






Universitat Autònoma de Barcelona

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <http://es.creativecommons.org/blog/licencias/>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>



**Universitat Autònoma
de Barcelona**

SINGLE SENSOR
MULTI-SPECTRAL IMAGING

A dissertation submitted by **Xavier Soria Poma**
at Universitat Autònoma de Barcelona to fulfil
the degree of **Doctor of Philosophy**.

Bellaterra, September 3, 2019

Director	Dr. Angel D. Sappa Centre de Visió per Computador
Thesis committee	<p>Dr. Arturo de la Escalera Hueso Dpto. Ingeniería de Sistemas y Automática Universidad Carlos III de Madrid</p> <p>Dr. Fadi Dornaika Dpto. de Ciencias de la Computación e Inteligencia Artificial Universidad del País Vasco</p> <p>Dr. Robert Benavente Vidal Dept. Ciències de la Computació and Centre de Visió per Computador Universitat Autònoma de Barcelona</p> <p>Dr. Domènec Savi Puig Valls Dept. Enginyeria Informàtica i Matemàtiques Universitat Rovira i Virgili</p> <p>Dr. Daniel Ponsa Mussarra Dept. Ciències de la Computació and Centre de Visió per Computador Universitat Autònoma de Barcelona</p>



This document was typeset by the author using $\text{\LaTeX} 2_{\epsilon}$.

The research described in this book was carried out at the Centre de Visió per Computador, Universitat Autònoma de Barcelona. Copyright © 2019 by **Xavier Soria Poma**. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the author.

ISBN: 978-84-948531-9-7

Printed by Ediciones Gráficas Rey, S.L.

Our intelligence is what makes us human, and
Artificial Intelligence is an extension of that quality
— Yann LeCun

To my family.

Acknowledgements

First of all, I would like to thank my tutor, who during this program, has been the key to me becoming a competent researcher. He, additionally to supporting me in the scientific field, became a friend to support me when I faced new challenges. Thank you Angel for being patient with me when I was a confused student trying to understand some definitions or when I started writing a paper. Even though I think this last one was a headache for him, he did not give up on me.

I have been fortunate to meet many inspiring people in the CVC. Especially, I would like to thank: to Pau Riba who, since the beginning of this journey in Barcelona, has always been ready to help and support in everything that I have needed; to Edgar Riba, who gave me feedback and advice on the technical level; to Arash Akbarinia for his insightful feedback on my research communications. I would also like to extend thanks to my great thesis committee: Dr Arturo de la Escalera Hueso, Dr Fadi Dornaika, Dr Robert Benavente Vidal, Dr Domènec Savi Puig Valls, and Dr Daniel Ponsa Mussarra.

I want to thank many people who made my stay in the Centre de Visió per Computador, Department of Computer Science of the Universitat Autònoma de Barcelona, pleasant and comfortable. Thanks to all CVC staff, much of my time there was like living at home. Moreover, I would like to thank friends and colleagues who make my stay in UAB an amazing experience: Albert Berengel, Arnau Barro, Yaxing Wang, Carola Figueroa, Lichau Zhang, Lorena Gimenez, Sounak Dey, Jose Luis Gomez, Pau Rodriguez, Laura López, Xiao Yi, Yang Fei, David Berga, Xim Cerda and many others.

The four amazing years that I spent in Spain as a PhD student are thanks to many people whom I met at the right time and location during my life. These marvellous people moulded me to always go further in any new project. Thank you to all my professors over the years: Angelica Urquizo, Patricio Humanante, Ruben Pazmino, Lexinton Cepeda, Diego Avila, Eugenia Solis, Roberto Villamarin, Alonso Alvarez, and Nicolay Samaniego during my undergraduate and master student years. Also, thanks to my professional colleagues Fabiola Rodas, Luis Mona, Marco Vasques, Carlos Guayas, Raul Losada, Victor Leon, Delfin Aucancela, and Diego Yumisa for their valuable support.

I am especially grateful to the Ministry of Education and SENESCYT, two Ecuadorian government institutions that supported me with the scholarship for the PhD program. Addi-

Acknowledgements

tionally, many thank to the Spanish government and CERCA Program/Generalitat de Catalunya that supported my research. All this would be impossible without their generous support.

And finally, last but by no mean least, I am very grateful to my family, Vanessa, Axel, Grimaneza, Vilma and my mom Manuela. I have not been with you when you need me, but you have always understood and supported me.

Bellaterra, September 2, 2019

Abstract

The image sensor, nowadays, is rolling the smartphone industry. While some phone brands explore equipping more image sensors, others, like Google, maintain their smartphones with just one sensor; but this sensor is equipped with Deep Learning to enhance the image quality. However, what all brands agree on is the need to research new image sensors; for instance, in 2015 Omnivision and PixelTeq (sensor manufacturers) presented new CMOS based image sensors, which are capable of capturing multispectral bands, these sensors are defined as multispectral Single Sensor Camera (SSC).

This dissertation presents the benefits of using a multispectral SSCs that, as aforementioned, simultaneously acquires images in the visible and near-infrared (NIR) bands. The principal benefits while addressing problems related to image bands in the spectral range of 400 to 1100 nanometres, there are cost reductions in the hardware setup because only one SSC is needed instead of two; moreover the cameras' calibration and images alignment are not required any more. Concerning to the NIR spectrum, even though this band is close to the visible band and shares many properties, the sensor sensitivity is material dependent due to different behaviour of absorption/reflectance capturing a given scene compared to visible channels. Many works in literature have proven the benefits of working with NIR to enhance RGB images (e.g., image enhancement, remove shadows in the RGB images, dehazing, etc.). In spite of the advantage of using SSC (e.g., low latency), there are some drawback to be solved. One of this drawback corresponds to the nature of the silicon-based sensor, which in addition to capture the RGB image, when the infrared cut off filter is not installed it also acquires NIR information into the visible image. This phenomenon is called RGB and NIR crosstalking. This thesis firstly faces this problem in challenging images and then it shows the benefit of using multispectral images in the edge detection task.

The RGB color restoration from RGBN image is the topic tackled in RGB and NIR crosstalking. Even though in the literature a set of processes have been proposed to face this issue, in this thesis novel approaches, based on DL, are proposed to subtract the additional NIR included in the RGB channel. More precisely, an Artificial Neural Network (NN) and two Convolutional Neural Network (CNN) models are proposed; as the DL based models need a dataset with a large collection of image pairs (RGB infected by NIR and target RGB image), a large dataset is collected to address the color restoration. The collected images are from challenging scenes where the sunlight radiation is sufficient to give absorption/reflectance

Acknowledgements

properties to the considered scenes. An extensive evaluation has been conducted on the CNN models, differences from most of the restored images are almost imperceptible to the human eye.

The next proposal of the thesis is the validation of the usage of SSC images in the edge detection task. Three methods based on CNN have been proposed. While the first one is based on the most used model, holistically-nested edge detection (HED) termed as multispectral HED (MS-HED), the other two have been proposed observing the drawbacks of MS-HED. These two novel architectures have been designed from scratch (training from scratch); after the first architecture is validated in the visible domain a slight redesign is proposed to tackle the multispectral domain. Again, another dataset is collected to face this problem with SSCs. Even though edge detection is confronted in the multispectral domain, its qualitative and quantitative evaluation demonstrates the generalisation in other datasets used for edge detection, improving state-of-the-art results. One of the main properties of this proposal is to show that the edge detection problem can be tackled by just training the proposed architecture one-time, while validating it in other datasets.

Key words: Multispectral single sensor camera, RGB-NIR, NIR, image processing, deep learning, color restoration, edge detection.

Resumen

Actualmente el sensor de imagen está normando la industria del teléfono inteligente. Mientras algunas marcas de telefonía exploran añadiendo más cámaras, otros como Google, le mantienen con un solo sensor a sus teléfonos inteligentes; pero este sensor está equipado con Deep Learning (DL) para mejorar la calidad de imagen. Sin embargo, en lo que todas las marcas están de acuerdo es en la necesidad de investigar en los nuevos sensores de imagen; por ejemplo, Omnivisión y PixelTeq (fabricantes de sensores de imagen) presentaron en el 2015 nuevos sensores basados en la tecnología CMOS denominado multispectral single sensors (SSCs).

Esta disertación presenta los beneficios de usar un SSC multiespectral que como se mencionó arriba, simultáneamente adquiere imágenes de las bandas visible e infrarrojo cercano (NIR). El principal beneficio cuando se trabajó con imágenes del rango espectral desde 400 a 1100 nanómetros, es la reducción de costo en la configuración del hardware. Solo se requiere una cámara SSC en vez de dos; además, la calibración de cámaras y el registro de imágenes ya no son requeridas. Con relación a la banda espectral NIR, aunque esta banda es la más cercana a la banda visible y comparte algunas propiedades, la sensibilidad del sensor depende del material de la escena debido a que el comportamiento en la absorción/reflejo capturada de una escena es distinta al canal visible. Muchos trabajos en la literatura han probado los beneficios de trabajar con NIR (por ejemplo para mejorar la calidad de imágenes RGB, remover sombras, quitar neblina, etc). A pesar de las ventajas de usar SSC (por ejemplo baja latencia) existen inconvenientes a ser resueltos. Uno de esos inconvenientes corresponde a la naturaleza del sensor, que además de capturar imagen RGB, cuando no tiene instalado en filtro NIR, también captura información del espectro NIR. Este fenómeno es conocido como RGB y NIR cruzado. Esta tesis primeramente aborda este problema en imágenes complejas y seguidamente muestra las bondades de usar imágenes multiespectrales en la tarea de detección de bordes.

La restauración de color desde una imagen RGBN es el tema relacionado al fenómeno RGB y NIR cruzado. Aunque en la literatura se propone un conjunto de procesos para resolver este problema, en esta tesis se proponen distintos enfoques, basados en DL, para sustraer la información NIR adicional que está en los canales RGB. Más precisamente, se propone una Artificial Neural Network (ANN) y dos Convolutional Neural Networks (CNN); como los métodos son basados en DL, se genera una base de datos con pares de imágenes (RGB

Acknowledgements

infectada con NIR y solo RGB). Las imágenes adquiridas son de escenarios complejos con suficiente radiación solar para estudiar las propiedades de absorción/reflejo a las escenas consideradas. Se ha llevado a cabo una evaluación profunda del modelo CNN, las diferencias de muchas de las imágenes restauradas son casi imperceptible al ojo humano.

La siguiente propuesta de esta tesis es la validación del uso de las imágenes obtenidas en SSC en la tarea de detección de bordes. Tres métodos basados en CCN son propuestos. Mientras el primero se basa en uno de los modelos más usados en la literatura, Holistically-nested edge detection (HED) denominado multispectral HED (MS-HED), los otros dos son propuestos luego de observar las limitaciones de MS-HED. Estas dos nuevas arquitecturas han sido diseñadas desde cero para usar solo esa configuración (entrenando desde cero); una vez que la primera arquitectura es válida en el dominio visible, un pequeño rediseño es propuesto al modelo original para abordar el problema multispectral. Al igual que en el caso anterior, una base de datos es generada para abordar el problema de la detección de bordes. Aunque la detección de bordes es abordada en el dominio multispectral, sus resultados cuantitativos y cualitativos demuestran la generalización en otros conjuntos de datos usados para detección de bordes, alcanzando resultados del estado del arte. Una de las principales propiedades de estas propuestas es mostrar que el problema de detección de bordes puede ser abordado entrenando el modelo propuesto una sola vez, mientras que puede ser validado en otras bases de datos que no fueron utilizadas para su entrenamiento.

Palabras claves: Camara multispectral de sensor unico, RGB-NIR, NIR, procesamiento de imagenes, deep learning, restauracion de color, detection de bordes.

Contents

Acknowledgements	i
Abstract (English/Spanish)	iii
List of figures	xi
List of tables	xv
1 Introduction	1
1.1 Research Objectives	3
1.2 Contributions and Outline	3
2 Related Work	5
2.1 Multispectral Imaging	5
2.2 Deep Learning	11
2.3 RGB Image Restoration from Single Sensor RGBN Image	16
2.4 Edge Detection from a Multispectral Framework	19
2.4.1 Low level Feature Based Methods	19
2.4.2 Biological Visual Perception Based Methods	20
2.4.3 Classical Learning Algorithm Based Methods	20
2.4.4 Deep Learning Based Methods	21

3	Multispectral Imaging	25
3.1	Introduction	25
3.2	Datasets Generation	27
3.2.1	Single Sensor Multispectral Image Datasets	28
3.2.2	Edge Detection Datasets	32
3.3	Conclusion	37
4	RGB Image Restoration from Single Sensor RGBN Image	39
4.1	Introduction	39
4.2	Methods	42
4.2.1	Neural Network Based Approach	42
4.2.2	Convolutional Neural Network Based Approaches	44
4.3	Experimental Results	45
4.3.1	Neural Network Based Approach	46
4.3.2	Convolutional Neural Network Based Approach	47
4.4	Conclusion	56
5	Edge Detection from a Multispectral Framework	57
5.1	Introduction	57
5.2	Proposed Approaches	60
5.2.1	Multispectral Holistically-Nested Edge Detection	61
5.2.2	Dense Extreme Inception Network for Edge Detection	63
5.2.3	Multispectral Dense Extreme Inception Network for Edge Detection	66
5.2.4	NIR Estimation Model	67
5.3	Experimental Results	68
5.3.1	Experiment Setup	69
5.3.2	Results from MS-HED	73

Contents

5.3.3 Results from DexiNed	75
5.3.4 Results from MS-DXN	82
5.4 Conclusion	94
6 Conclusions and Future Work	95
6.1 Conclusions	95
6.1.1 List of Contributions	96
6.2 Future Work	96
Bibliography	111

List of Figures

2.1	Bands of the electromagnetic spectrum.	5
2.2	Images in (a) and (b) are captured in a single shot by the SSC depicted in (d)—right side. The image in (c) is acquired by the SSC with ICF depicted in (d)—left side.	7
2.3	Color Filter Array patterns.	8
2.4	Sub-sampled RGBN data.	9
2.5	RGBN demosaicked samples from sub-sampled images in Fig. 2.4.	10
2.6	The signal interaction from n neurons and analogy to signal summing in the artificial neuron comprising the single layer perceptron, image from [9].	12
2.7	Convolutional Neural Network for image processing, e.g., handwriting recognition [65].	13
2.8	The VGG16 architecture [77].	14
2.9	The EfficientNet-B0 architecture [136].	15
2.10	SSC sensitivity in the RGB and NIR wavelength.	15
2.11	General scheme of RGBN imaging and color restoration process [97].	18
2.12	The first CNN architecture proposed for edge detection [34]. Architecture composed by three convolutional and three fully-connected layers.	21
2.13	RCF architecture [76]. The first column, from the stage 1 to the stage 5, are convolutional layers of VGG16 architecture.	22

List of Figures

3.1	(a) The modular single sensor camera used throughout the thesis. (b) Pair of single sensor cameras used for data collection together with the corresponding ground truth—the left side corresponds to the camera with ICF while the right side corresponds to the camera without ICF.	26
3.2	Multispectral imaging pipeline.	28
3.3	Results from each image processing technique applied to the RAW images provided by the SSCs. The image correspond to a sample from the OMSIV dataset.	29
3.4	Workflow followed during the generation of single sensor multispectral image datasets.	30
3.5	(<i>left</i>) single sensor cameras with and without ICF. The other images are samples from the OMSIV dataset in RGB (<i>top</i>) and RGB+NIR (<i>bottom</i>).	31
3.6	Miniaturized RGB+NIR samples from SSOMSI dataset. As appreciate in the figure, all the images are from the same scene at different day time.	31
3.7	Pipeline of the ground truth generation for the edge detection datasets.	32
3.8	Sample image from the BIPED dataset.	34
3.9	Result from the image fusion methods on a sample fused image from of MBIPED dataset.	36
3.10	The final fused color image from the RGB+NIR and NIR.	37
3.11	Sample image from the MBIPED dataset.	38
4.1	(a) RGB+NIR image (infected with near infrared); (b) NIR image acquired with the SSC of (a); (c) RGB image obtained with a SSC using an infrared cut off filter; resulting in a RGB image free of NIR infection.	39
4.2	Pipeline of multispectral imaging with color correction [139].	40
4.3	RGB+NIR image (infected with near infrared); \sim RGB is the image predicted by the proposed CCN based model for color correction; RGB is the target image used for training the DL model and also used for the evaluation.	41
4.4	Pipeline of the color restoration processes proposed in this Chapter. The rectangle "Color Restoration based on Deep Learning" stands for the different approaches proposed for the color restoration from an RGB+NIR image.	42
4.5	Illustration of the NN architecture used to learn the mapping function Ω	43

4.6	Illustration of the proposed deep learning architectures for RGB color restoration from multispectral SSC images: (a) CDNet and (b) ENDENet. CONV refers to convolution, DECONV to deconvolution and RELU is the non-linear function used for the layers in the respective illustration. The term "k3f32s2" refers to: k = kernel size (3 × 3), f = feature size (32) and s = size of stride (2,2), the same notation is used through the illustration.	44
4.7	Illustrations of: (a) Original RGB+NIR images; (b) Results obtained from the MSE color correction; (c) Results from the proposed approach; (d) RGB ground truth images obtained by using the ICF.	47
4.8	Snapshot of the Human Eye Perception test (HEP) performed over 15 images to evaluate the results from CDNet and ENDENet.	50
4.9	Four samples from the best results (see Table 4.1). Image numbers correspond to the values presented in Table 4.2	54
4.10	Four samples from the average and worst results (see Table 4.1). Image numbers correspond to the values presented in Table 4.2	55
5.1	BSDS images used for training DL models for boundary detections—the five provided annotations (All) and different level of consensus (2 and 3) are depicted. 59	59
5.2	Results from the state-of-the-art algorithms and the proposed methods (DexiNed). Note that HED [150], CED [143], RCF [75], and BDCN [44] have been trained with the BSDS500 [7] while DexiNed was trained just in our dataset (BIPED, see Chapter 3).	60
5.3	VGG16 [126] based multispectral holistically-nested edge detection CNN architecture.	62
5.4	Proposed Dense Extreme Inception Network (DexiNed), consists of an encoder composed by six main blocks (showed in light blue). The main blocks are connected between them through 1x1 convolutional blocks. Each of the main blocks is composed by sub-blocks that are densely interconnected by the output of the previous main block. The output from each of the main blocks is fed to an upsampling block that produces an intermediate edge-map in order to build a Scale Space Volume, which is used to compose a final fused edge-map.	63
5.5	Detail of the upsampling block that receives as an input the learned features extracted from each of the main blocks. The features are fed into a stack of learned convolutional and transposed convolutional filters in order to extract an intermediate edge-map.	64
5.6	Proposed multispectral architecture of DexiNed network (MS-DXN).	67

List of Figures

5.7	Hallu-net architecture, the images used in the figure are from OMSIV dataset ("Conv 5×5 " refers to a convolutional layer with a kernel size of 5×5 , Deconv refers to a deconvolutional layer).	68
5.8	(<i>left</i>) A DCD test image with the provided annotations [71]. (<i>right</i>) Contours after removing wrong annotations.	70
5.9	A sample of NYUD, note there are a large amount of missed edges in the ground truth.	71
5.10	RGB and NIR images from MBIPED dataset, GT, and all edge-maps predicted from both models considered for evaluation in Table 5.1.	74
5.11	Precision/recall curves on BIPED dataset. (a) DexiNed upsampling versions. (b) The outputs of Dexined in testing stage, the 8 outputs are considered. (c) DexiNed comparison with other DL based edge detectors.	75
5.12	DexiNed predictions from BIPED dataset.	76
5.13	Three resultant edge-maps from BIPED test dataset from the approaches presented in Table 5.2b.	78
5.14	Four edge-maps predicted by DexiNed for each tested dataset; they correspond to the best and worst F-measure for the given dataset. The 2nd and 5th columns correspond to the GT of the respective image.	80
5.15	(a) Precision/recall curves on MBIPED dataset, results from all methods considered for comparison (the multispectral images from MBIPED have been used just for MS-DXN, while the RGB images to all other approaches). (b) Precision/recall curves on MDBD dataset, the results is from the 20 test images. Note that the same images used for the proposed model are also evaluated in the other models.	84
5.16	Four samples from the results obtained with the models depicted in Table 5.5. The label in the images (1st row) corresponds to the datasets where they come from.	86
5.17	Four samples from the result obtained with the models depicted in Table 5.6. The label in the images (1st row) corresponds to the datasets where they are from.	89
5.18	Four samples from the result obtained with the models depicted in Table 5.7. The label in the images (1st row) corresponds to the datasets where they are from.	92
5.19	Four samples from the results obtained with the models depicted in Table 5.8. The label in the images (1st row) correspond to the datasets where they are from.	93

List of Tables

3.1	Evaluation of fused images from on MBIPED using four evaluation metrics: EN (Entropy measure), MI (Mutual Information), $Q^{ab/f}$ (edge transfer measure), and VIF (visual information fidelity). Larger values are better in each assessment metric.	36
4.1	Average and median values of the methods tested in the current work when the whole dataset (128 samples) was considered.	51
4.2	Results for a subset of 15 images from [132] (see details in Sec. 4.3.2). Underlined images are depicted in Fig. 4.9 and Fig. 4.10 for qualitative evaluation. Boldface values correspond to the best performance; #t corresponds to the number of times a given algorithm obtain the best performance. The section of HEP contains the qualitative evaluation (results from the survey), each value corresponds to the number of users that selects this result as the best one.	52
5.1	Quantitative evaluation conducted on BIPED and MBIPED datasets.	73
5.2	(a) Quantitative evaluation of the 8 predictions of DexiNed on BIPED test dataset. (b) Comparisons between the state-of-the-art methods trained and evaluated with BIPED.	76
5.3	Comparisons between DexiNed trained on BIPED with respect to the state-of-the-art methods in the literature trained with the corresponding datasets (values from other approaches come from the corresponding publications).	79
5.4	Quantitative evaluation of the different version of hallu-net. lr is the learning rate used on each experiment.	83
5.5	Comparisons between results obtained with DexiNed and MS-DXN with respect to the state-of-the-art methods in the literature. All methods were trained on MBIPED—in this case the multispectral images from MBIPED have been used just for MS-DXN, while the RGB images to all other approaches.	85

List of Tables

5.6	Comparisons of the results obtained with the proposed approaches (DexiNed and MS-DXN) and the by using RGB component of MBIPED, in the case of MS-DXN the NIR component has been obtained from hallu-net. In addition, SED (a non-learning based algorithm based edge detector) is considered for evaluation.	88
5.7	Comparisons of the results obtained in boundary datasets with the proposed approaches (DexiNed and MS-DXN) and the state-of-the-art methods proposed in the literature. All methods were trained by using the RGB componet of MBIPED, in the case of MS-DXN the NIR component has been obtained from hallu-net. In addition, SED (a non-learning based algorithm based edge detector) is considered for evaluation.	90
5.8	Comparisons of the results obtained in the scene or object segmentation datasets with the proposed approaches and the state-of-the-art methods proposed in the literature. All methods were trained with the RGB images from MBIPED. In addition, SED (a non-learning based algorithm based edge detector) is considered for evaluation.	91

1 Introduction

The usage of Artificial Intelligence (AI) has been increasing in recent year becoming almost ubiquitous in every computer system [2], from small devices like smartphones to factory automation machines [134]. Precisely, in the smartphones or mobile devices, which are increasing the usage per year around the world, each new version, releases hardware with increased computation capabilities enabling to Artificial Intelligence the functionally on such devices that a few years ago was just able for desktop or high-performance computers. Currently, those devices in addition to CPU (Central Processing Unit) are equipped with GPU (Graphics Processing Unit) and NPU (Neural Processing Unit) chips. For instance, as the consumers smartphones are marked by the quality of images those devices are capably of capturing, different phone market houses, like Google on its Pixel 3, are equipping their cameras with Deep Learning (DL) to enhance and capture the best possible scene, without compromising the time of processing [50].

