8. The anchor sizes and ratio need to be adapted to the application.

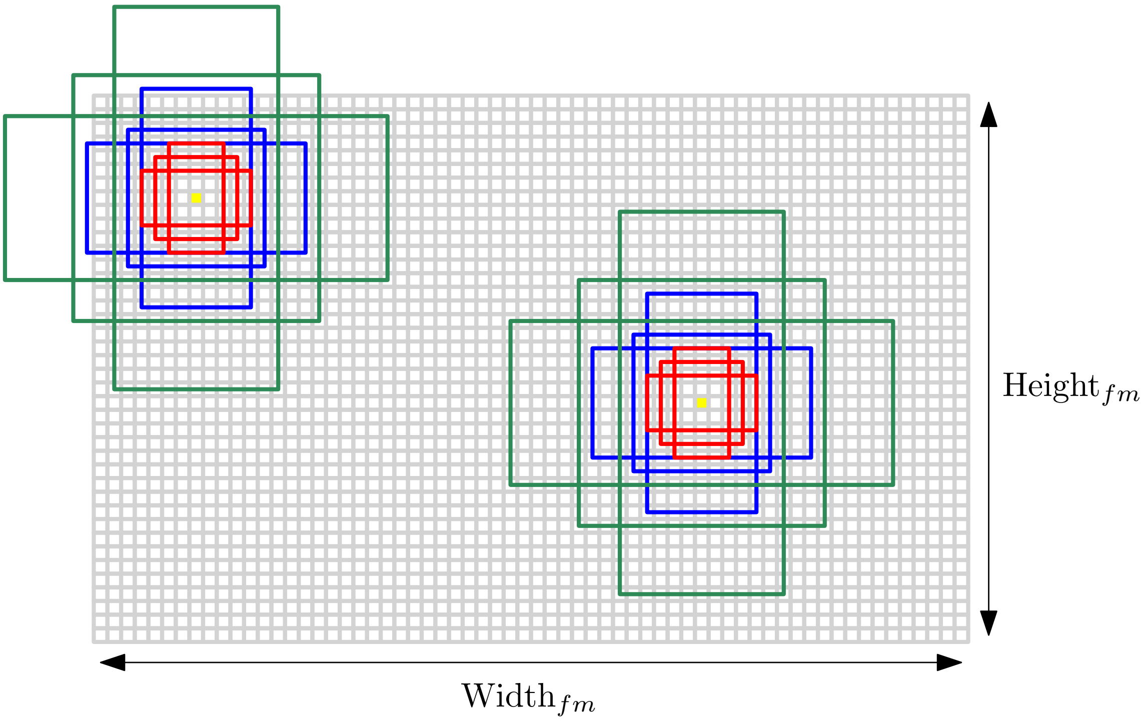


Figure 2.8: This sketch shows how the anchors are set up in a region proposal network at the example of two positions on the feature map within the RPN. A set of anchors with three ratios and scales slides over the feature map. At each pixel of the feature map, here shown in yellow, a classification score is computed and bounding box regression takes place with respect to each anchor.

During the training phase of the RPN the regression branch is only trained on anchors that have a high overlapping area with an object region. During training, an equal amount of anchors N_{cls} containing object regions and background are used. This ensures a balanced data set. A softmax-layer is applied on the binary classification scores to ensure that the scores of back-

and foreground for each anchor is summing up to 1. Hence, the binary classification score can be treated as a probability to belong to fore- or background. The loss to train an RPN is generated through the sum of the loss of binary classification $L_{cls}\left(p_i, p_i^*\right)$ and bounding box regression $L_{reg}\left(t_i, t_i^*\right)$:

$$L\left(p_i, t_i\right) = \frac{1}{N_{cls}} \sum_i L_{cls}\left(p_i, p_i^*\right) + \lambda \frac{1}{N_{reg}} \sum_i L_{reg}\left(t_i, t_i^*\right). \tag{2.1}$$

Here, i is the index of an anchor and p_i the predicted probability of anchor i. p_i^* is the binary ground truth label. The vector t_i is of length four to represent the deviation of the center of anchor i to the ground truth center box in x- and y-direction and the difference of width and height of the anchor i to the ground truth bounding box. The value N_{reg} is the amount of possible anchor locations which depends on the chosen number of different ratios and scales. The loss is then used during the backpropagation method to optimize the parameters within the network.

During feed forward, the anchors with the N highest foreground scores and the corresponding deviations from the regression branch are taken and fed to a classification branch. This classification branch can be in general any classification method. However, in the context of deep learning, the sharing of already extracted features showed fast and good results. Hence, a spatial region of interest pooling layer transforms each possible foreground box to a fixed sized tensor of the feature maps of the feature extractor network. This acts then as a fixed sized input to the fully connected classification branch, where the box is classified or rejected as background. The fully connected classification branch is computationally more expensive than the convolutional layers since the weights of the network are densely connected and it is highly recommended to ensure a low number of possible boxes for evaluation [21, 42, 43, 107].

The *single-shot detectors* such as YOLO [106] share the same idea of feature extraction and anchor boxes. However, to train the networks end-to-end, i.e. without any additional region proposal network in between, the binary classification is eliminated and instead, a direct classification to all classes is chosen. Since the classification uses a high amount of computational power, the single-shot detectors decrease the possible amount of anchors drastically by introducing a predefined grid on the feature extractor output. For each grid cell, several anchors are evaluated through one fully connected layer. The fully connected layer is then reshaped back to the size of the grid with a depth of the different anchor sizes times two for fore- and background confidence score and the number of different object classes. Each pixel on the feature map is computed based on several pixels on the input image. The amount of pixels used to generate one pixel value on the feature map is defined as the field of view of the network. Hence, one grid cell on the feature map corresponds to a certain field of view on the input image. Such a single shot detector allows a very fast classification and localization within one network since the possible locations are reduced drastically by design.

SSD [83] is comparable to YOLO, but, instead of using only the last feature extractor layer, it introduces additional layers, including down-sampled ones to allow region proposals at different scales. For each feature map scale, a dedicated set of convolutional layers is assigned to compute boxes and classification at different scales. With the HyperNet [71] in 2016, the idea of using not only the last output feature maps of the feature extractor network but also feature maps from lower layers, was introduced. The different feature maps are scaled to the same size and the feature maps are concatenated to form hyper feature maps.

2.3.3 Application on Small Object Detection

Since 2017, the performance gap of object detection with neural networks for larger objects compared to small object became visible. Zhang *et al.* [139] found that 50% of false-positive detections are objects of small scale. Wang *et al.* [134] studied the networks size for low resolution small object detection and found that, unlike for larger object detection, more layers stacked behind each other (deeper network) do not show the same improvement in detection rates compared to larger objects. Also the choice of larger kernels shows decreasing performance for small object detection at low resolution. Several work proposed to increase the input image for small objects even at low resolution, which showed increased detection performances [28,71,81,83,92]. However, this technique increases the computational costs significantly. In [4, 79], an intermediate super resolution step is proposed before the object proposal and classification step. The super resolution is induced by a generative adversarial network within the network architecture. The use of different scaled feature maps improves the detection performance for small objects within a large image. For small objects at low resolution a further down scaling is not beneficial since the information rather vanishes.

Finally, the application of CNN's on small objects depends strongly on the quality of the evolved features inside the network. Parameters like field of view, object specific, and global vs. local features determine the object region and detection performance of the network.

2.4 Feasibility of Super Resolution

To use the aforementioned localizers to detect small objects in automotive images one could increase the resolution of the images to increase the feature quality of the objects. As already stated in Section 1.3, using a better hardware to increase the image resolution is not feasible in the automotive sector at the moment. Hence, the resolution of an image could be enhanced with an algorithm. In literature super resolution is proposed to increase the resolution with the help of several low resolution images of the same or similar scene [126]. A prerequisite to apply super resolution is to have several low resolution images of the same scene that are either locally or time wise slightly different to each other. In this Section, the idea of super resolution is presented and a short feasibility study with a conclusion is given. The performance of the state of the art algorithms is only tested with subjective perception to get an insight into the possible use for small object detection at low resolution.

2.4.1 Concept of Super Resolution

The idea of fusing several low resolution (LR) images to one high resolution (HR) image was proposed at first by Huang and Tsai in 1984 [126]. In this first approach, the super resolution was performed in the frequency domain and could only register translational global motions. Any other motion degraded the performance. In the last decades, this idea was developed further and several different approaches have been tested. Super resolution on real images consists of three different steps:

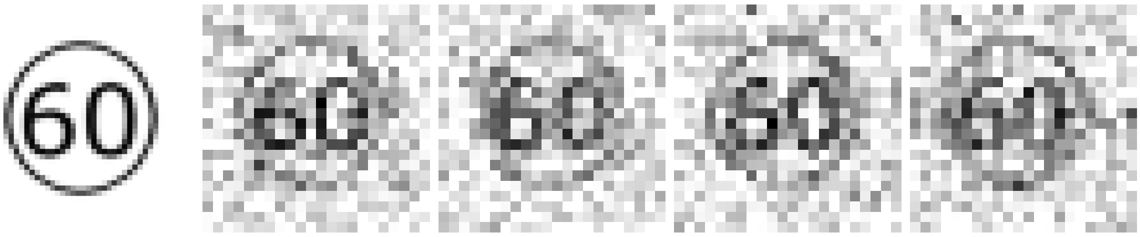
1. Finding of the Region of Interest (ROI), which contains solely objects with same dynamic parameters

2. Registration of all used images with respect to each other
3. Reconstruction to one HR Image.

The registration requires an accuracy of sub-pixels [113]. Research on all the named steps has been done in e.g. [41, 96, 130]. In the following Sections, a selection of proposed methods is explained and subject to evaluation for the use in the automotive environment using artificial data.

2.4.2 Experiment

The experiment is based on artificial data to determine the capability of state of the art super resolution methods. An artificial sign with the number 60 and a black circle around was created on 44 × 44 pixels. This original image is then down sampled to 22 × 22 pixels and Gaussian noise is added. The images are shifted in x-direction by -0.5, -0.25 and 0.5 low resolution pixels to create 4 low resolution images for testing of the registration and reconstruction methods (see Figure 2.9).



(a) High Resolution Original (b) Image 1, shift: 0 LR-Pixels (c) Image 2, shift: -0.5 LR-Pixels (d) Image 3, shift: -0.25 LR-Pixels (e) Image 4, shift: 0.5 LR-Pixels

Figure 2.9: These artificial image are used for Super Resolution tests. (a) shows the high resolution 60-sign without any gaussian noise of 44 × 44 px. The images to from (b)-(e) are down sampled test images of size 22 × 22 px with indicated subpixel shifts to each other in horizontal/ x-direction. b) no shift c) shift with respect to (b) in x-direction of -0.5 LR pixels c) of -0.25 LR pixels d) of 0.5 LR pixels.

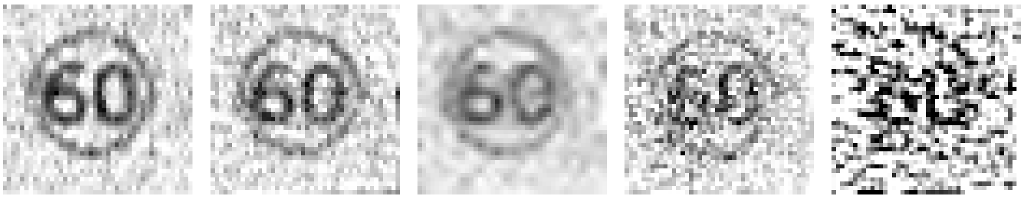
2.4.3 Registration Errors of Existing Algorithms

Given the region of interest, several registration algorithms are tested for suitability for small object regions at low resolution. Several registration methods make us of the frequency domain to detect shifts in the spatial domain [86, 89, 131]. In such methods, a shift in the spatial domain corresponds to a linear shift in the Fourier space. Additionally, rotations are visible in the amplitudes of the Fourier Transform (FT). Hence, the shifts in the Fourier space are determined through the amplitudes and phases (such as in Marcel *et al.* [89]). For increased robustness high-frequency components can be discarded (such as in Vandewalle *et al.* [131]). Lucchese *et al.* [86] uses a geometrical relation between a mirrored image and its original in Fourier space to determine rotational motion angles. Keren *et al.* [67] creates a Gaussian pyramid of down sampled images. For each scale, the registration is determined and refined on the next scale using a Taylor series. The registration of the four test-images is investigated with the previously presented registration algorithms from Vandewalle [131]), Marcel [89], Keren [67] and Lucchese [86]. The results are summarized in Table 2.1. The table shows that the method of Vandewalle exhibits closest results to the ground truth shifts as indicated in Figure 2.9. Nevertheless, even this method shows an absolute error to the real shift of up to 0.5 LR pixels and an relative error of up to 70 %.

	Image 1			Image 2			Image 3			Image 4		
Shifts	x	y	ϕ	x	y	ϕ	x	y	ϕ	x	y	ϕ
Ground Truth	0	0	0	-0.5	0	0	-0.25	0	0	0.5	0	0
Vandewalle	0	0	-	-0.061	-0.005	0	-0.296	-0.059	-	0.395	-0.020	-
Marcel	0	0	-	0	0	-	0	0	-	1	0	-
Lucchese	0	0	0	-2	1	-62.29	0	0	-7.2	1	0	-12.17
Keren	0	0	0	-0.061	-0.003	0.061	-0.34	-0.011	-0.25	-0.026	-0.015	-0.089

Table 2.1: Registration algorithms and the results compared to real shifts. In each field, the 4 values correspond to the low resolution image 1,2,3 and 4. Since the first image is seen as the reference, those values are all set to zero.

Given the registration results, it is of interest to understand the impact of errors within the registration for the final reconstruction of the four low resolution images to one high resolution image. For comparability reasons, the fairly simple bi-cubic interpolation reconstruction method is used. With this reconstruction, the low resolution images are placed on a high resolution grid according to their registration and the final high resolution gird is filled with the help of bi-cubic interpolation. In Figure 2.10, the reconstructed high resolution images by using bi-cubic interpolation based on the registrations values of the different registration algorithms are shown. It shows that errors in the registration degrade the quality of the image up to almost unrecognizable.



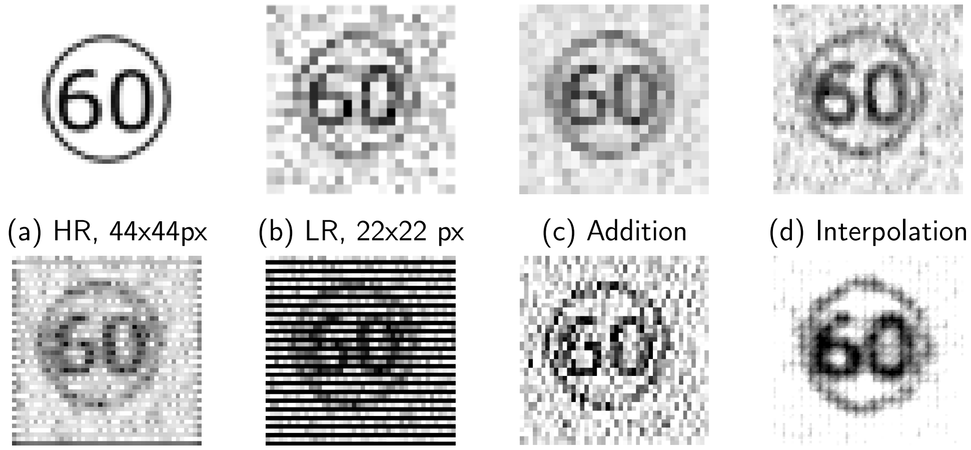
(a) (b) (c) (d) (e)

Figure 2.10: For better comparison of the impact of registration errors, the bi-cubic reconstruction interpolation is used for all experiments (b)-(e). In a) the registration was given, in b) the registration algorithm from Vandewalle in c) from Marcel in d) from Lucchese and in e) from Keren as shown in Table 2.1 is used to obtain the shifts.

2.4.4 Reconstruction Errors of Existing Algorithms

In the previous Section, different registration methods have been investigated. In this Section, different reconstruction methods are evaluated based on the visual appearance of the reconstructed high resolution image. The same test images as previously presented are used for the different reconstruction methods. For reconstruction, a naïve method is to use bicubic interpolation as explained previously or simple addition of the low resolution by placing the low resolution images according to the registration on a high resolution grid and adding the values. More sophisticated reconstruction methods have been proposed such as in [61, 65, 102, 143]. The reconstruction methods can be divided into two approaches, either projection onto convex sets or in iterated back projection. The Papoulis-Gerchberg method [65] as well as the POCS-method in [102], belonging to the former, make a first estimation of the pixels on the high resolution grid and uses high- or low-pass filters in the frequency domain until convergence or to approximate the point-spread-function of the camera. Belonging to the latter, Irani *et al.* [61] (Iterated Back Projection) and Zomet *et al.* [143] start with a rough estimation of the HR image, and iteratively add to it a "gradient" image. This gradient image is composed of the sum of errors between each LR image and the estimated HR image after registration is taken into account. Zomet *et al.* uses the median of all errors to create the gradient, which makes it more robust. This method is therefore called 'Robust Super Resolution'. Another reconstruction method is the Gaussian-carpet method. It is developed to suppress noise even more and to enhance contrasts. A detailed description of this method can be found in the Appendix A.

In Figure 2.11, tests on the reconstruction algorithms are shown. The



(a) HR, 44x44px (b) LR, 22x22 px (c) Addition (d) Interpolation

(e) POCS (f) P-G (g) Robust SR (h) Gauss-carpet

Figure 2.11: The reconstructed high resolution images of different techniques by given ground truth registration. The original and reconstructed images have a resolution of 44×44 pixels, while the four used LR image are 22×22 each (see Figure 2.9).

ground truth for registration is known to allow a direct comparison of the reconstruction methods. The same experimental set up as in Section 2.4.3 is used. With a visual assessment, it becomes visible that no reconstruction method has the capability of recovering the noise and sharpening the contrast at the same time. While the Interpolation, POCS and Robust SR could refine the edges within the 60, the noise is still present in all reconstructed images. The Gauss-carpet method could eliminate a majority of the noise and increase the contrast. However, it becomes visible that it tends to smear edges and is therefore, only for certain forms, a good reconstruction method. Nevertheless, it has to be stated that the Gauss-carpet approach takes approximately 50-60 sec while the other approaches take only up to 1 sec on the same hardware.

2.4.5 Discussion

Super Resolution is, at first glance, a possible idea as it could solve the problem of small object detection in a generic way. However, its calculation costs are high and the tests on the registration problem exhibit challenges for small objects at low resolution. Recent methods for super resolution with neural networks have been proposed [4, 27]. However, either those trained networks learn to abstract the high resolution image from a low resolution image through a set of trained data [76] or several low resolution images are fed into the network and one high resolution image is created through the network [59]. In both cases, the idea of using super resolution becomes obsolete as it is not a generic method for all kinds of situation anymore. Hence, one can ask the question, why not to detect the objects of desire directly and omit the expensive intermediate step of super resolution in between. For all this reasons, the method of super resolution is not followed up further as a reasonable approach for the detection of small and distant objects in automotive images. Even though Van Eekeren *et al.* [130] showed good results to enhance the resolution on small moving objects, the capturing camera was standing still and the movement of the object could be estimated as linear. Both environmental conditions are not given in an automotive environment. Super Resolution is well suited to enhance the image quality of an image sequence obtained by still standing cameras and well detectable and traceable objects.

2.5 Conclusion

In this Chapter, the main different approaches to detect small object regions within an image have been presented and discussed on the application of automotive images with distant objects.

Saliency maps show promising results for the distinction of fore- and background pixels, which qualifies for a generic object region detection. As of today, saliency maps have not been applied or evaluated in practice for the detection of small object regions. Frequency-based approaches show promising results, but fail to enclose all object region pixels. In this thesis, in Chapter 3, a saliency inspired voting scheme is proposed to fill this gap.

Hand-engineered region proposal methods suffer from the locality of chosen features to detect object regions. As presented, merging methods lead to weak localizers for small objects, while window scoring methods can show more robust detection performance. In Chapter 3, a window scoring state-of-the-art method is evaluated on a test set for distant objects in an automotive environment.

Region Proposal Networks show promising performances for object region detection. However, they fail to show the same performance on small scale objects as expressed by Zhang *et al.* [139]. Moreover, two-stage detectors are more suitable for the detection of small object regions as they allow to evaluate dense windows in various scales without adding extensive computational costs. Additionally, two-stage detectors exhibit higher localization accuracy than single-shot detectors [58] due to the choice of architecture. For these reasons, single-shot detectors not are considered in this thesis. For the use-case of automotive applications, only small neural networks are feasible as otherwise the computational costs may be too high. State of the Art network architectures are ZF-Net [136] and VGG16-Net [116]. Both architectures are built with stacked convolutional layers with different kernel sizes and depths. The computational costs during one execution are 53 GMACs for the ZF-Net while the VGG16-Net net consumes 209 GMACs for the convolutional layers including pooling layers. Hence, the small ZF-net consumes four times less computational costs and is therefore well suited for automotive applications. Due to a low number of kernels and layers, the representing features within the network are locally restricted. A global field of view can only be reached with more layers and large kernels, which results in higher computational costs. It is, however, possible to incorporate global priors into locally restricted feature extractors, as described in Chapter 4. The features within a neural network are learned end to end through labeled data. A high quality of the developed features in a network gives also a high accuracy in localizing object regions or object classification. To understand the diversity and influence of certain features for the special use case of small objects, apart from visual assessment, a neutral feature evaluation method is desired. With such method, it is possible to determine robust, strong, and important features within a network. In Chapter 5, a novel net-surgery method is proposed, which allows to determine such features. To decrease the size of a network, only those robust features can be maintained without degrading the performance significantly.

Chapter 3

Anomaly Detection / Hypothesis Generation

The present Chapter is dedicated to the generic detection of object regions within an automotive image. Figure 3.1 shows the structure of the thesis.

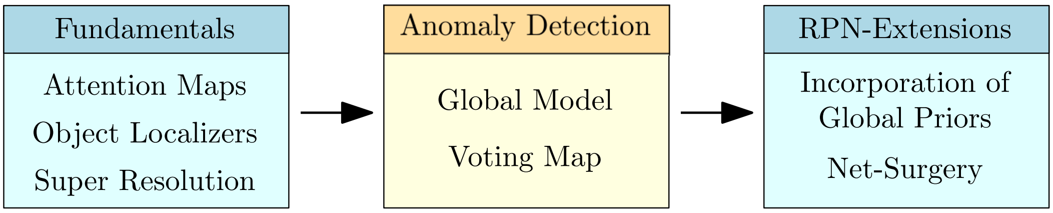


Figure 3.1: Following the discussion of existing approach and their feasibility on the detection of distant objects for the automotive environment, this Chapter presents an approach to detect small object regions through the idea of anomaly detection.

A saliency inspired Voting Map is formed to determine all pixels belonging to objects in the image. The proposed method is class independent and consequently, generic object regions can be detected. At first, related work for generic object detection is discussed shortly, followed by the concept of environment modeling in the image through few homogeneous regions. Those homogeneous regions are used to form the Voting Map, which highlights object regions. The used metric and data set are shown before the method is compared to a state of the art generic object region detector. The Chapter concludes with the key aspects of the proposed method and a summary. In this Chapter, excerpts have already been published by the author in [5].

3.1 Motivation

The motivation of the following Chapter is the development of a method which allows to detect small objects on motorway scenes without the need of training data set to optimize parameters for the different object classes. Especially for small objects on motorways the generation of a feasible and well balanced data set is expensive as often close objects occlude distant objects. Hence, a hand engineered method generating generic object regions within the image of a motorway is highly interesting. Generic object region detectors can then be used not only for the detection of vehicles in the distance but also traffic signs, trucks, motorbikes and construction site lights. Such a construct allows to reduce the image information down to the most interesting or salient regions within the image. Driving on motorways can lead through tunnels, under bridges or large traffic signs. Such changes in illumination can lead to overexposure due to the slow adaption within the camera. Therefore, the detection method needs to be highly adaptive on single frames with such rapid illumination changes. Different weather conditions such as sun, rain, fog as well as day and night times lead to a variety of image properties. While rain as well as fog reduces the contrast of object regions, sunny conditions let edges appear more distinct within the image. Especially cast shadows comprise strong edges within images and hence the object region method is ideally lighting independent. Here to summarize, hypothesis generation faces several challenges such as complex environments, different lighting and weather conditions, very small objects and objects being far away from the camera. Depending on the physical properties of the camera system used, automatic hypothesis generation of small and distant objects may be difficult due to poor contrast images and the very few pixels that those objects occupy in images.

The majority of existing approaches to object detection in literature, however, address hypothesis generation of objects with sufficient resolution only. Here, a variety of approaches make use of symmetry operators that measure the symmetry of patches about a defined axis [73, 90, 132]. Symmetry operators can use intensity only, but often evaluate image edges as well in order to increase their robustness to noise [7]. While hypotheses of near objects can be generated very robustly, low resolution and missing edge information of distant objects may make the generation of hypotheses impractical for driver assistance systems.