On the one hand, since a couple of years ago, most of the smartphone manufacturers increase the number of camera sensors to acquire as much information as possible from a scene (e.g., Samsung, Huawei, Apple). However, Google on the other hand, using just one camera sensor, according to DXOMARK [63], still got the state of the art image quality with its Pixel 2 in 2017. One of the key features to reach such distinction, even with the current generation (Pixel 3), is the inclusion of diverse learning algorithms [23, 121] together with the functionality of its camera sensor. For instance, the front camera of Pixel phone is capable of capturing marvelous selfies thanks to DL model [50] used for recognizing facial expressions and shot just on the right time.

Following the trends of the technology in developing new and small cameras, recently a new family of multispectral Single Sensor Cameras (SSC) has been developed. This thesis is focused on the processing of images from these new sensors under the umbrella of Deep Learning (DL), which will result in the new generation of the image processing research. The main characteristic of the SSCs [20, 79] is that the captured image contain registered information from different spectral bands. In the particular case of the cameras used throughout this dissertation, the image sensor is able to capture information in the spectral bands ranging

from of 400 till 1100 nanometers (nm). In other words, images from the visible and near infrared (NIR) spectral bands are acquired in a single shot.

As the images captured from the visible band (RGB images) have abundant contextual information from the human visual world, because of this image provides texture details with high definition perceptible from the human visual system, the RGB images have been enough to be used in the different fields without the necessity of any additional band. In other words, the information provided from the visible band was enough to reach the desired work. By contrast, the multispectral image has been largely used in a narrow domain (mainly for remote sensing applications) [19]; in recent years its usage has been proliferated in different fields like medicine or defense applications. Precisely, infrared band images have been frequently used after the visible images due to their radiation difference that allows the information acquisition in all weather or day/night conditions [85]. From the infrared band, a sub part, near infrared, has also been used frequently, which shares many properties visible spectrum due to its proximity to the visible band [104].

The NIR information acquired by a camera is material dependence. Therefore, fields like photography [118], agriculture [148], remote sensing [151], medicine [33] and many others, have been considered the usage of the NIR image sensors to accomplish the desired objective. Generally, working with the RGB and NIR images have accomplished collecting images with two cameras, the one for visible band and the other for the NIR band. Later, those images are aligned or registered before any further task. While using the SSC mentioned above, these processes are avoided reducing the time of image processing and collection, and more importantly, reducing the cost of the project and removing errors in the processing pipeline related with the registration. The proposal of this thesis is to study the benefits of the combined use of information. In other words how to improve image processing results by using information from another band. This combined usage of information has been studied in the literature in problems related with image enhancement, dehazing, or noise removal. This thesis will take as case study the color restoration and edge detection problems.

The usage of SSC, although presents many advantages, also introduces some problems that need to be tackled. One of this problem is related with the fact of the single shot capturing of two spectral bands. It makes that the CMOS sensor's cells of R, G, B, also be perceptible of the NIR light. Hence, the RGB image also capture an unknown percentage of NIR information, this phenomenon is termed in the literature as NIR crosstalking in RGB image or NIR infection on RGB channels [19, 51, 96, 104, 139]. The process to solve this infections, is generally referred to as RGB image restoration. Although, several methods have been proposed in the literature, Chapter 4 presents three new approaches based on Artificial Neural Network (NN) and Convolutional Neural Network (CNN). The proposal for RGB image restoration is addressed by using images from outdoor scenarios, where the sunlight is sufficient to make the NIR sensors capture different information than the visible one. One of the key components in the training of DL based approaches are the target value (in this case a target image); therefore, two datasets have been collected, which are presented in Chapter 3. For the target image purpose (with is free of NIR infection), another SSC has been considered, but this time, the

camera is set with an Infrared Cut off Filter (ICF termed in the literature also as IRCF) in front of the single sensor, this setting is detailed in Chapters 2 and 3. The ICF is a NIR filter that enable the pass only to the visible light.

The edge detection task, as mentioned above is the second case study that is going to be tackled in this thesis with images from SSC, to shows the advantages and benefits of these camera sensors. The edge detection has a long history and a large number of proposed models [10, 35]; it is still being used in applications such as image to image translation [53], photo sketching [71], optical flow estimation [144] and so on. The proposed approaches are based on the usage of Convolutional Neural Networks [67, 68] and the multiscale learning with a deep supervision proposed by HED [150]. As the images captured by SSC comes from two spectral bands, the Ground Truth (GT)—edge map annotated by human subjects—is obtained in a fused representation of images from that spectral bands. Even though, the images are from a SSC, the Chapter 5 quantitatively and qualitatively demonstrate that the proposed methods are able to generalize the edge detection in the state-of-the-art datasets used for this task. To tackle this problem, another dataset has been collected and presented in Chapter 3, which has target edge-maps annotated in the edge level.

1.1 Research Objectives

To achieve the goals of this thesis, the research objectives have been formulated as follow:

- **Multispectral image collection by using multispectral single sensor cameras.** in order to fulfill the next objectives a large dataset of multispectral images need to be acquired; furthermore, in order to count with ground truth information two SSCs have been used to acquire clean RGB images (SSC with NIR filter) together with RGB+NIR. Several outdoor scenarios have been considered in order to have a large variability of scenes.
- **RGB image restoration from a RGB+NIR images.** As the image acquisition with the SSC derive in RGB crosstalking with NIR information, with the purpose to give the same human visual world to the image from such a camera, the RGB color correction problem is tackled.
- **Edge detection from a multispectral domain.** With the purpose of validating the usage of SSC in the computer vision tasks, the edge detection problem is considered. In this case the objective is to show that information from one spectral band can help to improve results when image from another spectral band are processed.

1.2 Contributions and Outline

The outline and contributions of this dissertation are:

- **Multispectral Imaging (Chapter 3).** The multispectral single sensor cameras have been

used since the last decade, different demosaicking techniques have been proposed to get a high resolution RGB and NIR images captured in a single shot. This chapter, with the purpose to go further and by using two Single Sensor Cameras (SSC), collect different datasets in visible, and visible and near infrared spectral bands. That is, one camera is used to capture just the visible domain and the other is used to capture in the multispectral domain; this setting will provide deep learning based models with sufficient data to feed and then study their results.

- **RGB Image Restoration from Single Sensor RGBN Image (Chapter 4).** Once the single sensor camera is used without ICF, the RGB channels acquire a percentage of NIR component on their respective channels. Lastly, this problem is tackled from different approaches, which have many processes to correct such RGB and NIR crosstalking. The majority of the approaches proposed in the literature have been evaluated on images captured by SSCs in controlled indoor scenes. This chapter 4 tackles the problem with outdoor images acquired with sufficient sunlight to evaluate the RGB and NIR crosstalking in different conditions and correct the NIR infected RGB images by deep learning based models. The proposed models make the color restoration without any further processes, unlike to the state-of-the-art techniques do.
- **Edge Detection from a Multispectral Framework (Chapter 5).** In this chapter one of the recurrent task in the computer vision community is considered to validate the usage of this image sensor. This chapter presents DL based models carefully designed to detect edges in the multispectral domain by using images acquired from this type of camera. Although the images used by the model are from multispectral domain (visible and near infrared spectral bands), the proposed multispectral model can generalize its prediction task in the most used datasets for edge detection. Even though the model is trained one time just with the dataset for edge detection presented in Chapter 3, whenever the evaluation is performed in other most used datasets for the same task, it reaches the state-of-the-art results.

The aforementioned contributions have been presented at conferences and published in a scientific Journal. More details of such publications are included in the Chapter 6.

The thesis is organized as follows. In chapter 2 relevant works related to the proposed approaches are summarized into four sub-fields: *i* multispectral imaging, *ii* deep learning, *iii* RGB image restoration from single sensor RGBN image, and *iv* edge detection from a multispectral framework. Chapter 3 presents the multispectral single sensor image processing and the different datasets collected for the following chapters. Chapter 4 tackles the RGB image restoration with the proposal of three deep learning based models—the one based on NN and the other based on CNN. Chapter 5 focuses on the edge detection problem training by using images collected with the SSCs but also evaluating its performance in the most used datasets for edge, contour and boundary detection; in addition, two object/scene segmentation datasets have been considered. Finally, in Chapter 6 the thesis is concluded and presented the future works.

2 Related Work

The algorithms developed in this thesis are mainly focused in color image restoration and edge detection in the multispectral domain. However, a key parts to accomplish the development of these methods is image acquisition and processing to train Deep Learning models. Therefore, before reviewing of the state-of-the-art methods in color restoration and image edge detection, an introduction to multispectral image acquisition and processing, together with a summary of deep learning based approaches, are considered.

2.1 Multispectral Imaging

The human eyes, on the one hand, are capable of perceiving in its youth and special scene conditions, from 310 nm (ultraviolet) till 1100 nm (near infrared) wavelength band [127] but normally the range of the vision is from 400 nm to 700 nm. On the other hand, bees perceive from 300 to 600 nm (ultraviolet till most parts of visible spectrum), and the mantis shrimp photoreceptor can detect waves from deep ultraviolet till far-red (300-720 nm) bands. As in the nature, the image sensor technology can acquire most of the bands illustrated in Fig. 2.1 [30]. The most used camera sensors are in visible spectrum, but the technology of their sensors is

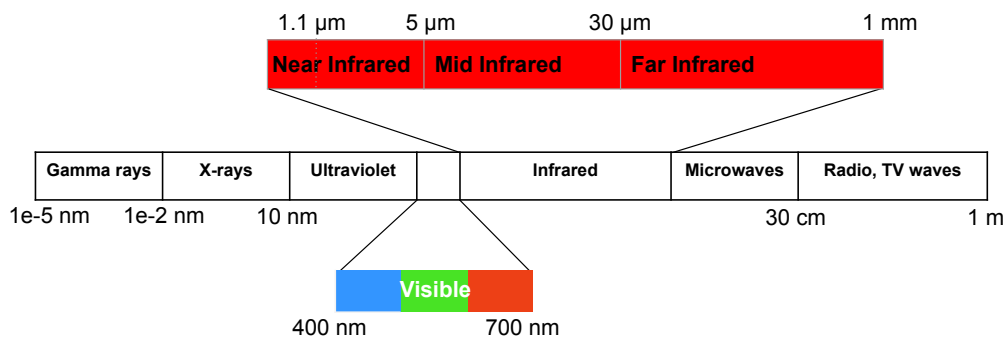


Figure 2.1: Bands of the electromagnetic spectrum.

capable of capturing lights from ultraviolet till near infrared (NIR) (380 nm - 1100 nm) [19]. The current silicon CMOS sensors are not sensitive to the ultraviolet band but they are still capable of detecting spectral bands from 400 nm till 1100 nm. Normally, whenever the NIR band is captured together with visible bands there are desaturations in the visible channels (RGB). Generally, to avoid this problem, an Infrared Cut off Filter (ICF) is used to block the wavelengths greater than 700 nm and perceives just the visible lights like the human eyes see in its daily life.

During the last decades camera sensor technology has drop his price allowing new and breakthrough developments; for instance in recent years a new type of multispectral Single Sensor Camera (SSC) for simultaneous acquisitions of color and near-infrared images in the CMOS system has been proposed. The usage of a camera from this technology is one of the main motivations on this thesis. Although it is an appealing technology, new problems have to be tackled in order to solve the band overlap problems discussed below. These SSCs do not use ICF (see, Fig. 2.2d—right side), thus its sensor is capable of capturing all the visible bands and a part of the NIR (till 1100 nm). The images in (a) and (b) of Fig. 2.2 are samples of the images acquired in one shot by such a camera, which are the visible and near infrared images. Note how the RGB representation (Fig. 2.2a) is affected by the NIR content on the scene—image desaturation problem—, this affection is more appreciated in vegetation areas.

The image desaturation problem, which appears when using a SSC without ICF, is tackled in the literature by means of the usage of color restoration approaches. Generally, to make a color correction, a target image is needed, hence in this thesis two single sensor cameras, rigidly attached in a rig with a small baseline, are considered. One of these cameras has an ICF (see the image depicted in Fig. 2.2d—left side), which allows to acquire RGB images without the desaturation problem. A sample image acquired with this camera is depicted in Fig. 2.2c. Through the work of this thesis, since two single sensor cameras with and without ICF filters are used, in addition to the multispectral imaging, in some sections the corresponding RGB image processing is going to be tackled.

In the SSCs the time exposure and focus are manually set. Generally, in the digital cameras these operations are on-camera tasks, which permits a sufficient light absorption to capture natural scenes. Once such processes are set, a SSC will save a mosaic image array to later imaging with the image processor tool. This tool has a set of image processing techniques to apply to a raw image data before its representation in a display. These operations are termed as image processing or imaging, which in the multispectral domain is termed **multispectral imaging**. Basically, the processes involve image normalization, white balance, contrast equalization, chromatic adaptation, gamma correction, and demosaicking [39, 109]. The sequence of operations vary according to camera.

In general, images obtained by cameras with SSC technology are an array of RAW data, which corresponds to an image processed with the minimum number of operations. These operations are performed by the sensor to the information perceived from a given scene. To acquire as much information as possible from different spectral bands by using just a single

2.1. Multispectral Imaging



(a)



(b)



(c)



(d)

Figure 2.2: Images in (a) and (b) are captured in a single shot by the SSC depicted in (d)—right side. The image in (c) is acquired by the SSC with ICF depicted in (d)—left side .

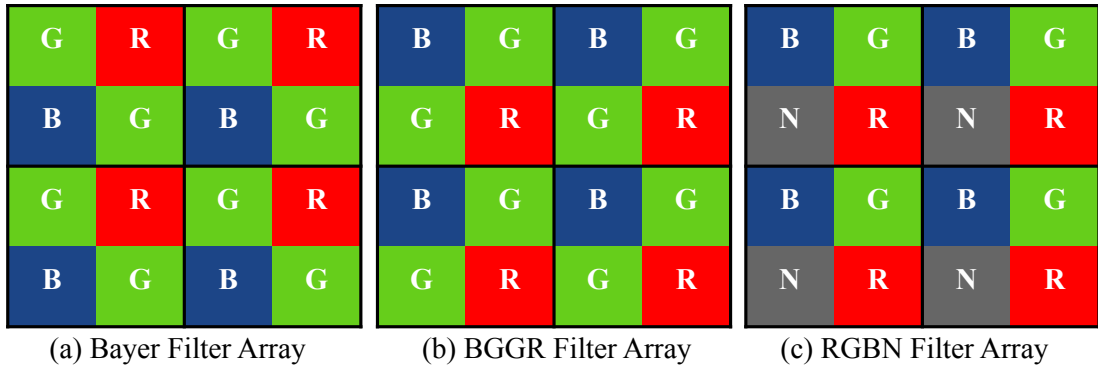


Figure 2.3: Color Filter Array patterns.

image sensor, three important factors need to be considered [8, 56]: the light source $e(\lambda)$, surface reflectance $r(\lambda)$ and camera sensitivity $c(\lambda)$, which in turn $c(\lambda) = \{R(\lambda), G(\lambda), B(\lambda)\}$, λ refers to spectral band delivered from different factors. As the image is required in color (RGB¹) and the sensor just gives mosaic data, a demosaicking process is needed. The Demosaicking process is an interpolation process to estimate missing color values at each pixels [39]. The previous step to demosaick the raw data, for instance the data represented in the illustration of Fig. 2.3a, is the design of a Color Filter Array (CFA) pattern, which in the multispectral domain is referred to as Multispectral Color Filter Array (MFA) pattern. Basically, CFA patterns are designed by the camera sensor makers; for example, the CFA pattern mostly used in the commercial cameras is GRBG termed as Bayer filter array [13], which is illustrated in Fig. 2.3a. The multispectral sensor used in this work has a BGNR MFA pattern (N stands for NIR component) and a BGGR when an ICF is used, see Fig. 2.3b,c. In other words, when the ICF is considered, the sample ratio for R, G, B bands corresponds to 1/4, 1/2 and 1/4, respectively; when the SSC does not use ICF the ratio corresponding to the green band (G) is shared with the NIR channel, therefore, the four bands are divided into equal proportions.

The CFA or MFA [115] pattern selection influences the accuracy of the image processing pipeline; the selected pattern define the usage of the demosaicking technique. That is, depending on the CFA pattern a specific demosaicking process will be performed. As aforementioned, the design of CFA depends on sensors makers; however, many CFAs have been proposed [81] with the aim of enhancing the image quality. Furthermore, when the sensors are capable of capturing multispectral data (more than three bands) the design of MFA also vary resulting in different representations. For example, in Fig. 2.3c there is a pattern formed by 4 channels in the following order BGNR, 2×2 (N stands for NIR channel). Other MFA patterns have been proposed in the literature [59], which are an hybrid implementation from classical CFA (Bayer CFA), but in this case the dimension of the pattern is 10×10 , the additional band (N) is extracted from the R and B cells. In [139] two MFA have been proposed with the patterns size of 8×8 . These patterns are sampled from Bayer filter array. The sample density of the

¹Through this thesis the terms RGB, color image and visible spectrum images will be indistinctly used.

2.1. Multispectral Imaging

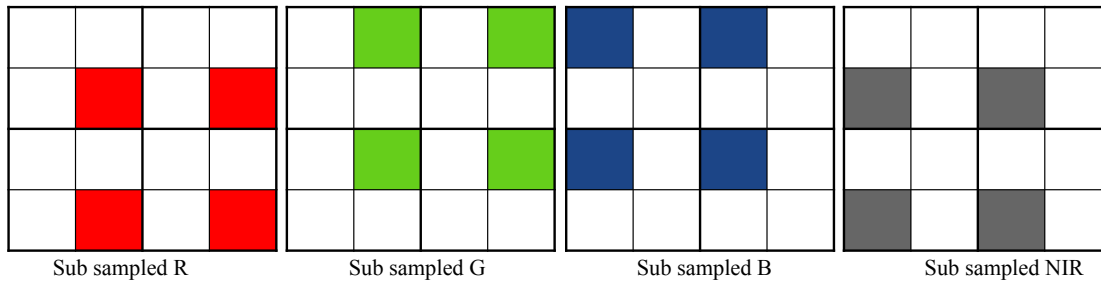


Figure 2.4: Sub-sampled RGBN data.

first MFA for RGBN is $1/5$, $1/2$, $1/5$, and $1/10$, respectively; the sample density for the second MFA proposal is $1/8$, $1/2$, $1/8$, and $1/4$. The corresponding near infrared cell is extracted from the R and B cells in the first proposal, in the second one, the N is extracted just from the R cells. Lastly, in [98] two variants of RGBN MFA are proposed, 4×4 and 10×10 sizes. While the sampling density of the first one is $1/8$, $1/2$, $1/8$ and $1/4$ for the R, G, B, N, respectively; the second one is $1/5$, $1/2$, $1/5$ and $1/10$ in the same order as the first one. In [98] the authors also demonstrate that it is not necessary to design high sized MFAs because the best results are obtained from MFA of 4×4 followed by the 2×2 , and finally the MFA sized by 10×10 .

Once the mosaic raw image is sub-sampled by the CFA, the result is the input for the demosaicking algorithm; Fig. 2.4 shows input sub-sampled images for the pixel estimation (i.e., blank neighbors). In recent years an extensive set of demosaicking algorithms have been proposed in the literature, this chapter describes just a few of them; for an extensive review on demosaicking techniques see overviews in [74, 93]. The most basic demosaicking algorithms are linear and bicubic interpolations [105, 158] of the nearest neighbor. While the linear interpolation takes four closest neighbors, the bicubic interpolation uses 16 closest neighbors and the convolution is accomplished by cubic convolution—it is derived from a set of conditions imposed on the interpolation kernel, this kernel is composed by piecewise cubic polynomials [55].

When the computational cost is not a problem for the demosaicking task, different methods can be considered [93]. For instance, [4] and [3] present methods based on the Laplacian operator. The main characteristics of these methods lie on the reconstruction of green pixels as the luminance information, these algorithms start from the Bayer CFA; then, the interpolation is applied in the horizontal and vertical directions with the Fourier spectrum analysis. Another elaborated demosaicking technique is proposed in [160]; it is termed adaptive intraband interpolation, which is composed by the conjunction of two demosaicking approaches based on support vector regression and linear minimum mean square estimation. While the first one achieves the interpolation by intraband operation, the second one uses the interband operation, which make a more robust results. Iterative residual interpolation is proposed in [156] that is the iterative refinement process for the missing pixels in green channel, once the green channel is reconstructed, the other channels (blue and red) are interpolated with the residual interpolation [57]. This method is mainly based in the green channel reconstruction, which

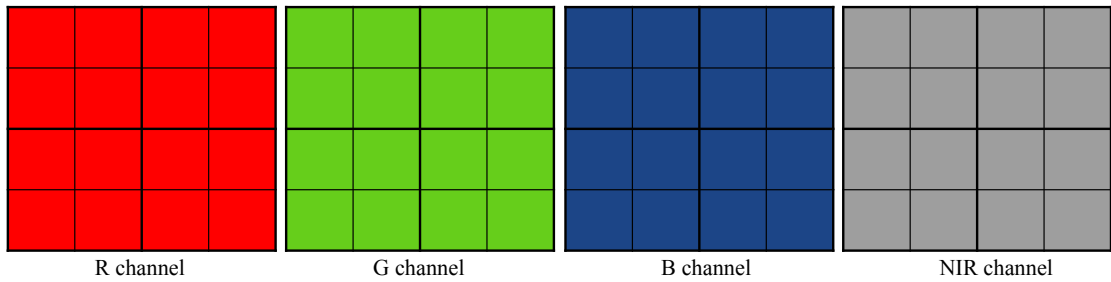


Figure 2.5: RGBN demosaicked samples from su- sampled images in Fig. 2.4.

is processed two times (horizontal and vertical), the obtained results are merged by a soft decision scheme; adaptive residual interpolation [95], adaptively selects a suitable iteration number and combine two different types of residual interpolation at each pixel. That is, the green interpolation at the red lines is performed (horizontal and vertically) by residual interpolation [57] and minimized Laplacian residual interpolation [58] with the iterative framework. Lastly, convolutional neural network based demosaicking approaches have been proposed in [137] and [135]. These DL based models, before fed the CNN architecture, interpolate from the 2D mosaic image a 3D input image. While in [137] a bilinear interpolation is used, in [135] a gradient-corrected bilinear interpolation [87] is considered to reach the 3D input image. The peculiarity in [135] is that the input image fed three models (with the same architecture) and the outputs are fused forming the color image. An illustration of the output from any demosaicking algorithm is depicted in Fig. 2.5, in this particular case there are four channels representing the channels of the sensor used for this thesis.

As mentioned above, the key steps in the image processing are sub-sampling based in a given CFA and demosaicking. However, there are other processes that imply the image enhancement, mainly if the image provided by the camera is in RAW data. The most important methods in this field are image intensity adjustment, gamma correction, and white balance. The intensity adjustment or contrast equalization maps from the image intensity values in gray-scale to a new one saturating the pixel values. The resultant image increases the contrast, therefore, it is more suitable for the white balance and the gamma correction. The white balance (WB) is the method used to remove unnatural color captured from a scene and as each pixel depends on the illumination, the image acquired in an illumination under a low color temperature will appear reddish in the captured image. Therefore, by using WB, the image will be corrected till it becomes to an image captured as in a sufficient natural light [146]. The gamma correction is a point wise non-linearity [32], this can be summarized as $g(I) = I^\gamma$, the value of γ can be determined by passing a calibration target (e.g., a Macbeth chart). Based on experiments performed in [32], γ can also be estimated by the higher order correlation in the frequency domain.

The results of the aforementioned image processing techniques are the images a and c in Fig. 2.2. The image in Fig. 2.2a is captured from the SSC without ICF, as can be appreciated, such an image looks more reddish than the one in Fig 2.2c since the image in Fig. 2.2a has

spectral crosstalking. The spectral crosstalking (or the NIR infection in the RGB channels) is visible because of the desaturation of the visible band; generally this issue is originated by the lack of a multispectral filter, hence the spectral sensitivity propagate to the neighbor spectral bands as will be presented next in this Chapter (see Fig. 2.10 for this issue). To solve the problem mentioned above, a color correction or restoration is need to obtain RGB images from a SSC.

2.2 Deep Learning

Deep Learning (DL) is a sub class of machine learning that is based on the usage of Artificial Neural Network (ANN) for learning multiple levels of representations in order to model complex relationships among data from experience with respect to some class of task [24, 36]. That is, the nature of DL comes from ANN that can extract features from a different types of data, supervised and unsupervised. The main objective of ANN based algorithms is to develop mathematical models that will enable ANNs to learn by mimicking information processing and knowledge acquisition in the human brain, specially in the human nervous system [9]. The learning process on an ANN, specially in the common form of DL (supervised learning), is the iterative process that by computing the loss (objective) function that measure the distance between the predicted value and the target value, trains to map a non-linear function of a specific task. In the iterative process, the model modifies its internal adjustable parameters to reduce the distance [66].

The active research on ANN started back in 1940s with the computer model based on the neural networks of the human brain [101]. In the end of 1950s, a single artificial neuron with the introduction of perceptron is proposed in [9] (see illustration of perceptron in the right side of Fig. 2.6); later in 1990 the necessity of additional neural layer between the input and the output layers was highlighted in [46], which carry out multilayer perceptron (MLP) term. The MLP incursion in the ANN architecture improved the performance on the non-linear classification problems. However, by increasing the hidden layers in the network, the training process becomes unstable, which derived the investigation in new training strategies like by back propagating (BackPro) errors [114].

One of the key concepts of the modern deep network learning procedure was proposed in the mid of 1980s, the proposal is the backpropagation method applied to the network after the forward process (the non-linear connection from the input layer to the output layer, through the whole of hidden layers) [114]. The backpropagation, contrary to its predecessors, where the feature analyzers in the hidden layers were fixed by hand, repeatedly adjusts the weights of the connections in the network's layers from the output layer till the input one [9]. This procedure, later, was applied in a real world problem recognizing handwritten zip codes from U.S. mail [67]. This model used the BackPro in the network architecture of input, output and three hidden layers size. Since then, the BackPro has been successfully applied in Convolutional Neural Network (CNN) architectures [65].

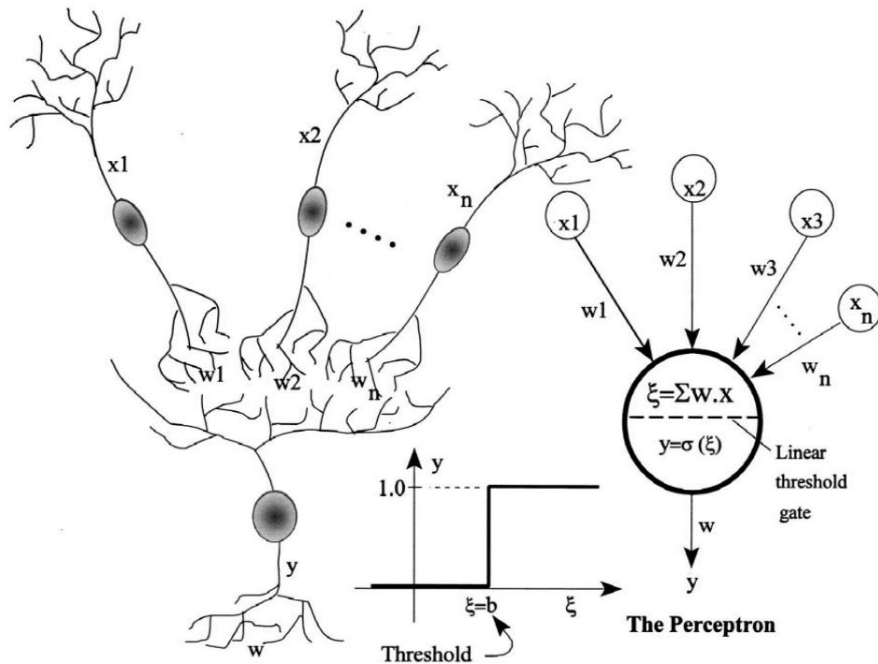


Figure 2.6: The signal interaction from n neurons and analogy to signal summing in the artificial neuron comprising the single layer perceptron, image from [9].

The modern DL models are generally based in Convolutional Neural Networks. The CNN means that the neural network employs a mathematical operation termed convolution [36], which is a specialized kind of linear operation for an image I , $I_{conv} = I * k$, where k is a kernel, generally, is sized in 2D by odd numbers (e.g., 3×3) containing values generated by hands or by a Gaussian function, $*$ is the symbol of convolutional operation. The first CNN based architecture deeply defined was proposed in 1995 by [65]. Figure 2.7 shows the convolution neural network architecture from [65], where the input of the architecture is a 28×28 gray-scale image, then, such data is forward propagated through the 5 convolutional neural layers till get the size of $26 \times 1 \times 1$. The data in each CNN layer is termed **feature map** and its size, for example, in the first CNN hidden layer, " $4@12 \times 12$ " is interpreted such as 4 filters with the feature map size of 24×24 , actually, this notation is ordered such as " $24 \times 24 \times 4$ ". Even though, the CCN architectures were proposed in the last century, the success of this approach was noticed, principally, with the AlexNet [62] proposal. This CNN architecture is composed by five convolutional and three fully-connected layers, such a design was sufficient to win the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2012) competition achieving the test error rate of 15.3%, the second best competitor scored 26.2%.

Since AlexNet many CNN variants have been proposed for different computer vision applications, for instance, edge detection [150], super resolution [26], segmentation [161], color restoration [131], image recognition [126], and many others. However, just a few of them

2.2. Deep Learning

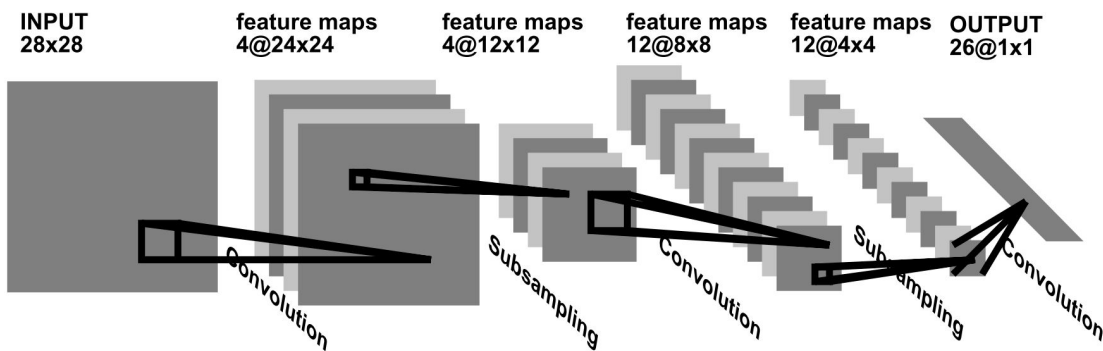


Figure 2.7: Convolutional Neural Network for image processing, e.g., handwriting recognition [65].

have been efficiently designed to be replicated in other computer vision tasks. They are, VGG [126], U-Net [112], ResNet [45], Inception v3 [133], Xception [21], efficientNet [136], among others.

The CNN architectures, proposed from the Visual Geometry Group (VGG) at the University of Oxford, were six versions, from them, the VGG16 and VGG19 are the most used in different computer vision and image processing tasks. For instance, the VGG16 version can be appreciated in Fig 2.8. This architecture is designed by blocks of convolutional and fully-connected layers, specifically, composed by 13 convolutional and 3 fully-connected (MLP) layers. VGG architecture scored the second best mark in the classification task of ILSVRC-2014 competition. One year later, a Deep Residual Learning for Image Recognition (ResNet) has been proposed. Like VGG, ResNet approach also has different version according to the depth of its layers (ResNet50, ResNet100, ResNet152, and so on), the ResNet50 has been frequently used in the literature. Although the deeper neural network is more difficult to train, the most insidious part of such a network is the skip connection (termed in the paper as "shortcut connections"), which make to the deep layers of ResNet do not lose features extracted in the different levels of the ResNet blocks (each block has two convolutional layers). The skip connection is the process that sums the feature map outputs from the previous blocks with the output of the active block, so the input of the next block is the result from the skip connection, this process is repeated until the last block of ResNet. As aforementioned, this process makes that deeper CNN architectures turn straightforward to train but also improve the state-of-the-art results. Actually, ResNet won the 1st place in the ILSVRC-2015 classification competition.

U-Net [112] is another widely used CNN architecture that was proposed in the same year as ResNet [45]. This time, U-net (23 convolutional layers) is designed with an encoder-decoder approach intended for image segmentation with a small dataset for the training process. The encoding part is accomplished by the CNN model while the decoding part consists of an upsampling process followed by a convolutional layer with a kernel of 2×2 . U-Net won the International Symposium of Biomedical Imaging (ISBI) cell tracking challenge with a large margin. Since the 2016, most of researchers started designing optimal CNN architectures,

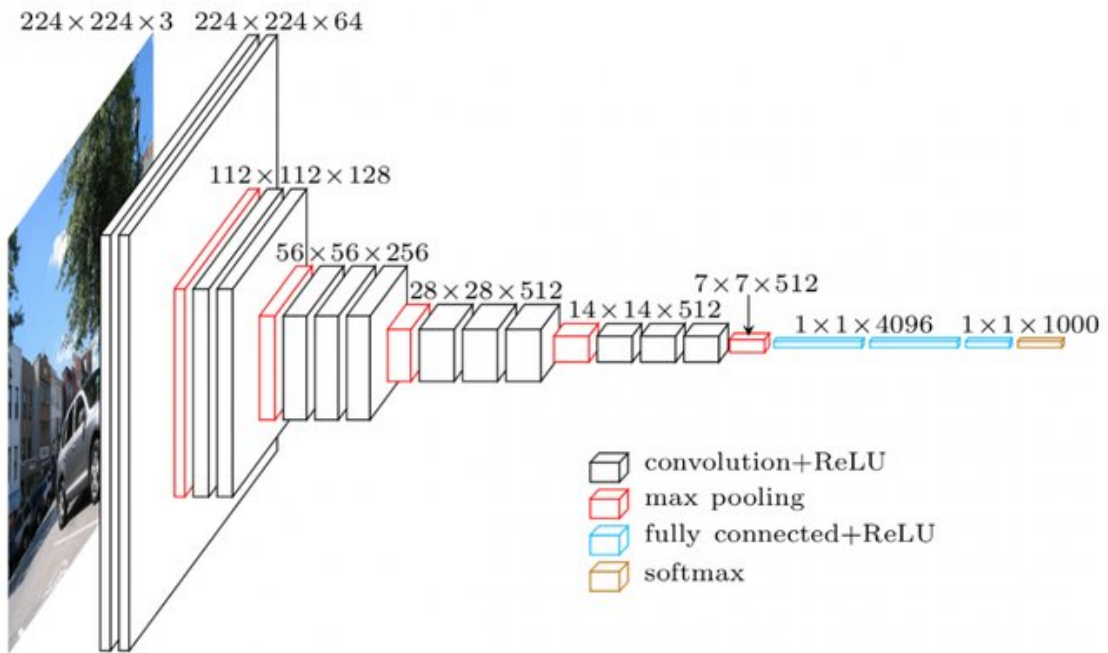


Figure 2.8: The VGG16 architecture [77].

which have fewer number of parameters but still overcome the results from VGG and ResNet. The example of such methods are Inception v3 [133] and Xception [21], both CNN architectures proposed by Google researchers. While Inception v3 has around 23M (millions) parameters, Xception has around 22M (the parameter count reported on ImageNet 1000 classes without counting fully-connected layers [21]). The similarity of those architectures resembles in the usage of depthwise separable convolution (DWSC) instead of the traditional convolutional layers. The special characteristic of DWSC lies on that it uses spatial and depth dimensions of the input image. With those neural network tuning, such methods can train faster than their predecessors but also improving the scores of the ILSVRC validation set.

Although, several CNN architectures have been proposed in the literature, which are not tackled in this review, the last CNN architecture considered is termed EfficientNets [136], proposed in 2019. Actually, EfficientNets is a family of architectures proposed on the manuscript [136]. Comparing to the previous CNN models tackled above, these methods goes much further in the architectural design optimization. Precisely, the variant of EfficientNets, B7, achieves state-of-the-art score on ImageNet, while being 8.4x smaller and 6.1x faster on inference than the best existing CNNs [136] at this time. Figure 2.9 presents the B0 variant of EfficientNets family; The architectures are highlighted as efficient, since they consider in the given network layer the length (L_i), width (C_i) and resolution (H_i, W_i) without changing the particular network layer defined in the baseline network. This architecture has been developed based on an extensive set of experiments performed before reaching the final setup on the three network's scaling dimensions (width, depth and resolution), which makes an

2.3. RGB Image Restoration from Single Sensor RGBN Image

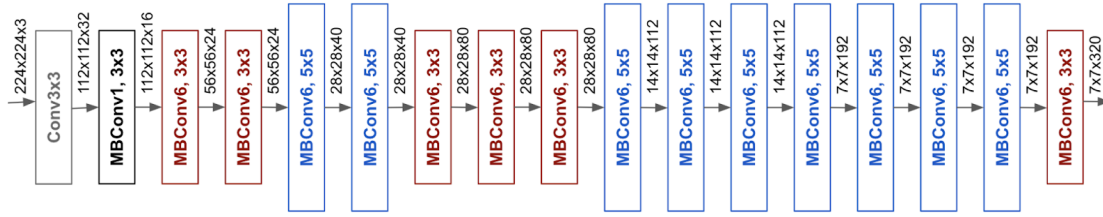


Figure 2.9: The EfficientNet-B0 architecture [136].

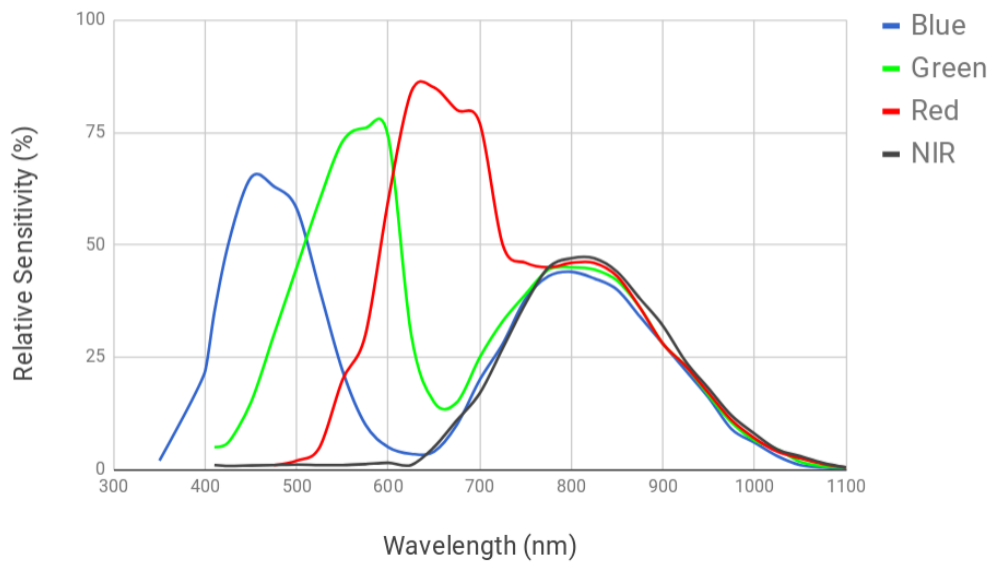


Figure 2.10: SSC sensitivity in the RGB and NIR wavelength.

straightforward derivations or re-use in other tasks.

CNN based approaches have beat classical methods in much of the computer vision tasks, in some cases even better than the human visual perception. However, CNN architectures alone (even with backpropagation) cannot converge to achieve competitive results. Therefore, different training optimizers and non-linear functions have been proposed to produce a fast convergence. The training optimizes, like Stochastic Gradient Descent (SGD), helps to stabilize the training; while non-linear functions, like RELU, help to correct the predictions in the layer level of CNN architectures. In addition to the previous elements, batch normalization, instance normalization, pooling and dropout are usually considered to regularize the training of CNN architectures.

2.3 RGB Image Restoration from Single Sensor RGBN Image

As briefly mentioned in section 2.1, a RGBN single sensor camera is able to capture information in RGB channels (visible bands) till 1100 nm wavelength (near infrared) [19, 104]. Whenever the sensor acquires images from NIR band, the information provided from the scene is different to the visible spectrum due to the surface reflection in the NIR band is material dependent, see the spectral sensitivity graph of Red, Green, Blue, and NIR from a SSC in Fig. 2.10. For instance, most dyes and pigments used for material colorization are somewhat transparent to NIR. This means that the difference in the NIR intensities is not only due to the particular color of the material, but also due to the absorption and reflectance of dyes [116].

The absorption/reflectance properties mentioned above are used for instance in remote sensing applications for crop stress (water and nutrient stress being the most common) and weed/pest infestations. These applications are based on the fact that NIR is not absorbed by any pigments within a plant, it travels through most of the leaf and interacts with the spongy mesophyll cells. This interaction causes about half of the energy to be reflected and the other half to be transmitted through the leaf. In plants with turgid and healthy mesophyll cell walls and in dense canopies, more NIR energy will be reflected and less transmitted. This cell wall/air space interaction within these cells causes healthy vegetation to look very bright in the NIR spectral band. In fact, much more NIR is reflected than visible. By monitoring the amount of NIR and visible energy reflected from the plant, it is possible to determine the health of the plant. This absorption or reflectance properties make that the RGB+NIR images appear desaturated, specially, in the areas with much reflectance. This type of desaturation is formally termed as NIR overlaying/crosstalking in RGB image.