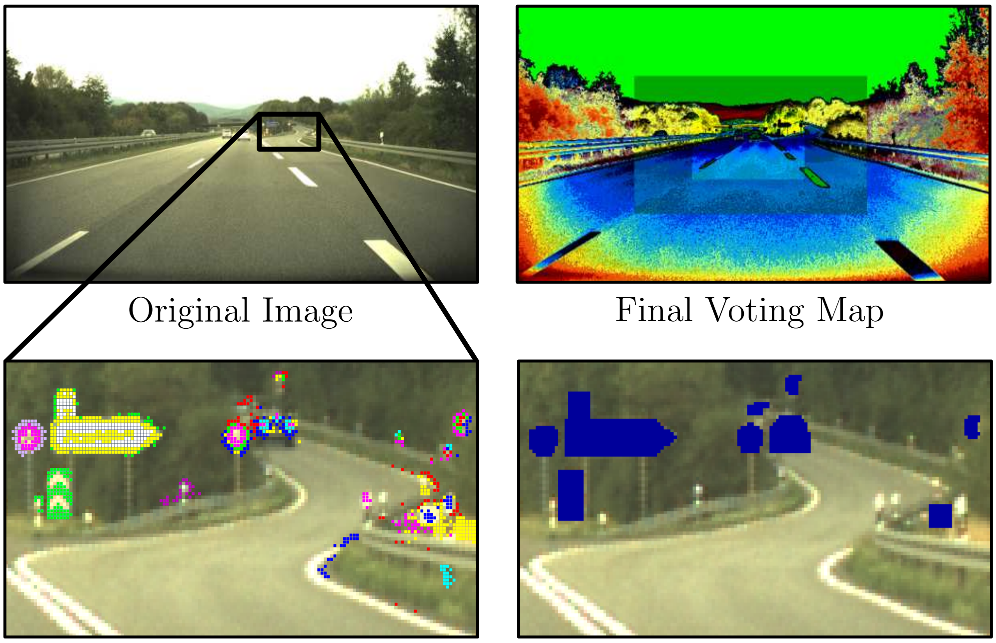
In principle, the effect of image noise on symmetry can be weakened by making additional use of an object's cast shadow such as proposed in [14, 70, 97]. Cast shadow may be a useful feature for close objects such as vehicles, but may not be visible or existent for distant objects at all. In addition, other approaches exploit color information as a powerful cue for hypothesis generation [33, 46, 78, 135]. The drawbacks to color-based hypotheses, however, include *a-priori* knowledge about the appearance of objects being detected and sensitivity to light as well as weather conditions.

In order to address the challenges, there has been an increasing interest in more robust and complex features such as HOG or Haar-features [118] recently. Such features allow for direct classification of objects such as humans [23] or vehicles [18, 119]. Adding additional features to HOG features can further increase the robustness of detection. As an example, Sun *et al.* [123] evaluated the performance of principal component analysis, wavelets, and Gabor filters for rear-view vehicle detection. He *et al.* [51] proposed a framework that combines appearance, structural and shape features. The framework detects parts of objects individually and imposes structural constraints among the parts for detecting the entire object, and can be extended by probabilistic approaches [133].

3.2 Outline

In summary, while such approaches have been proven to be extremely robust for hypothesis generation of close objects, their reliance on edge extraction, HOG, structural and shape features may make them impractical for hypothesis generation of small and distant objects due to low resolution and insufficient edge information. The approach of Zhang *et al.* [137] attempted to address this issue by artificially increasing the resolution of potential regions of interests, but the reliance of *a-priori* knowledge about the regions to zoom in may limit the approach to a well-defined set of scenarios only. As such, a generic approach to hypothesis generation for detecting small and/or distant objects in images for the purpose of driver assistance, that does not rely on *a-priori* knowledge, and can cope with low-resolution areas and missing information, would be of great interest.

The main contribution in this Chapter is the introduction of a simple yet generic approach to hypothesis generation of small and distant objects for



Original Image

Final Voting Map

Zoomed Result

Zoomed Ground Truth

Figure 3.2: Proposed approach to generic hypothesis generation. Top: Original image and final Voting Map. Bottom: Hypotheses of distant candidate objects and ground-truth.

driver assistance based on a novel voting scheme (see Figure 3.2). The voting scheme is inspired by salient region detection that aims to emphasize regions in natural images that appear visually distinct and attractive when compared to the rest of an arbitrary scene [111, 112]. Following this concept, the key idea is to generate hypotheses of small and distant objects since candidate objects usually differ from their environment. Hence, in this work, the environment is modeled as a composition of very few, large areas with homogeneous appearance. These areas are computed by using image statistics extracted only from the image information available, without *a-priori* knowledge about the embedding environment needed. Regions that cannot be assigned to one of these areas are potential candidate locations and form, hence, hypotheses for further processing. In addition, we assume that close objects with high resolution are already detected by other image-based object detectors.

In contrast to other approaches that attempted to detect distant objects by increasing the resolution of images [137] artificially[1], the here presented approach can work on original images based on the concept of saliency-guided Voting Maps for candidate distant region detection. Due to the single image based independent saliency method, this approach is robust to

---

[1] Similar to the methods presented in Section 2.4.

the variation in the environment such as different scale, lighting, weather and traffic scenarios. Experimental results based on a variety of test data and metrics demonstrate promising performance.

3.3 Proposed Approach

Human drivers can easily classify regions as belonging to road users or background due to a number of attributes such as color, intensity and size among others across single and multiple scales [110]. This makes road users appear more salient when compared to the rest of the scene or background. Here, we want to make explicit use of this property and take advantage of inherent characteristics in natural images for generating hypotheses of small and distant objects for driver assistance in an efficient manner. As shown in Figure 3.3, the overall architecture of the proposed approach can be broken down into four main stages: (i) patch selection in image zones, (ii) generation of Voting Maps, (iii) combination of Voting Maps and (iv) retrieval of hypotheses. A detailed description of each stage is provided in the following Sections.

3.3.1 Patch Selection in Image Zones

We now wish to define a model to represent the different characteristics of the image's background in a global manner. Here, we model the environment as being composed of a few regions with homogeneous appearance, where each region is described by a set of patches p representing its global characteristics. As such, the extraction of such patches is the first step towards modeling the overall characteristics of a natural image. Moreover, we assume that the appearance of road users depends on their distance to the camera, and hence, to their position in images. More specifically, for general road scenes, the assumption is made that the appearance of objects changes from the outside of an image to the image center. Thus, the image is subdivided into different zones $z^{(m)}$ with $m \in [1...n]$, $n \in \mathrm{N}$ first.

In this work, we consider the environment as being composed of regions with homogeneous appearance as opposed to regions containing small or distant objects, with using the inner statistics of each patch $p_j^{(m)}$ to classify

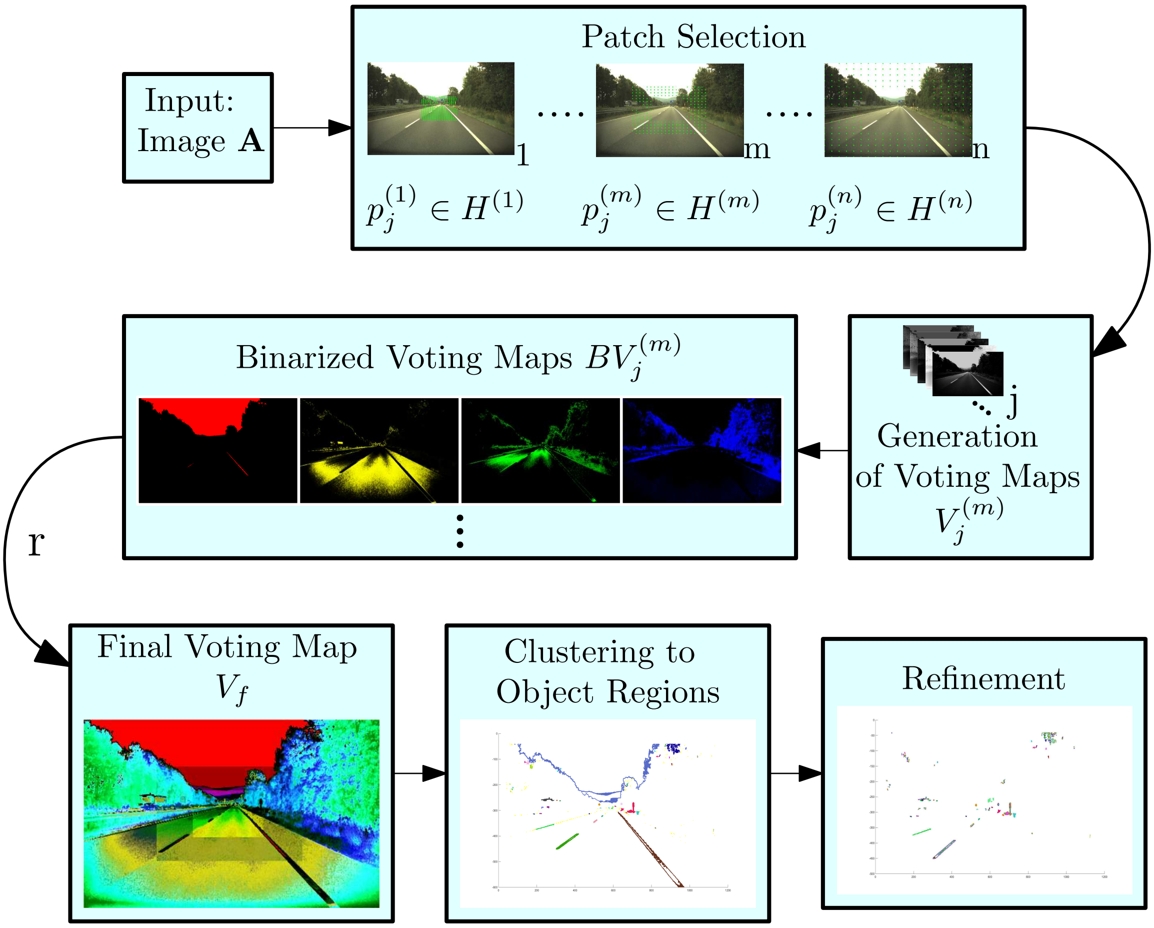


Figure 3.3: Sketch of the overall architecture of the proposed approach. The patch selection in image zones is followed by a generation of Voting Maps and a binarization. After accumulation of binarized Voting Maps to a complete Voting Map the final object hypotheses are retrieved and refined.

a patch as belonging to background or to an object region. We model the inner statistics of each patch $p_j^{(m)}$ by a Gaussian distribution $\mathbf{P}(p_j^{(m)})$ computed over the intensities $\mathbf{I}$ of all pixels x within $p_j^{(m)}$. Given the definition of homogeneity, it can be assumed that the Gaussian distribution within an homogeneous patch p_h can be modeled using low variances $\sigma(p_j^{(m)})$, while in-homogeneous patches comprise a wide distribution, resulting in large values of $\sigma(p_j^{(m)})$. Hence, the standard deviation $\sigma_j^{(m)}$ is chosen as the parameter to dismiss in-homogeneous patches and to produce a set of homogeneous patches $\mathrm{H}^{(m)}$ for further processing as follows:

$$\mathrm{H}^{(m)} = \{p_j^{(m)} \mid p_j^{(m)} \left(\sigma(p_j^{(m)}) \leq \sigma_T^{(m)}\right)\}. \tag{3.1}$$

For environments with homogeneous background content, the assumption can be made that the distribution $\mathbf{P}(\sigma(p_j^{(m)}))$ computed across all standard deviations of $p_j^{(m)}$ in $z^{(m)}$ is Gaussian. The value $\sigma_T^{(m)}$ is then the apex of the Gaussian distribution, classifying the patches as homogeneous or inhomogeneous.

To ensure that only background patches are included in $\mathrm{H}^{(m)}$, the patch size has to be large enough to ensure that no area on a large object is considered as background due to its homogeneity. Otherwise a patch may be placed inside an object. Hence, the size of the patches $w_p \times h_p$ is chosen in such a way that objects smaller than the patch are still detected within each zone. Internal experiments within this work showed that the smallest patch contains ideally at least 100 pixels to ensure statistical examination.

3.3.2 Generation of Voting Maps

Each homogeneous patch $p_j^{(m)} \in \mathrm{H}^{(m)}$ is belonging to a non-object region according to our model. Hence, it is possible to generate a set of Voting Maps $\mathrm{V}^{(m)}$ for each zone m,

$$\mathrm{V}^{(m)} = \{V_j^{(m)} \mid j \in [1, \cdots, N]\} \tag{3.2}$$

for each $p_j^{(m)}$. The Voting Maps $V_j^{(m)}$ have the size of an input image $\mathbf{A}$, and the value $V_j^{(m)}(x)$ of each pixel x represents the pixel's similarity as to belonging to background patches. To determine $V_j^{(m)}$, we compute a feature descriptor $\bar{\mathbf{I}}_j^{(m)}$ for each patch $p_j^{(m)} \in \mathrm{H}^{(m)}$ and a descriptor $\mathbf{d}(x)$ for each pixel in $\mathbf{A}$. Following [2], the Voting Map can be computed as:

$$V_j^{(m)}(x) = \|\bar{\mathbf{I}}_j^{(m)} - \mathbf{d}(x)\|. \tag{3.3}$$

In the proposed method, we use the intensity values of the LAB color space as features, and determine $\bar{\mathbf{I}}_j^{(m)}$ by computing the mean feature vector across all pixels of a patch $p_j^{(m)}$. $V_j^{(m)}$ is the Voting Map corresponding to patch $p_j^{(m)} \in \mathrm{H}^{(m)}$. Figure 3.4 shows two example Voting Maps.

To determine which pixels belong to the same background region as the patch, a thresholding on the Voting Map takes place. During thresholding, pixels with a high similarity in comparison to the whole rest of the pixels are considered as belonging to the same background region as the patch.

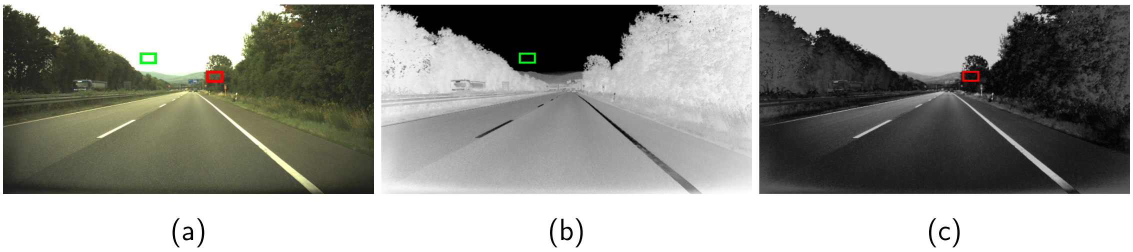


Figure 3.4: The position of two homogeneous patches in the sky (green box) and in the trees (red box) is shown in the original image (a). The Voting Maps are computed for the patch shown in green in (b) and for the patch shown in red in (c).

In this context

$$BV_j^{(m)}(x) = \begin{cases} 1 & \text{if } V_j^{(m)}(x) \geq T_{a,j} \\ 0 & \text{if } V_j^{(m)}(x) < T_{a,j} \end{cases}$$

is the binarized Voting Map computed for a patch $p_j^{(m)} \in \mathrm{H}^{(m)}$, and

$$T_{\mathrm{a,j}} = \rho \cdot \mathbf{Dyn}\left(V_j^{(m)}\right) \tag{3.4}$$

an adaptive threshold value. A closer analysis of the Voting Maps $V_j^{(m)}$ showed that the distribution of similarity values follows a Gaussian mixture model. As such, we determine $\mathbf{Dyn}(\cdot)$ by applying Otsu's scheme [100] on $V_j^{(m)}$ with $\rho = 0.2$ being a control parameter. Finally, each binarized Voting Map $BV_j^{(m)}$ contains pixels belonging to environment $BV_j^{(m)}(x) = 0$ or regions $BV_j^{(m)}(x) = 1$ forming potential candidate objects.

3.3.3 Combination of Voting Maps

By combining now all binarized Voting Maps, it is possible to obtain only those pixels which belong to object regions. Therefore, all pixels that are not voted to any background patch are identified. Thus, all binarized Voting Maps are combined according to which zone m the corresponding patch did belong. The final Voting Map $V^{(m)}$ for zone m is a weighted combination of the binarized map $BV_j^{(m)}$ and the adjacent binarized map $BV_j^{(m+1)}$. The combination of different zones is introduced since the edges between different zones are difficult to model due to different patch sizes within different zones:

$$V^{(m)} = \sum_j r \cdot BV_j^{(m)} + (1 - r) \cdot BV_j^{(m+1)} \tag{3.5}$$

$\forall\ x \in Z^{(m)}$ and a factor $r \in [0, 1]$. We then compute the final Voting Map V_f by accumulating the individual and binarized Voting Maps $V^{(m)}$ as follows:

$$V_f = \sum_{m=1}^{n} V^{(m)}. \tag{3.6}$$

3.3.4 Generating Hypotheses

After accumulating all binarized Voting Maps, there may be pixels left in the final Voting Map that have not been assigned to the surrounding environment. These pixels may belong to the regions of potential candidate objects and form, hence the regions for generating hypotheses. In order to extract relevant regions for hypothesis generation, we first cluster the pixels according to their location in images using *density based spatial clustering for applications with noise* DB-Scan proposed by Ester *et al.* [29]. After clustering, a refinement stage is implemented to avoid generating single hypotheses containing multiple close-by candidate object regions or noise, with assuming a different appearance and, hence, different features of different but close candidate objects. Thus, the deviation of all features $\mathbf{I}_c$ within one cluster to its mean value $\bar{\mathbf{I}}_c$ should not exceed a certain value dev_c computed by

$$\mathrm{dev}_c = \left| 1 - \frac{\mathbf{I}_c}{\bar{\mathbf{I}}_c} \right|. \tag{3.7}$$

If several object regions form one cluster, it is sub-clustered iteratively with a k-means clustering scheme on all features, followed by re-clustering to avoid mini-clusters. Since we expect only objects with a certain size, clusters which do not show good compactness are dismissed. The compactness criterion includes the elongation in x- and y-direction of the cluster as well as the total size in square pixels. A cluster is dismissed if it contains more than 300 pixels or if it is spread in x- or y- direction by more than 150 pixels. Table 3.1 summarizes the implementation details.

Zones (rectangular)	$n = 3$
Patch Sizes	14-29 x 10-20 pixels
ρ	0.2
r	0.66
dev_c	0.1

Table 3.1: Implementation details: Experimental evaluations show best performance for the above reported parameters.

3.4 Evaluation

The approach is evaluated on a motorway traffic data set newly created for this work and described in the following Subsection. It is of interest whether all vehicles and objects in the distance are included in the final object hypotheses. This is evaluated by a set of metrics and by a qualitative analysis of the objects in the detected regions.

3.4.1 Data set and Labeling

For evaluation of the approach, a data set of 80 different motorway scenes was recorded. The selection of images ensures that each scene environment differs significantly. Figure 3.10 shows several examples of the data set. The size of the images is 640×1176 pixels. All relevant objects are labeled pixel-wise as object regions of interest in each scene, see Figure 3.10b. Hereby, only traffic signs which point against the direction of movement are considered as an object region. Likewise objects which are covered by another object by at least 70% are considered as a single combined object region. The database contains in total 618 objects as ground-truth. Table 3.2 summarizes the relevant data set characteristics.

3.4.2 Evaluation Metrics

The evaluation is subdivided into measuring the performance based on pixel-level metrics and object-level metrics. As quality criterion recall, precision and the F_β score are computed. The recall is used to determine the quality of a method if high costs are associated with False Negatives (FN), while precision is a good measure if the costs of False Positives (FP)

No. of different scenes	80
Image Size	640 px × 1176 px
Average No. of object-pixels per image	538
Average percentage of object-pixels per image	0.071%
No. of objects smaller than 200 pixels	618
... Cars	344
... Trucks	83
... Traffic Signs	166
... SOS Telephones	14
... Others (flags, gates, advertisement)	11
No. of objects occluded by 40-70 %	265

Table 3.2: Data set description

is high. These both measures are given by

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}, \qquad \text{Precision} = \frac{\text{TP}}{\text{TP+FP}},$$

$$F_\beta = \left(1 + \beta^2\right) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot (\text{Precision} + \text{Recall})},$$

where TP is the number of relevant records retrieved, FN the number of relevant records not retrieved and FP the number of irrelevant records retrieved. For the pixel-level evaluation, a record is the overlap of ground truth and detection of a single pixel. For object-level evaluation, a record is counted as TP when at least 10% or 20% of the total ground truth object region pixels are detected. For the generation of object hypothesis, the object-wise recall is ideally 1 in order to ensure that all object candidates are included. Nevertheless, the pixel wise amount of FP records should be as low as possible. Hence, the used F-measure is twice as sensitive to recall than precision ($\beta = 0.5$).

3.4.3 Results

The performance of the approach without refinement is evaluated and followed by an evaluation of the refined regions. Finally the proposed approach is compared to the objectiveness measure.

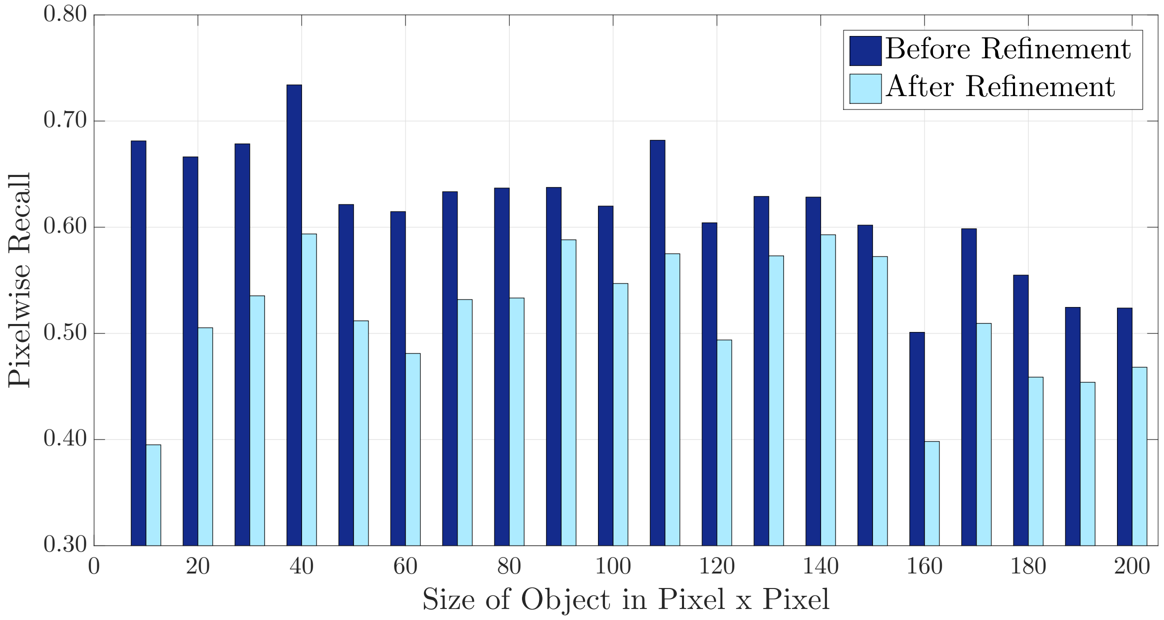
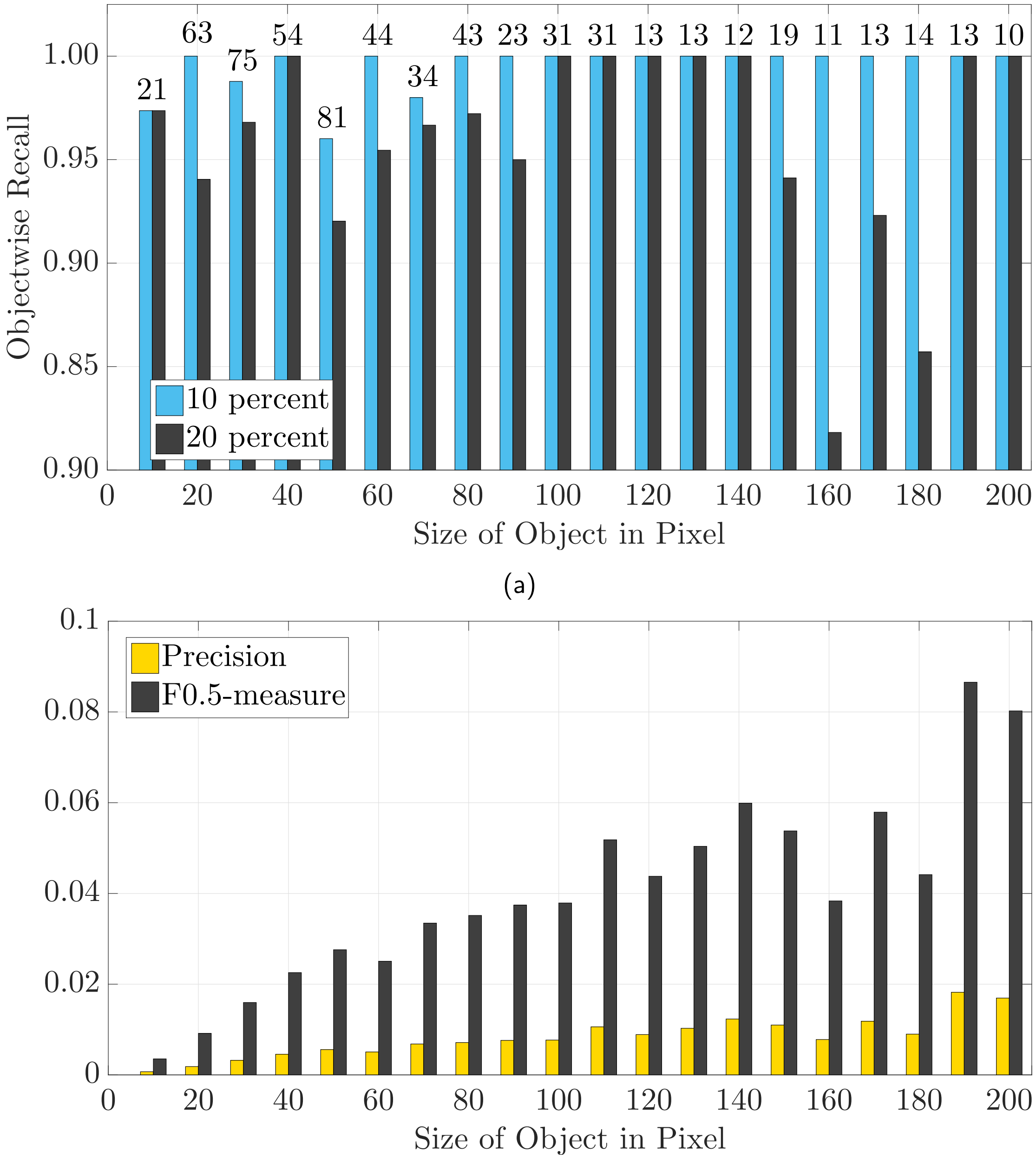


Figure 3.5: Pixel-wise recall before (dark blue) and after refinement (light blue). On average the object regions are covered by 61.9% before refinement and 51.6 % after refinement.