The NIR crosstalking in RGB is tackled as a color correction or restoration process to subtract the percentage of NIR information infected in each visible channel (R,G,B). Most of the approaches for multispectral color restoration in the literature are focused on interpolation and spectral decomposition (signal processing level) to perform the color restoration process. In [19], the authors propose an approach based on color correction method using a color-checker chart as calibration target. Initially, a white balance process is applied, then chromatic adaptation is considered to finally make color correction with a linear transformation matrix (Color Correction Matrix, CCM). Although interesting results are presented it should be mentioned that this approach has been tested in a particular set of RGBN images obtained in laboratory conditions with fluorescent lamp. The obtained color corrected images were evaluated using a X-rite color checker chart. It should be noted that in the evaluation scenario, the visible spectrum was not affected by sunlight. Furthermore, the tested scenario does not contains vegetation where the NIR influence plays a considerable role. This color correction in constrained scenario has been also considered in [104], although the authors mention the sunlight presence in the indoor tested scenes. In these works the authors tackle the complex problem of restoring the three channels (R_{vis} , G_{vis} and B_{vis}), which are contaminated with an unknown quantity of NIR information. The authors propose a method that is based on a spectral decomposition. It implies that the spectral response of each channel will correspond to the RGBN values. Each one of the channels contains a NIR and a VIS spectrum part, which

2.3. RGB Image Restoration from Single Sensor RGBN Image

results in the following formulation: $N = NIR_{vis} + NIR_{nir}$, $R = R_{vis} + R_{nir}$ and so on; hence, the corrected colors are obtained as follows:

$$(\hat{R}_{vis}, \hat{G}_{vis}, \hat{B}_{vis})^T = \mathbf{M} (R, G, B, N)^T, \quad (2.1)$$

where \mathbf{M} is the decomposition matrix obtained by modeling the sensor sensitivity and spectral band correlation; it is a scaling factor coefficient that relates the visible and NIR spectral bands. Authors describe that the additional NIR information infected in the RGB channels maybe an unknown value. In other words, the spectral sensitivity curves presented in Fig. 2.10 depends on the sensor and are needed to solve eq. (2.1). Note that the amount of NIR information will depend on both, the light in the scene and the type of material present on it. In other words, just by using a demosaicing technique, with different color filter array patterns, do not generate images like those obtained with an infrared-cutoff filter lens. The NIR information needs to be removed from the RGB channel [51, 104].

Another color image restoration technique has been proposed in [91]. In this case the visible image is obtained by subtracting at each visible channel a different amount of NIR information, according to coefficients previously computed. These coefficients are obtained from the sensor sensitivity curves (an example of such curves is presented in Fig. 2.10). This particular formulation is only valid for a NIR wavelength range of [650 – 819nm], since the camera used in that work is only sensible to the aforementioned range values. Although the results are quite good and the algorithm is efficient and fast to compute, its main drawback lies in the short wavelength validity as well as in the coefficient estimation, which once estimated are used as constant values for the whole set of images; then, [96] using different Multispectral Filter Arrays (MFA), linear mappings and demosaicing techniques like [91] improved results for the crosstalked problem for images acquired by a SSC. In [139], the authors propose a demosaicing and color correction matrix for improving the quality of acquired RGB and NIR images. The performance of this approach has been only evaluated using indoor images; when it is used in outdoor scenarios, the color correction does not work properly so that the obtained results do not look like natural colors provided by a RGB sensor camera. Usually, the pipeline followed to the color restoration process is tackled as shown in Fig. 2.11.

Nowadays, a large number of classical computer vision problems are being solved under the Deep Learning (DL) framework [36, 66]. In the image enhancement field several DL based approaches have been proposed [28, 43], in particular focused on the super-resolution (SR) [43] and image denoising (ID) [140] problems. Although not focused in color correction, in image enhancement domain, high quality images are intended to be obtained from the given low-resolution or corrupted ones. Somehow the current work is related with those approaches in the sense that a better color image representation is sought from the given NIR infected images. Hence, in the rest of this section ID and SR representative techniques, based on the DL framework, are briefly described.

A few years before the proliferation of CNN based approaches, authors of [54] developed

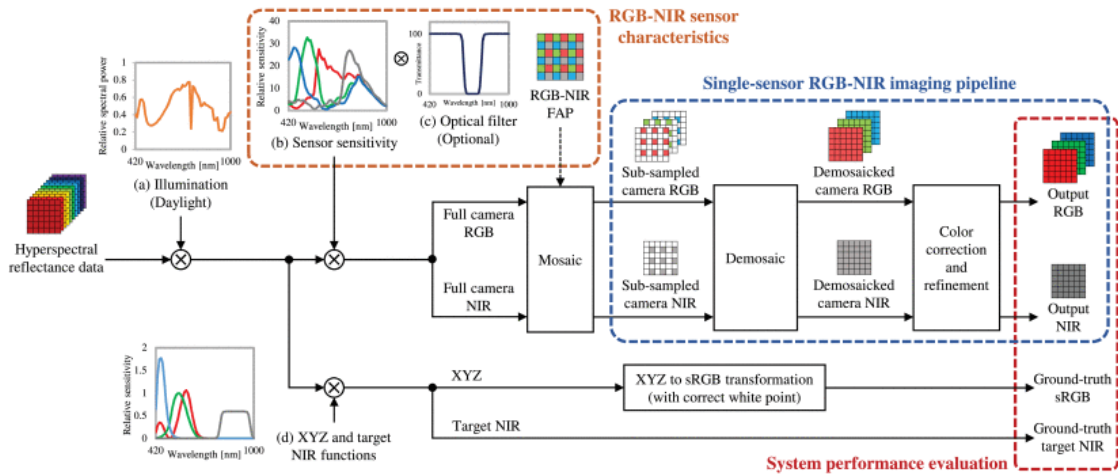


Figure 2.11: General scheme of RGBN imaging and color restoration process [97].

an architecture based on ANN for natural image denoising. The proposed model has 4 hidden layers with 24 units each one; to optimize the training process they used online (Stochastic) gradient descent and its results overcame the state of the art at that time, with just a few processes. In [16] instead, authors propose a multi layer perceptron to map the input vector via two hidden layers (2000 and 1000 units each one) with the corresponding output layer with 10 units, and similarly to [54], stochastic gradient descent was used for the loss optimization.

On the other hand, in the SR field there is a large number of recent contributions also based on the DL framework [43]. In this section, just the most representative models, based on CNN [67] are summarized. In [27], which is an extended version of conference paper, the authors propose to use Deep CNNs to restore high-resolution images; this approach is referred to as Super-Resolution Convolutional Neural Network (SRCNN). This method, made a nonlinear mapping function by using effective feature maps. The approach has three stages: patch extraction (convolution by a set of filters), non-linear mapping (high resolution patches using a 3×3 filter) and reconstruction (averaging to produce final high resolution). The SRCNN architecture has two hidden layers with 64 and 32 units respectively, the used non-linear operator was RELU [100] and mean square error (MSE) as the loss function for the training stage. Recently, [69] presents a Generative Adversarial Network (GAN) with the aim to get finer texture details in the restored images; the authors propose a DL based model termed Super-Resolution GAN (SRGAN). The main contribution was its discriminative architecture (a Residual Network: ResNet) and the loss functions. The perceptual loss function (called content loss) was the key for the generator architecture. Another deep learning based model has been presented in [22], it is obtained as the combination of ResNet and pixel recursive network (PixelRec). The ResNet is designed similar to SRGAN discriminative part and the PixelRec is a feed forward convolutional network that takes a low resolution image through a series of 18 – 30 ResNet blocks. In these works, in addition to PSNR and SSIM [49], human evaluations are considered, since the quality evaluation of enhanced images highly depends on human

perception.

In the last years new color restoration techniques have been proposed based on new CFA patterns [97]. In addition to CFA (three types), a new demosaicking technique has been proposed to interpolate blanks left by the CFAs. This study also demonstrates that it is not necessary to develop large size of CFA because the 2×2 and 4×4 overcome the 10×10 pattern size. In addition, learning algorithm based techniques also have been proposed to correct the NIR crosstalk, which on the one hand, convolutional sparse coding model corrected RGB+NIR outdoor images without sufficient sunlight radiation [51]. On the other hand, a deep CCN is used to make a demosaicking and parallelly correct images desaturated by SSC [123]. The backbone of this deep learning model is U-Net architecture [112].

2.4 Edge Detection from a Multispectral Framework

One of the applications that will be tackled in this thesis with the Multispectral Single Sensor Camera is the edge detection. The goal behind this approach is to improve the edge detection process by using information from one spectral band into the other band. In other words, the objective is to enrich visible spectrum edge detection with features extracted from the NIR band. The edge, contour and boundary detection, some times, are assumed as a synonym task. However these tasks are different since in contour/boundary detection just objects' contour/boundary need to be detected, but in edge detection all edges present in the given image need to be detected [35].

Due to the fact of the novelty of this framework (multispectral edge detection) there are not that much approaches available in the literature. However, there is a large collection of techniques proposed for RGB or Gray-scale image edge, contour and boundary detection, which are deeply reviewed in [10, 35, 165]. The review in this section has been organized as follows: Low level feature based method, biological visual perception based methods, classical learning algorithm based methods and deep learning based methods.

2.4.1 Low level Feature Based Methods

The early edge detection algorithms such as Sobel [128], Robert and Prewitt [113] are based on first order derivative, where the input images are smoothed by the linear local filters termed also linear operators, normally set for two orientations, horizontal and vertical, then edges are detected by thresholding. The usage of more sophisticated operators has been later on proposed; for instance, some of these approaches are based on the usage of second order derivative where edges are detected by the extraction of zero crossing points. In these approaches Gaussian filters and Laplacian of Gaussian [88] are used to smooth the image to later localize edges. In the mid-80s Canny [18] proposes an edge detector by grouping three key processes: *i*) filter out the noise from the given image by performing an approximation to the first derivative of Gaussian filter; *ii*) find out edge directions by using image gradients; and finally *iii*) apply a non-maximum suppression and then hysteresis thresholding for the

final edge map. In spite this approach has been proposed long time ago, it is still nowadays used in some applications [102]. Different variants of the aforementioned methods have been proposed in the literature [119, 120]; for instance, in [106] a combination of multiple steps, together with non-linear filters, has been proposed to detect discontinuities even in challenging scenarios. Other approaches have been proposed for edge detection based on the usage of 2-D Hilbert Transform [61] and Wavelet transform [122]. In [141] a Laplacian based approach has been proposed in the attempt to avoid false zero-crossing detections originated by Gaussian and Laplacian filters. In most of the aforementioned approaches Gaussian kernels are used, which is also widely used in brain-inspired models as will be presented in next section.

2.4.2 Biological Visual Perception Based Methods

With the aim to understand the human visual system, since the 60s of the last century, experiments on monkeys and cats started given important advances on the understanding of Primary Visual Cortex (PVC). For instance, it was discovery the edge formation in the simple cells (V1) [157]. The authors in [157] develop a mathematical model to simulate the human retinal vision by Gaussian derivatives; and inspired on such a model in [11] a special line weight function is introduced for edge enhancement, which consists of a combination of zero and second order Hermite function. Later on, in [38] Gabor energy maps have been proposed to recreate non-classical receptive field of PVC for contour detection. This proposal has been evaluated with a 40 image dataset, which contains for each of the given image its corresponding contour annotation—this dataset could be considered as the first annotated ground truth proposed in the literature to validate edge detection algorithms. This approach has been later on improved by [138] by modeling spatial facilitation and surround inhibition with local grouping functions and Gabor filters respectively. More recently, in [154] a modulation of double opponent cells is performed to extract more complex edge properties from color and texture. This work has been then extended in [6] by modeling a secondary visual cortex, in addition to the model of: the retina, lateral geniculate nucleus and PVC. Another study of psychophysical research, this time focusing on boundary detection, is proposed in [92]. The authors propose to use 3 Gabor filter with different scales, in addition to the usage of principal component analysis and a final classifier. In other words, although it is a brain-inspired model based approach, it could be considered as a machine learning based approach like those presented in next sub-section.

2.4.3 Classical Learning Algorithm Based Methods

The challenge of edge detection in natural scenes has motivated the development of learning based algorithm. One of the first supervised learning (SL) approaches for boundary detection has been presented in [89], which uses different filters to extract gradients from brightness, color and texture. The authors then train a logistic regression classifier to generate a Probabilistic boundary (Pb) edge map. Later on, different variants of Pb detectors have been proposed (e.g., [7, 86]). These new proposals focus on global information besides using local informa-

2.4. Edge Detection from a Multispectral Framework

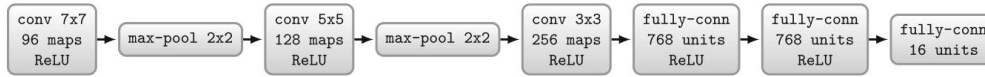


Figure 2.12: The first CNN architecture proposed for edge detection [34]. Architecture composed by three convolutional and three fully-connected layers.

tion like the original approach. They are referred to as globalized Probability of boundary (gPb). In these approaches the gradients are computed by three scales for each image channel individually. In [110], a conditional random field based approach has been proposed. Its maximum likelihood parameters are trained given the image edge annotations. This model captures continuity and frequency of different junction by a continuity structure. On the contrary, the usage of a sparse code gradient has been proposed in [149]. In this case, the patches are classified by a support vector machine. Lastly, the usage of a Bayesian model has been considered in [147]. Edge maps have been predicted by a Sequential Monte Carlo based approach using different gradient distributions. The random forest framework is also considered for edge detection in SL, where each tree is trained independently to capture complex local edge structures [25].

2.4.4 Deep Learning Based Methods

In recent years the usage of CNN has been converted almost ubiquitous in the computer vision fields. Precisely, the first edge detection based on CNN has been proposed in [34]. In this approach, like in aforementioned machine learning models, given an ground truth (GT) edge-map, the neural network with the nearest neighbor (NetNN) classifier gives an output with edge predictions from the input image patch. This model is composed by two stages; firstly, an input image path is propagated through the neural network, see the proposed architecture in Fig 2.12, secondly, by using nearest neighbor search, the output of the CNN (low-dimensional vector) is transformed in the edge-map that correspond to the GT.

Since NetNN many methods have been proposed [35]; most of such approaches use as their backbone architecture the VGG16 [126], before the edge detection procedure. The Holistically Nested Edge Detection (HED) [44] approach was the first one in using this architecture, see Fig. 2.8 for a general overview of that network. From this architecture just the convolutional layers are used (13 convolutional layers). The parameters of this architecture are obtained by training the network in the ImageNet dataset [62] as a classification problem. Then, the network is fine tuned by training it on the datasets conceived for edge detection. Another key contribution of HED was the loss function derived from cross-entropy, which is used in the edge-map predictions from each convolutional block.

Following the proposal of HED, RCF [76], CED [142] and BDCN [44] used VGG16 with some modifications. RCF [76] uses the same configuration as HED, but instead of getting edge predictions from each VGG block, it extracts edges from each layer on the blocks. Figure 2.13 shows the RCF design, the first column of that image contains the VGG16 convolutional layers

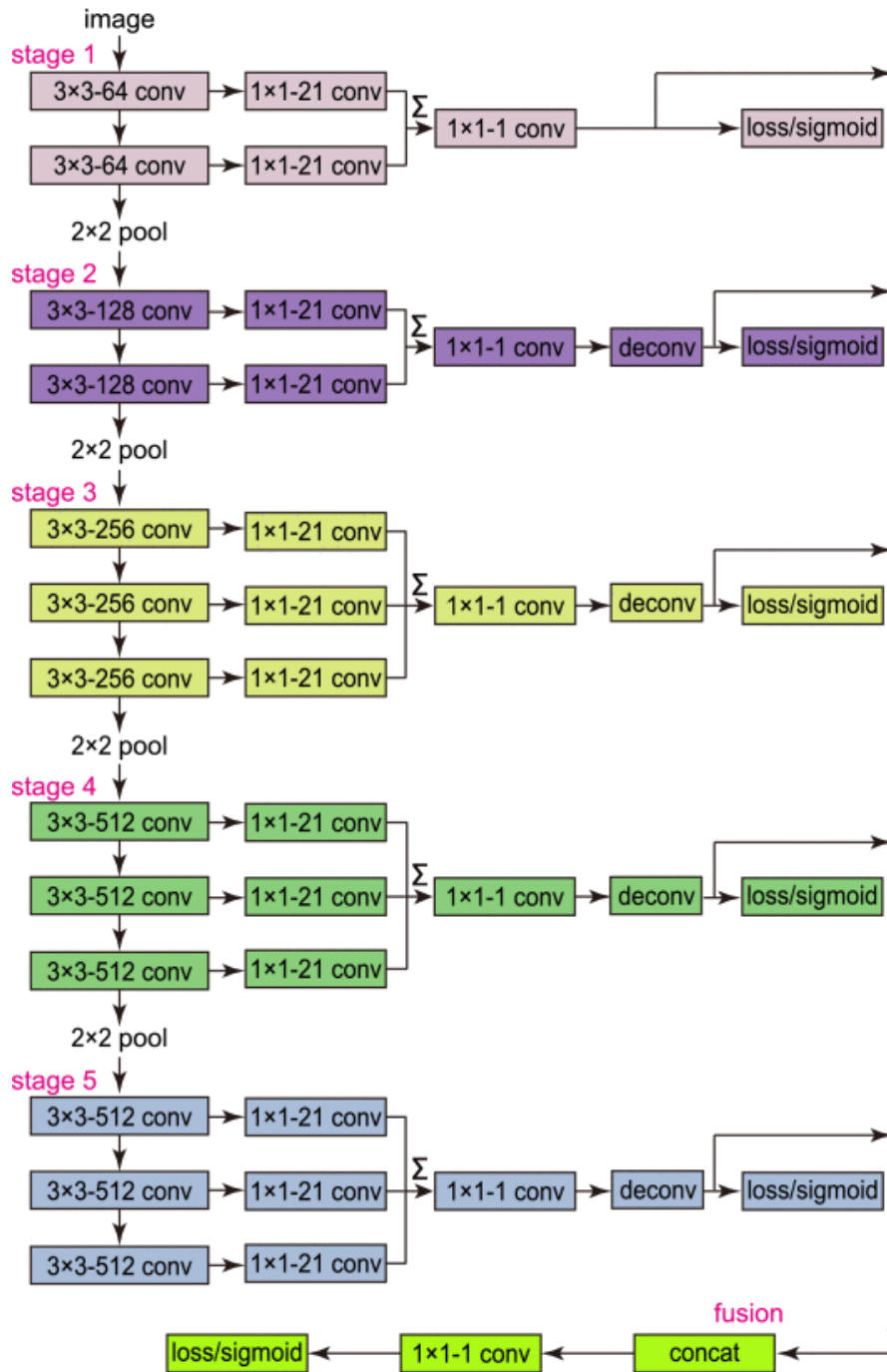


Figure 2.13: RCF architecture [76]. The first column, from the stage 1 to the stage 5, are convolutional layers of VGG16 architecture.

2.4. Edge Detection from a Multispectral Framework

split up in five stages (or blocks). This proposal outperform HED since better defined edges are estimated. On the contrary to previous approaches, in CED refinement blocks (RB) are used to merge outputs from every single block. Even though the quantitative performance of the aforementioned methods overcomes the state-of-the-art of classical learning algorithms, the predicted edge maps are not as sharp as expected, somehow detected edges are coarse. Hence, Generative Adversarial Networks (GAN) [37] have been considered to sharpen those thick edges (e.g., [153] and [159]). Lastly, another more elaborated CNN based edge detector, BDCN, is proposed in [44]. Although the backbone of BDCN is VGG16, the proposal combines BDCN with scale enhancement module, which utilizes dilated convolution to generate multi scale features instead of using deeper CNNs. Actually, the edge-map prediction from BDCN are the state-of-the-art, with only one drawback, which is the time consumption in the training process.

Even though the VGG16 architecture is the most widely used as base model, other architectures have been also considered; for instance, ResNet50 [45] has been used on the variant versions of [76] and [142]. With these variants tiny improvements, of about 1%, have been obtained. In spite of that, VGG16 is going to be assumed as the standard architecture used on the edge detection approaches.

3 Multispectral Imaging

This chapter tackles the visible and near infrared single sensor camera setup and the image acquisition tasks for collecting different datasets that will be used throughout the thesis. The problem is addressed by using a specific single sensor technology to acquire images in outdoor scenarios. This chapter provides benchmarks to be used in the evaluations of the different contributions of this dissertation.

3.1 Introduction

Nowadays, while the digital camera industry sales are dropping [47], the world smartphone markets sales are increasing. One of the key features for the success of mobile phone markets is its camera's capabilities. For instance, according to DXOMARK [64], two brands with the best smartphone cameras are Samsung and Huawei and these brands are at the top of the list in the phone market. While Huawei P30 pro has 5 camera sensors (one in the front and 4 in the back), Samsung Galaxy S10 plus also has 5 but 2 in the front and 3 in the back. Each camera sensor from these brands has its own features, which helps to enhance or capture some particular information from the scenes that could not be acquired if just a single sensor is installed.

This chapter, following the active research on new image sensors, exploits the additional features provided by multispectral Single Sensor Cameras (SSC) [19]. The SSC is the silicon based Complementary Metal-Oxide Semiconductor (CMOS) sensor [29], which is capable of capturing multi-spectral bands in a single shot. Precisely, the perception of the wavelength bands by this sensor are from 400 till 1100 nanometers (nm); that is, Visible (Red, Green, Blue—RGB) and Near Infrared (NIR—N) spectral bands, termed herein as RGBN. The proximity of the NIR band to the visible one shares many properties; for instance, an image captured from a NIR sensor can be perceptible for human eyes. However, as the surface reflectance in the NIR band is material dependent (if there is enough sunlight infrared radiation), the captured image by such a camera will deliver two type of information from the same scene. Such information, in the domain of scene understanding is very welcomed.

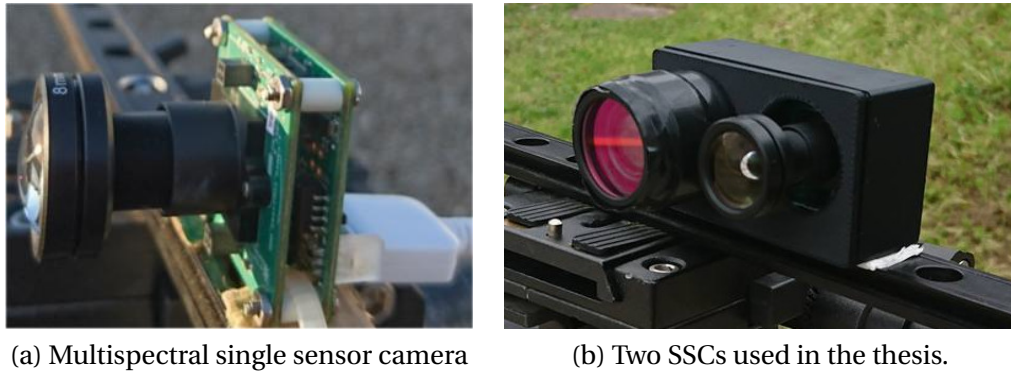


Figure 3.1: (a) The modular single sensor camera used throughout the thesis. (b) Pair of single sensor cameras used for data collection together with the corresponding ground truth—the left side corresponds to the camera with ICF while the right side corresponds to the camera without ICF.

The RGBN single sensor setup can be done, for example, in the customer digital camera by removing the infrared cut off filter (ICF or IRCF) as in [78, 90]. In the current work, for a more suitable manipulation of the information, the sensor developed by the OmniVision (OV4682 CMOS image sensor) is considered, which is assembled in a modular camera by [29], see the camera in Fig. 3.1a. This camera has 4 megapixels with a maximum of 2k resolution. Due to the camera is manufactured in a modular way, the time exposure and the lens focus are manually set, and the mosaic data is organized with four color filter arrays [19]. Although there are a lot of opportunities by capturing the visible and near infrared spectral bands in a single shot from a SSC, for instance in the scene understanding applications, image enhancement, object recognition, etc. (there are much information than just using a single band), there are also new problems that needs to be tackled; for instance when the images are acquired without ICF an important color desaturation effect is appreciated, as highlighted in sections 2.1 and 2.3.

The color desaturation produced by the visible and near infrared spectral bands, captured in a single shot, is termed as overlap or crosstalking between the NIR and RGB bands [29]. Which means, the visible channels also have NIR additional data, termed here in as RGB+NIR. Therefore, a color correction or restoration is needed. To tackle the color correction, a target image captured only on the visible range is needed. In order to obtain such a RGB image without NIR infection, another SSC is used. This second camera is used with an infrared cut off filter (ICF) (see Fig. 3.1b) to only record information from the visible spectrum. As the problem of color correction is tackled with deep learning, a large dataset with sufficient sunlight infrared radiation is generated. It should be noted that most of related works present indoor image datasets. On the contrary, all the datasets acquired in the context of this thesis corresponds to outdoor scenarios and contains two pairs of images acquired with the pair of SSCs shown in Fig. 3.1b; one of these cameras provides visible spectrum images (RGB) while the other multispectral images (RGB and NIR bands). In other words, from a given scene,

there will be three images RGB, RGB+NIR and NIR. These images have been collected as RAW data—a file that contains a minimal processed data, to later make the multispectral and visible spectral image processing.

Once the hardware and the software for the RAW image data acquisition are set, the multispectral image processing or imaging is the next step. The multispectral imaging is a set of processing techniques [80] used to enhance a RAW image data. The resultant image, usually, should be similar to that perceived by the human eye from a given scene. This task is accomplished separately to each image delivered by these two cameras. When the image processing is done, as the issues tackled in this thesis are based on deep learning, pixel wise registered images are needed. Hence, the target image (visible band) need to be aligned to the RGB+NIR and NIR images. This alignment or registration process is performed by Matlab Image Alignment Toolbox [31]. Note that since a SSC is used, the RGB+NIR and NIR are already registered in the same reference system.

The contributions of this chapter are as follow:

- Different single sensor multispectral image datasets are collected as RAW data and shared with the community [132]. Precisely, two datasets have been collected: *i*) 150 images, from just a single scenario, acquired at different time of the day; and *ii*) 500 images from different urban areas containing buildings and vegetation. The datasets are basically oriented for color correction when the RGB image is cross-talking with NIR.
- A dataset termed Multispectral Barcelona Image for Perceptual Edge Detection (MBIPED) is also collected and shared with the community. The MBIPED dataset is intended for edge detection purposes; hence it contains 250 high resolution images together with the corresponding edge annotations.

3.2 Datasets Generation

This section presents details on the datasets generated with the single sensor cameras. Firstly, a brief description of the multispectral imaging pipeline is described. From now on, the notation of the RGB images delivered from the SSC with ICF will be I_{rgb} , the RGB+NIR and NIR images acquired with the SSC without ICF will be referred to as $I_{rgb+nir}$ and I_{nir} , respectively. Where I means a variable with image intensity data, and the sub indexes, for example rgb , refers to the spectral band where it came from.

Figure 3.2 shows the multispectral and visible spectral pipeline to get the human perceptual level image from a RAW image data. Once a scene is captured by the SSC, the spectral data is sub sampled by using Color or Multispectral Filter Array (CFA/MFA) patterns, which are BGGR for color and BGNR for the RGBN bands, such representation is illustrated in Fig. 3.2 (sub-sampled MFA image), in the case of image captured with the camera setup with ICF, the color filter array pattern is BGGR. A representative illustration of the image data before the subsampling process (mosaic RGBN data in Fig. 3.2) can be appreciate in Fig. 3.3a. This data

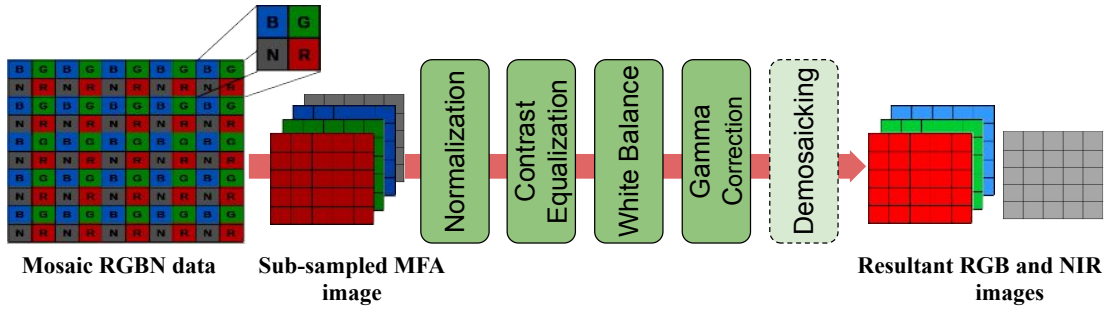


Figure 3.2: Multispectral imaging pipeline.

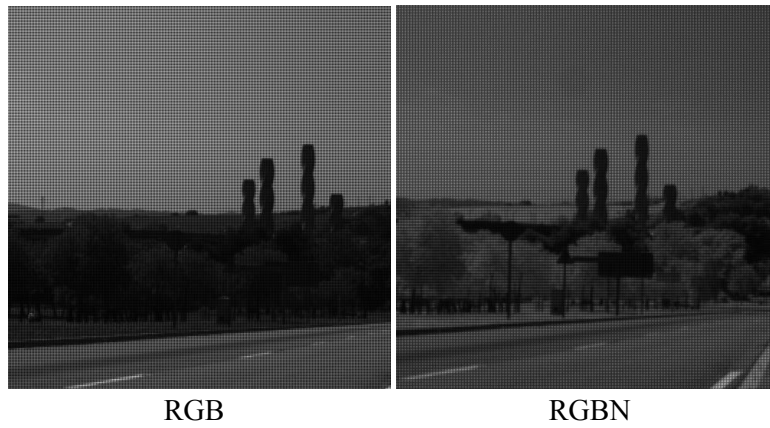
is normalized in the range of $[0,1]$; then, the normalized data I_{norm} is saturated by about 2% in the bottom and the top of all pixels, this process is termed contrast equalization. As a result, a more contrasted image is generated, see Fig. 3.3c. When the raw image data is acquired from the camera another image processing technique is required to adjust the color, white balance, in this case a simple white balance (SWB) is applied. The SWB starts with the mean computation to each image channel $\mu_{rgb} \in \mathbb{R}^{3 \times 1}$; then, the $I_{rgb} = I_{rgb} \times \left(\frac{\max(\mu_{rgb})}{\mu_{rgb}} \right)$, the results is illustrated in Fig. 3.3d. Finally, the gamma correction is applied setting the γ with 0.4545; in other words the gamma correction is applied such as $I_{rgb} = I_{rgb}^\gamma$.

As illustrate in Fig. 3.2, the last process before the output image is demosaicking. However, as demosaicking implies the pixels interpolation, which means that after this process both RGB+NIR ($I_{rgb+nir}$) and NIR (I_{nir}) images will loss information, the resultant data does not have demosaicking in order to have as much information as the sensor provides from each spectral band. However, whenever the demosaicking is needed, the bicubic interpolation is applied. Figure 3.3 shows the results from all the processes followed in the multispectral image processing pipeline.

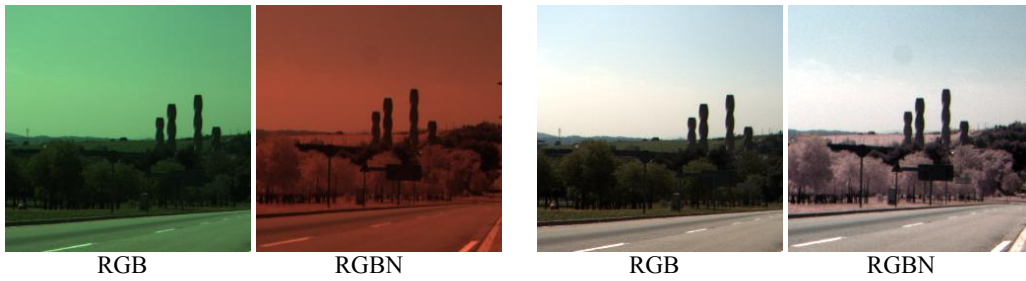
3.2.1 Single Sensor Multispectral Image Datasets

In recent years the number of datasets containing RGB and NIR images **has** increased (e.g., [15, 83]). Nevertheless, usually their images are acquired at least by two cameras and not by a single sensor. Hence, the camera calibration and image registration is needed before the corresponding image manipulation. To the best of our knowledge, there are only a few outdoor single sensor multispectral datasets publicly available and there is a lack of a benchmark dataset to evaluate color correction or other classic computer vision tasks. Gathering a large set of images not only is necessary to quantitatively evaluate different algorithms, but also it allows approaches based on learning algorithms to have sufficient data for their training procedures. The datasets collected for color correction are described below, which are: *i*) Outdoor MultiSpectral Images with Vegetation (OMSIV); and *ii*) Single Scene Outdoor MultiSpectral Images (SSOMSI).

3.2. Datasets Generation

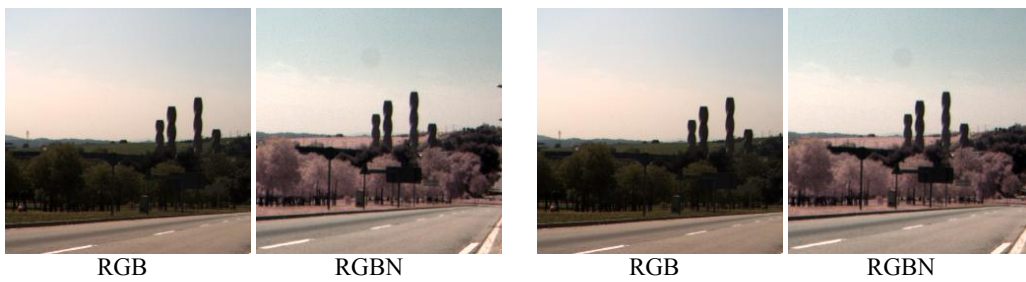


(a) Mosaic RAW images



(b) Subsampled normalized images

(c) Contrast equalized images



(d) White balanced images

(e) Gamma corrected images

Figure 3.3: Results from each image processing technique applied to the RAW images provided by the SSCs. The image correspond to a sample from the OMSIV dataset.

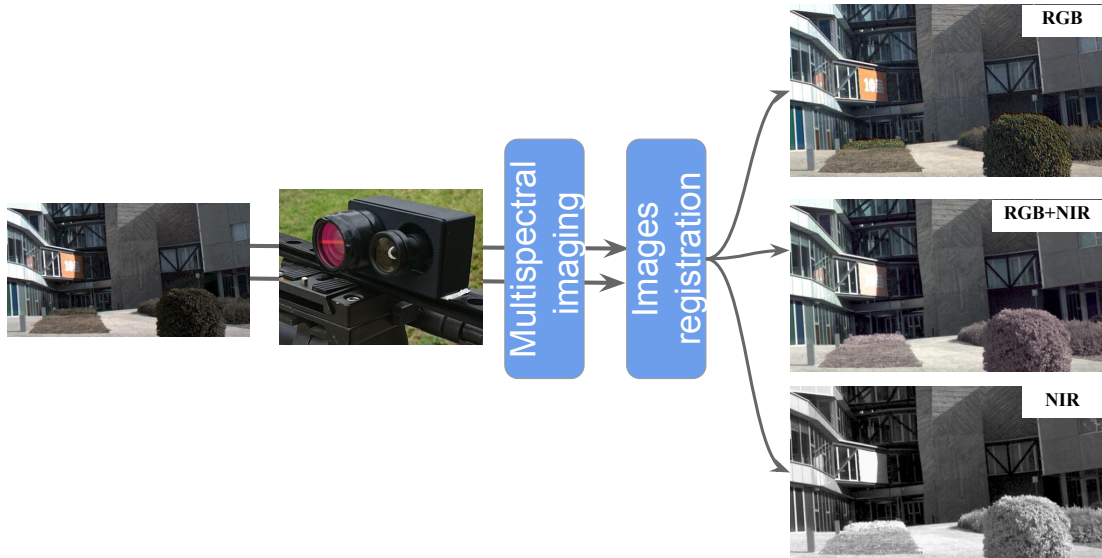


Figure 3.4: Workflow followed during the generation of single sensor multispectral image datasets.

Outdoor MultiSpectral Images with Vegetation (OMSIV)

Most of the color correction methods are evaluated with images captured in fully controlled indoor scenarios, without the presence of any vegetation or specular materials. As mentioned in Sec. 2.1, datasets containing images from different conditions and materials are needed to solve the crosstalking issues in the SSCs, mainly when learning based approaches are considered. Figure. 3.5 illustrate the image desaturation problem (see bottom images in the second and third column). Therefore, a set of Outdoor MultiSpectral Images with Vegetation (OMSIV) is collected. The dataset contains images from outdoor scenarios and most of them contain vegetation.

The entire OMSIV database was captured with the camera introduced in Sec. 3.1. Each multispectral image is supplemented with its corresponding “visible” image captured with the same camera but with an ICF, which blocks the IR light radiation. The original images are in a RAW format and they have to be converted into RGBN images. The size of each RAW image is 1280x720 pixels, and after its channels are separated, the size becomes 640x360x4 for the I_{rgbn} and 640x360x3 for the I_{rgb} (visible channels). In order to guarantee overlap between I_{rgbn} images and their corresponding I_{rgb} (obtained with ICF), the images have been cropped to the size of 580x320x4. These cropped regions have been registered using [31], in order to have a target of the MSI that is overlapped by NIR on its visible band.

The OMSIV dataset contains 500 pairs of multispectral (MS) images (including both the $I_{rgbn+nir}$ and their corresponding I_{rgb} for ground truth). All images were acquired in urban outdoor scenarios with sun light. The OMSIV images were captured from January to April. As

3.2. Datasets Generation



Figure 3.5: (*left*) single sensor cameras with and without ICF. The other images are samples from the OMSIV dataset in RGB (*top*) and RGB+NIR (*bottom*).

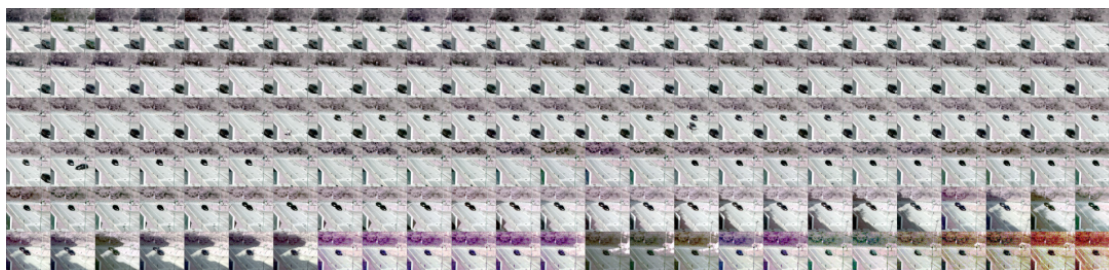


Figure 3.6: Miniaturized RGB+NIR samples from SSOMSI dataset. As appreciate in the figure, all the images are from the same scene at different day time.

mentioned above, the dataset is fully available for downloading to develop and evaluate RGB color restoration approaches, as well as other MS image processing algorithms (e.g., image enhancement, colorization, etc.).

Single Scene Outdoor MultiSpectral Images (SSOMSI)

The dataset Single Scene Outdoor MultiSpectral Images has been collected with the same cameras than the previous one and its RAW images have been also similarly set. The difference appears in the final cropped size, which is $256 \times 256 \times 4$ instead of $580 \times 320 \times 4$ of OMSIV. As shown in Fig. 3.6, this dataset contains images of the same scenario obtained throughout the whole day.

The SSOMSI dataset contains 150 pairs of images. All images were acquired from a scene with sufficient sun infrared radiation. The images are collected in the university campus area in March, from 10:00 till 18:00. About 18 images per hour have been acquired. The images from the last hours, with little sunlight during the sunset, are not that much affected by the desaturation problem but are affected by the lack of NIR illumination. As OMSIV, this dataset

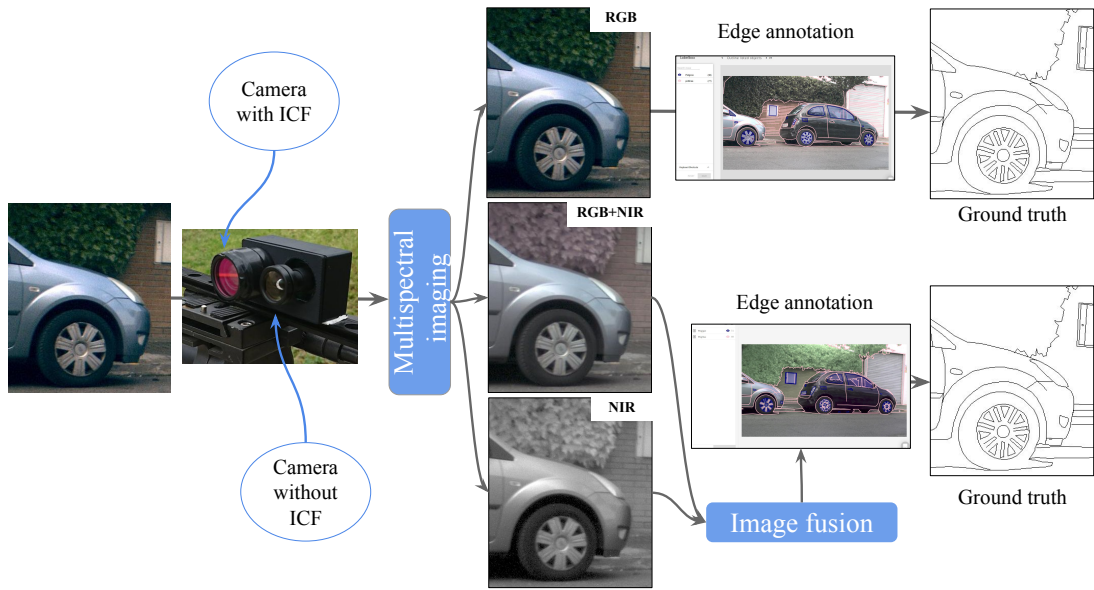


Figure 3.7: Pipeline of the ground truth generation for the edge detection datasets.

is fully available for downloading.

3.2.2 Edge Detection Datasets

This section presents the datasets generated for edge detection purposes. Figure 3.7 presents the workflow followed to collect both Barcelona Images for Perceptual Edge Detection (BIPED) and Multispectral Barcelona Images for Perceptual Edge Detection (MBIPED) datasets. Once a scene is captured by the single sensor cameras, these images are processed in Multispectral Imaging (Sec. 3.2), the ground truths generation for the datasets are independent generated from the images delivered by each camera. The dataset with images and their corresponding GT from the SSC with ICF is termed BIPED; while the dataset with images and GTs from RGB+NIR and NIR images, from the SSC without ICF, is termed MBIPED. Note that before the MBIPED image annotation an image fusion process is performed to RGB+NIR and NIR images.

Similarly to OMSIV dataset, 250 images have been collected by the SSCs in high definition. As illustrated in Fig. 3.7, each scene on the dataset has $(I_{rgb}, I_{rgb+nir}, I_{nir})$. Once the images are captured from the respective spectral band, the RAW images are processed in the multispectral imaging stage, the resultant images, I_{rgb} from the camera with ICF, and $I_{rgb+nir}$ and I_{nir} images from the SSC without ICF, are annotated in the edge level. Later, these annotations are recorded as edge-map GTs for the DL based edge detectors.

As mentioned above, the GT annotation was accomplished slightly different for each spectral band image. For a preliminary evaluation on edge detection models, the BIPED

dataset was firstly annotated since it just contains RGB images due to a ICF has been used during the image acquisition.

Barcelona Images for Perceptual Edge Detection (BIPED)

As mentioned above, high resolution images were selected to have accurate edge annotation, BIPED has been acquired with the same camera used for OMSIV, but this time, the final image size after registration is 1280×720 . This dataset contains 250 urban and university campus scenes. The GT (annotated edge-maps) has been generated by using a crowdsourcing based tool¹. Figure 3.8 shows the GT generation process; the image in Fig. 3.8a is annotated in the edge level (see Fig. 3.8b), then, such annotations are organized to make the edge-map GT, see Fig. 3.8c. This ground truth generation process is applied over the whole dataset.

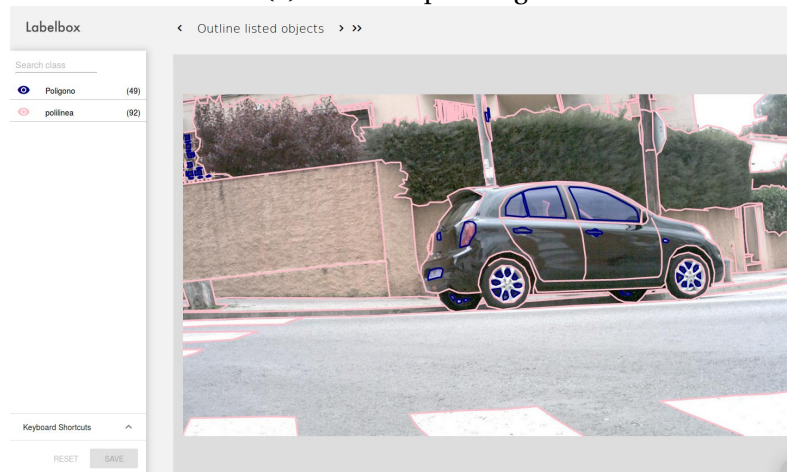
Concerning the annotation strategy, the acquired images have been carefully annotated by experts in the computer vision, hence no redundancy has been considered. In spite of that, all results have been cross-checked in order to correct possible mistakes or wrong edges. The annotation process was divided into four steps. *i*) Initially, all RGB images were annotated by expert collaborators, just one annotation per image; *ii*) At the end of the first annotation process, the results were checked by the administrator; *iii*) Then, the HED model [150] was trained and validated with the first version of the obtained GT; *iv*) With the results of HED, the whole dataset was double checked and carefully re-annotated by correcting few mistakes, again, just by the administrator. The second and fourth steps have been performed by just one subject in order to follow the same criteria in all annotated images. This dataset is publicly available as benchmark for evaluating edge detection algorithms. The generation of a new dataset is motivated by the lack of image's edge datasets; actually, the most popular datasets for edge detection are the ones intended for boundary detection and segmentation.