Before Refinement

First, the detected object hypotheses are analyzed on the object level. Figure 3.5 shows the pixelwise recall before and after refinement. In average, 61.9% of the object region pixels are detected. In particular small object regions with less than 40 pixels are detected with a pixelwise recall above 65%. Hence, the algorithm is well-suited for the detection of small object candidates that differ in their appearance from their environment. Figure 3.6a shows the object-level recall for object candidates that are covered by at least 10% or 20% of the total ground truth object region pixels. The proposed method detects in average 99.5% for the 10% pixel-count threshold and 96% for the 20% pixel-count threshold of objects before refinement. The average pixel-wise precision is here at 0.8% and the average F-measure is at 4.1% as shown in Figure 3.6b and Figure 3.7b.

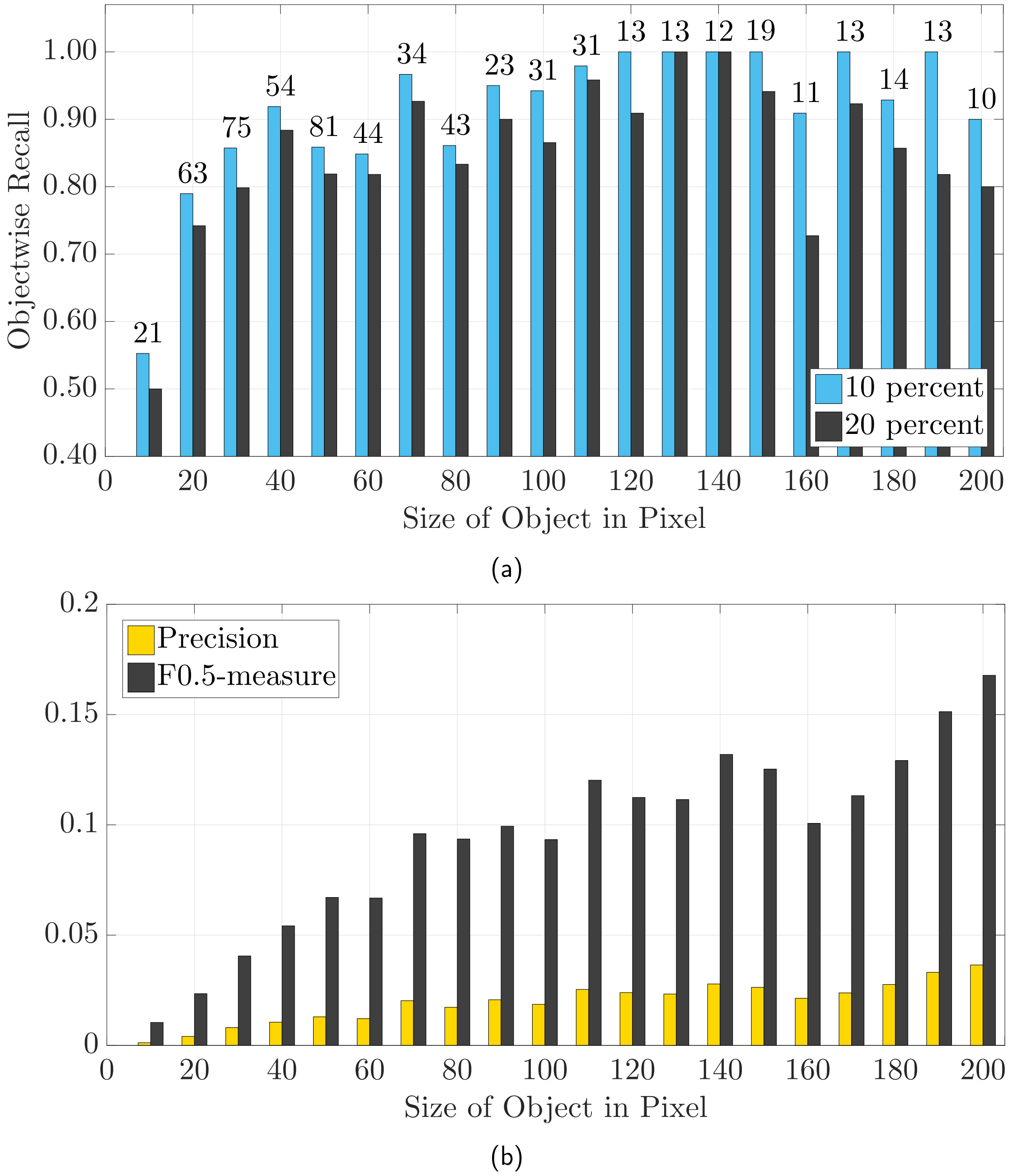
A high number of irrelevant records retrieved is mainly due to false-positives on street markings, transition pixels between sky and trees as well as false-positives on bushes or grass on the side of the street. Especially the false-positives bush/grass-pixels are often not covered by any patch since this area is cluttered by objects or are too small stripes. The search area to find and classify small object candidates is now diminished from 752640 pixels for the complete image to on average of 10716 pixels found before refinement. In summary, high recall on the object-level and



(a)

(b)

Figure 3.6: In (a) the recall of object regions which are covered by at least 10 % (light blue) and 20 % (grey) of records before refinement is displayed. (b) shows the same measurements as in (a) after refinement. In (c) the pixel-wise precision values (orange) and F-measure values (grey) before refinement are shown. In (d) the same values as in (c) but after the refinement stage are shown. The number on top of the bars in (a) and (b) displays the amount of objects within the data set for each object size.



(a)

(b)

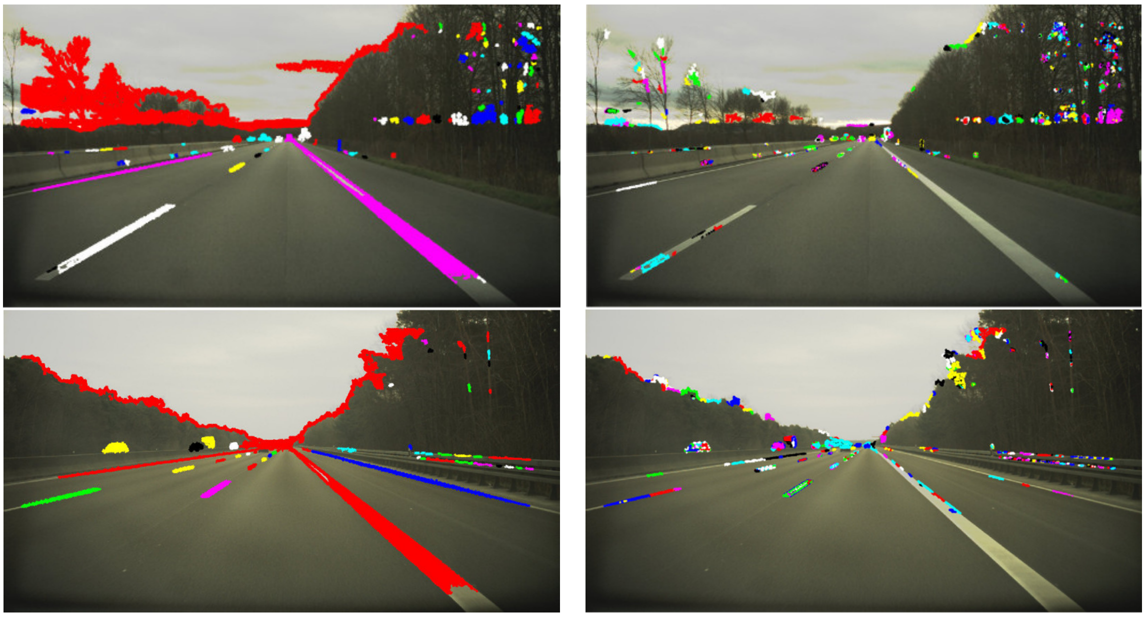
Figure 3.7: The results as explained in Figure 3.6 are shown when refinement is applied. The number on top of the bars in (a) and (b) displays the amount of objects within the data set for each object size.

pixel-level is achieved before refinement. However, a considerable amount of false positive pixels on larger areas such as the transition from sky to environment can be further reduced by an additional refinement step. Target of the refinement step is to exclude larger, connected areas from the pixel candidates for small objects in the distance.

After Refinement

After refinement, we expect an increase in precision and F-measure values while the recall remains approximately the same. Nevertheless, the overall detection rate of object regions which are covered by more than 10% and 20% for object regions smaller than 200 pixels drops to 91.3% and 85.1%, respectively, as shown in Figure 3.7a. The object regions smaller than 200 pixels are on average covered by 51.6% of pixels as shown in Figure 3.5. However, the pixel-wise precision increases to 2.0 % and the F-measure increases to 9.5 % (Figure 3.6b and 3.7b).

In Figure 3.8 two examples illustrate how the refinement step dismisses false-positive records, in particular, transition pixels. The drop in all recall values - especially for object regions smaller than 60 pixels - is due to occluded cars which are close to the horizon where mostly the front-shields are visible and the rest of the cars is hidden behind road boundaries or other cars. As the windshield mirrors the sky, it shares the image characteristics of the sky and/or tree. Such a car is detected before refinement but is associated with the cluster describing the transition pixels between sky and trees by DB-Scan [29] clustering. The following object region refinement is not able to distinguish the windshield-pixels from sky-tree-transition and due to the compactness criteria the car may be missed.



(a) (b)

Figure 3.8: Two scenes before (a) and after (b) refinement. Many false positives records of transition pixels and in trees are identified and dismissed during the refinement stage.

In Figure 3.9 two scenes with different weather conditions and the result of the proposed method are shown. Shown empirical on these examples, the advantage of image dependent detection methods only becomes visible, as the detection of anomalies is independent of weather conditions.

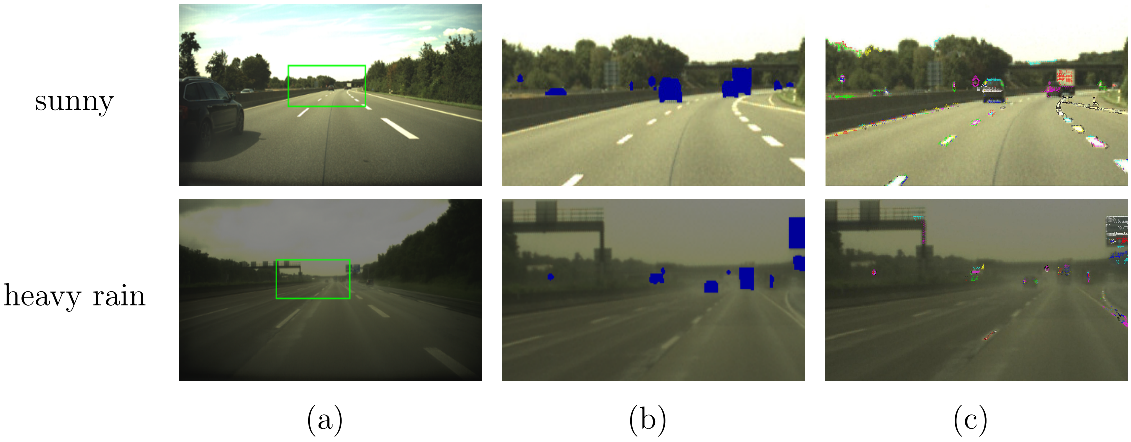
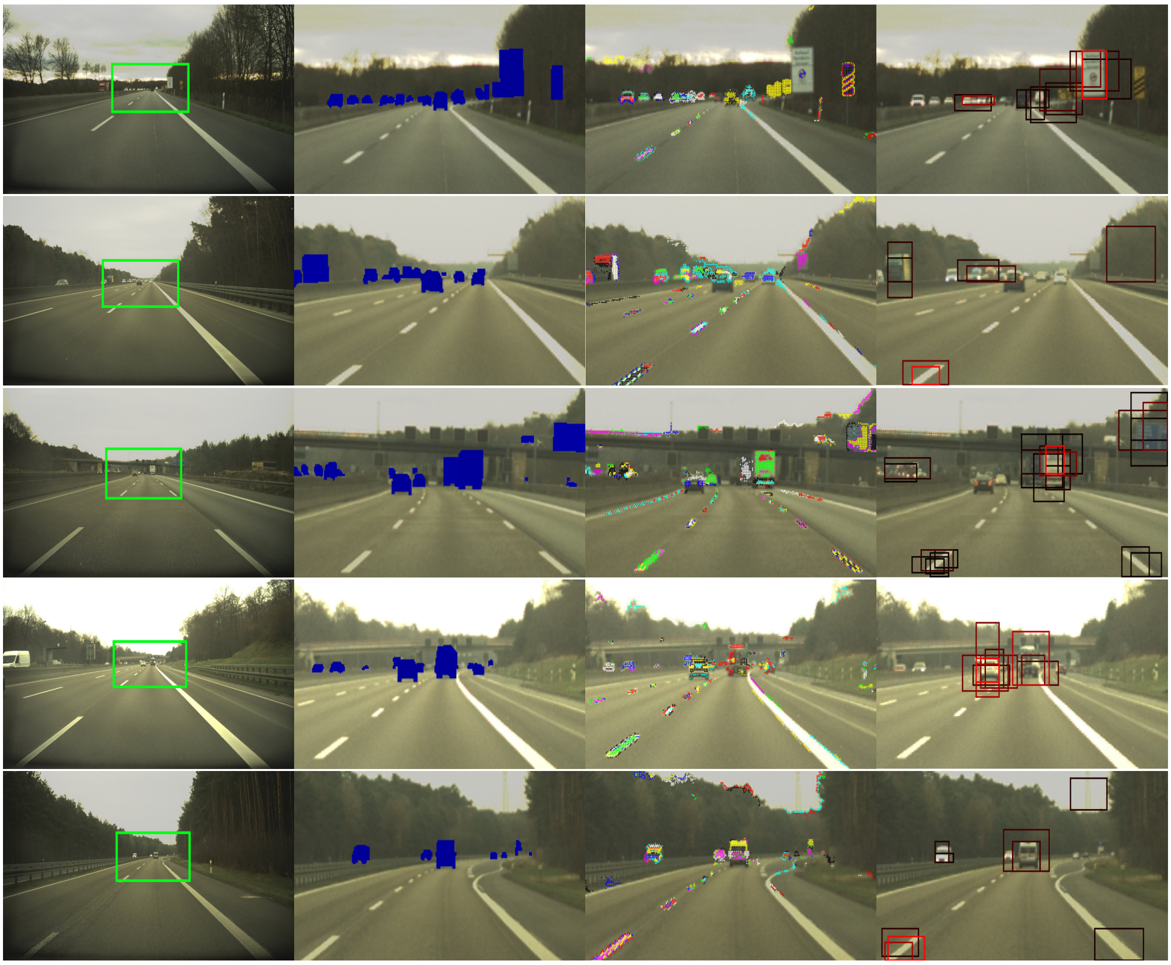


Figure 3.9: (a) shows the original image for two different wheather conditions: sunny (top row) and heavy rain (bottom row). (b) and (c) show the ground truth labels and the computed distant candidate object regions.

On average 538 pixels of the whole image are belonging to an object which is smaller than 200 pixels. For object regions smaller than 20 pixels, the average pixel number is 24 pixels per image. The maximum precision value of 3.6% is observed for objects smaller than 200 pixels. Considering the fact that small target objects cover less than 0.07 % of the full image, high precision values are difficult to reach and even a small false-positive rate contributes to a significant loss in precision. Since the portion of object pixels is so small compared to the whole number of pixels in the image, a double-digit precision region is already hard to reach.

The proposed approach detects pixels that are associated with vehicles or traffic-relevant object candidates in the distance. An additional classification stage is required for hypothesis verification. Often a HOG-based support vector machine (SVM) classifier is used for this purpose. Feature computation and classification is especially expensive for small objects leading to a need for efficient hypothesis generation. For objects smaller than 200 pixels, on average 3991 pixels are retrieved after refinement as possible object candidates. The improvement in sliding-window scanning required for hypothesis verification can be approximated by the ratio of the total number of pixels and the detected small object pixels. On average, the number of pixels that need to be considered for classification is $752640/3991 = 188$ times less than full-image scanning.



(a) (b) (c) (d)

Figure 3.10: (a) shows the original image. The green rectangle illustrates the zoomed region in (b)-(d). The zoomed versions are 293x159 px and show distant objects. In (b) the zoomed ground truth data is shown. (c) shows the here presented method after refinement and (d) the objectiveness measure where the red rectangular gives the most probable object within the image.

3.4.4 Comparison to Objectiveness Measure

We wish to estimate the quality of the proposed method in comparison to another method. Alexe *et al.* [3] developed an algorithm for generic object region detection by a set of objectiveness cues. We trained their algorithm on 30 images with 219 bounding boxes from our data set. Figure 3.10d shows the performance on distant object candidates and illustrates the difficulty to detect distant salient object regions in complex environments. The images given to the objectiveness algorithm are cropped to the size of 293 x 159 pixels and only the first 70 object boxes with the size of maximal 50 x 50 pixels are shown (see Figure 3.10d). The objectiveness cues are well-designed for large objects; yet small objects in the distance are more difficult to be detected by the same cues. The multi-scale saliency can not be applied on whole motorway scenes since the color distribution within such scenes is too wide to obtain relevant results. Since distant objects occupy few pixels on the image, the edges density criteria as well as superpixel straddling cannot give significant evidence for an object hypothesis. For distant objects, local or semi-local indicators are more suitable. To compare the bounding box algorithm of Alexe *et al.* with our proposed pixel wise method, all pixels within a bounding box are counted as a record. Those records are then evaluated as described in Section 5.5.3. Figure 3.11 shows the average pixel-wise recall, precision and F-measure values over all object sizes smaller than 200 px for the objectiveness-measure and our method before/after refinement.

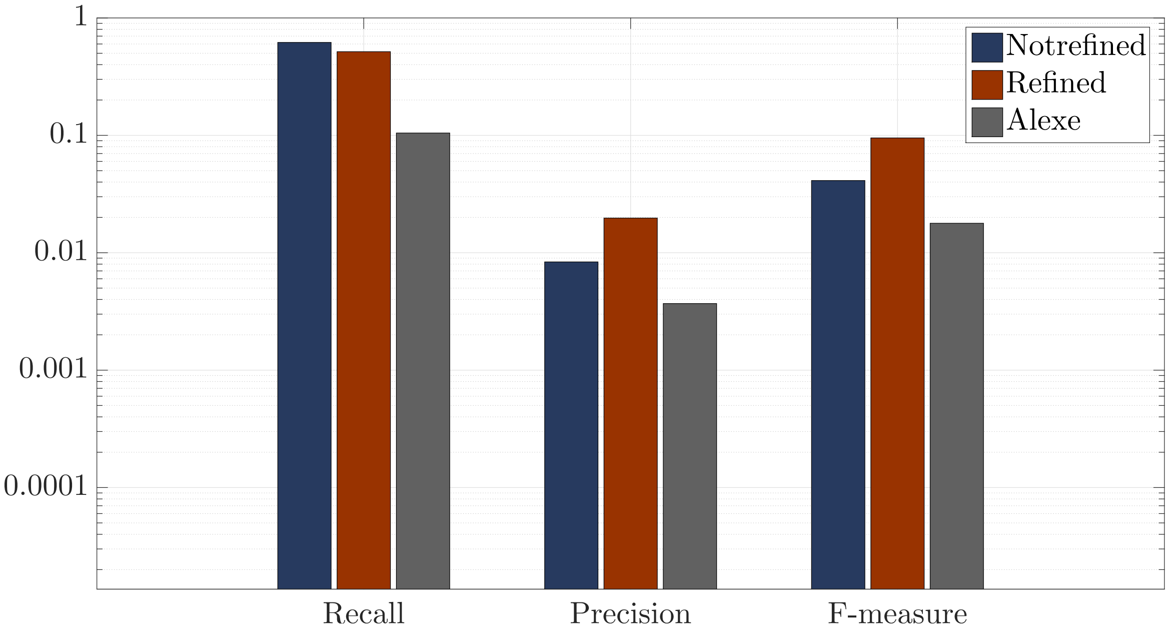


Figure 3.11: The average recall, precision and F-measure values over all object sizes are shown for the objectiveness measure [3] and our method. The values are displayed in logarithmic scale to be able to compare recall, precision and F-measure values within one plot. It can be seen that our approach outperforms the approach to hypothesis generation based on [3] on all metrices chosen.

It shows that the here introduced method scores significantly better in recall values while the precision and F-measure values are roughly within the same range. Hence, the proposed method can be used to increase the detection rate of object regions for small object candidates.

3.5 Summary and Conclusion

A novel generic hypothesis generation for objects in large distance to the camera or low pixel coverage on the image is proposed. The proposed approach is tailored to the detection of small and/ or distant objects within the image through the proposed saliency inspired Voting Maps. The approach applies a frequency-tuned saliency generator [2] and uses distinct patches to create zone-based Voting Maps. The selection of usable patches for the Voting Map creation remains highly adaptive to each image and ensures hereby a high adaptability to different weather and illumination condition during motorway drives. The Voting Maps lead directly to generic object region detection, thereby outperforming existing hypothesis generation based on the objectiveness measure [3]. The effectiveness of the Voting Maps is demonstrated on motorway scenes using only color channels, while other features such as orientation and local distinctiveness are also applicable.

An extensive parameter optimization such as needed for convolutional neural network based approaches is hereby not required. Hence, a large data set for parameter optimization is not needed and can be implemented in different camera set ups without any additional extensive parameter adaption. As so the proposed method is class independent and is able to detect traffic sign, truck, car and motorbike regions concurrent.

Temporal filtering and incorporating prior knowledge such as road models can increase the performance of our approach further. It is possible to classify the homogeneous patches used to generate Voting Maps in simple classes such as sky or trees to achieve a semantic segmentation with all Voting Maps belonging to the same class. With such information a further improvement by dismissing false-positive object hypotheses enclosed in regions such as sky or trees is possible.

Applied on motorway scenes, the approach shows promising results in generating hypotheses of automotive objects for ADAS algorithms as shown in Section 3.4.3. Through its high adaptability on single-frames the object region detection gives a promising base for following processing of the hypothesis regions. The reduction of image information down to regions of interest for following inspection through subsequent classification algorithms allows to reduce the data flow within automotive car architectures from camera units to processing units.

As motivated in the beginning of this chapter the proposed approach can be used for generic object region detection with the advantage of being able to fully understand the information retrieval via signal transformation of this method. The global feature extraction in form of anomalies with respect to larger homogeneous image areas can be retraced simply during calculation. For safety engineers in the automotive industry such a retrace-ability is favored to methods based on machine-learned methods, where the computation steps within the detector exhibit for humans abstract features or computations. However, in several recent publications [71, 83, 106, 107, 114] convolutional neural network based concepts showed a strong capability for applications in computer vision due to its kernel-based feature extraction. As opposed to the method presented in this chapter, the following chapters present a technique to increase the detection for small automotive object regions with a machine learned feature extractor, while taking a deeper look into the computational steps within the network.

Chapter 4

Incorporation of Global Priors

In this Chapter, the incorporation of priors with global features to local CNN-based object region detectors is proposed, see Figure 4.1.