The level of detail of the annotations generated in the current BIPED can be appreciated in Fig. 3.8c, see cars and street signals. As BIPED is collected to train deep learning based edge detectors, 50 images have been randomly selected for testing and the remaining 200 for the training and validation. In order to increase the number of training images, a data augmentation process has been performed as follow: *i*) since the BIPED images are in high resolution they are split up into half of image width size; *ii*) similarly to the augmentation process in HED, each of the resulting images is rotated by 15 different angles and cropped by the inner axis oriented rectangle; *iii*) images are horizontally flipped; and finally *iv*) two gamma corrections have been applied (0.3030, 0.6060). This augmentation process resulted in 288 images per each of the given image. 10% of this augmented data has been taken for the validation process.

¹<https://labelbox.com/>



(a) BIED sample image.



(b) Screenshot of the annotation process.



(c) Annotated edge map (i.e., ground truth).

Figure 3.8: Sample image from the BIPED dataset.

Multispectral Barcelona Images for Perceptual Edge Detection (MBIPED)

The MBIPED dataset is acquired at the same time as BIPED, but by the SSC without ICF; hence, each scene has $I_{rgb+nir}$ and I_{nir} images. The main difference with BIPED is that before the image annotation process, the $I_{rgb+nir}$ and I_{nir} are fused to later annotate edges in the resultant image. The MBIPED shares the same characteristic than BIPED such as, image size, number of scenes, training and evaluation samples, and the data augmentation for the DL training stage.

The edge level annotation in MBIPED, as highlighted above, is accomplished in the fused images. The image fusion process aims to combine $I_{rgb+nir}$ and I_{nir} images obtained to generate a much informative representation [85]. In other words, the fusion is carried out with $I_{rgb+nir}$ and I_{nir} , each channel of $I_{rgb+nir}$ was fused with the NIR component, therefore, the $I_f \in R^{w \times h \times 3}$, w and h are image width and height size, respectively. Since there is a large collection of image fusion methods [85, 117], the techniques presented in [85] are considered to choose the most robust fusion approach for visible and near infrared image fusion. Seven representative methods have been selected and evaluated in MBIPED test dataset from the 19 deeply studied in the survey [85]. The selected methods are Laplacian Pyramid (LP) [17], Hybrid Multi Scale Decomposition (HMSD) [163], Guided Filtering based Fusion (GFF) [72], Gradient Transfer Fusion (GTF) [84], Two Scale Image Fusion based on Visual Saliency (TSIFVS) [12], Infrared Feature Extraction and Visual Information Preserving (IFEVIP) [162], and the fusion method by Averaging (AV) $I_{rgb+nir}$ and I_{nir} .

The assessment metrics used for evaluating the quality of the fused images are the ones generally used in the literature. They are: Entropy (EN) that measures the amount of information contained in a fused image [111]; Mutual Information (MI) that is the quality index that measures the amount of information transferred from $I_{rgb+nir}$ and I_{nir} to the I_f (fused image) [108]; $Q^{ab/f}$ that measures the amount of edge information that is transferred to the fused image [152]; and finally, the metric based on Visual Information Fidelity (VIF) [41]. The VIF metric is developed to assess fusion performance objectively, it exhibits good prediction performance in natural images. Table 3.1 presents the evaluation of the seven methods mentioned above in these four fusion performance evaluation metrics. These methods have been evaluated in the test dataset of MBIPED (50 images). Concerning to the subjective evaluation, five users have evaluated the results of fused images from the seven methods. The images have been presented as shown in Fig. 3.9.

According to the results from the evaluations (Table 3.1), GFF obtains the best result when MI and $Q^{ab/f}$ evaluation metrics are considered; in the contrary, if the MI metrics is considered the best result is reached by AV; finally, if the $Q^{ab/f}$ metrics is considered the best result is obtained by the fusion perormed by the LP algorithm. However, when EN and VIF are considered GTF and IFEVIP, respectively, reach the best result. As a conclusion from the results depicted in Table 3.1 we could say that a similar performance is obtained by the different approaches (highlighted values correspond to the best performance for a given metric); hence, a subjective evaluation is considered to select the best one for the proposed dataset. The

Methods	EN	MI	$Q^{ab/f}$	VIF
AV	7.411	6.787	0.639	0.801
LP	7.289	5.503	0.668	0.811
GFF	6.165	7.273	0.711	0.847
GTF	7.434	5.789	0.619	0.749
HMSD	7.320	5.413	0.642	1.025
IFEVIP	7.041	4.175	0.554	1.097
TSIFVS	7.412	5.744	0.635	0.856

Table 3.1: Evaluation of fused images from on MBIPED using four evaluation metrics: EN (Entropy measure), MI (Mutual Information), $Q^{ab/f}$ (edge transfer measure), and VIF (visual information fidelity). Larger values are better in each assessment metric.



Figure 3.9: Result from the image fusion methods on a sample fused image from of MBIPED dataset.

images fused (I_f) from the seven methods have been evaluated by five researchers to later on judge the best approach. Figure 3.9 presents just a sample of the fused images by the seven fusion techniques. The five evaluators selected GTF as the best one, see Fig. 3.9 (bottom right). As it can be appreciated, the fusion resulting from the Gradient Transfer Fusion (GTF) method depicts more details and texture of the image than other methods considered during the evaluation. For example, image fused with GFF presents perceptual noise in several regions of the image. The key of GTF to deliver good perceptual fused images, is that this method takes into account the details from the two source images and selects as a guided image the source image that has more details; in this case, the visible band image ($I_{rgb+nir}$) is selected as guide. As mentioned above, the image fusion is conducted to each channel of $I_{rgb+nir}$ independently with I_{nir} . Figure 3.10 shows an illustration of the resulting fused image, this type of image is used to generate the GT for MBIPED.



Figure 3.10: The final fused color image from the RGB+NIR and NIR.

Figure 3.11a shows a sample of MBIED fused image, which is used to annotate edges from the scene. As appreciated in Fig. 3.11b, the edges are carefully annotated. The edge-map resulting from the annotation is depicted in Fig. 3.11c.

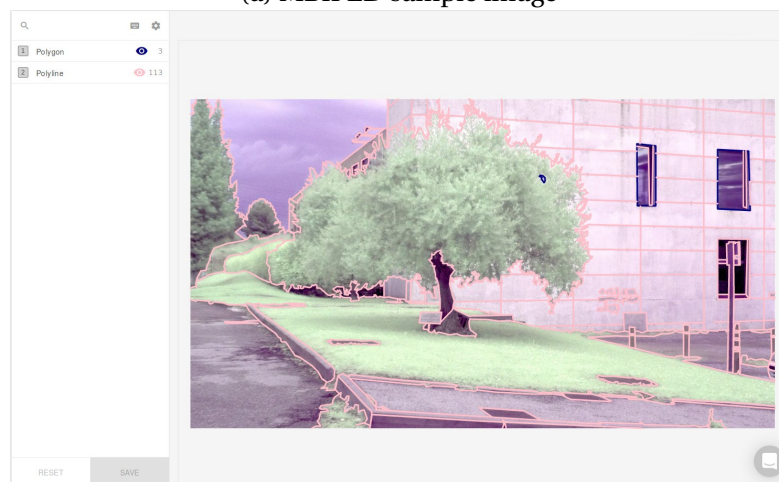
3.3 Conclusion

Multispectral images from Single Sensor Cameras offer a large number of opportunities for improving image processing algorithm (e.g., enhancement, dehazing, edge detection, etc.) as well as applications based on information from that spectral bands (e.g., image vegetation index estimation). The main attractiveness of the information provided by these cameras lies on the fact that the provided images are registered in the same reference system, avoiding the classical registration problem and focusing the research just on the combined use of the provided information. In such an interesting context, and following the trends in the computer vision community of developing learning based approaches, the generation of large dataset is a requirement to develop robust solutions. In this chapter two datasets (i.e., OMSIV and SSOMSI datasets) have been collected to tackle the desaturation problem that appears in the multispectral SSCs. All images from these dataset are from outdoor scenes with sufficient sunlight condition. In addition, much of such images have vegetation; therefore, the color correction or restoration can be deeply studied. The images from both datasets are saved in RAW format and each scene has RGB, RGB+NIR and NIR images registered.

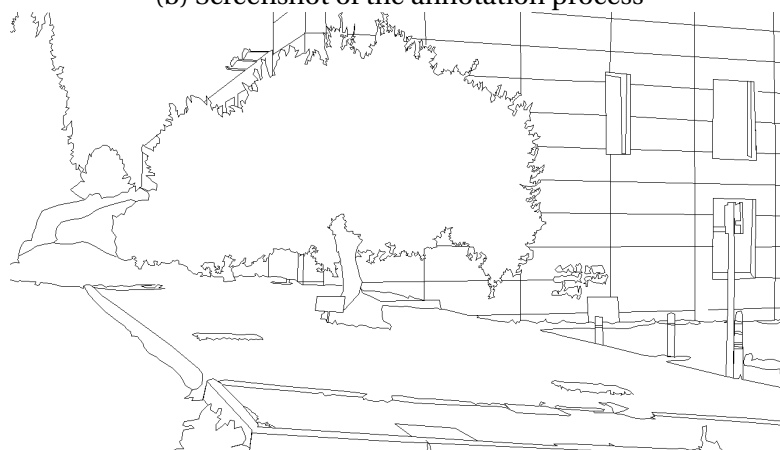
In addition to the datasets acquired for the color correction problem, two datasets (i.e., BIPED and MBIPED), containing images with the corresponding annotations, have been collected for tackling the edge detection problem. In these datasets each scene contains RGB, RGB+NIR and NIR images registered in the same reference system. As these images were collected from two SSCs, the ground truth generation (annotated edges) was separately performed, one for the visible spectrum and one for the fused images resulting from ($I_{rgb+nir}$ and I_{nir}).



(a) MBIPED sample image



(b) Screenshot of the annotation process



(c) Annotated edge-map (i.e., ground truth)

Figure 3.11: Sample image from the MBIPED dataset.

4 RGB Image Restoration from Single Sensor RGBN Image

This chapter presents two approaches for color image restoration of single sensor RGBN images. The first one consists of a multilayer perceptron based approach, which was intended to evaluate the challenge and drawback of this problem. On the contrary, the second proposed method is a more elaborated one based on the usage of convolutional neural networks. Both approaches are detailed below.

4.1 Introduction

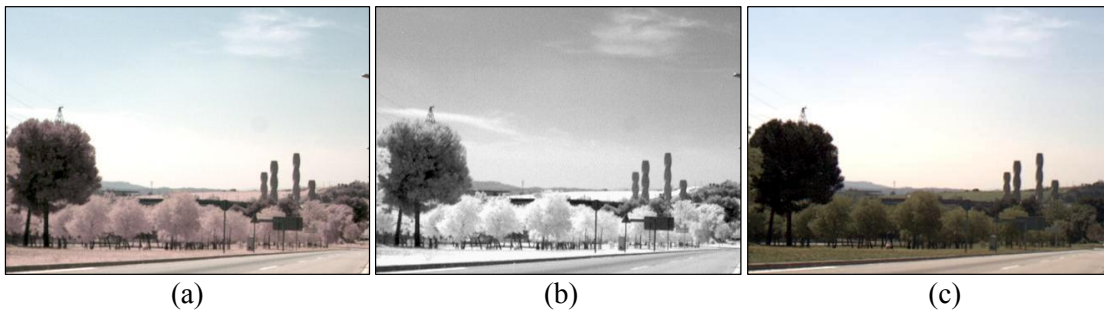


Figure 4.1: (a) RGB+NIR image (infected with near infrared); (b) NIR image acquired with the SSC of (a); (c) RGB image obtained with a SSC using an infrared cut off filter; resulting in a RGB image free of NIR infection.

Through this chapter the same notation used throughout the thesis will be adopted; RGB refers to a three channels RGB image; RGBN refers to a four channel image where the RGB channels are free of NIR infection and N stands for the NIR component; RGB+NIR refers to a three channel RGB image where each channel is infected by the NIR component; finally, RGBN+NIR refers to a four channel image where each RGB channel is infected by the NIR component. The RGB or color image restoration is also referred to as color correction. This process or set of processes allow to obtain color images near to the target or the human perception level from a scene acquired by the image sensor. The most common color correction method

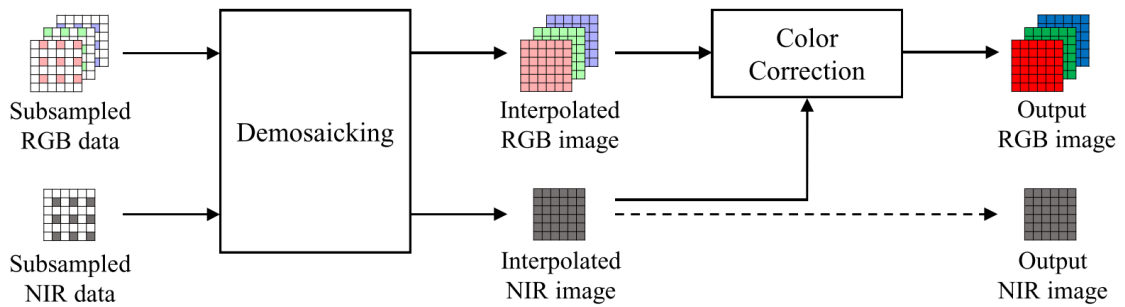


Figure 4.2: Pipeline of multispectral imaging with color correction [139].

is by using a color correction matrix (CCM) obtained with the Macbeth color checker chart. However, in the case of images from a multispectral (MS) Single Sensor Camera, which can capture just the visible information from a scene if the Infrared Cut Off Filter (ICF) is installed in front of the sensor, the macbeth chart based color correction approach cannot be used. The color correction process cannot be performed with the Macbeth chart because of the absorption/reflectance of the surface captured from a scene are different in the visible and near infrared wavelength bands (see details in Sec. 2.3).

Figure 4.1a shows an illustration of an image where the desaturation problem can be easily appreciated when compared to the image presented in (c); this effect is perceptible for human eyes. These details are more perceptible in the vegetation area, and the difference can be much notorious if there is enough sunlight radiation in the scene. The claim of the difference can be more perceptible focusing in the Fig. 4.1b, which is the NIR part of the image in (a), captured in a single shot. Therefore, in images with different reflectance to visible band, the color restoration from the RGB+NIR (RGB image infected by the NIR component) cannot be accomplished by using a single operation like CCM. Most of the state-of-the-art methods follow the workflow of the illustration in Fig. 4.2 [139].

As mentioned in Chapter 2, several sets of processes are required to accomplish the RGB restoration. Figure 4.2 shows the standard workflow to correct the RGB images corrupted by NIR component in images from a SSC. Firstly, new Color or Multispectral Filter Array CFA/MFA patterns are set to sub sample the multispectral image, the sub sampled image is demosaicked to the whole multispectral data. After the MS interpolation, the images in RGB channels still have NIR information, then, color correction process needs to be applied. There are different color correction approaches according to the literature, for instance, CCM, spectral decomposition, percentage of NIR estimation and substitution, among others. After such processes the resultant images are free of the NIR component. Generally, this set of processes have been evaluated by using images from controlled scenarios and some times with the NIR percentage of crosstalking provided by the sensor manufacturers. Contrary to the state-of-the-art methods proposed in the literature for color restoration, this Chapter tackles the color restoration by using outdoor images with sufficient sunlight radiation (like images presented in Fig. 4.1), without any information provided about the NIR percentage in the RGB

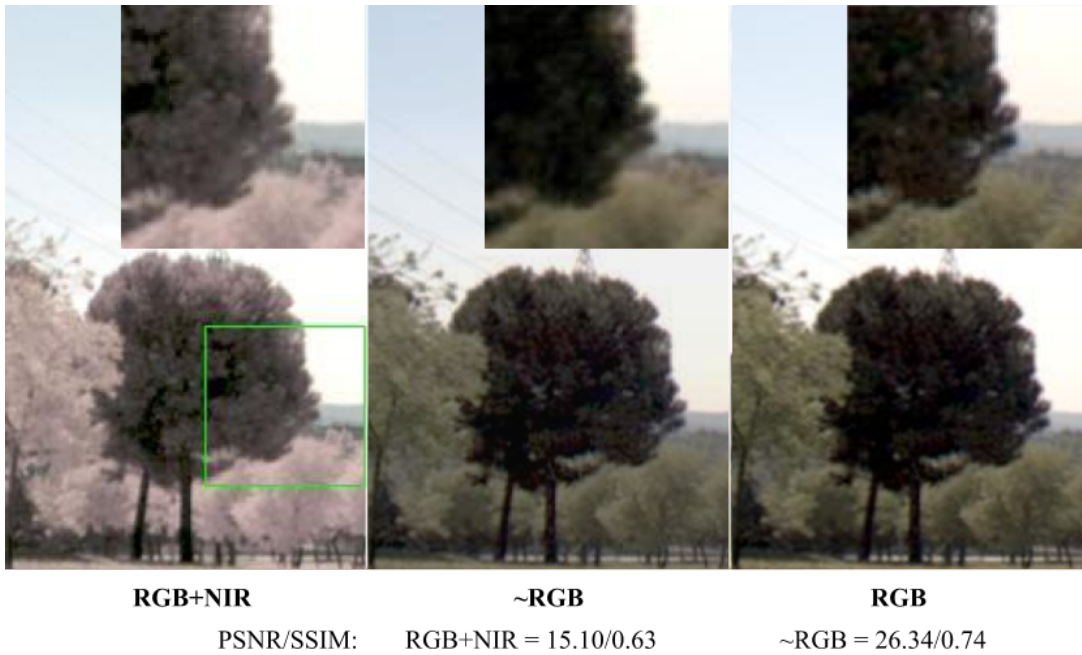


Figure 4.3: RGB+NIR image (infected with near infrared); ~RGB is the image predicted by the proposed CCN based model for color correction; RGB is the target image used for training the DL model and also used for the evaluation.

channels, but more importantly, by only using Neural Network (NN) or Convolutional Neural Network (CNN) approaches to subtract the additional NIR information in the visible image. In other words, the task is accomplished by using just deep learning (DL) techniques, avoiding the set of processes proposed in most of the state-of-the-art techniques.

For the purpose of color correction based on DL techniques, a large dataset of RGB+NIR with its corresponding target RGB image is needed. To the best of our knowledge, OMSIV dataset (see Section 3.2.1) is the only one with the characteristic detailed above (i.e., outdoor images, sufficient sunlight radiation, multispectral paired image data). Therefore, the evaluation results from the CNN based approach is deeply study in OMSIV test dataset by using quantitative and qualitative image similarity metrics. A sample of the result obtained from the proposed CNN method is presented in Fig. 4.3, $\sim RGB$ is the RGB image free of NIR component predicted by the CNN model. As appreciated in the zoom-in images, the difference between these images ($\sim RGB$ and RGB) is almost imperceptible.

In summary, the main contribution in this chapter is two-fold:

- A multilayer perceptron model is proposed to evaluate the possibility of removing NIR information from RGB+NIR images.
- Two custom CNN architectures are proposed to generate a RAW RGB image from its

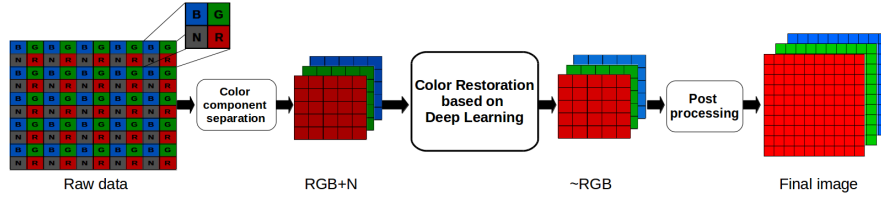


Figure 4.4: Pipeline of the color restoration processes proposed in this Chapter. The rectangle "Color Restoration based on Deep Learning" stands for the different approaches proposed for the color restoration from an RGB+NIR image.

corresponding RAW RGB+NIR. The code of these fully trained networks is available through https://github.com/xavysp/color_restorer.

4.2 Methods

The methods proposed to reach a \hat{I}_{rgb} from an $I_{rgb+nir}$ are described in this section. Firstly, a multilayer perceptron based approach is introduced and then two Convolutional Neural Network based approaches are described. Figure 4.4 shows the pipeline followed to perform the color restoration processes. It consists of the following stages. Firstly, the mosaic RAW data is sub sampled into the respective channels $I_{rgbn} = [R_{vis+nir}, G_{vis+nir}, B_{vis+nir}, N_{nir}]$ (vis: visible band), $I_{rgbn} \in \mathbb{R}^{w \times h \times 4}$, where w and h are image width and height size, and 4 the number of image's channels. Then, the RGB channels, affected by NIR spectral information ($I_{rgb+nir}$), are selected as input data ($I_{rgb+nir} = [R_{vis+nir}, G_{vis+nir}, B_{vis+nir}]$) for the proposed models. The NIR channel, I_{nir} is not considered in the CCN model but it is used in the NN method; that is, the inputs of the NN based method are four pixels, three from the RGB+NIR and one pixel from the corresponding NIR component. Next, the color restoration process is applied. Finally, the output (\hat{I}_{rgb}) from the proposed approaches is post-processed with the image processing techniques tackled in Sec. 3.2. Note that after the raw mosaic data separation the image becomes half size; this image is kept like this by two things: i) just the information provided by the camera sensors is considered (without any interpolation estimation); ii) when information from the NIR channel is obtained, it corresponds to the half image size, hence to make a study of the NIR behavior in the RGB channel we do not want pixel estimations, therefore we keep the RGB image size as the NIR image size.

4.2.1 Neural Network Based Approach

The removal of the NIR information from the RGB channels of a multispectral single sensor camera is formulated as a regression problem. This regression problem is solved by using a neural network (NN) defined by two hidden layers with ten neurons each. The network is trained to learn a mapping function $\Omega : \mathbb{R}^4 \rightarrow \mathbb{R}^3$ that maps a pixel color from an $I_{rgb+nir}$ to an Y_{rgb} ($Y_{rgb} = I_{rgb}$) pixel value of the same scene, but obtained without NIR information.

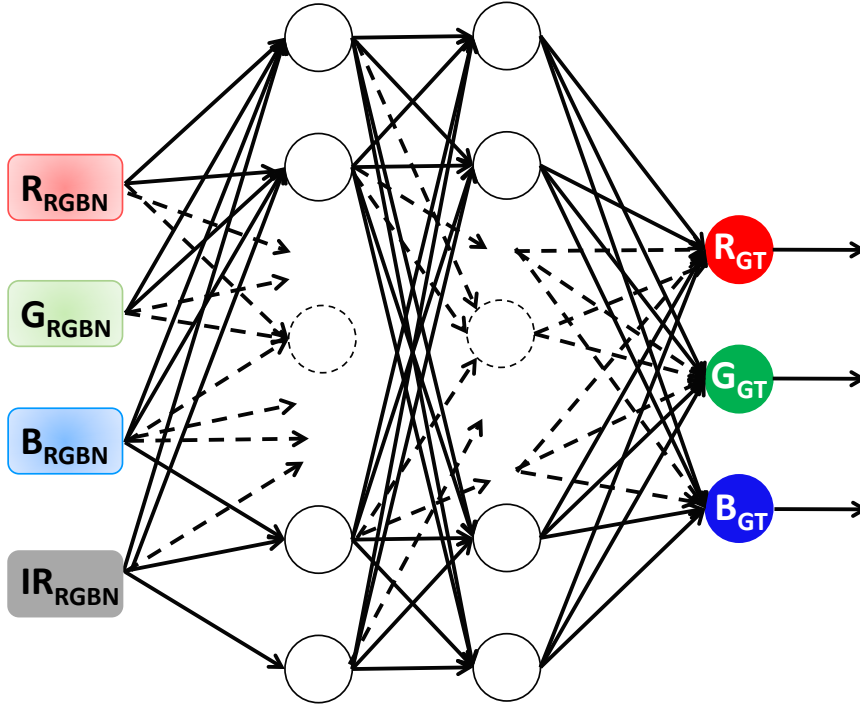


Figure 4.5: Illustration of the NN architecture used to learn the mapping function Ω .

The network model's input consists of a tuple $I_{rgb} = \{R_{vis+nir}, G_{vis+nir}, B_{vis+nir}, N_{nir}\}$ that represent 4 pixel values. The model has two hidden layers of ten neurons each and produces an output tuple $\hat{I}_{rgb} = \{\hat{R}_{vis}, \hat{G}_{vis}, \hat{B}_{vis}\}$ that estimates the R,G,B values of the RGBN SSC where the NIR information has been filtered.

To train the architecture the smooth L1 loss function is considered, which is defined as:

$$loss(x, y) = \frac{1}{P} \sum \begin{cases} 0.5 * (x_i - y_i)^2 & \text{if } |x_i - y_i| < 1 \\ |x_i - y_i| - 0.5 & \text{otherwise} \end{cases} \quad (4.1)$$

where, x correspond to the set of all $I_{rgb+nir}$ pixel values in the training set, and y represents the set of the corresponding pixels from the ground truth (I_{rgb}).

The network has been trained using Stochastic Gradient Descent with the following parameters: learning rate 1, momentum 0.17, weight decay 1e-5, and batch size of 128. Figure 4.5 presents just an illustration of the proposed architecture.

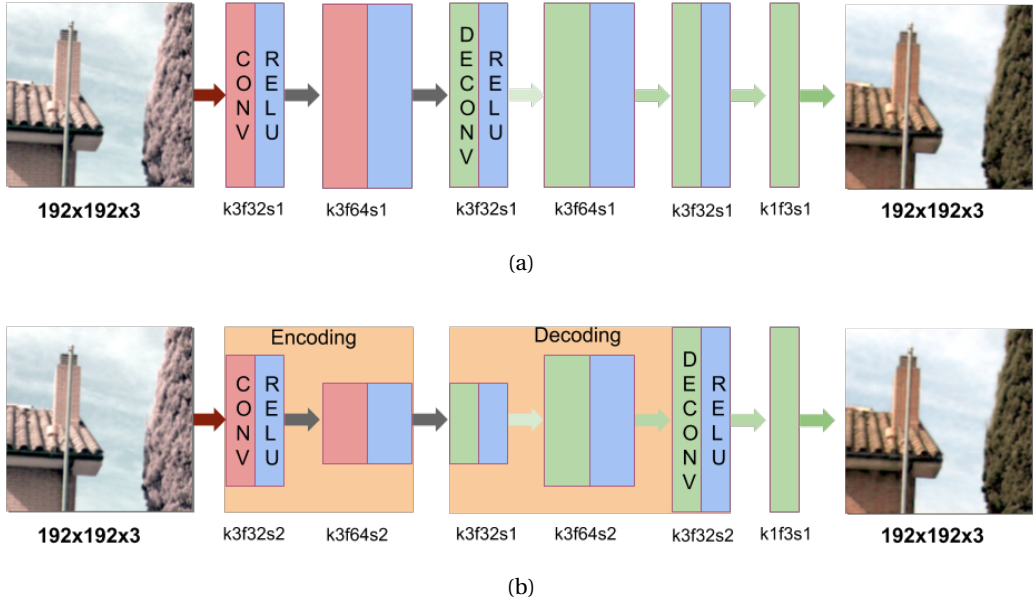


Figure 4.6: Illustration of the proposed deep learning architectures for RGB color restoration from multispectral SSC images: (a) CDNet and (b) ENDENet. CONV refers to convolution, DECONV to deconvolution and RELU is the non-linear function used for the layers in the respective illustration. The term "k3f32s2" refers to: k = kernel size (3×3), f = feature size (32) and s = size of stride (2,2), the same notation is used through the illustration.

4.2.2 Convolutional Neural Network Based Approaches

The previous NN based approach, although was able to remove NIR component given $I_{rgb+nir}$ images, it has some limitations due to the reduced size of layers and neurons. Whenever the layers and neurons are augmented in this NN approach the hardware consumption increase. To overcome this limitation, two more evolved CNN based approaches are presented below.

The first one consists of a Convolutional and Deconvolutional Neural Network (CDNet) that is formed by two and four layers respectively, see Fig 4.6a. As in the Sec. 4.2.1, CDNet attempts to clean the NIR infection in the RGB channels ($I_{rgb+nir}$), therefore, in the first two convolutional layers, the input $I_{rgb+nir}$ is filtered with the aim to clean NIR information, then, in the Deconvolution part, it reconstructs the whole characteristics of the input data but without NIR spectral information. Therefore, the output layer gives predicted image (\hat{I}_{rgb}) supervised with the ground truth image (Y_{rgb}), summarizing, $\hat{I}_{rgb} = CDNet(I_{rgb+nir}, Y_{rgb})$. Where $I_{rgb+nir} = [R_{vis+nir}, G_{vis+nir}, B_{vis+nir}]$, $Y_{rgb} = [R_{vis}, G_{vis}, B_{vis}]$ and $\hat{I}_{rgb} = [\hat{R}_{vis}, \hat{G}_{vis}, \hat{B}_{vis}]$. In the proposed CDNet architecture, five layers have RELU activation functions and the last one is just a fully deconvolutional layer. The architecture remains the same *height* \times *width* size but not in its numbers of feature maps which are as follow [32, 64, 32, 64, 32, 3], 64 and 32 units are used as recommended in [26] for efficiency with hardware limitation. To estimate the network parameters and minimize the loss in \hat{I}_{rgb} , the Mean Square Error (MSE) was considered as

in the state-of-the-art techniques for restoration. Moreover, to optimize the CDNet training, AdamOptimizer with learning rate 0.0003 was considered. In terms of the number of layers, two convolutions followed by four deconvolutions have been used; these values have been obtained after evaluating different configurations trying to avoid image deterioration during the convolution stages and trying to get the best result after deconvolution stages.

In addition to the CDNet presented above, a second architecture has been proposed to evaluate the capability of NIR removal from the given color infected images. This second architecture is similar to the CDNet but with encoder and decoder in the convolution and deconvolution layers, respectively. This new architecture is referred to as Encoder-Decoder Neural Network (ENDENet). Figure 4.6b presents an illustration of this network architecture and gives details of the set up of the layers. As depicted in Fig. 4.6b, the first, second, fourth and fifth layers have a (2, 2) stride, hence image size is encoded from 192×192 to 48×48 and decoded to the same size as in the $I_{rgb+nir}$ size. The proposal of ENDENet lies on the usage of fewer parameters than CDNet but still score competitive results in the quantitative evaluation. This comparison is presented in Sec. 4.3.2.

Loss Function

Like in other kinds of models based on DL [36], either for CNN or NN, it is necessary to prepare the data for the training, and regularize and optimize the architecture. In addition to the architecture, the loss function or cost function (\mathcal{L}) is an essential part of supervised learning. It measures the errors between the prediction made by the model with respect to the given ground truth. In other words, after every iteration, the computation of the loss function is performed to evaluate how well is going through the training process. Within the context of our application (color restoration), for the minimization of \mathcal{L} the Mean Square Error (MSE) is considered, which is computed as follow:

$$\mathcal{L} = MSE(Y_{rgb}, \hat{I}_{rgb}) = \frac{1}{P} \sum_{i,j}^{w,h} (Y_{i,j} - \hat{I}_{i,j})^2, \quad (4.2)$$

where (w, h) are the image width and image height respectively, P is the total number of image pixels, (i, j) are the pixel indices, \hat{I}_{rgb} is the image predicted by the network and Y_{rgb} is the ground truth corresponding to the given $I_{rgb+nir}$. Finally, the MSE value, which is used in the loss function, is considered to minimize or optimize with an objective function [60].

4.3 Experimental Results

This section presents results obtained with the two approaches described above. Firstly, qualitative results from the NN based approach are summarized. Secondly, results from the CNN based methods are presented.

4.3.1 Neural Network Based Approach

The NN presented in Sec. 4.2.1 has been trained and evaluated with the OMSIV dataset acquired with the pair of cameras presented in Fig. 3.1 (see dataset details in Sec. 3.2.1). Note that although the cameras have been rigidly attached, trying to place their optical axis as parallel as possible, the obtained images need to be registered. The registration process is needed in order to guarantee a pixel-to-pixel correspondence. Differences due to camera disparity are neglected since cameras' optical axis are quite near in comparison to the depth of the objects in the scene (actually, the dataset has been created having in mind this assumption, so that scenes containing objects far away from the camera have been considered). The Y_{rgb} and $I_{rgb+nir}$ images have been registered using the Matlab Image Alignment Toolbox (IAT)¹. From the 89 pairs dataset, 71 pairs have been used for both the network training and for computing color correction matrix (M_{CC}); the remainder 18 pairs were used for validating the results. A qualitative validation (just 18 pairs) has been performed comparing the results from the neural network with those obtained by using a naïve color correction based approach.

In order to qualitatively evaluate the obtained result, a naïve approach, based on the usage of color correction matrix is performed. The MCC is obtained by minimizing the square error between the RGB and the corresponding RGBN+NIR images; M_{CC} is obtained by minimizing the following error function:

$$E = \sum_{i=1}^{i=P} (Y_{rgb} - \mathbf{M}_{CC} \times I_{rgb+nir})^2, \quad (4.3)$$

where Y_{rgb} corresponds to the ground truth values (elements from the camera with infrared cut off filter); $I_{rgb+nir}$ corresponds to the pixel values obtained with the SSC; \mathbf{M}_{CC} is the color correction matrix (a square matrix of 3×3 elements); and P represents all the pixels from all the images used to estimate (E). In order to do a fair comparison the same set of images (71 pairs of images of 256×256 pixels) used to train the network has been considered; in other words (E) is computed by considering $P = 71 \times 256 \times 256 = 5,439,488$ elements. Actually, the images used in the NN training and evaluation processes were a subset from the OMSIV dataset (see Sec. 3.2.1).

Figure 4.7c presents just three illustrations of the results obtained with the proposed neural network, which can be compared with the corresponding ground truth presented in Fig. 4.7d. As it can be appreciated, although the network architecture is a perceptron with 2 hidden layers and each one with 10 neurons, the results obtained with the proposed approach look quite similar to those from the ground truth. On the contrary, the result obtained from the color correction matrix (see Fig. 4.7b) are not so similar to those from the ground truth. As a conclusion from these results we have identified two ways to improve them. The first one is related with the dataset. A larger dataset, with more color variability, obtained at different daylight time is required to learn a more accurate mapping function by the network. The

¹<http://iatool.net/>

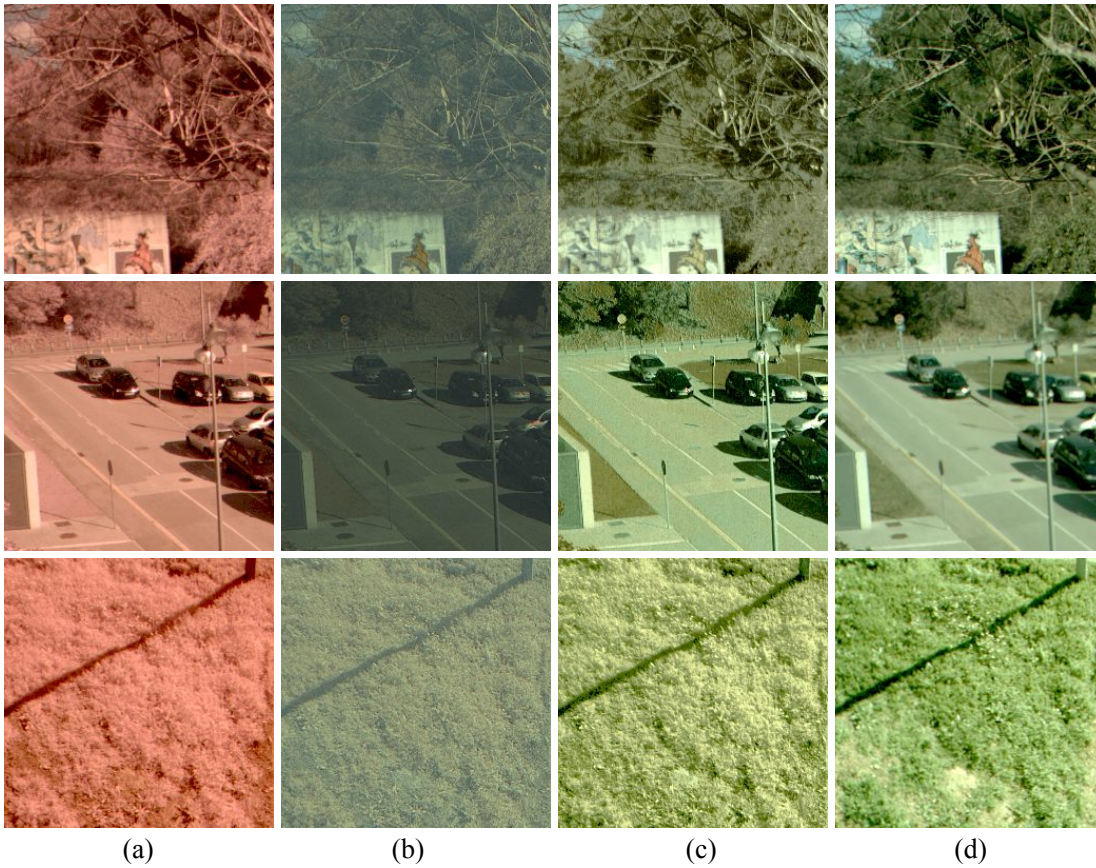


Figure 4.7: Illustrations of: (a) Original RGB+NIR images; (b) Results obtained from the MSE color correction; (c) Results from the proposed approach; (d) RGB ground truth images obtained by using the ICF.

second improvement is related with the used architecture, other configurations should be evaluated (more hidden layers and more neurons per layer). In Fig. 4.7 just three results from the 18 pairs used for validation are presented, the whole set of images used for validation can be downloaded from: http://www.cvc.uab.es/~asappa/Results_for_Validation.rar.

4.3.2 Convolutional Neural Network Based Approach

System Setup

The effectiveness of the different approaches is assessed using the whole OMSIV dataset. OMSIV dataset has been obtained with a SSC that captures at the same time visible and near infrared bands, just outdoor scenes with vegetation have been acquired in different sunlight conditions as well as challenging environments in urban scenarios; for the purpose of the ground truth, another SSC with ICF is used, then pairs of acquired images are registered [31],

for a more detailed description of OMSIV see Chapter 3. Since there is a high bands crosstalk in images, a big volume of training data is required to train the proposed models. From the OMSIV dataset 500 images have been used for training and validation, and 32 for testing; the given images are of 256×256 pixels. A data augmentation process has been used to increase the number of patches to train both networks. From this data augmentation process 12599 patches have been obtained, 12371 were taken for training and 128 for validating; for the model evaluation purpose the remainder 32 images were augmented to 128. A subset of 15 images is presented and analyzed in Sec. 4.3.2. During the data augmentation process, patches of (192×192) were obtained from the original images.

In order to obtain the best performance of the architectures presented in Fig. 4.6, different configurations were tested by adding and removing layers of the encoding and decoding stages [140]. Additionally, filters with sizes (3×3) , 5×5 and 7×7 were evaluated in both architectures in order to find the best one. The filter with size (3×3) was selected since when large filter sizes were considered the texture of \hat{Y} is smoothed. A 1×1 filter was added to the last layer in ENDENet as well as CDNet. Tensorflow library on a TITAN X GPU was considered to the training stage. Note the MSE values are normalized in a $[0-255]$ scale; the loss function converged to a MSE value of 50 in both models. On average, the training stage took about 6 days. The ENDENet, due to its different layer sizes, took about 1.45 minutes for each epoch, while the CDNet took 2 minutes.

Evaluation Metrics

The general metrics used for the images similarly evaluation are Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) [145]. However, for further comparison, Angular Error (AE) and the Color Difference of the version of 2000 (ΔE_{00}) [82] have been also considered. On the one hand, PSNR is defined as follow:

$$MSE = \frac{1}{P} \sum_{i=1}^P (Y_{rgb}(i) - \hat{I}_{rgb}(i))^2, \quad PSNR = 10 \times \log_{10} \left(\frac{max^2}{MSE} \right), \quad (4.4)$$

where P is the number of pixels in Y_{rgb} and \hat{I}_{rgb} , i is the index of pixel location, MSE is the mean square error. On the other hand, SSIM considering \hat{I}_u and \hat{I}_σ , which are the average and standard deviation of \hat{I}_{rgb} , respectively, compute the luminance $lumi(Y_{rgb}, \hat{I}_{rgb})$, contrast $cont(Y_{rgb}, \hat{I}_{rgb})$, and structural $stru(Y_{rgb}, \hat{I}_{rgb})$ measurements, where the final $SSIM = [lumi]^\alpha \times [cont]^\beta \times [stru]^\gamma$ is computed as:

$$lumi = \frac{2 \times \hat{I}_u \times Y_u + c1}{\hat{I}_u^2 + Y_u^2 + c1}, \quad cont = \frac{2 \times \hat{I}_\sigma \times Y_\sigma + c2}{\hat{I}_\sigma^2 + Y_\sigma^2 + c2}, \quad stru = \frac{\sigma_{\hat{I}_Y} + c3}{\hat{I}_\sigma + Y_\sigma + c3}, \quad \text{where} \quad (4.5)$$

$$\sigma_{\hat{I}_Y} = \frac{1}{P-1} \sum_{i=1}^P [\hat{I}(i) - \hat{I}_u(i)][Y(i) - Y_u(i)].$$

$\alpha = \beta = \gamma = 1$, while c_1, c_2, c_3 are constants to avoid instability.

The Angular Error is defined as follow:

$$AE = \cos^{-1} \left(\frac{\text{dot}(\hat{I}_{rgb}, Y_{rgb})}{\text{norm}(\hat{I}_{rgb}) * \text{norm}(Y_{rgb})} \right), \quad (4.6)$$

as other similarity measures [145], AE is computed over every single pixel of a pair of images improving MSE values; even though PSNR, SSIM and AE give results for the pixel level, the definition above assume to be for the image, this value is the average of the results from all the pixels.