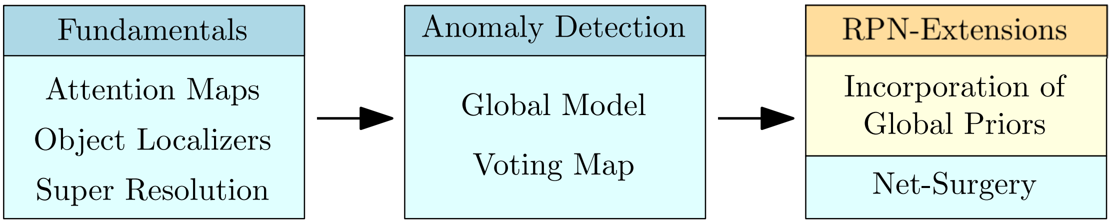


Figure 4.1: In this Chapter, the incorporation of global priors to locally restricted region proposal networks is described and evaluated.

Through such incorporation, only a comparable small data set is needed to increase the object region detection performance of the network. Following a short introduction of specific region proposal methods, the architecture for incorporating priors is presented. In an extensive evaluation new metrics for the specific case of region proposal networks are applied on a test data set specifically labeled for small objects. The Chapter concludes with the key aspects of the proposed method and a summary. The major results of this Chapter have previously been published in [35].

4.1 Motivation

Given the availability of powerful processing units and sensors such as Radar-, Lidar- and Camera-based systems, advanced safety functions at high speed can not only be implemented on embedded systems efficiently, but also leverage the increased computing power to enhance the reliability of detecting the targets. As already motivated previously, one example comprises the detection and classification of distant objects, where robust detection is crucial to increase the use-case and robustness of assistance functions on motorways beyond today's level [5]. This came along with the introduction of deep learning and convolutional neural networks, with the latter resulting in a dramatic increase of the detection performance. Recent detection approaches based on deep learning combine three stages in one pipeline for efficiency and run-time reasons: 1) feature extraction, 2) region proposals and 3) object classification. The objective of feature extraction is to learn a meaningful set of features representing the target objects. Region proposal methods make use of these features to identify regions potentially containing target objects and to reduce the number of regions fed to an expensive classification stage. Approaches to proposing regions in deep learning architectures are of great importance due to the need of detecting regions containing distant objects very quickly for the purpose of robust detection for autonomous driving. The majority of recent approaches to region proposals make efficient use of a network structure and leverage convolutional layers to extract features for both region proposal determination and object classification. Those so called region proposal networks (RPNs) rely on sets of predefined anchors, which are rectangular boxes of different scales and sizes. The network used as an RPN slides over the feature maps of a convolutional layer, computes scores for each anchor at each location and refines the boxes by bounding box regression. The score shows how likely an anchor contains an object or background. The bounding box proposals with the highest scores are then proposed as possible object regions within the image and passed to the classification branch [83, 106, 107, 114]. However, the size of the areas chosen to judge the score of an anchor is relatively small when compared to the size of original images, and global information may be missed. As a result, existing RPNs use features extracted from local surroundings to determine the presence of an object in a local fashion and tend to ignore global relationships between interesting objects and their surrounding. A more detailed explanation on RPNs can be found in Section 2.3.

Prior work on RPNs has attempted to address these issues by several approaches. Ohn-Bar and Trivedi [99] proposed an integrated framework based on a multi-scale scheme in which features around a region of interest are extracted at different scales to determine the object's location and presence.

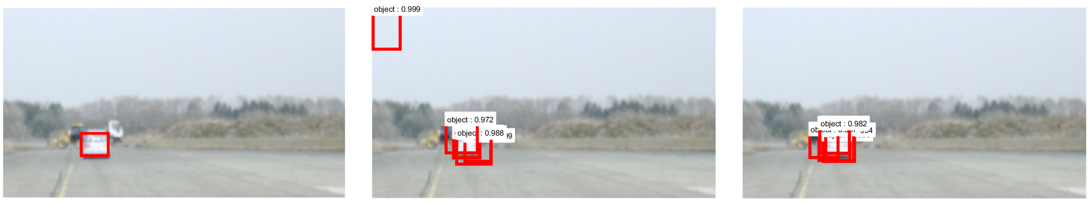
Hoang *et al.* [53] make use of the approach of Ren *et al.* [107] and utilize several feature maps from different convolutional layers as additional input to the RPN. By using this architecture, information extracted at lower layers contain a rich set of fine low-level features that can be used to detect small objects. Zhang *et al.* [138] integrates the *à trous* trick [15] to increase the resolution of the feature maps in order to enhance the region proposals for small objects. However, these approaches do not incorporate any global knowledge about the presence of objects within an image and tend to propose regions containing background only. Since the object score per anchor is computed based on the local representation of the inspected feature maps, relationships between distant objects or background information are not modeled explicitly.

First methods for incorporating global context in CNN architectures propose to use semantic segmentation as input [13,91]. Shrivastava *et al.* [115] built a framework that provides top-down segmentation information to the RPN. They propose a network in a network architecture to compute the segmentation as a primary step, and to consider segmented regions at different layers in the RPN. This two-network strategy requires pixel-wise labeling and bounding box labeling for the training set, and results in high annotation effort required to enable training and in a huge computational overhead. Also Zhang *et al.* [140] proposed to add additional networks which learn to guide the network for the special use case of pedestrian detection. However, also here the computational costs are very high. Therefore, an efficient method for incorporating global context into a region proposal network would be much desired that can improve the proposals of regions and does not rely on additional training.

4.2 Outline

The main contribution of this Chapter is to extend the architecture of a RPN by incorporating global information that can be computed with no or only some additional training data. Here, global knowledge is introduced to guide the RPN towards interesting areas and provide region proposals for those regions only. Inspired by the concept of saliency, this Chapter leverages the global nature of saliency detection to compute the global information supporting the RPN in an unsupervised fashion. Given that salient regions have higher probability to contain objects of interest as background regions [5,64,110,111], saliency-inspired priors can help a RPN better suppress false positives proposed by layers using local information only as shown in Figure 4.2. This results in an increased recall of RPNs based on a global prior, and reduces the number of bounding boxes to be evaluated by an often time-consuming classification branch in a neural network architecture.

Moreover, high recall for RPNs is of particular importance for detecting distant and small objects because the number of possible object candidates in an image increases with decreasing object size. A highly relevant application is the detection of distant vehicles on motorways for advanced driver assistance systems and autonomous driving. First evaluations based on a variety of metrics and a distant vehicle data set demonstrated a promising performance of the proposed approach.



(a) Ground truth (b) With no prior (c) With global prior

Figure 4.2: Incorporation of priors into a region proposal network (RPN). From left to right: Ground truth, the first five bounding boxes proposed by an RPN with no prior, and the improved proposals when using a global prior.

4.3 Incorporation of External Data

In todays network architectures, the lower layers extract several local features such as edges, motifs, parts of objects and finally object descriptors to create feature maps, that are commonly used by both a region proposal and a classification branch for efficiency reasons. However, feature maps considering the entire image and leveraging global information have the potential to better emphasize global distinctions in the entire image and suppress less relevant local features. As such, the use of global information to amplify local features in a global fashion and to help the RPN better determine object regions is desired. In this work, the approach aims to combine local and global features within one architecture in order to in-

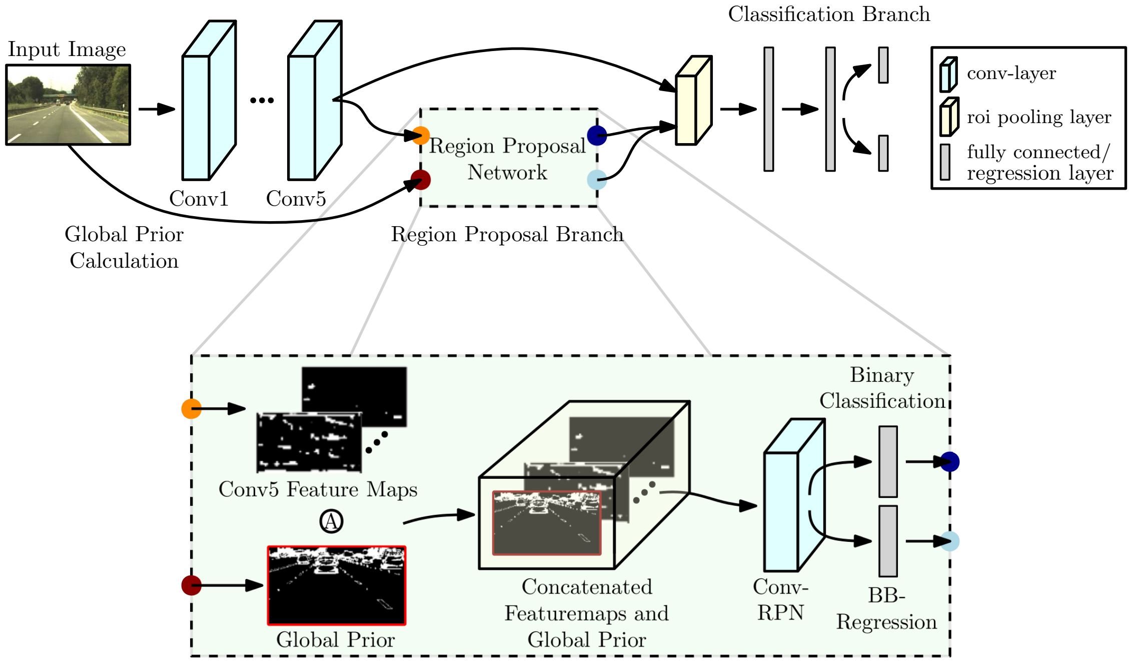


Figure 4.3: The architecture of a prior-based RPN: A convolutional network is used which consists of a region proposal branch and a classification branch. Here, the region proposal branch as shown in the highlighted area is extended by incorporating global priors. The overall network bases on the work of [107].

crease the performance of the RPN significantly. Moreover, keeping the computational complexity as low as possible despite the introduction of additional functionality in the architecture is of high interest. Figure 4.3 shows the entire architecture of such a network, where the region proposal branch has been extended to consider global information, potentially extracted in an unsupervised fashion. The regions detected can then be sent to a classifier to either identify objects or reject wrongly detected regions. By using this scheme, the proposed approach does not require additional labeling and can leverage pre-trained networks and considerably small data sets. The following Sections describe the proposed architecture and three global location priors in more detail.

4.3.1 Faster R-CNN with ZF Net

The following two contributions are build upon the Faster R-CNN architecture [107] which is a commonly used object detector that uses a common feature representation for the region proposal (RPN) and binary classification branch (BCN) as shown in Figure 4.3. The Faster R-CNN 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 [87]. Such an two-stage detector achieves higher detection accuracies than single-shot architectures like Yolo or SSD [58] (further details can be found in Section 2.3.2).

Different core network architectures can be used for feature extraction. Here, as a small network architecture the ZF-Net is chosen 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.* [32] 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.

The RPN consists of one convolutional layer followed by a box regression layer, and uses feature maps extracted in the fifth convolutional as an input to propose the corners of the bounding boxes and the corresponding object scores. After applying non-maximum suppression, the proposed bounding boxes are transferred through a roi-pooling-layer to the fully connected classification branch of the ZF-Net. The roi-pooling-layer takes as input the feature maps from the last shared convolutional layer and four coordinates of the top left and bottom right corners of the region of interest. In the Faster R-CNN the coordinates are computed by the region proposal network. The roi-pooling-layer converts the output of the last shared convolutional layer within the proposed regions of the RPN to a same sized feature map through max pooling to fit to the following fully connected classification step.

4.3.2 Incorporation of Global Prior

To overcome the local restriction in the RPN when generating feature maps, it is proposed to integrate global knowledge as a prior in order to consider global dependencies within an image. Here, distinctive features are computed in a global fashion and provided as a prior map to better emphasize local but important regions. The global prior map is combined with the local feature maps provided by the five convolutional layers and fed to the subsequent RPN. It is proposed to include the prior map after the fifth convolutional layer because of the low overall impact of the prior map on the feature maps produced by the first layers representing low level features only [136]. The same observation was made by Shrivastava *et al.* [115] who showed no significant improvements of an extended RPN over a regular RPN when incorporating segmentation at an earlier stage. As such, the extraction of low and high-level features is an important step towards proposing object regions and can be supported best by including the global prior after the last shared feature maps. Finally, the prior maps are normalized and pooled to fit to the size of the fifth convolutional feature maps, and serve as an input to the RPN to provide both global features and their location for the task of object detection. It is worth to mention that prior maps can be computed by any approach considering global context such as unsupervised saliency, and is independent of any prior knowledge of the scene. Moreover, incorporating the prior after the last layer allows to use pre-trained off-the-shelf network models and prevents over-fitting, especially, when small data sets are available only.

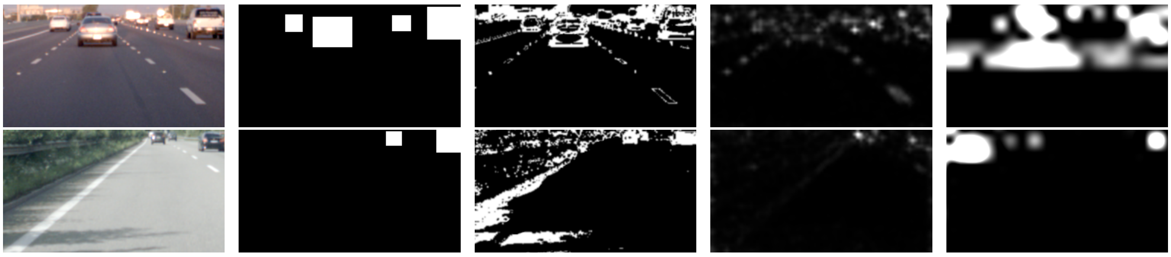

(a) Source (b) Ground Truth (c) Voting Map (d) SR (e) VA

Figure 4.4: Different global priors shown for two example images. From left to right: Input image a), ground truth boxes b), Voting Map [5] c), Spectral Residual [57] d) and Visual Attention [64] e).

4.3.3 Prior Maps

It is desired to leverage the global nature of saliency approaches to compute prior maps based on the concept of saliency. Saliency maps are designed to highlight discontinuities within the entire image by comparing the global and local appearance of features and regions, and can provide a global focus of attention. Given the use-case of detecting small vehicles in images where the vehicles are unique and distinct from the rest of the scene, these properties make saliency computation a desired global prior for improving the region proposals in a RPN. Moreover, saliency maps can be computed in an unsupervised fashion without using prior knowledge and are, therefore, applicable to any type of input data.

In this work, three approaches to visual attention, i.e., Visual Attention [64], Spectral Residual [57] and Voting Maps [5] are selected. Each approach is described in the following Section, and the global prior maps computed for the use-case of distant vehicle detections are shown in Figure 4.4.

Visual Attention Map

The visual attention map is one of the earliest approaches to highlight possible objects within an image and is inspired by the human visual attention process [64]. The visual attention map is composed of nine spatial scales of a given image, and provides a global prior considering local influences. At each location within the scaled image, the surroundings and the features color, intensity variance and edge orientation are computed and compared against the rest of the image. Finally, the saliency values of the maps computed for each feature are accumulated and combined to form the final attention map. The consideration of the different spatial scales, the local influences in a global framework and its tendency to highlight regions in a compact manner make this visual attention map a good candidate for a global prior for the purpose of proposing regions for small objects of different sizes and colors.

Spectral Residual

Hou *et al.* [57] create saliency maps by analyzing anomalies in the frequency spectrum of an image. The assumption is made that each image exhibits a highly predictable spectral distribution that is modeled by the log spectrum of a down-sampled image. Hence, a log Fourier spectrum of the input image is computed and subtracted by the log spectrum of the down-sampled image afterwards. The Spectral Residual is then transformed to the spatial domain and represents the final saliency map. Since regions containing small objects exhibit high spatial frequencies when compared to background, Spectral Residual can be beneficial for guiding the RPN towards regions containing small objects and is a good approach for providing prior maps. In addition, the transformation from the spatial to the frequency domain and the computation of the spectral residual can be done very cost efficiently by using hardware-accelerated components on an embedded device.

Voting Map for Distant Objects

The approach to computing Voting Maps [5] is tailored to the task of highlighting small object regions in motorway scenarios and has, therefore, been chosen as a possible global prior. The underlying idea comprises a modeling of the background within the image with very few homogeneous areas. Each homogeneous region is described by several patches which are fulfilling an adaptive homogeneous criteria. For each homogeneous patch, an individual Voting Map is created by subtracting the average color feature value within the patch from the image, followed by an adaptive threshold to compute a binary representation of the individual Voting Maps. The final Voting Map is computed by accumulating all individual Voting Maps. A detailed description of the Voting Map can be found in Chapter 3.

4.4 Metric and Results

The proposed approach is evaluated on a motorway traffic data set including many distant cars and trucks. The evaluation includes the results for the baseline algorithm and for the extension of the RPN with three different saliency priors. The performance is measured by the recall and the localization error. In addition, the computational costs of the different approaches are compared.

4.4.1 Data set and Data Augmentation

A data set for distant vehicle detection has been recorded and includes 1035 different motorway scenes. The data set contains over 2.000 distant cars and trucks with object sizes smaller than 30×30 px and is subdivided into training, validation, and test set as illustrated in Table 4.1.

Data set	Nr. of Images
Train	827
Validation	104
Test	104

(a)

Object Size (px)	Nr. of Objects
$[1-10]$	469
$]10-20]$	680
$]20-30]$	1135
$]30-200]$	903

(b)

Table 4.1: Details of the used data set. (a) illustrates the number of images used for training, test and validation. (b) shows the number of differently sized objects in the entire data set. In average two objects smaller than 30×30 px are included in each image.

The image size is 1176×640 px. To fine-tune pre-trained networks which are initially trained for larger objects, each image is cropped to a size of 160×290 px and upscaled to the input size of the CNN. In addition, data set augmentation is used for both training and testing (see Figure 4.5). To avoid over-fitting during training, each image is flipped vertically and used additionally for training. For the test set, 20 crops are taken from each image using a random jittering around the image center, ensuring that each crop includes at least one object. Cropping may result in truncated objects that are excluded from the evaluation. Slightly occluded objects are included in the data set.

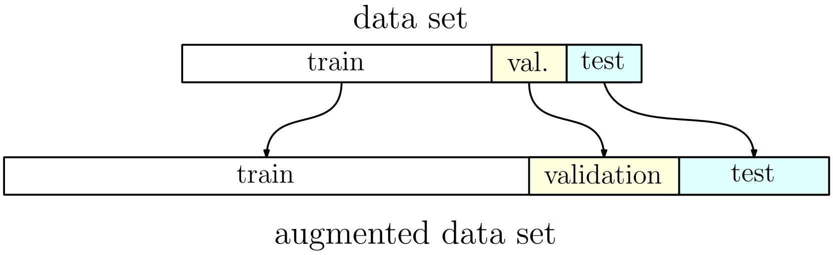


Figure 4.5: Data augmentation of the used data set for each subset. The Train part is to optimize/generalize parameters in the network. The validation set is used during training to obtain a feedback on the choice of parameters. This gives the possibility to stop training early in case of inappropriate parameter choice. The test data set is only used for the final evaluation to determine the performance of the model. Data Augmentation is used to increase the data set size.

4.4.2 Training

A ZF-net architecture with weights pre-trained on the ImageNet data set is used as the core network for the Faster R-CNN, as motivated in Section 2.5. The ZF-net architecture is first proposed by Zeiler and Fergus [136] for the purpose of classifying whole images. Hence, it consists of 5 convolutional layers and two fully connected layers. Within the Faster R-CNN set up the fully connected layers are not considered and instead the RPN is built on the 5th convolutional layer. In Table 4.2, the architecture parameters of the ZF-net are presented. The architecture of the RPN is summarized in Table 4.3.

ZF-net					
Layer	H	W	D	Kernel FxF	Stride
Input	1000	553	3	-	-
Conv1	500	277	96	7x7	2
Pool1	250	140	96	3x3	2
Conv2	126	70	256	5x5	2
Pool2	64	36	256	3x3	2
Conv3	64	36	384	3x3	1
Conv4	64	36	384	3x3	1
Conv5	64	36	256	3x3	1

Table 4.2: The table shows the parameters in the ZF-net architecture of the first 2 pooling and 5 convolutional layers and the RPN architecture. The first column shows the layer name while H, W and D are the height, width and depth of the output matrix of the layer with an example input of H = 1000, W = 553 and D = 3. The kernel size during convolution is given with the values FxF. The stride shows how many entries are not considered between two positions of the kernel during convolution.

Following [107], an alternated training is used to train the common feature representation, the RPN branch, and the classification branch. Therefore, both branches are trained end-to-end by backpropagation and stochastic gradient descent (SGD) [74]. The loss to train an RPN is generated through the sum of the loss of binary classification $L_{cls}\left(p_i, p_i^*\right)$ and bounding box regression $L_{reg}\left(t_i, t_i^*\right)$:

$$L\left(p_i, t_i\right) = \frac{1}{N_{cls}} L_{cls}\left(p_i, p_i^*\right) + \frac{\lambda p_i^*}{N_{reg}} L_{reg}\left(t_i, t_i^*\right). \qquad (4.1)$$

Here, i is the index of an anchor and p_i the predicted probability of anchor i to belong to foreground. p_i^* is the binary ground truth label. The vector t_i is of length four to represent the deviation of the center of anchor i to the ground truth center box in x- and y-direction and the difference of width and height of the anchor i to the ground truth bounding box. The number of randomly evaluated regions per image N_{cls} is here set to 256, while the number of possible anchor locations N_{reg} is 2304. The control parameter λ is used to balance both losses and is consequently set to 10. An anchor belonging to background p_i^* is set to 0, hence, the loss for regression is ignored for anchors belonging to the background.

RPN without external data					
Layer	H	W	D	Kernel FxF	Stride
Input	64	36	256	-	-
Conv-proposal	64	36	256	3x3	1
Cls-scores	64	36	18	1x1	1
Regression	64	36	36	1x1	1

Table 4.3: The table shows the parameters in the RPN architecture. The layers Cls-scores and Regression are fully connected layers which propose binary classification and regression targets for each anchor. For further explanation, see Section 2.3.2.

Regions proposed by the RPN, which have an IoU of more than 0.7 with an object region, are considered during training of the classification branch including the shared layer Conv4 and Conv5. The alternating training starts with the shared convolutional layers and the RPN. During the second stage of training, the weights of the RPN are frozen and the classification branch including the shared convolutional layers are trained. In a third stage, only weights of the RPN are subject to training. In the last stage, the fully connected layers of the classification branch only are trained. Such training procedure ensures an optimization of all weights within the different branches of the whole Faster R-CNN. Within this contribution, the fourth and fifth convolutional layers are fine-tuned on the pre-trained ImageNet ZF-net, and the RPN weights are randomly initialized with a Gaussian deviation of 0.01. The global learning rate is set to 0.001 for the first 60k iterations and 0.0001 for the last 20k iterations. In Figure 4.6, the loss and error on the validation set during training of the first stage, shared layer and RPN optimization, is shown. The loss of the regression is hereby given through

$$L_{reg}\left(t_i, t_i^*\right) = \sum_i \mathrm{smooth}_{L1}\left(t_i - t_i^*\right), \tag{4.2}$$

where

$$\mathrm{smooth}_{L1} = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise.} \end{cases}$$

The loss for classification is the log loss over two classes (foreground vs. background). In Figure 4.6a, the loss for binary classification and in Figure 4.6b, for regression during training of the first stage for different iterations on the validation set are shown. These curves show that the parameter within the network are optimized for the binary classification and regression task after 80k iterations as the curves reaches a stable value.

The input size for the Faster R-CNN using the ZF-net is 553 × 1000 px. The image crops are upscaled by a factor of 2.4, which was shown by Fan *et al.* [32] to improve the performance especially for the task of small object detection. For distant vehicle detection, a single aspect ratio of 1:1 is sufficient for the anchors. The anchor scales are chosen to $[16, 64, 128]$ px. The small number of chosen anchors prevents overfitting for small data sets. The roi-pooling-layer has a 6x6 output to the fully connected layer of the classification branch.

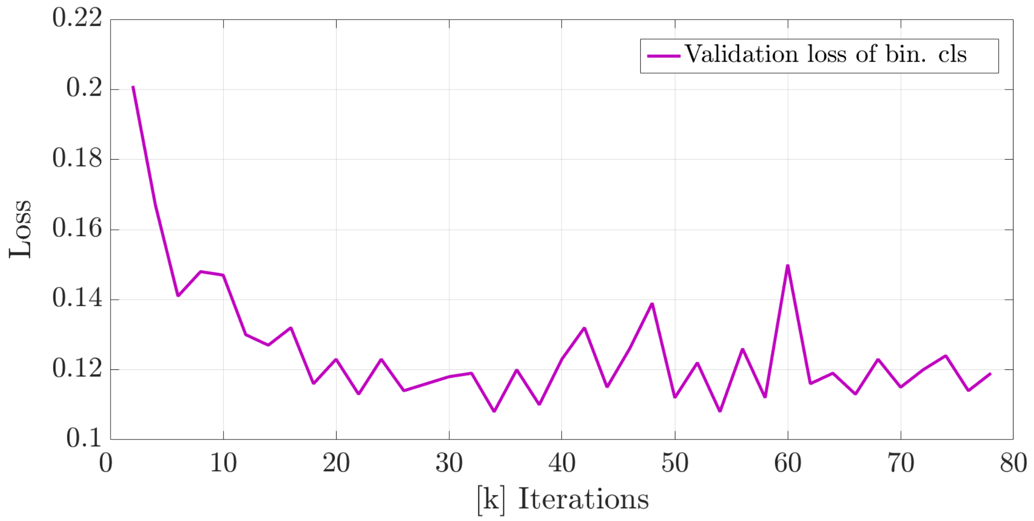
4.4.3 Evaluation Metrics

The performance is evaluated using a set of metrics: 1) the *recall* measures how many objects are proposed by the RPN and 2) the *mean area* and the *mean normalized distance* measure the localization accuracy of the proposed bounding boxes. The metrics depend on the intersection over unit (IoU).

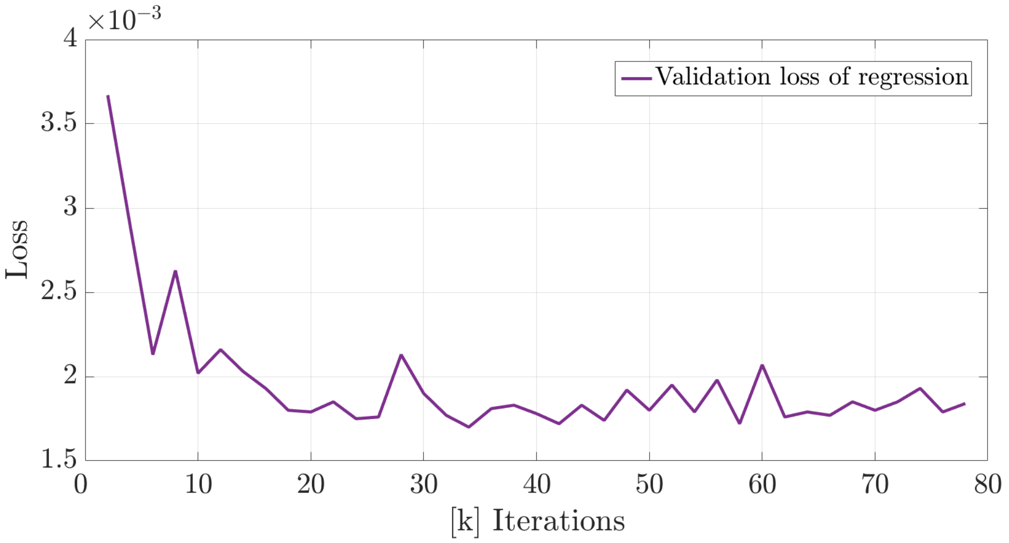
Intersection over Unit

The IoU measures the overlap between the area of the ground truth bounding box $\mathrm{A_{gt}}$ and the area of the detection bounding box $\mathrm{A_{dt}}$:

$$\mathrm{IoU} = \frac{\mathrm{A_{gt}} \cap \mathrm{A_{dt}}}{\mathrm{A_{gt}} \cup \mathrm{A_{dt}}}.$$



(a) Loss of binary classification



(b) Loss of regression

Figure 4.6: The loss for binary classification and regression of the RPN on the validation set during first stage training: shared convolutional layer and RPN. Fig.(a) shows the loss of the binary classification and Fig.(b) the regression loss of the multitask loss within the RPN.

When evaluating the performance of object detectors, an object is counted as a true positive when the IoU between ground truth and detection is larger than 0.5. For evaluating the performance of object proposal generators, lower thresholds can be considered because classification networks can refine the boxes by an additional bounding box regression. For this reason, the experiments are computed using the thresholds 0.25 and 0.5.

Recall

An important property of a region proposal method is a high recall. A high recall means that most regions containing desired object regions are in the list of region proposals. To accomplish this, the list of region proposals may end up containing a higher number of false positive regions that do not contain any object region. A high precision describes a proposal list with only true-positive regions at the possible cost that some true-positives are missing in the list. While the subsequent classification stage is optimized to discard false positives, it cannot retrieve false negatives. Hence, even though a longer region proposal list and therefore a lower precision uses higher calculation costs, a high recall is needed for a region proposal method since missing object regions impact the whole object detection chain directly severely. When all object regions in the test set are detected, the optimal recall of 1 is achieved as

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

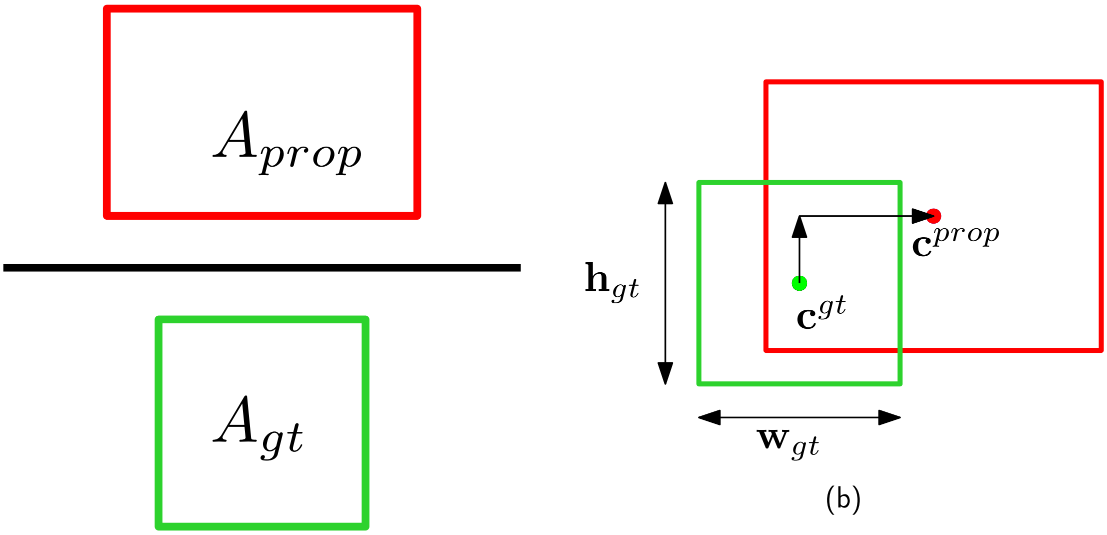
where TP is the number of relevant matches retrieved and FN the number of relevant matches not retrieved. The recall is plotted over the number of selected bounding boxes. To compare the algorithms by a single score, the normalized area under curve (AuC) is computed for the first 150 and for the first 600 bounding box proposals. The bounding box proposals are sorted by their score.

Mean Area and Mean Normalized Distance

To evaluate the localization error, the mean area and the mean normalized distance are introduced (see Fig. 4.7). The mean area $\overline{\text{mA}}$ measures how well the size of the bounding box is estimated by the RPN within a test set. Only bounding boxes that are counted as true positives are considered for computing the metric

$$\overline{\text{mA}} = \langle \text{mA} \rangle = \langle \frac{\text{A}_{\text{prop}}}{\text{A}_{\text{gt}}} \rangle,$$

where A_{prop} is the area of a proposed bounding box, A_{gt} is the area of the ground truth box and $\langle \cdot \rangle$ denotes the average value. The optimum value is 1 when the proposed bounding box size matches the ground truth. Values



(a) (b)

Figure 4.7: Sketch of the (a) mA and (b) mD measure. The green rectangle depicts Ground Truth (*gt*) and the red one a proposed bounding box (*prop*).

smaller than 1 indicate that the box size is underestimated and larger than 1 indicate that the box size is overestimated in average.

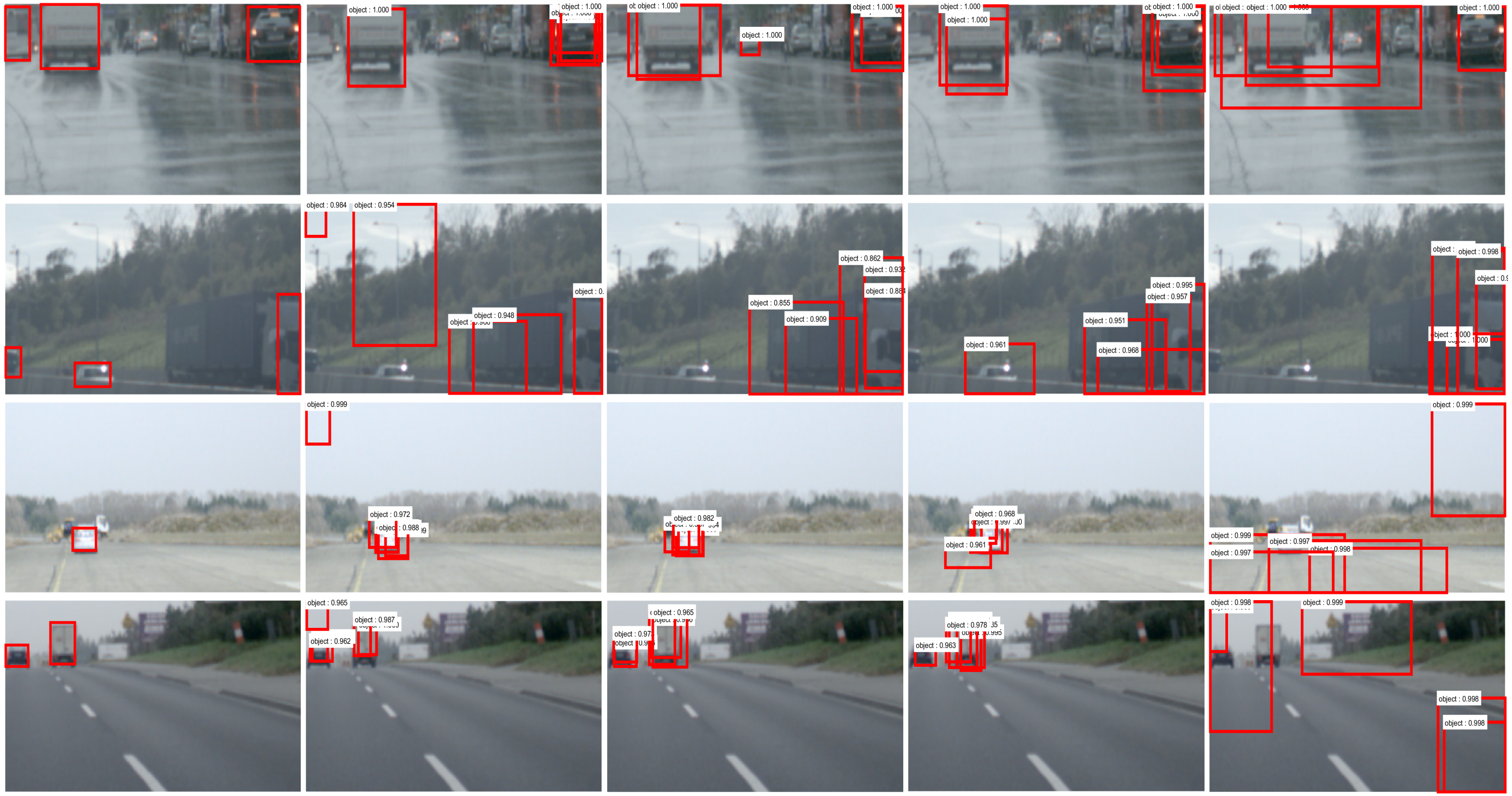
To evaluate the displacement of proposed boxes, the mean normalized distance $\overline{\mathrm{mD}}$ over a test set is introduced. The value mD is a measure to evaluate the localization error in x- or y-direction by comparing the centers of the ground truth box with the proposed box. It is normalized by the ground truth bounding box elongation. When the proposed box is centered exactly on the ground truth, the mD would be zero. We obtain the mean normalized distance

$$\overline{\mathrm{mD}} = \langle \mathrm{mD} \rangle = \langle \sqrt{\left(\frac{c_x^{\mathrm{prop}} - c_x^{\mathrm{gt}}}{\mathrm{w_{gt}}}\right)^2 + \left(\frac{c_y^{\mathrm{prop}} - c_y^{\mathrm{gt}}}{\mathrm{h_{gt}}}\right)^2} \rangle,$$

where $\mathbf{c}^{\mathrm{prop}}$ is the center of the proposed bounding box, and $\mathbf{c}^{\mathrm{gt}}$ the center of the ground truth bounding box. The parameter $\mathrm{w_{gt}}$ corresponds to the elongation of the ground truth box in y-direction while $\mathrm{h_{gt}}$ is the elongation of the ground truth box in x-direction.

4.4.4 Results

Figure 4.8 illustrates the results for four selected image crops.



(a) Ground Truth (b) No Prior (c) Voting Map (d) SR (e) VA

Figure 4.8: Visual assessment of the regions proposed: Ground truth (a), and the results provided by an RPN using no priors (b) and priors based on Voting Maps (c), Spectral Residual (d) and Visual Attention(e). The five boxes with the highest scores are shown only for illustrative purposes.

The five bounding box proposals with the highest scores are plotted for the RPN of the baseline algorithm and the RPN with prior information. The baseline algorithm tends to propose bounding boxes with high scores in background areas. When computing a saliency map using the Voting Map approach or the spectral residual approach and incorporating the saliency map as prior information in the RPN, false bounding box proposals can be suppressed. The saliency maps of these two approaches are sensitive to distant objects and the performance of the RPN can be improved, see Figure 4.4 for illustration of the saliency maps.
When computing the saliency map using the visual attention approach, broad regions are considered as salient and distant objects may be missed. Using such a saliency map as prior information for the RPN increases the size of the proposed bounding boxes and shifts the attention of the RPN to other areas (see Figure 4.8e). These findings are analyzed in the following in more detail using the metrics introduced in Section 5.5.3. Figure 4.9 and 4.11 show the recall curves for the first 10 to 600 proposed bounding boxes and for the IoUs of 0.25 and 0.5. Figure 4.10 and 4.12 zoom in the range of 500 to 600 proposed bounding boxes. The plots show that the recall increases fast until approximately 150 proposed bounding boxes and then slowly converges to its maximum value. For Faster R-CNN applications, the number of proposed bounding boxes that are forwarded to the classification branch is fixed. The larger the number, the higher the computational effort for the classification stage.

IoU of 0.25

In Figure 4.9, it can be seen that for an IoU of 0.25 the incorporated saliency prior with the Voting Maps and spectral residual method show for the first 150 proposed bounding boxes a higher recall than without any prior (a detailed view on the results is shown in Table 4.5). For this case, the best approach is the spectral residual prior, which results in an increase of over 2.0% for the first 150 proposed boxes. Figure 4.10 illustrates the convergence of the recall to its maximum value on the data set. For 600 selected bounding boxes, the recall for the RPN with Voting Maps or spectral residual as saliency prior is higher than for the RPN with no prior. Here, the Voting Map prior as well as the spectral residual increases the overall recall for the first 600 proposed boxes by more than 1.2%, while the visual attention map shows worse results than with no

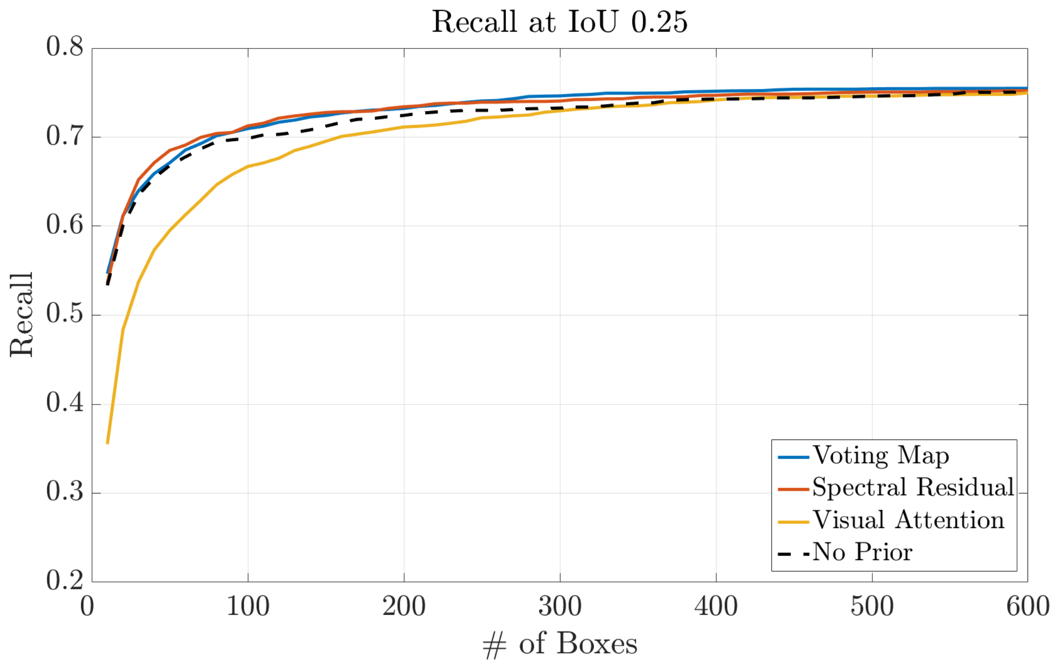


Figure 4.9: The mean recall for IoU $= 0.25$ over the number of proposed boxes with highest scores is shown

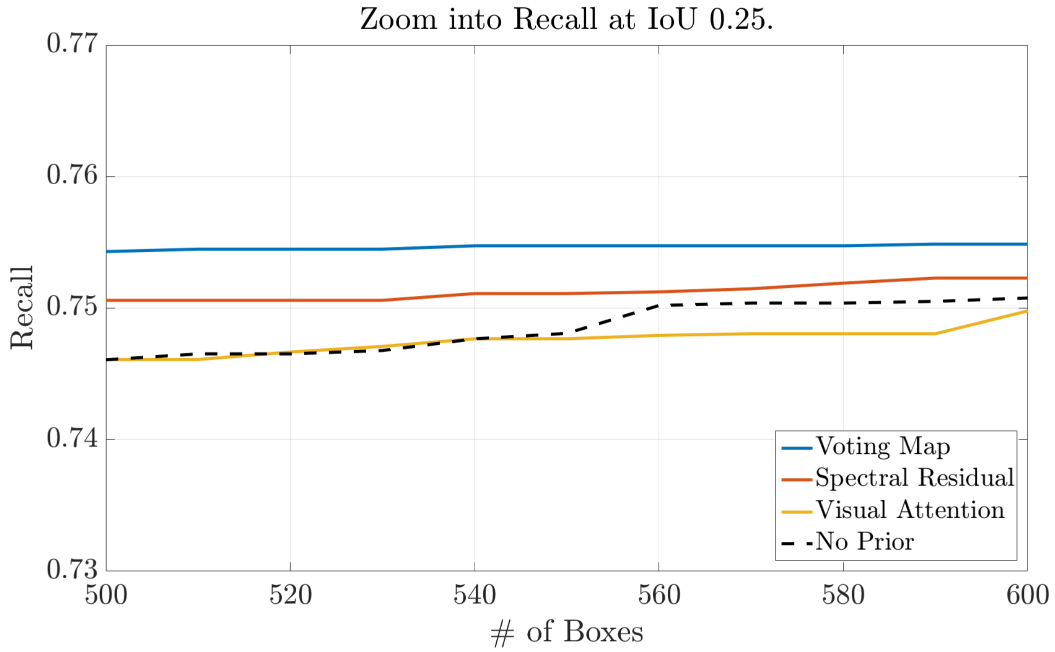


Figure 4.10: A zoom of the recall for the first 500 to 600 boxes is shown for IoU $= 0.25$.

prior. For a more compact comparison, Table 4.4 summarizes the AuC for the different approaches.

The RPN including the Voting Map approach achieves higher performance than the RPN without prior information. The AuC measures the abso-

Nr. of Boxes	150		600	
Global Prior	IoU 0.25	IoU 0.5	IoU 0.25	IoU 0.5
None	0.681	**0.545**	**0.731**	**0.619**
Spectral Residual	**0.686**	0.528	0.730	0.601
No Prior	0.672	0.540	0.722	0.612
Visual Attention	0.612	0.438	0.703	0.575

Table 4.4: Area under the curve (AuC) for an IoU $= 0.25$ and IoU $= 0.5$. Each AuC is calculated for the first 150 and 600 proposed boxes.

lute performance of the methods while Table 4.5 shows the relative performance of the methods to the baseline RPN without prior information.

Nr. of Boxes	150		600	
Global Prior	IoU 0.25	IoU 0.5	IoU 0.25	IoU 0.5
None	0	0	0	0
Voting Map	1.401	**1.121**	**1.279**	**1.145**
Spectral Residual	**2.039**	-1.994	1.089	-1.740
Visual Attention	-8.955	-18.724	-2.531	-6.059

Table 4.5: Percentual mean difference in recall of the methods compared to the method with no prior for the first 150 and 600 proposed boxes. A positive value shows a better performance of prior-based method over the prior-less method.

IoU of 0.5

When choosing a stronger criterion for matching the detection box with the ground truth box, here using an IoU of 0.5, the recall values are smaller. Proposed bounding boxes that overestimate the object size or include only object parts are not included in the evaluation. The RPN using the Voting Map as prior information shows an increase over 1.1% of recall for the first 150 proposed boxes compared to the RPN without prior information. For the first 600 proposed boxes, the best saliency prior is the Voting Map approach with an overall increase over 1.1%.

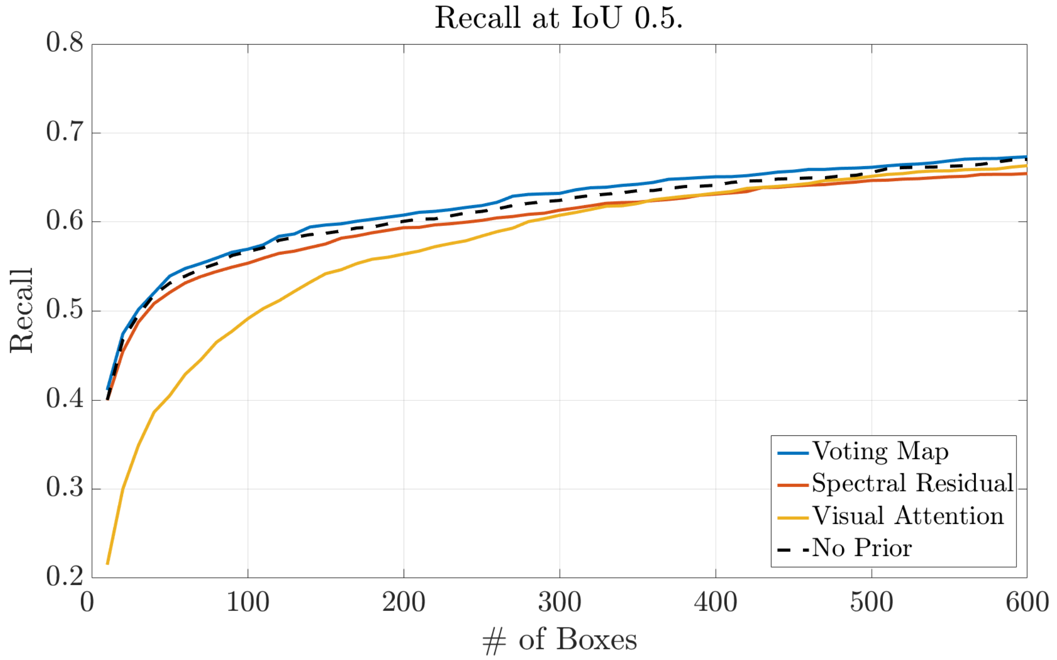


Figure 4.11: The mean recall for IoU $= 0.5$ over the number of proposed boxes with highest scores is shown.

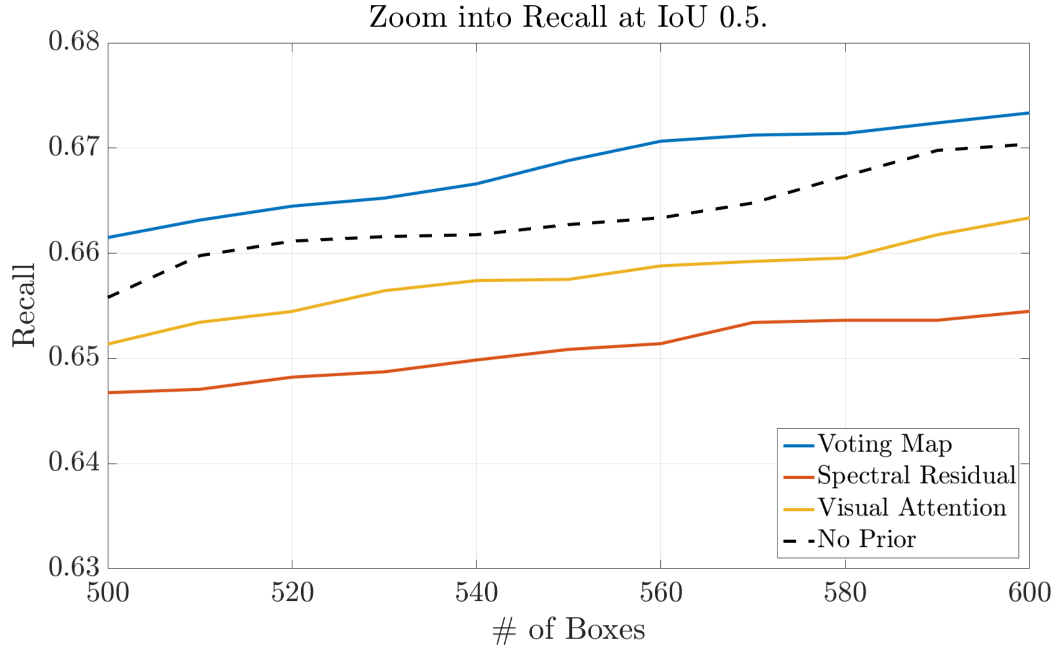


Figure 4.12: A zoom of the recall for the first 500 to 600 boxes is shown for IoU $= 0.5$.

Mean normalized Distance and mean normalized Area

Table 4.6 summarizes the mean normalized area for the four approaches. For the IoU of 0.5, the size of the proposed bounding box matches approximately the size of the proposed bounding box. When a lower IoU threshold is chosen, it can be seen that the RPN with and without prior information overestimates the object size.

Global Prior	$\mathrm{mA}_{\mathrm{IoU}\,=\,0.25}$	$\mathrm{mA}_{\mathrm{IoU}\,=\,0.5}$
None	**1.45**	**1.099**
Voting Map	1.46	1.107
Spectral Residual	1.46	1.114
Visual Attention	1.58	1.108

Table 4.6: Mean normalized area (mA) of the boxes accepted by an IoU of 0.25 and 0.5. Mean areas larger than 1 indicate an over-estimation of the box sizes.

Figure 4.13 illustrates the mean normalized distance over an increasing number of proposed bounding boxes for each method. For a weak match criterion such as IoU of 0.25, the displacement of the proposed bounding box is strongly influenced over the number of proposed boxes, while for a strong criterion of IoU 0.5 the displacement stays approximately the same. It can be seen that the mean displacement of the proposed boxes is growing with increasing number of proposed boxes. Hence, proposed bounding boxes that match the object well in size and position usually have a high score and are often included already in the first 150 bounding boxes.

The mean number of boxes per object is summarized in Table 4.7. The RPN proposes in average 7 matches for a ground truth box when the IoU is 0.25. These numbers confirm the results in Figure 4.13 and a trend to including more false-positives proposals when increasing the number of selected bounding boxes can be observed.

Global Prior	IoU 0.25	IoU 0.5
Voting Map [5]	**7.7**	3.6
Spectral Residual [57]	7.6	**3.7**
None	7.3	3.5
Visual Attention [64]	6.4	3.1

Table 4.7: Mean number of boxes matching an object, computed for 600 bounding boxes for different matching criteria, i.e., IoU $= 0.25$ and IoU $= 0.5$. The larger the number of matches is, the more boxes cover an object.

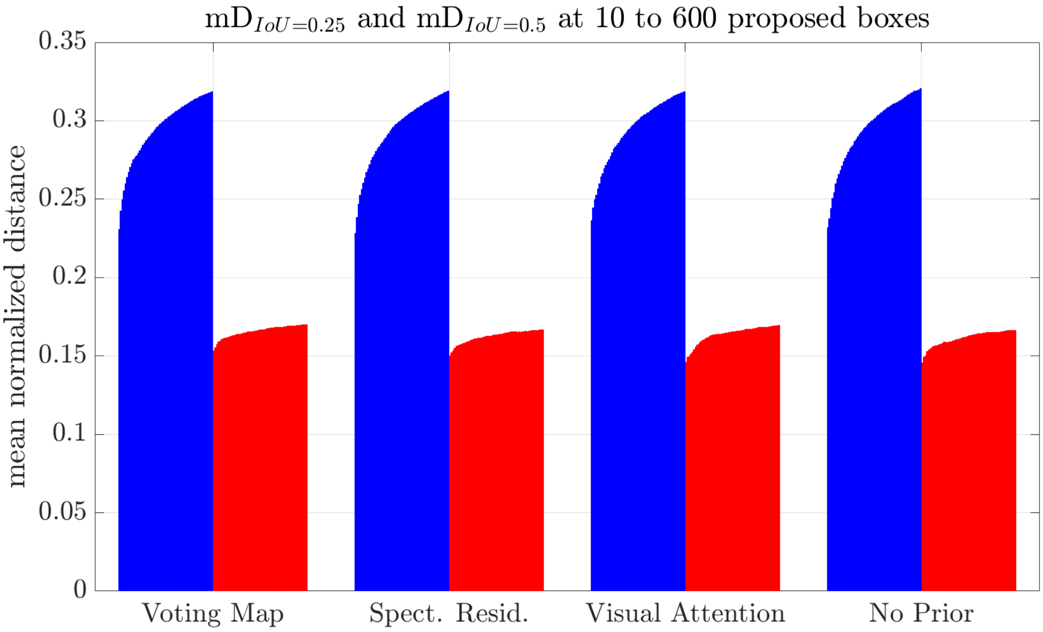


Figure 4.13: Mean normalized distance $\mathrm{mD}_{\mathrm{IoU}}$ of the proposed boxes at IoU of 0.25 in blue and 0.5 in red. Each bar shows the development of the mean distance over the first 10 to 600 boxes.

Execution Time

In this Section the execution time of the approaches to incorporating saliency-guided priors into a RPN is evaluated. All run-time experiments are conducted on an Intel Core i7-4800Mq at 2.70 GHz and a NVIDIA Quadro K1100M graphic card. The RPN is optimized for CUDA and executed in *Matlab*. The computation of the prior maps is performed in *Matlab* with no hardware-optimized code. Table 4.8 illustrates the execution time for computing the global prior, and the execution times for a conventional RPN and the RPNs incorporating the different prior maps. Here, it is desired to distinguish between the computation of the prior maps and the computation of the region proposals to study the execution times individually. Table 4.8 shows that executing the computation of the additional prior based on Spectral Residuals increases the runtime by just 2%, whereas the computation of Visual Attention within a prior-based RPN increases the overall execution time of up to 44%. In addition, the consideration of the prior in the region proposal network increases the overall execution time of up to 3ms only, indicating the high efficiency of the architecture chosen.

Finally, it is interesting to evaluate the computation time of the method

Global Prior	Prior Calc.	RPN + Prior	% Increase
None	-	0.2837	+0
Spectral Residual	0.0044	0.2855	+2.0
Voting Map	0.0920	0.2865	+33.4
Visual Attention	0.1195	0.2856	+42.8

Table 4.8: Execution time [sec] of a prior-based RPN for 600 proposed boxes using images of 291×161 px. The right column shows the percentual increase of the execution time of a prior-based RPNs compared to a RPN with no prior.

proposed by Shrivastava *et al.* [115] against our method as it shares the same idea of incorporating an additional feature map to the network. The approach of [115] segments the feature maps of the last shared convolution layer (conv5), and the segmented maps are added to the RPN as an additional input. Here, according to the used Parse Net [84] relying on the VGG16-Net, the segmentation based on three convolutional layers and four fully-connected layers need to be computed to obtain the final segmentation. In addition, the classification branch of the Faster R-CNN consisting of two fully-connected layers takes approximately twice the time of the RPN for 600 proposed boxes. Hence, for the work of [115], it is possible to estimate an execution time that is about 200% larger than the execution time of our approach due to expensive segmentation and computation of four fully-connected layers.

4.4.5 Discussion

The proposed architecture with the incorporation of a global prior shows improvements in the performance of the RPN based on the chosen evaluation metrics. The incorporation of the global prior at the last shared convolutional layer can reduce the number of region proposals that need to be evaluated by a classification branch. In this approach, the region proposal network incorporating the Voting Map as a global prior gained the best overall performance. This may be due to the fact that the Voting Map is tailored to the task of drawing attention to small objects in automotive scenarios.

In order to understand the lower performance when incorporating other priors, it is beneficial to take a look at the feature maps provided by the last convolutional layer. Studies show that interesting areas in the feature maps are represented as blobs and soft corners as extracted by early convolutional layers focusing on edges and corners within an image [136]. As such, the global priors relying on Voting Maps and Spectral Residual share similar properties with the maps provided by the last convolutional layer, and benefit the region proposal branch in identifying interesting regions in the network. In addition, prior maps relying on Visual Attention do not show consistent results and may, therefore, not be appropriate for the use of localizing small objects in the automotive context.

With the incorporation of the global prior at the last convolutional layer, non-linearities within the global priors can be taken into account within the following convolutional layer of the region proposal network. However one might argument that more layers will increase the non-linearity and therefore increase a possible representation of the global saliency features. This increases the chances of overfitting and due to the considerably small data set available gives no meaningful representation.

Finally, it is shown that incorporating the global prior in the chosen architecture increases complexity and the computational costs of a RPN only marginally, while increasing the performance of the underlying RPN. It is expected that the computation of the priors can be optimized further to decrease their execution time. Since computing the priors and performing the convolutions are data-independent, the execution time can be decreased further by parallelizing the two calculations.

4.5 Summary and Conclusion

This chapter presents an approach to incorporating a global prior into a locally restricted region proposal network to better guide the proposal of regions towards interesting regions containing objects.

The use of saliency-inspired methods for computing the global prior in this network increases the overall performance of the RPN and improves the recall of proposed regions. It is found that certain characteristics within the different global saliency maps contribute better to the localization of object region than others. Since the priors are computed in an unsupervised fashion, no additional labeling is required. Moreover, the simple architecture is very efficient in terms of runtime and allows to use pre-trained networks. In addition, the technique can also be applied to analysis tasks in medical image analysis [19, 20].

The detailed metric for the special use-case of small objects allows a more precise evaluation of which tasks within the region proposal networks are still subject to further improvement. It is shown that especially for small objects the proposed regions of the RPN are slightly too large while the localization of the proposed regions shows best performance within the very first proposed boxes. This finding shows that the regression head of the RPN is not sufficiently adjusting the regions size. In the following chapter this finding is also subject to closer investigation and possibilities to improve its performance.

Chapter 5

Relevance Based Net-Surgery

This chapter presents a novel method to determine relevant features inside a convolutional network, see Figure 5.1. Further, it is demonstrated how possible additional features could further improve the networks detection performance. For these purposes, the developed feature maps are analyzed with a newly proposed *post-trained net surgery*, which allows to exclude feature maps during execution and to evaluate the relevance of feature maps. The main features for different object sizes are determined and the performance of the network is analyzed. A data set with high emphasis on small objects down to 8 px width for the need to train an entire Region Proposal network from scratch was created. In comparison to the used data sets in the previous chapters, it exhibits higher challenges as the fraction of objects smaller than 10 px is larger. Excerpts have already been published by the author in [36].

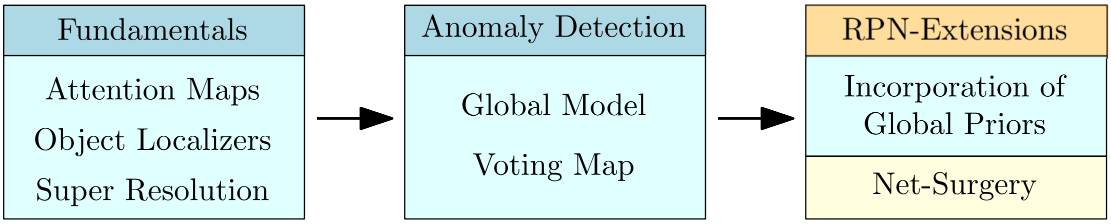


Figure 5.1: In this Chapter, the relevance based net-surgery is described, which allows to determine the most relevant features within an region proposal network.

5.1 Motivation

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.

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.

As explained in detail in Chapter 2 different network architectures such as in [42, 83, 106, 107, 114] make use of features to propose object regions. In [77], a first concept for bounding box regression using features solely from the CNN was proposed. The detection of small objects in low-resolution images, however, remains a challenge for all approaches presented. Li *et al.* [80] use feature maps 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 a 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.* [58] evaluate different CNN architectures and show the highest accuracy of Faster R-CNN [107] over a variety of convolutional object localizers. Faster R-CNN [107] 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, is 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.* [58] 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.* [139] 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 [71] and the Feature Pyramid Network [81] use feature maps at higher resolution with more low level features to detect small objects. However, in the case of low resolution images, using different feature maps is difficult as object regions provide weak local features. B. Cheng *et al.* in [16] investigated the multi-task problem of the regression and classification head within the RPN. They found that the common feature representation for the different tasks is showing less performance than a decoupled feature representation, mainly due to the different object sizes present within one picture. However, their used shared network is computationally several times more expensive than the desired small networks for automotive hardware constraints and therefore their findings are hardly comparable with the here desired application.

5.2 Outline

In this chapter the role of different features for the different tasks within the Faster R-CNN [107] is investigated. The results can be used to improve the overall network performance with focus on very small objects in low-resolution for driver assistance systems. As shown in Figure 5.2, vehicles in 200m 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.

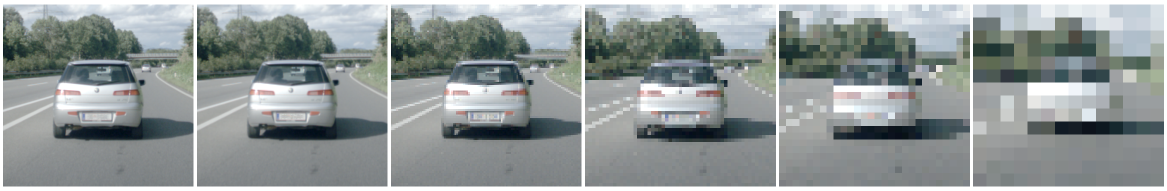


Figure 5.2: The feature quality depends highly on the image resolution (from left to right: 443×421px, 222×211px, 111×106px, 55×53px, 28×27px, 14×14px)[1]. For this reason, detection of low resolution objects is a challenging task.

Furthermore, the impact of the feature maps and priors on bounding box regression and binary object classification for very small objects is studied. At first the importance of the feature maps by clustering them based on their correlation to the task specific ground truth data is evaluated. 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, different external priors are incorporated into the RPN chain before bounding box regression and binary classification and their contributions to the overall performance of the RPN is studied. In difference to the previous Chapter 4, the prior is incorporated after the RPN specific convolutional layer and therefore the impact of the evolved features can be studied directly.

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 object detection in low-resolution images. The saliency map showing best feature representation is taken as input for the RPN only and the performance on a single map input is computed. Finally, evaluations demonstrate the need for incorporating priors into the network architecture to increase the recall for small object detection significantly.

---

[1] For the resolution in (a) and (b) the license plate of the shown vehicle is blurred due to personal data protection

5.3 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.

Region Proposal Network (RPN)

As described in Section 4.3.1 in this contribution, the Faster R-CNN based on the ZF-Net is used as base as well, which includes a region proposal Network (RPN) and a binary classification network (BCN). An RPN proposes bounding boxes for the subsequent binary classification network. 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 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 5.1 summarizes the adapted parameters for small object detection.

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**

Table 5.1: 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.

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.

5.4 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.

5.4.1 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.* [136] 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. It is of interest how many feature maps include class-relevant information, redundancy of the class-relevant information, and whether or not the information content is depended on object attributes such as size. For this reason, post-trained, relevance-based clustering of feature maps is introduced, where only a subset of feature maps is used during inference.

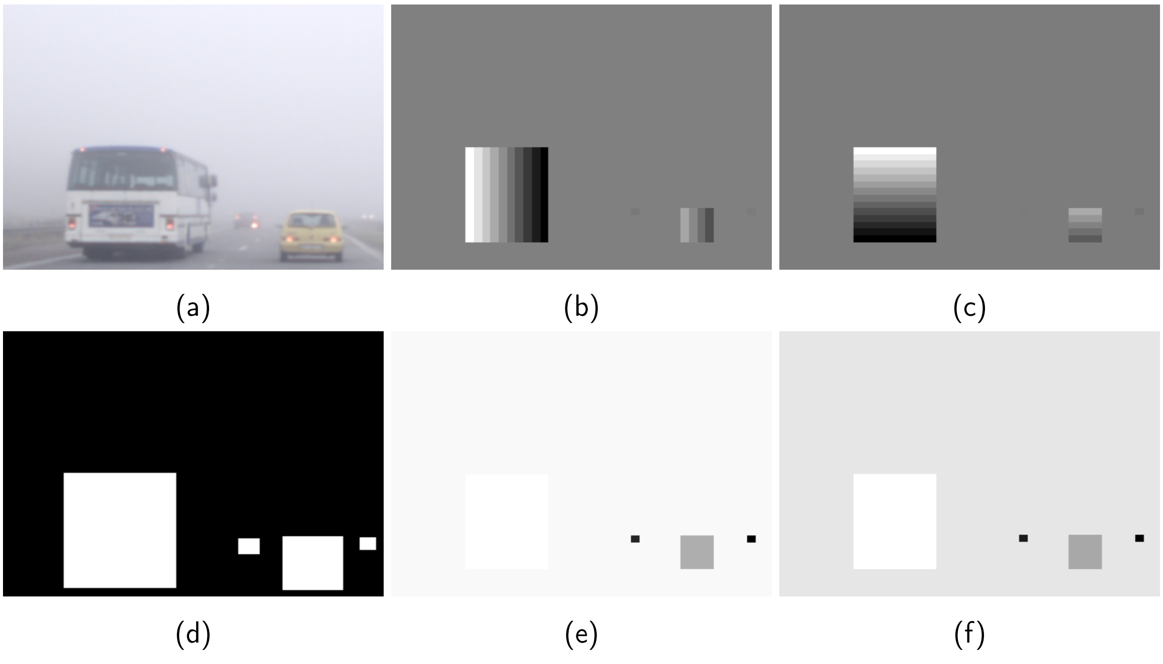

(a) (b) (c)
(d) (e) (f)

Figure 5.3: 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

Clustering of RPN Feature Maps

Given that several feature maps evolve similar features [136], 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 5.3.

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 $\mathrm{fm}_{i,j}$ and ground truth BB-rect$_i$ as shown in Figure 5.4.

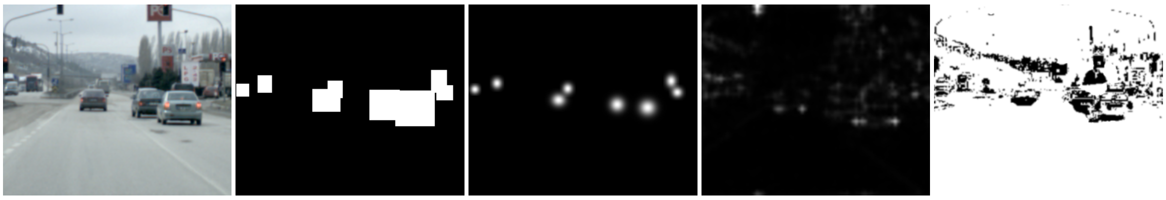

(a) Input (b) BB-rect (c) BB-rect.-gauss (d) SR (e) Voting Map

Figure 5.4: (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 [57] and (e) the Voting Map [5].

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. The pseudo code in Algo.1 summarizes the steps for feature map clustering procedure.

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

Algorithm 1 Pseudo code of the feature clustering

```
procedure CLUSTERING($\mathrm{fm}_{i,j}, \text{BB-rect}_i$)
    loop Over all test-images $i$
        loop Over all $\mathrm{fm}_{i,j}$
            $V_{i,j} = abs|\mathrm{Corr}\,[\mathrm{fm}_{i,j}, \text{BB-rect}_i]|$
        $S_i = sort \downarrow (V_i)$
        divide $S_i$ in $Q$ subsequent equal long vectors
        $S{*}_i = [S_{i,1}, ...S_{i,q}, ...S_{i,Q}]$
    $a_{i,j} = z_q$ if $V_{i,j} \in S *_{i,q}$ and $z_q \in [1, .., q, .., Q]$
    $\mathrm{cl}_j$ = most frequent $z_q$ in $a_{i,j} \forall i$
    return $\mathrm{cl}_j$
```

clusters 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 5.5 visualizes the net surgery.

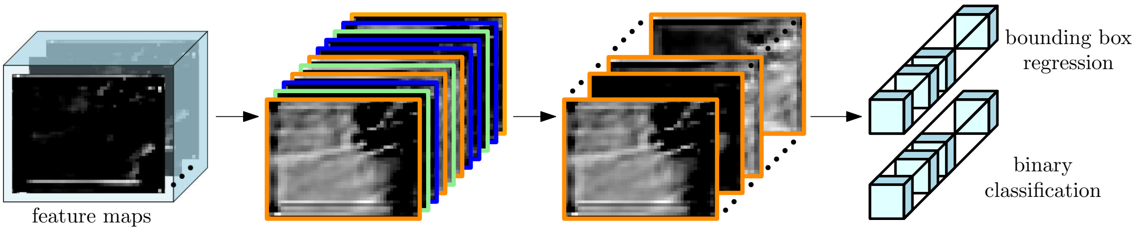


Figure 5.5: The feature maps of the last layer of the RPN are clustered using the similarity between the feature map activation and ground truth as cluster criterion. We analyzed the features maps within each cluster regarding: their contribution to the bounding box regression and binary classification.

5.4.2 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 5.6 shows three different ways of incorporating of prior information during feed-forward and training phase into the last layer of the region proposal network.

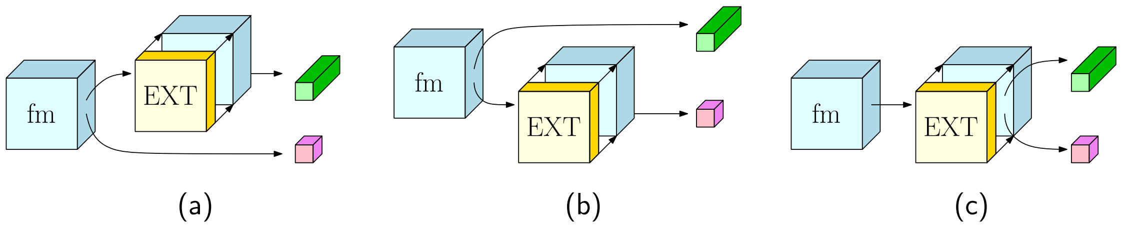


Figure 5.6: Prior information is added as external data (EXT) to the output of the last convolutional layer (fm) of the RPN for either (a) bounding box regression, (b) binary classification, or (c) both bounding box regression and binary classification. The blue boxes represent data from the feature map, the yellow boxes external data and green and pink boxes the 1×1 kernels for regression (green) or classification (pink).

Motivated by improving the performance of the RPN by incorporating prior feature maps as in [35], 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 5.6a), prior to classification (Figure 5.6b), and prior to both classification and bounding box regression (Figure 5.6c). 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 5.6b, are created by setting all pixel values inside an object region to one and zero otherwise (see Figure 5.4b). 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 5.6a follows another pattern 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 5.3 (c)-(f) [43]. The ground truth data $t_{x,y,w,h}$ for the regression is then given as following

$$\begin{aligned} t_x &= (x - x_a)/w_a & t_w &= log(w/w_a) \\ t_y &= (y - y_a)/h_a & t_h &= log(h/h_a) \end{aligned}$$

"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 an point of origin laying on the center of an object region, the parameters can be defined to $d = 2$, $\mu = 0$ and

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

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 real-time are Spectral Residual [57] and a Voting Scheme [5]. 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.

5.5 Metric and Results

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.

5.5.1 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 5.2. Objects that are occluded by less than 50% are included in the data set. The image size is 1024×640 px. Table 5.2 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[2]. 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.

---

[2] In Section 4.4.1 data augmentation is explained further

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

Table 5.2: 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.

5.5.2 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 architecture for the used net is described in Section 4.4.2. 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.* [32] 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 R-CNN was trained as proposed in [107] with original parameter set-up as mentioned in Section 4.4.2. The trained weights of this base model are here 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 batch size for the RPN is reduced drastically to 20 to generate a balanced foreground/background set of possible anchors during training (Table 5.3).

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

Table 5.3: Details of the training parameters to train the RPN and classification head of the Faster R-CNN for a data set with small objects

5.5.3 Evaluation Metrics

The performance is evaluated using the recall metric based on the Intersection over Unit (IoU = 0.5) as described earlier in Section 4.4.3. 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.

5.5.4 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 5.4b) is used. The clustering uses the absolute correlation values, and 3, 5 and 10 clusters are formed. In Figure 5.7, 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. In Figure 5.8, an example of the strongest feature maps within one cluster and

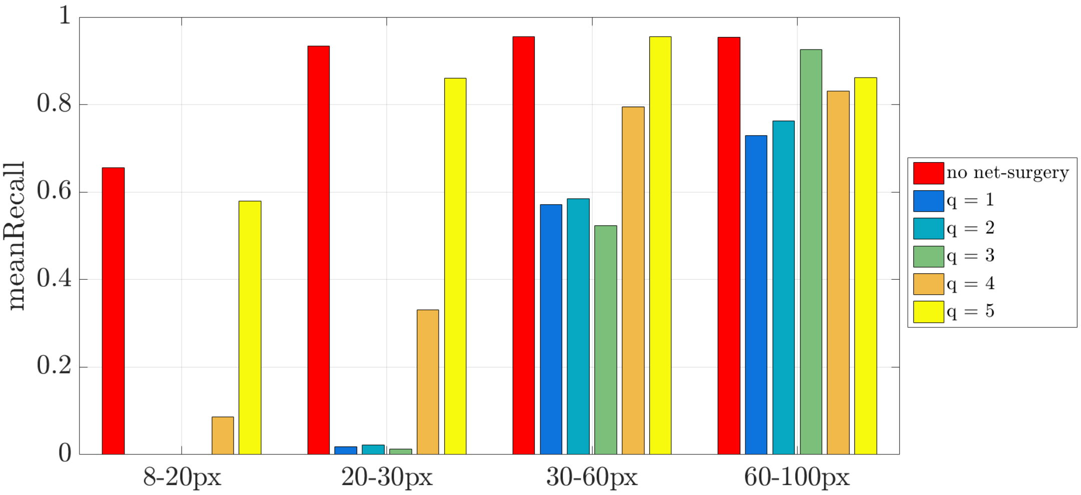


Figure 5.7: The recall for cluster-based net surgery for 600 boxes and 5 different clusters is shown. For each cluster q, the feature maps with highest range of absolute correlation to BB-rectangular are activated. The red bar corresponds to an architecture without net-surgery, while each bar to the left represents a different activated cluster with increasing correlation. We can see that different features impact the performance for detecting differently sized objects.

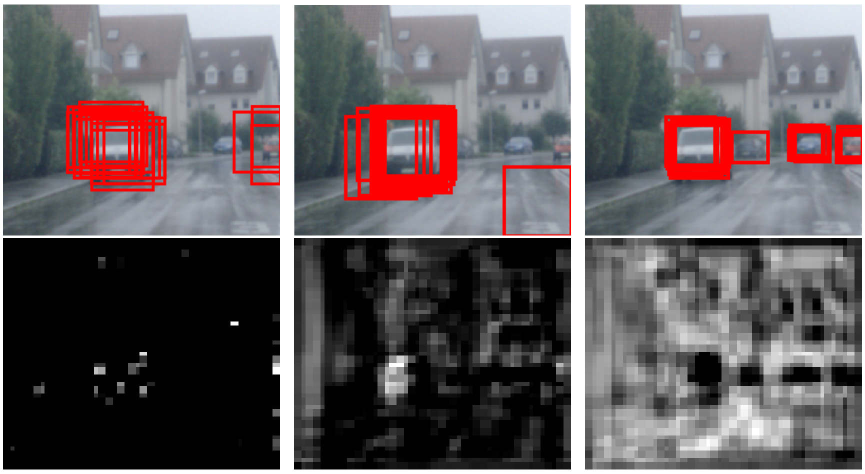
its corresponding region proposals is shown. The strongest feature map is determined by highest weight corresponding to the feature map within one cluster.

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 5.4 it is shown that the branch size can be reduced by e.g. 66%

Q	Δq_{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%

Table 5.4: The table shows the potential of branch size reduction Δbranch-size and reduction of recall Δq_{max} for the different object sizes and different number of clusters Q. The baseline is the model without any net surgery.



(a) 0-20% (b) 40-60% (c) 80-100%

Figure 5.8: In the top row the first 20 proposals for each cluster to which the featuremaps displayed in the second row belong. The second row shows the feature maps with the strongest weights in each cluster. The cluster belonging to the feature maps has the top x% of correlation score as indicated in the caption. It can be seen that the features for small and larger objects differ. For small objects a higher correlation score to BB-rect data is more useful.

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 [94].

5.5.5 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 5.6, the different architectures are shown.

Comparison of only Regression or Classification Branch

Figure 5.9 shows the recall for different external data composed with ground truth knowledge in the regression branch only (architecture as

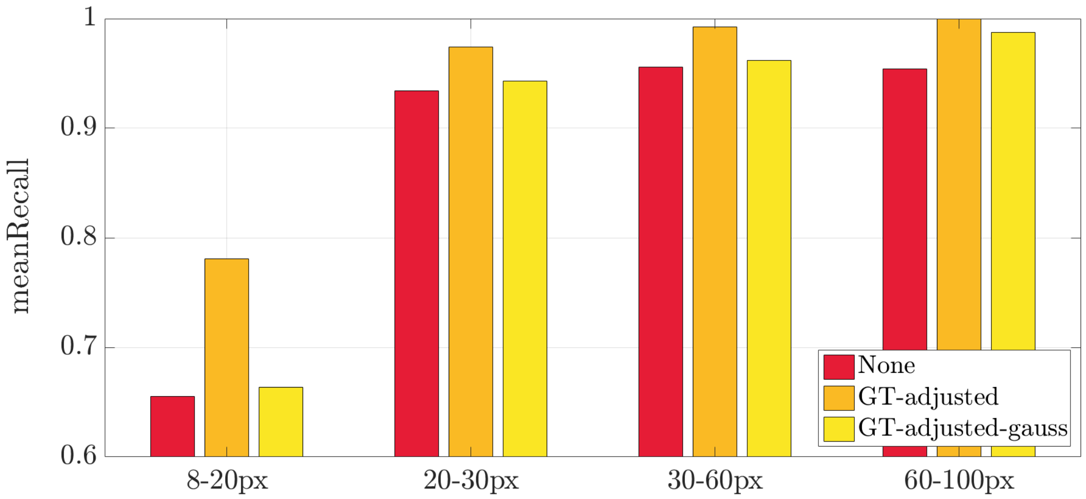


Figure 5.9: Mean recall with ground truth external data only for the bounding box regression branch of the RPN. GT-adjusted includes all ground truth data regression (4 maps). BB-rect consists of binary classification ground truth only (1 map). The flag "gauss" denotes a degradation of the input data as described in Section 5.4.2.

in Figure 5.6a). 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 5.10, 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.

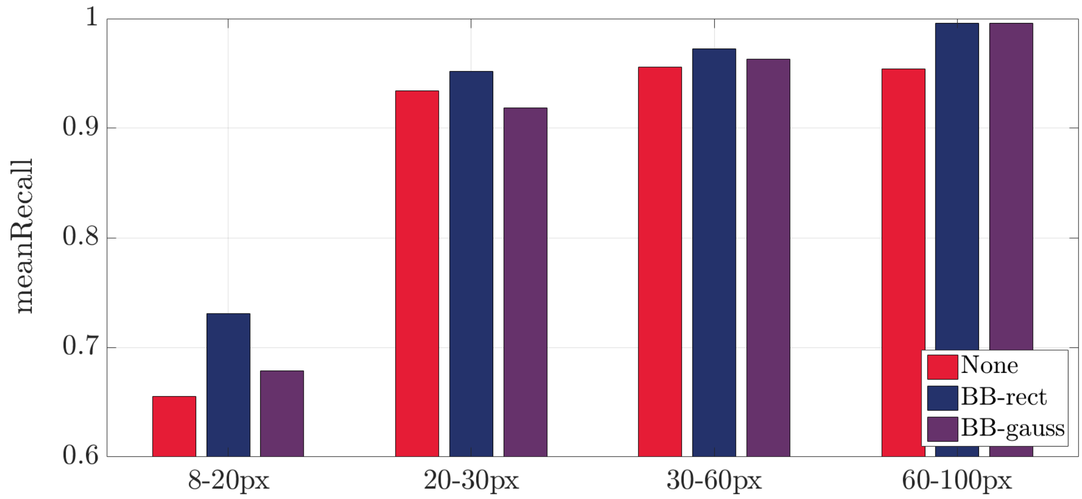


Figure 5.10: Mean recall with BB-rect external data only for the binary classification branch of the RPN. GT-adjusted includes all ground truth data regression (4 maps). BB-rect consists of binary classification ground truth only (1 map). The flag "gauss" denotes a degradation of the input data as described in Section 5.4.2.

Bounding Box Regression and Classification Branch

In Figure 5.11, 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-adjusted), 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 increase the recall for larger objects up to 30 px.

Figure 5.12 visualizes the recall in detail for objects of a width between

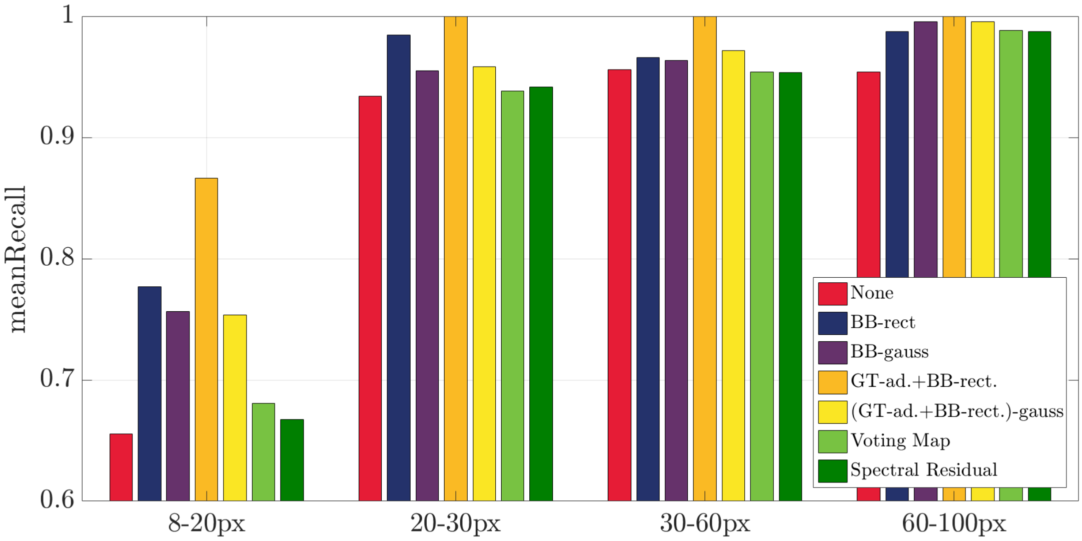


Figure 5.11: The recall for low resolution objects is significantly lower than for objects with a minimum size of 20 px. Adding prior maps as input to bounding box regression and binary classification improves the recall and the localization of low resolution objects dramatically, where ground truth prior maps estimate upper bounds. GT-adjusted + BB-rect includes all ground truth data for binary classification and regression (5 maps). BB-rect consists of binary classification ground truth only (1 map). The flag "gauss" denotes a degradation of the input data as described in Section 5.4.2.

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. In Figure 5.13, the statistics of the test set, which is used for evaluation, is shown. Hence, it becomes visible how challenging the used data set is due to the high occurrence of objects with widths between 8-25px.

Especially in this challenging region, saliency maps as external data improve the recall. The reason 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.

Only-Saliency-Map Input

Following the finding that especially in the challenging region of small objects, saliency maps improve the recall when incorporated as external data, the RPN is trained and tested with a saliency map as input only. Both regression and binary classification head, are optimized during train-

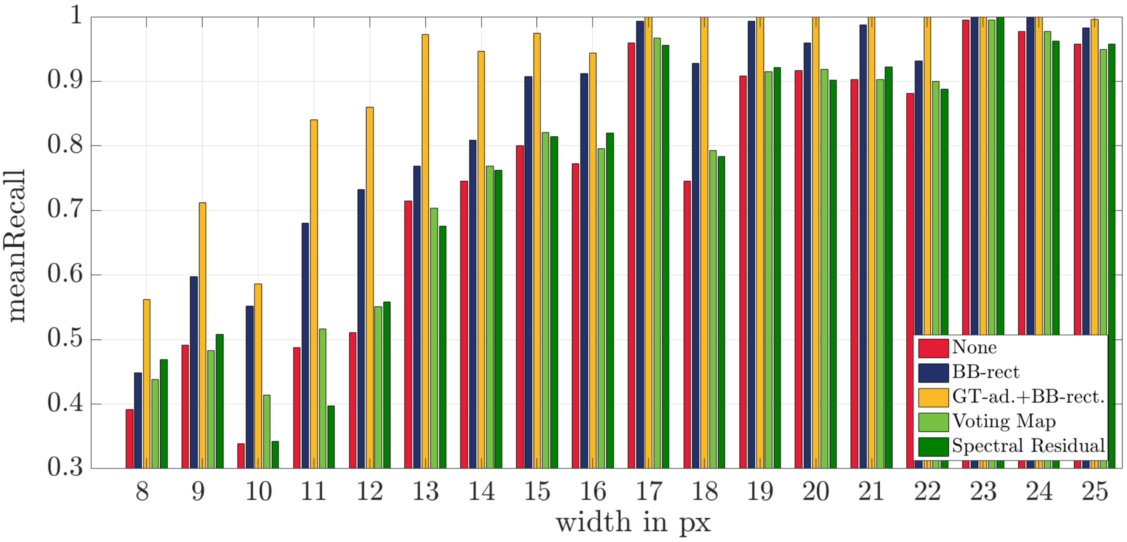


Figure 5.12: Detailed visualization of the mean recall for small objects of sizes 8-25 px width.

ing and evaluated jointly during testing. Hence, the RPN is composed of one convolutional layer with one saliency map as input. As such the variety of possible features is reduced to force the network to optimize only on the five most important feature maps, as displayed ideally in Figure 5.3. This reduction is chosen only for this experiment as no additional features from other maps need to be taken into account. Exemplary the Spectral Residual map is used as saliency map as it showed promising results especially for small objects in Figure 5.12. Figure 5.14 shows the recall for a network with only the Spectral Residual map as input and five evolved feature maps within the RPN. It can be seen that the recall increases over increasing object sizes until objects of width 60 px. For larger objects, the recall decreases again. A closer look on the computationally costs of this system, shows that with fairly small computational costs, it is possible to obtain a high recall. Instead of the previously used 256 feature maps, only 5 feature maps are used. The input is reduced from 256 to 1 dimension. Since all other parameters remain the same, the computation cost for the neural network is only 0.0076% of the network using all feature maps as input and output. As it was already shown in Section 4.4.4, the additional costs to calculate the Spectral Residual compared to an execution of the RPN are summing up to 2%. Hence, the total computational costs for this set up is only 2.0076%. Simultaneously, the recall decreases for objects of width 30-60 px by only 6%. Within this experiment, the difference of desired features for each object size becomes more clear: for objects of

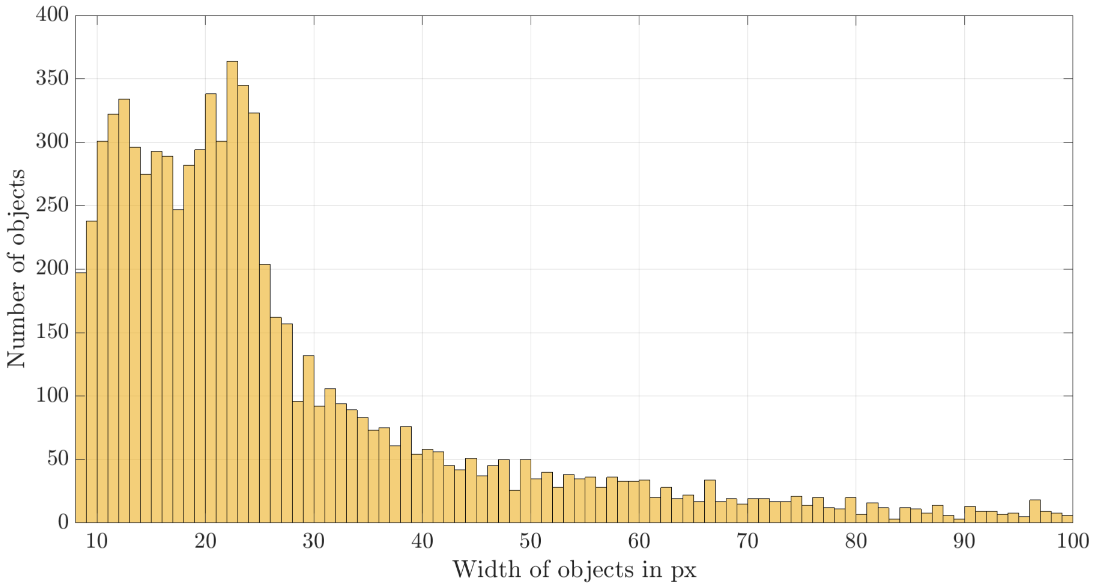


Figure 5.13: Number of objects within the evaluation data set per object size.

sizes 30-60 px the recall reaches 90.9 % while it decreases drastically for smaller and larger objects (see Figure 5.14). Depending on the application for middle sized objects, this finding shows promising results. For the detection of small objects, the use of only saliency maps as input is not recommended. The combination of extracted features from the CNN and the saliency map increases the overall performance for small objects.

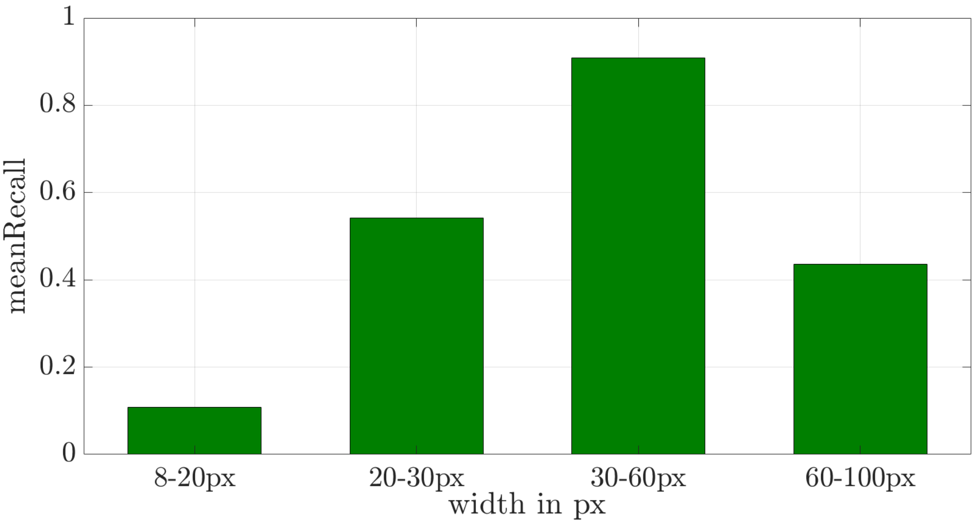


Figure 5.14: Visualization of the mean recall for small objects of sizes 8-100 px width. Only one saliency map is used as input to the RPN and only five feature maps are formed during training. Here, the Spectral Residual map is chosen.

5.6 Summary and Conclusion

This work studies the final feature maps of a region proposal network 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. It is possible to identify the most prominent features useful for either regression or binary classification and to evaluate the impact of a network pruning of irrelevant features.

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.

4) For the special use-case of small objects, only a combination of external data and CNN extracted features improves the performance of a RPN. Using one saliency map as feature map for the RPN input only gives the chance to reduce the computational costs drastically, by narrowing down good performances on limited object sizes only. Depending on the application, such behavior is possible or even desired and allows to reduce hardware requirements. For very small objects however, only a combination with CNN extracted features showed good results.

Finally, this chapter showed that the feature representation for the different tasks within the RPN demands different characteristics. Increasing the feature quality only for one of the tasks shows only a slight increase of performance, while a combination increases the performance non-linearly. Hence, to increase the region detection for small objects further, a jointly improvement of all extracted and external features needs to be considered.

Chapter 6

Conclusion and Outlook

6.1 Conclusion

In this thesis, different computer vision approaches have been newly developed, evaluated, and extended for the localization of object regions in images for the special case of distant and/or small objects in an automotive environment. Object region localization faces several challenges for small or distant objects with an automotive camera, as the resolution and imaging properties are limited. Hence, only few information per object region is present and localizers based on mainly local features comprise reduced capabilities.

In contrast to state-of-the-art localizers, the proposed Voting Map based object region detection emphasizes global distinctiveness of object regions. The effectiveness of the Voting Maps is demonstrated for small object region detection on motorway scenes using color channels. Other features such as orientation and local distinctiveness are also applicable. The generation of object regions based on the Voting Map is highly adaptable to illumination changes, even rapid changes impose no significant decrease of performance as the method allows for single-frame evaluation. Environmental changes such as rain or fog do not affect the localization since global distinctive features are used for determining object regions. The developed method can be easily accustomed for different camera set-ups without the need of an extensive parameter search or optimization. Additionally, the proposed method supports the reduction of data for subsequent information retrieval methods, i.e. high resolution images can be reduced in data size to the most interesting regions. Such reduction of data size gives the possibility to decrease the data flow within the vehicle system. A decreased data flow is improving the safety of a system due to lower loads within the cables and CPUs.

Convolutional network based region proposal methods for the use in vehicles suffer from mainly locally learned features, which fail to map global distinctiveness of small object regions. The incorporation of global priors, allows to use small locally restricted networks than large and computationally expensive architectures. The global priors are designed in such a way that the calculation is small e.g. Spectral Residual map, which is computed by two Fourier Transforms. Such computation can be implemented efficiently with hardware acceleration. The global priors are calculated without a-prior knowledge and do not impose additional expensive data labeling for further parameter optimization. All possible optimization parameters have been evaluated and adjusted to small object detection. Furthermore, a metric designed to evaluate small object region proposals shows the following insights: the proposed regions are generally too large, while the localization demonstrates robust results. This finding implies that the object region is generally detected, whereas the regression to the final object region shows limitations.

Inspired by this finding, a method for further detailed investigation has been developed: the post-trained net-surgery. It allows to evaluate the importance and significance of features within a region proposal network. Main contributing features can be determined, which enables to reduce the network size significantly by maintaining the performance. Different incorporated priors are evaluated based on these findings, which enhance the performance further. The proposed method is suitable for any applications, where a variety of features is used such as e.g. speech recognition or big data statistics. Specifically, for autonomous cameras with low computational power, such network reduction method is of high interest.

To reduce the computational costs within the system even further, a saliency map, which comprises suitable features for the detection of small object regions, is used as input for the region proposal network. With only a minimal computation of five convolutional filters of size 3×3×1 a region proposal network comprises a high recall performance for a defined range of object sizes. This finding shows that a relevance based feature selection allows to reduce computational costs drastically by focusing on limited application ranges.

The performance of the region proposal network was improved through an extensive evaluation of the parameters during the training phase. The main performance gain was reached by the adjustment of input scaling and batchsize reduction. Both parameter adjustments improved the mean recall by more than 10% each.

To evaluate and optimize all parameters sufficiently, an automotive data set was created, which includes object regions of minimal 8 px width. The data set contains 7960 objects with a width of maximal 100 px. Small objects of 8-30 px width make 67% of the whole data set. Only very few data sets address such small objects, e.g. the KITTI data set contains only labeled vehicles down to a minimal size of 25 px height. The data set includes scenes for different weather conditions as well as tunnel and bridge scenes.

Given an object region, tests on super resolution show low capabilities and chances to improve the significance of features for a following classification step for small objects at low resolution. Latest state of the art super resolution methods involve convolutional networks, which are either learned to recover sharp edges for predefined classes or map several low resolution input images to one high resolution image. In the first case, the idea of using super resolution becomes obsolete as it is not a generic method for any kind of situation anymore. For the second case, one can ask the question, why not detecting the objects of desire directly and omitting the expensive intermediate step of super resolution in between. For these reasons and the experiments in this thesis, the method of super resolution is less recommended for the detection of small and distant objects in automotive applications.

6.2 Outlook

The present thesis achieved the objective to improve the detection of distant objects and reducing fatal accidents on high speed roads such as high- or motorways. Several methods are proposed for the localization of small object regions. However, to make the detection more powerful, as future directions, the given methods can be extended by some further features and details: the Voting Map generation can be expanded by using e.g. temporal features, global, and local texture features. Such extension may increase the performance even further, however, with the cost of an increased demand of calculations. Additionally, weak semantic segmentation with the help of classified patches, can reduce the false-positive rates even more through logical exclusion of possible object region locations in certain segments such as e.g. the sky. For further development on region proposal networks, the integration of global priors into the classification branch and their impact on classification motorway scenarios with small objects may be considered. False-positive proposed regions may be filtered out by the improved classification head. The post-trained net-surgery method showed strong capabilities in understanding convolutional networks to enhance feature representation. At the current state, it is designed layer-wise, however, it could be considered to stretch the clustering of feature maps over the whole network, omitting layer borders. In such a fashion, a network could be reduced in size even more, by maintaining the detection performance.

Appendix A

Gauss-Carpet Super Resolution Method

To suppress noise during the reconstruction of Super Resolution, a Gauss-Carpet method was developed. The main concept of the gauss-carpet reconstruction method is a special form of interpolation, where the low resolution value is smoothed with a gauss distribution before interpolation takes place. In the following the method is described in detail. For better clarity only only registration shifts of $n \cdot 0.5$ LR pixels are considered, however, the method can be extended to any registration shifts. To reconstruct one HR image out of 4 LR images the gauss-carpet method, scans the LR image in quadratic blocks of 2×2 pixels at first. Each block is then mapped on a higher resolution grid of 8×8 pixels. Since it can be assumed that the camera has a point spread function (PSF) in a Gauss shape, each LR pixel value is modeled in a Gauss distribution over the HR block. The Gauss distribution is then discretized on the 8×8 pixels and the contributions of each LR pixel value according to the Gauss distribution are added. The used Gauss distribution is as following:

$$g = \frac{1}{2\pi} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{A.1}$$

with $\mu = 0$ and $\sigma^2 = 1$. The parameters are chosen in such a way, that the LR pixel value addition on the HR pixel with the highest distance to the LR-pixel value is negligible. This prohibits too much smoothing and can be adjusted according to the assumed PSF. Due to the addition of the contribution of the LR pixel values a carpet shaped distribution of the measured LR pixels arises which is shown in figure A.1. As an example the value of the LR pixels of a 2×2 patch L is distributed as following on the 8×8 patch H

$$L = \begin{bmatrix} L_1 & L_2 \\ L_3 & L_4 \end{bmatrix}, \vec{g} = \frac{1}{2\pi} e^{(0.5\vec{x}-0.5)^2}, \vec{x} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{pmatrix}^T \qquad \text{(A.2)}$$

$$H = \begin{bmatrix} g_1L_1 + g_8L_2 + g_8L_3 + g_8L_4 & g_2L_1 + g_7L_2 + g_8L_3 + g_8L_4 & \dots & g_8L_1 + g_1L_2 + g_8L_3 + g_8L_4 \\ g_2L_1 + g_8L_2 + g_6L_3 + g_8L_4 & g_2L_1 + g_7L_2 + g_7L_3 + g_7L_4 & \dots & \vdots \\ \vdots & g_3L_1 + g_7L_2 + g_6L_3 + g_6L_4 & \dots & \vdots \\ \vdots & \dots & \dots & \vdots \\ \vdots & \dots & \dots & g_8L_1 + g_6L_2 + g_8L_3 + g_2L_4 \\ g_8L_1 + g_8L_2 + g_1L_3 + g_8L_4 & \dots & \dots & g_8L_1 + g_8L_2 + g_8L_3 + g_1L_4 \end{bmatrix}$$

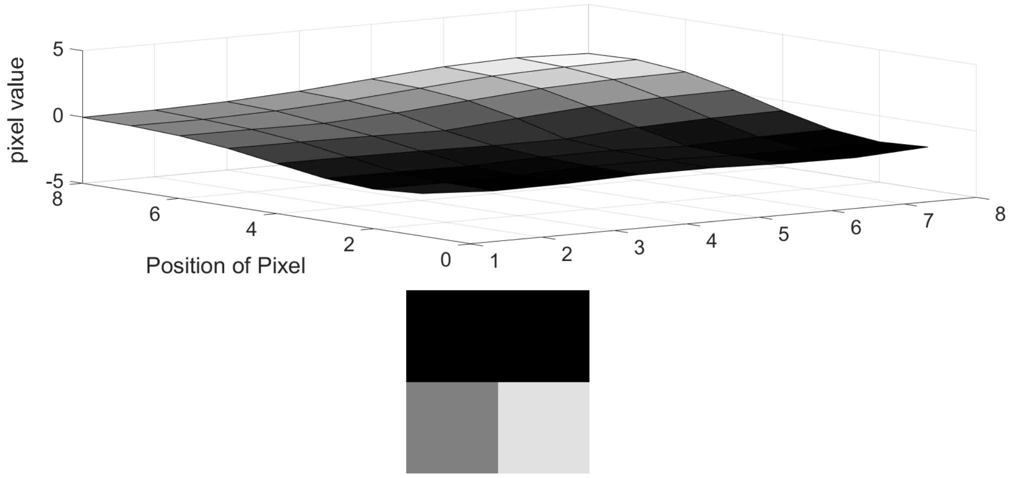


Figure A.1: Formation of the HR blocks with a Gauss carpet based on the LR pixel values. The 8th pixel is not shown as a full pixel in this graph but the value at the edges on the carpet represents the values for the 8th row or column.

This procedure is computed on all used LR images for super resolution. For the actual reconstruction only the inner part of 4×4 sub-pixels of the gauss carpet is utilized. To include the shift of the images the inner part is shifted according to the registration, see figure A.2. As a final step an average is taken from all inner parts of the gaussian carpets on the same location of the image.

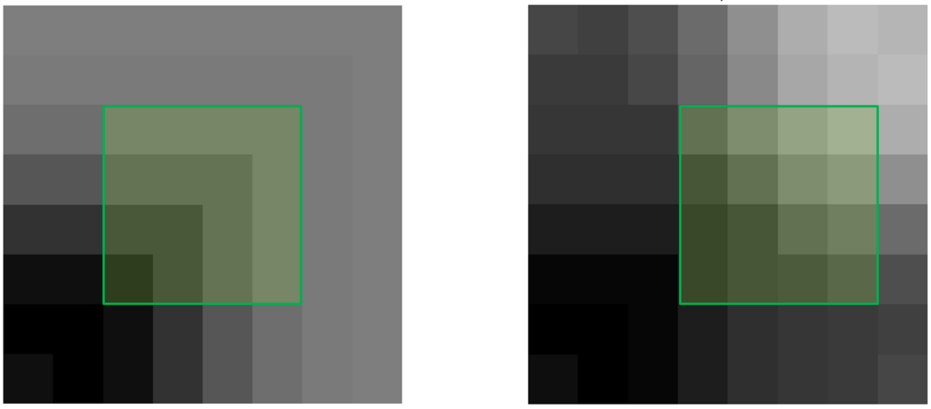


Figure A.2: Two Gauss carpets of the same object which are shifted by 0.25 LR pixels in x-direction. The green area denotes the inner part of the gauss carpets.

The developed gauss carpet method allows to reduce noise sufficiently as for the reconstruction the surrounding of each LR Pixel is taken into account. The major disadvantage is the high computational costs due to the scanning of all participating images.

References

[1] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk. Salient region detection and segmentation. In *International Conference on Computer Vision Systems*, pages 66–75. Springer, 2008.

[2] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009.

[3] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2189–2202, 2012.

[4] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem. Finding tiny faces in the wild with generative adversarial network. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[5] A.-K. Batzer, C. Scharfenberger, M. Karg, S. Lueke, and J. Adamy. Generic hypothesis generation for small and distant objects. In *Proc. of the IEEE Conference on Intelligent Transportation Systems*, pages 2171–2178, 2016.

[6] B. E. Bayer. *Color Imaging Array.* US Patent 3,971,065, 1976.

[7] M. Bertozzi, A. Broggi, and A. Fascioli. Vision-based intelligent vehicles: State of the art and perspectives. *Robotics and Autonomous Systems*, 32(1):1–16, 2000.

[8] P. Bian and L. Zhang. Biological plausibility of spectral domain approach for spatiotemporal visual saliency. In *International Conference on Neural Information Processing*, pages 251–258. Springer, 2008.

[9] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015.

[10] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

[11] G. Brazil, X. Yin, and X. Liu. Illuminating pedestrians via simultaneous detection and segmentation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4950–4959, 2017.

[12] R. Brinkmann. Region of interest. In *The Art and Science of Digital Compositing: Techniques for Visual Effects, Animation and Motion Graphics*, pages 340–341. Morgan Kaufmann, 2008.

[13] C.-A. Brust, S. Sickert, M. Simon, E. Rodner, and J. Denzler. Convolutional patch networks with spatial prior for road detection and urban scene understanding. *arXiv preprint, arXiv:1502.06344*, 2015.

[14] Y.-M. Chan, S.-S. Huang, L.-C. Fu, and P.-Y. Hsiao. Vehicle detection under various lighting conditions by incorporating particle filter. In *Proc. of the IEEE Intelligent Transportation Systems Conference*, pages 534–539, 2007.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint, arXiv:1412.7062*, 2014.

[16] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang. Revisiting rcnn: On awakening the classification power of faster rcnn. In *Proc. of the European Conference on Computer Vision*, pages 473–490, 2018.

[17] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3286–3293, 2014.

[18] M. Cheon, W. Lee, C. Yoon, and M. Park. Vision-based vehicle detection system with consideration of the detecting location. *IEEE Transactions on Intelligent Transportation Systems*, 13(3):1243–1252, 2012.

[19] A. Chung, X. Y. Wang, R. Amelard, C. Scharfenberger, J. Leong, J. Kulinski, A. Wong, and D. A. Clausi. High-resolution motion-compensated imaging photoplethysmography for remote heart

rate monitoring. In *Proc. of the Conference on Multimodal Biomedical Imaging X*, volume 9316, page 93160A. International Society for Optics and Photonics, 2015.

[20] A. G. Chung, C. Scharfenberger, F. Khalvati, A. Wong, and M. A. Haider. Statistical textural distinctiveness in multi-parametric prostate mri for suspicious region detection. In *Proc. of the International Conference on Image Analysis and Recognition*, pages 368–376, 2015.

[21] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, pages 379–387, 2016.

[22] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[23] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[24] J. G. Daugman. Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):1169–1179, 1988.

[25] F. de Ponte Müller. Survey on ranging sensors and cooperative techniques for relative positioning of vehicles. *Sensors*, 17(2):271, 2017.

[26] S. B. Destatis. Verkehrsunfälle. In *Verkehr*, volume 7, 2018.

[27] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016.

[28] C. Eggert, D. Zecha, S. Brehm, and R. Lienhart. Improving small object proposals for company logo detection. In *Proc. of the ACM Conference on Multimedia Retrieval*, pages 167–174, 2017.

[29] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[30] H. Eum, C. Yoon, H. Lee, and M. Park. Continuous human action recognition using depth-mhi-hog and a spotter model. *Sensors*,

15(3):5197–5227, 2015.

[31] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[32] Q. Fan, L. Brown, and J. Smith. A closer look at faster r-cnn for vehicle detection. In *Proc. of the IEEE Intelligent Vehicles Symposium*, pages 124–129, 2016.

[33] C.-Y. Fang, C.-S. Fuh, P. Yen, S. Cherng, and S.-W. Chen. An automatic road sign recognition system based on a computational model of human recognition processing. *Computer Vision and Image Understanding*, 96(2):237–268, 2004.

[34] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar. Fast and robust multiframe super resolution. *IEEE Transactions on Image Processing*, 13(10), 2004.

[35] A.-K. Fattal, M. Karg, C. Scharfenberger, and J. Adamy. Saliency-guided region proposal network for cnn based object detection. In *Proc. of the IEEE Conference on Intelligent Transportation Systems*, pages 1–8, 2017.

[36] A.-K. Fattal, M. Karg, C. Scharfenberger, and J. Adamy. Distant vehicle detection: How well can region proposal networks cope with tiny objects at low resolution? In *European Conference on Computer Vision Workshops*, pages 289–304. Springer, 2018.

[37] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[38] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun. Salient object detection by composition. In *Proc. of the IEEE Conference on Computer Vision*, pages 1028–1035, 2011.

[39] W. T. Freeman and M. Roth. Orientation histograms for hand gesture recognition. In *Proc. of the IEEE Workshop on Automatic Face and Gesture Recognition*, volume 12, pages 296–301, 1995.

[40] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[41] L. M. Gevrekci. *Super Resolution and Dynamic Range Enhancement of Image Sequences*. PhD thesis, Louisiana State University, 2009.

[42] R. Girshick. Fast r-cnn. In *Proc. of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[43] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[44] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[45] C. Guo and L. Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing*, 19(1):185–198, 2010.

[46] D. Guo, T. Fraichard, M. Xie, and C. Laugier. Color modeling by spherical influence field in sensing driving environment. In *Proc. of the IEEE Intelligent Vehicles Symposium*, pages 249–254, 2000.

[47] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, pages 545–552, 2007.

[48] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, volume 15, pages 10–5244, 1988.

[49] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Proc. of the European Conference on Computer Vision*, pages 346–361. Springer, 2014.

[50] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[51] L. He, H. Wang, and H. Zhang. Object detection by parts using appearance, structural and shape features. In *Proc. of the IEEE International Conference on Mechatronics and Automation*, pages 489–494, 2011.

[52] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[53] T. Hoang Ngan Le, Y. Zheng, C. Zhu, K. Luu, and M. Savvides. Multiple scale faster-rcnn approach to driver's cell-phone usage and hands on steering wheel detection. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 46–53, 2016.

[54] J. Hosang, R. Benenson, P. Dollár, and B. Schiele. What makes for effective detection proposals? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):814–830, 2016.

[55] J. Hosang, R. Benenson, and B. Schiele. How good are detection proposals, really? In *the Proc. of the British Machine Vision Conference*, pages 1–12. BMVA Press, 2014.

[56] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):194–201, 2012.

[57] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[58] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 4, 2017.

[59] Y. Huang, W. Wang, and L. Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *Advances in Neural Information Processing Systems*, pages 235–243, 2015.

[60] N. Imamoglu, W. Lin, and Y. Fang. A saliency detection model using low-level features based on wavelet transform. *IEEE Transactions on Multimedia*, 15(1):96–105, 2013.

[61] M. Irani and S. Peleg. Improving resolution by image registration. *Graphical Models and Image Processing*, 53(3):231–239, 1991.

[62] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, 2000.

[63] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194, 2001.

[64] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1254–1259, 1998.

[65] D. Jain. Superresolution using papoulis-gerchberg algorithm. *Digital Video Processing, Stanford University*, 2005.

[66] Y. Kang and X. Li. A novel tiny object recognition algorithm based on unit statistical curvature feature. In *Proc. of the IEEE European Conference on Computer Vision*, pages 762–777. Springer, 2016.

[67] D. Keren, S. Peleg, and R. Brada. Image sequence enhancement using sub-pixel displacements. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 742–746, 1988.

[68] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

[69] J. Kim. Detection of traffic signs based on eigen-color model and saliency model in driver assistance aystems. *International Journal of Automotive Technology*, 14(3):429–439, 2013.

[70] Z. Kim. Realtime obstacle detection and tracking based on constrained delaunay triangulation. In *Proc. of the IEEE Intelligent Transportation Systems Conference*, pages 548–553, 2006.

[71] T. Kong, A. Yao, Y. Chen, and F. Sun. Hypernet: Towards accurate region proposal generation and joint object detection. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 845–853, 2016.

[72] Kraftfahrt-Bundesamt. Bestand in den jahren 1960 bis 2018 nach fahrzeugklassen. In *Fahrzeugklassen und Aufbauarten*, 2018.

[73] A. Kuehnle. Symmetry-based recognition of vehicle rears. *Pattern Recognition Letters*, 12(4):249–258, 1991.

[74] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[75] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural Networks: Tricks of the Trade*, pages 9–48.

Springer, 2012.

[76] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, page 4, 2017.

[77] K. Lenc and A. Vedaldi. R-cnn minus r. *arXiv preprint, arXiv:1506.06981*, 2015.

[78] D. Li and Q. Xu. An efficient framework for road sign detection and recognition. *Sensors and Transducers*, 165(2):112, 2014.

[79] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan. Perceptual generative adversarial networks for small object detection. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1951–1959, 2017.

[80] W. Li, M. Breier, and D. Merhof. Recycle deep features for better object detection. *arXiv preprint, arXiv:1607.05066*, 2016.

[81] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, page 4, 2017.

[82] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. In *Proc. of the IEEE International Conference on Computer Vision*, pages 2999–3007, 2017.

[83] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *Proc. of the European Conference on Computer Vision*, pages 21–37. Springer, 2016.

[84] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *arXiv preprint, arXiv:1506.04579*, 2015.

[85] D. G. Lowe. Method and apparatus for identifying scale invariant features in an image and use of same for locating an object in an image, 2004.

[86] L. Lucchese and G. M. Cortelazzo. A noise-robust frequency domain technique for estimating planar roto-translations. *IEEE Transactions on Signal Processing*, 48(6):1769–1786, 2000.

[87] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 4898–4906, 2016.

[88] S. Manen, M. Guillaumin, and L. Van Gool. Prime object proposals with randomized prim's algorithm. In *Proc. of the IEEE Conference on Computer Vision*, pages 2536–2543, 2013.

[89] B. Marcel, M. Briot, and R. Murrieta. Calcul de translation et rotation par la transformation de fourier. *TS. Traitement du Signal*, 14(2):135–149, 1997.

[90] G. Marola. Using symmetry for detecting and locating objects in a picture. *Computer Vision, Graphics, and Image Processing*, 46(2):179–195, 1989.

[91] C. C. T. Mendes, V. Frémont, and D. F. Wolf. Vision-based road detection using contextual blocks. *arXiv preprint, arXiv:1509.01122*, 2015.

[92] Z. Meng, X. Fan, X. Chen, M. Chen, and Y. Tong. Detecting small signs from large images. In *Proc. of the IEEE Conference on Information Reuse and Integration*, pages 217–224, 2017.

[93] T. Michalke, J. Fritsch, and C. Goerick. Enhancing robustness of a saliency-based attention system for driver assistance. In *International Conference on Computer Vision Systems*, pages 43–55. Springer, 2008.

[94] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz. Pruning convolutional neural networks for resource efficient transfer learning. *Proc. of the Conference on Learning Representations*, 2016.

[95] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. of the IEEE Conference on Machine Learning*, pages 807–814, 2010.

[96] N. X. Nguyen. *Numerical Algorithms for Image Superresolution.* PhD thesis, Stanford University, 2000.

[97] J. Nuevo, I. Parra, J. Sjöberg, and L. M. Bergasa. Estimating surrounding vehicles' pose using computer vision. In *Proc. of the IEEE Intelligent Transportation Systems Conference*, pages 1863–1868, 2010.

[98] H. Nyquist. Certain topics in telegraph transmission theory. *IEEE Transactions of the American Institute of Electrical Engineers*, 47(2):617–644, 1928.

[99] E. Ohn-Bar and M. M. Trivedi. Looking outside of the box: Object detection and localization with multi-scale patterns. *arXiv preprint, arXiv:1505.03597*, 2015.

[100] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[101] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy, et al. Deepid-net: Deformable deep convolutional neural networks for object detection. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2403–2412, 2015.

[102] S. Panda, M. Prasad, and G. Jena. Pocs based super-resolution image reconstruction using an adaptive regularization parameter. *International Journal of Computer Science Issues*, 8(5):155, 2011.

[103] E. Rahtu, J. Kannala, and M. Blaschko. Learning a category independent object detection cascade. In *Proc. of the IEEE Conference on Computer Vision*, pages 1052–1059, 2011.

[104] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä. Segmenting salient objects from images and videos. In *Proc. of the European Conference on Computer Vision*, pages 366–379. Springer, 2010.

[105] P. Rantalankila, J. Kannala, and E. Rahtu. Generating object segmentation proposals using global and local search. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2417–2424, 2014.

[106] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[107] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[108] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533,

1986.

[109] C. Scharfenberger, S. Chakraborty, and G. Färber. Robust image processing for an omnidirectional camera-based smart car door. *ACM Transactions on Embedded Computing Systems*, 11(4):87, 2012.

[110] C. Scharfenberger, A. Jain, A. Wong, and P. Fieguth. Image saliency detection via multi-scale statistical non-redundancy modeling. In *Proc. of the IEEE Conference on Image Processing*, pages 4294–4298, 2014.

[111] C. Scharfenberger, A. Wong, and D. A. Clausi. Structure-guided statistical textural distinctiveness for salient region detection in natural images. *IEEE Transactions on Image Processing*, 24(1):457–470, 2015.

[112] C. Scharfenberger, A. Wong, K. Fergani, J. S. Zelek, and D. A. Clausi. Statistical textural distinctiveness for salient region detection in natural images. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 979–986, 2013.

[113] R. R. Schultz, L. Meng, and R. L. Stevenson. Subpixel motion estimation for super-resolution image sequence enhancement. *Journal of Visual Communication and Image Representation*, 9(1):38–50, 1998.

[114] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint, arXiv:1312.6229*, 2013.

[115] A. Shrivastava and A. Gupta. Contextual priming and feedback for faster r-cnn. In *Proc. of the European Conference on Computer Vision*, pages 330–348. Springer, 2016.

[116] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint, arXiv:1409.1556*, 2014.

[117] A. Singh, C. H. H. Chu, and M. A. Pratt. Visually salient features for highway scene analysis. In *Proc. of the IEEE Conference on Machine Vision Applications*, pages 357–360, 2015.

[118] S. Sivaraman and M. M. Trivedi. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior

analysis. *IEEE Transactions on Intelligent Transportation Systems*, 14(4):1773–1795, 2013.

[119] S. Sivaraman and M. M. Trivedi. Active learning for on-road vehicle detection: A comparative study. *Machine Vision and Applications*, 25(3):599–611, 2014.

[120] I. Sobel and G. Feldman. A 3x3 isotropic gradient operator for image processing. *In the Stanford Artificial Project*, pages 271–272, 1968.

[121] R. N. Strickland and H. I. Hahn. Wavelet transform methods for object detection and recovery. *IEEE Transactions on Image Processing*, 6(5):724–735, 1997.

[122] M. Strolz, Q. Mühlbauer, C. Scharfenberger, G. Färber, and M. Buss. Towards a generic control system for actuated car doors with arbitrary degrees of freedom. In *Proc. of the IEEE Intelligent Vehicles Symposium*, 2008.

[123] Z. Sun, G. Bebis, and R. Miller. Monocular pre-crash vehicle detection: Features and classifiers. *IEEE Transactions on Image Processing*, 15(7):2019–2034, 2006.

[124] H. Süße and E. Rodner. Abtasttheoreme. In *Bildverarbeitung und Objekterkennung*, pages 123–146. Springer, 2014.

[125] H. Süße and E. Rodner. Momente, matching und merkmale. In *Bildverarbeitung und Objekterkennung*, pages 515–587. Springer, 2014.

[126] R. Tsai and T. S. Huang. Multiframe image restoration and registration. *Advances in Computer Vision and Image Processing*, 1(2):317–339, 1984.

[127] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[128] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *In Proc. of the International Conference on Computer Vision*, pages 1879–1886. Springer, 2011.

[129] M. Van den Bergh, G. Roig, X. Boix, S. Manen, and L. Van Gool. Online video seeds for temporal window objectness. In *Proc. of the IEEE Conference on Computer Vision*, pages 377–384, 2013.

[130] A. W. Van Eekeren. *Super-Resolution of Moving Objects in Under-Sampled Image Sequences.* PhD thesis, TU Delft, Delft University of Technology, 2009.

[131] P. Vandewalle, S. Süsstrunk, and M. Vetterli. A frequency domain approach to registration of aliased images with application to super-resolution. *EURASIP Journal on Applied Signal Processing*, 2006:233–233, 2006.

[132] W. von Seelen, C. Curio, J. Gayko, U. Handmann, and T. Kalinke. Scene analysis and organization of behavior in driver assistance systems. In *Proc. of the IEEE International Conference on Image Processing*, volume 3, pages 524–527, 2000.

[133] C. Wang, H. Zhao, C. Guo, S. Mita, and H. Zha. On-road vehicle detection through part model learning and probabilistic inference. In *Proc. of the IEEE International Conference on Intelligent Robots and Systems*, pages 4965–4972, 2014.

[134] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang. Studying very low resolution recognition using deep networks. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4792–4800, 2016.

[135] X. Yang. Enhancement for road sign images and its performance evaluation. *Optik-International Journal for Light and Electron Optics*, 124(14):1957–1960, 2013.

[136] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proc. of the European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[137] D. Zhang, X. He, and H. Li. Data-driven street scene layout estimation for distant object detection. In *Procc. of the IEEE Conference on Digital lmage Computing: Techniques and Applications*, pages 1–7, 2014.

[138] L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection? In *Proc. of the European Conference on Computer Vision*, pages 443–457. Springer, 2016.

[139] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection? In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1259–1267, 2016.

[140] S. Zhang, J. Yang, and B. Schiele. Occluded pedestrian detection through guided attention in cnns. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6995–7003, 2018.

[141] Z. Zhang, J. Warrell, and P. H. Torr. Proposal generation for object detection using cascaded ranking svms. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1497–1504, 2011.

[142] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *Proc. of the European Conference on Computer Vision*, pages 391–405. Springer, 2014.

[143] A. Zomet, A. Rav-Acha, and S. Peleg. Robust super-resolution. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 630–645, 2001.

CURRICULUM VITAE

M.Sc. Civ.Ing. Ann-Katrin Fattal
née Batzer

PERSONAL DATA

Date & place of birth	5th of November 1990 in Frankfurt am Main
Nationality	german

EDUCATION AND WORK EXPERIENCE

since 2018	**Systems Engineer, Systems and Technology** Continental AG, Frankfurt Am Main, Germany
2015 - 2018	**Research Associate in Electrical Engineering and Information Technology** Institute for Control Methods and Robotics, Technische Universität Darmstadt, Germany Advanced Engineering, Continental AG, Frankfurt am Main, Germany
2012 - 2015	**Master of Science in Engineering Physics** Specialization: Technical Optics Technische Universität Darmstadt, Germany Master Thesis: *Optical Alignment Method for Stacking High Resolution Zone Plates*
2013 - 2015	**Civilingenjör in Engineering Physics** Specialization: Medical Engineering Double Degree Program Royal Institute of Technology (KTH), Stockholm, Schweden
2009 - 2012	**Bachelor of Science in Physics** Technische Universität Darmstadt, Germany Bachelor Thesis: *Compound Refractive Lenses for X-Rays*