ΔE_{00} is computed by transforming \hat{I}_{rgb} and Y_{rgb} to the CIELAB color space (\hat{I}_{LAB}, Y_{LAB}); this color similarity measure is commonly used in the color research field and it is a key evaluator to see the results of restoration in outdoor RGB+NIR images with vegetation, because as mentioned in Sec. 4.1 the presence of NIR in vegetation desaturate the normal human color vision. For each pair of image pixels $\Delta E_{(i,j)}$ is obtained as follows.

$$\begin{aligned} \Delta E_{(i,j)} = & \left[\left(\frac{\Delta L'}{k_L S_L} \right)^2 + \left(\frac{\Delta C'}{k_C S_C} \right)^2 + \left(\frac{\Delta H'}{k_H S_H} \right)^2 \right. \\ & \left. + R_T \left(\frac{\Delta C'}{k_C S_C} \right) \left(\frac{\Delta H'}{k_H S_H} \right) \right]^{\frac{1}{2}}, \end{aligned} \quad (4.7)$$

where $\Delta L', \Delta C'$ and $\Delta H'$ are the CIELAB lightness, chroma and hue differences, respectively; S_L, S_C , and S_H are weighting functions for the lightness, chroma, and hue components. k_L, k_C and k_H are factors to be adjusted according to different viewing parameters; the R_T function is intended to improve color-difference equation for fitting chromatic differences in the blue region (for a detailed description see [82]). Once computed $\Delta E_{(i,j)}$ for the whole set of pairs of pixels, the ΔE_{00} represents the percentage of pixels with a ΔE value in between two given thresholds [a,b]. More specifically, in the current work ΔE_{00} represents the percentage of pixels in the range between [0,10]. ΔE values higher than 10 correspond to pixels with a color difference easily detected by human eyes (see [82] for more details). The ΔE_{00} is obtained as follow:

$$\Delta E_{00} = \frac{sp_{[a,b]} \times 100}{W \times H}, \quad (4.8)$$

where (W, H) values correspond to the image width and image height; $sp_{[a,b]}$ is the total

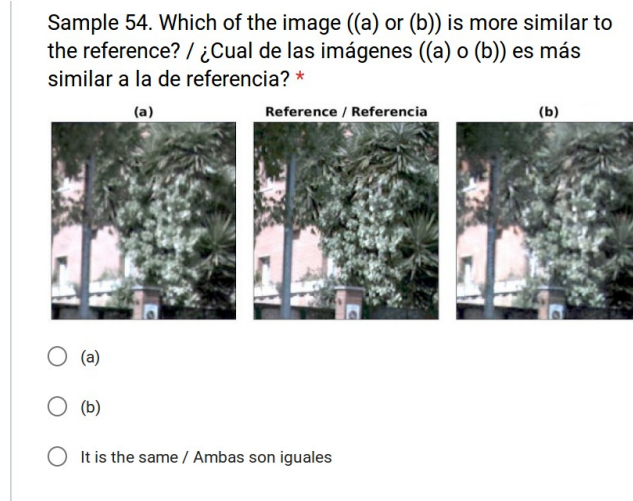


Figure 4.8: Snapshot of the Human Eye Perception test (HEP) performed over 15 images to evaluate the results from CDNet and ENDENet.

number of pairs of pixels whose $\Delta E_{(i,j)}$ values are in the range of [a,b]; it is obtained as:

$$sp_{[a,b]} = 0, \quad sp_{[a,b]} = \sum_{i,j}^{W,H} \begin{cases} sp_{[a,b]} + 1 & \text{if } a \leq \Delta E_{(i,j)} \leq b \\ sp_{[a,b]} & \text{otherwise} \end{cases} \quad (4.9)$$

as mentioned above, in the current work [a,b] have been set to [0,10].

For further evaluation purposes in [69] and [22], in addition to the quantitative metrics, the authors consider evaluations from human perception to quantify the ability of different approaches to perceptually reconstruct an image. In the current work, qualitative evaluations have been also considered over a subset of images. This subset of images was selected for human perceptual judgment, referred to as Human Eye Perception (HEP). Basically, HEP is a survey made by different users to know which of the images, obtained by CDNet and ENDENet, are more similar to the ground truth. For this HEP test, 15 images have been selected from the best (5 images), the average (5 images) and the worst (5 images) cases, according to ΔE_{00} in CDNet results; ΔE_{00} was selected since it is generally used in the color related literature to evaluate results by human perception. It should be mentioned that the same set of images would result if the ENDENet were considered. These 15 images were evaluated by 50 subjects. They were asked "which of the images ((a) or (b)) is more similar to the reference?"; in case the images were perceptually the same the subjects could give this answer, "It is the same". Figure 4.8 shows the snapshot of the GUI used to evaluate the performance of CDNet and ENDENet.

4.3. Experimental Results

Methods	PSNR		SSIM		AE		ΔE_{00}	
	average	median	average	median	average	median	average	median
<i>I_{rgb+nir}</i>	13.530	13.613	0.563	0.566	32.069	32.024	24.13	16.395
[96]	20.284	20.508	0.621	0.661	8.691	8.737	76.726	85.227
[139]	18.007	17.674	0.547	0.553	8.048	8.281	64.558	73.717
SRCNN [26]	20.928	21.324	0.684	0.708	8.271	7.723	75.182	82.122
ENDENet	24.069	25.409	0.777	0.809	5.561	5.303	86.563	95.893
CDNet	23.544	24.447	0.767	0.792	5.171	5.098	86.008	93.313

Table 4.1: Average and median values of the methods tested in the current work when the whole dataset (128 samples) was considered.

Quantitative and Qualitative Results

This section firstly details the state-of-the-art approaches (i.e., [96], [139] and [26]) used for comparing results with the proposed approaches. Then, quantitative and qualitative results from the proposed approaches together with the corresponding comparisons are presented.

In [96] the authors presents three methods, which where evaluated with different demosaicing algorithms and linear mappings. In the current work, following the suggestion from [96] and according to the type of sensor used in our experiments [132], a uniform RGB-NIR pattern [19] is used, then demosaicing is performed using [91] and finally the color correction is obtained through a polynomial linear mapping [48], since this combination produces the best performance. The second algorithm used to compare the proposed approach has been presented in [139]. This algorithm has been described in section 2.3; from the two MFA patterns presented in [139] the one with best result has been selected. Finally, the third comparison has been performed with respect to a super-resolution CNN based approach [26], termed as SRCNN. This super-resolution approach has been selected for comparisons since it shares similar purposes when trying to preserve features and edges in the reconstructed images. This SRCNN approach has been detailed in section 2.3 and the pipeline followed the same process as the proposed approaches (see Fig. 4.4). It should be mentioned that just for comparison purposes all resulting images are resized to the same size; it mean that images resulting from [96] and [139], which are full size images, are down-sampled with bicubic interpolation up to half their size in both height and width. Regarding images resulting from SRCNN and our approaches (CDNet and ENDENet), they are already half size in both height and width. The down-sampling process was performed to the final resulting image. It should be also mentioned that similar comparative results are obtained in the other way around; it means, if the comparisons were performed at full resolution by up-sampling results from SRCNN and our approaches (CDNet and ENDENet).

Image #	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>	<u>13</u>	<u>14</u>	<u>15</u>	#t
Method	PSNR															#t
$I_{rgb+nir}$ [96]	5.27	18.39	13.96	18.29	14.42	15.73	17.41	17.10	5.72	11.22	13.41	15.73	17.12	9.12	8.10	-
[139]	20.69	29.69	29.03	28.30	21.03	20.50	16.77	15.49	16.00	15.80	30.64	20.61	10.14	22.91	20.24	3
SRCNN [26]	28.02	26.99	29.38	20.09	15.09	19.05	14.00	15.99	15.37	15.98	18.39	17.73	8.47	20.27	17.98	0
ENDENet	17.95	25.17	23.59	23.97	25.54	20.68	17.70	23.55	18.25	19.54	9.63	17.79	12.45	24.57	11.17	2
CDNet	20.56	31.51	31.90	30.05	32.08	22.86	25.76	22.87	23.58	14.59	8.15	18.02	8.76	24.84	9.73	6
	31.89	30.96	33.67	26.21	30.32	23.55	22.88	20.76	19.21	22.04	9.68	19.17	9.73	11.38	10.86	4
Method	SSIM															#t
$I_{rgb+nir}$ [96]	0.61	0.67	0.77	0.59	0.63	0.45	0.49	0.62	0.38	0.41	0.30	0.43	0.63	0.55	0.44	-
[139]	0.79	0.86	0.78	0.79	0.67	0.59	0.33	0.42	0.48	0.57	0.84	0.61	0.63	0.77	0.71	3
SRCNN [26]	0.80	0.76	0.73	0.59	0.54	0.50	0.25	0.42	0.46	0.46	0.53	0.45	0.52	0.71	0.64	0
ENDENet	0.87	0.81	0.73	0.64	0.79	0.64	0.42	0.74	0.54	0.65	0.37	0.59	0.68	0.78	0.44	2
CDNet	0.93	0.88	0.82	0.77	0.86	0.61	0.68	0.85	0.74	0.57	0.26	0.57	0.58	0.85	0.47	8
	0.93	0.88	0.82	0.77	0.86	0.61	0.68	0.79	0.63	0.64	0.29	0.58	0.64	0.65	0.52	2
Method	AE															#t
$I_{rgb+nir}$ [96]	33.10	29.90	29.10	31.09	36.19	29.35	33.11	32.32	31.91	29.18	28.91	27.80	34.82	33.58	31.51	-
[139]	3.55	8.06	9.11	9.54	10.66	8.26	12.49	10.53	8.78	4.53	6.83	9.50	5.46	7.33	7.31	0
SRCNN [26]	1.62	6.29	8.40	9.91	10.50	7.45	11.98	9.05	7.31	3.74	5.74	9.53	5.03	5.96	8.44	1
ENDENet	2.88	7.14	18.33	6.55	7.84	8.50	9.92	9.46	8.85	5.55	7.30	11.47	4.68	6.12	8.75	0
CDNet	1.82	5.27	8.40	4.91	4.15	6.87	8.31	5.46	7.94	4.13	4.56	9.19	2.71	3.56	4.65	3
	1.41	5.34	8.22	4.37	4.10	7.27	6.58	5.29	6.78	4.44	4.61	8.84	2.21	3.36	4.44	11
Method	ΔE_{00}															#t
$I_{rgb+nir}$ [96]	3.88	30.65	84.48	5.83	1.92	3.78	34.71	38.91	4.54	0.68	1.00	3.74	2.92	19.12	9.68	-
[139]	99.43	99.45	99.82	94.91	95.64	85.16	77.36	66.46	57.30	47.85	99.65	70.49	18.61	95.33	49.76	5
SRCNN [26]	99.56	99.38	99.53	89.77	22.01	78.15	44.73	56.82	40.98	66.50	58.87	50.95	8.29	87.80	47.94	0
ENDENet	94.48	98.88	96.89	89.11	95.65	50.59	60.12	84.17	67.29	77.54	1.46	21.38	21.38	92.53	20.76	0
CDNet	99.47	99.51	99.91	99.05	99.34	84.34	99.32	93.10	87.42	60.50	1.61	33.42	13.03	98.20	22.25	4
	99.82	99.41	99.95	99.07	99.37	90.37	89.62	88.89	83.08	88.01	3.28	52.54	18.04	39.40	26.36	6
Method	HEP															#t
SAME	32	8	5	9	15	8	22	16	11	13	5	6	28	29	16	3
ENDENet	13	36	40	22	5	10	4	15	4	4	37	12	11	4	12	4
CDNet	5	6	5	19	30	32	24	19	35	33	8	32	11	17	22	8

Table 4.2: Results for a subset of 15 images from [132] (see details in Sec. 4.3.2). Underlined images are depicted in Fig. 4.9 and Fig. 4.10 for qualitative evaluation. Boldface values correspond to the best performance; #t corresponds to the number of times a given algorithm obtain the best performance. The section of HEP contains the qualitative evaluation (results from the survey), each value corresponds to the number of users that selects this result as the best one.

4.3. Experimental Results

The OMSIV dataset presented in Chapter 3 has been used to evaluate the performance of the two proposed architectures, as well as to compare the results with the aforementioned state-of-the-art approaches. As can be appreciated in Table 4.1, both network architectures have similar results. However, when PSNR, SSIM, and ΔE_{00} metrics are considered, ENDENet gets the best results. CDNet outperforms ENDENet just when AE is considered. In all the cases the proposed architectures outperform the three algorithms from the state of the art evaluated in the current work.

In order to analyze more in detail results from the two proposed architectures, a subset of 15 representative images, has been picked up as mentioned above. Table 4.2 presents results from the quantitative and qualitative metrics SSIM, PSNR, AE, ΔE_{00} and HEP. The values of AE are like MSE, the smallest the better; while with the other metrics, the largest the better. Additionally, in the first row of each section in the table, similarity values computed between the original image ($I_{rgb+nir}$) and the Y_{rgb} are provided to be used just as a reference. Figures 4.9 and 4.10 present qualitative results for the images underlined in the table.

Since the images in the subset correspond to the best, worst and average cases, they cannot be used to identify which network has the best performance. These quantitative results are presented just to appreciate the improvement with respect to the state of the art. Actually, as can be appreciated in Table 4.2, ENDENet and CDNet have similar results. Taking into account the number of times a given network architecture obtains the best result, out of the 15 images, CDNet outperforms ENDENet if AE or ΔE_{00} evaluation metrics are considered. On the contrary, if PSNR and SSIM are considered the ENDENet network obtains the best results.

As mentioned above, in addition to the quantitative evaluation, a test using 50 subjects has been performed over these 15 images. From the 15 evaluations, CDNet generates the best results in 8 cases, while ENDENet in just 4 cases, in the reminder 3 cases both networks produce a similar result.

Figure 4.9 and 4.10, for illustration purposes, present color correction results from the methods evaluated in Table 4.1. Figure 4.9 shows four best results from CDNet according to ΔE_{00} ; while Fig. 4.10 shows a pair of images with an average value and a pair of images where CDNet gets the worst results. The last three rows in Figs. 4.9 and 4.10 correspond to the results obtained with [96], [139] and [26] respectively. The four first rows correspond to original given image ($I_{rgb+nir}$), ground truth, results from CDNet architecture and results from ENDENet architecture.

As can be appreciated in Fig. 4.9 and Fig. 4.10, most of the images in the dataset contain vegetation. According to Table 4.1 and Table 4.2 results from the proposed network architectures achieve highest score in most of the evaluation metrics. For instance, according to the median values (Table 4.1), more than half of \hat{I}_{rgb} samples get a ΔE_{00} value above 90%. In other words, more than 90% of pixels have been restored closely to the corresponding ground truth values, their difference is almost imperceptible by human eyes. In Fig. 4.10, Image #12 corresponds to a challenging scenario where none of the proposed approaches get good results. This bad results are mainly due to the materials present in the scene. More samples

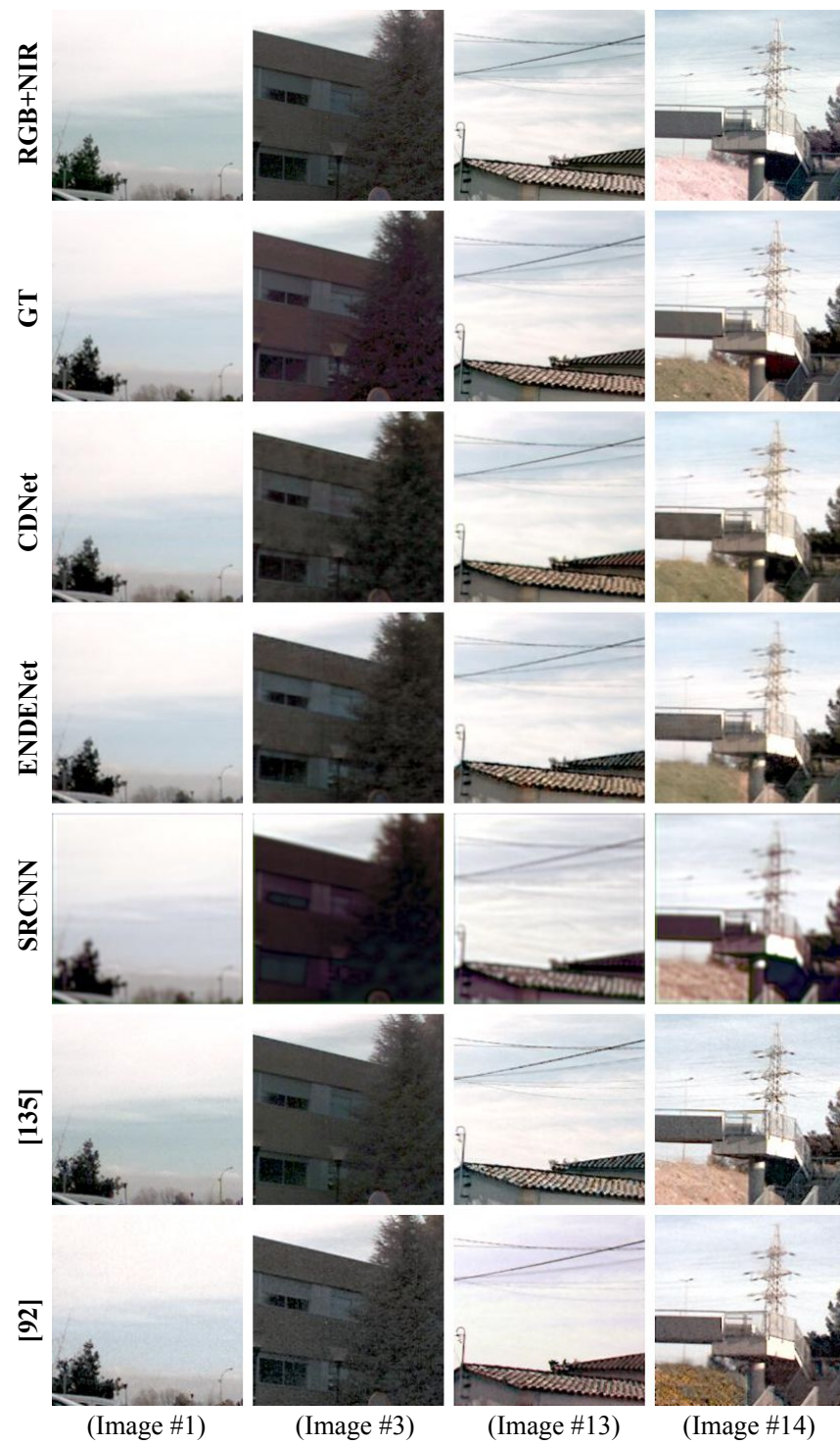


Figure 4.9: Four samples from the best results (see Table 4.1). Image numbers correspond to the values presented in Table 4.2

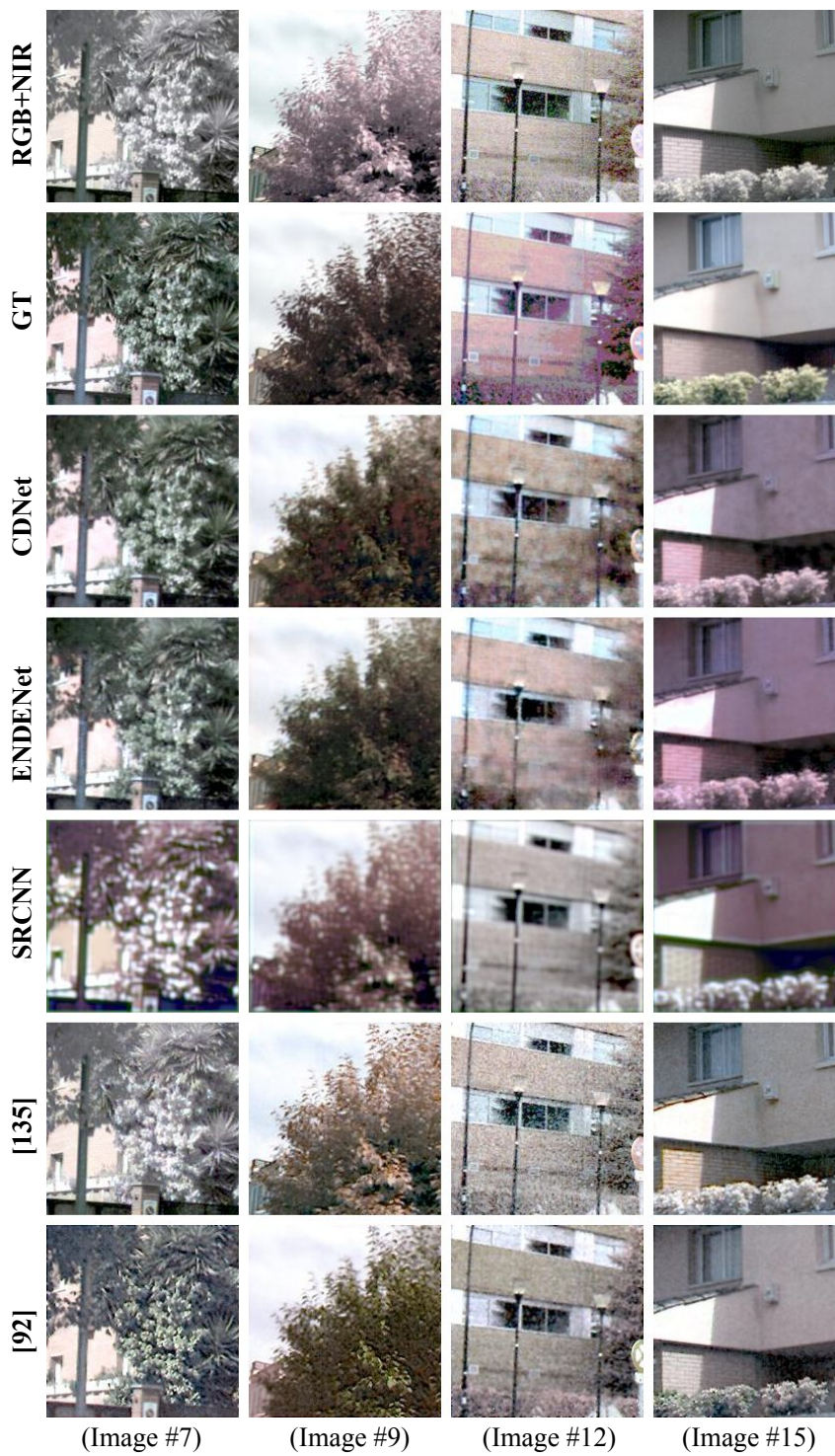


Figure 4.10: Four samples from the average and worst results (see Table 4.1). Image numbers correspond to the values presented in Table 4.2

containing these features and materials will be required to improve the performance.

4.4 Conclusion

This chapter proposes three variants of learning algorithms for RGB color restoration from RGB+NIR multispectral images, specifically using images from the domain of visible and near-infrared wavelength. The first proposal is based on neural networks, although being a small size architecture, it outperforms method based on CCM. The second proposal is based in CNNs. This proposal has two variants of CNNs, although both have the same number of layers, the major difference is on the encoder-decoder layers. The ENDENet is based on the encoder-decoder approach, which even though has a fewer number of parameters outperform the state-of-the-art methods used for comparison. The CDNet, although scored as the second-best one, its results preserve edges better than ENDENet in the RGB predicted images, which are free on NIR component; such edge preservation is noted when a subjective evaluation is conducted because of that CDNet scored the best result.

The results, furthermore, suggest that a multispectral image with sufficient sunlight will be restored to the visible band keeping its main features and it will be almost imperceptible for human eyes. The proposed approaches, either ENDENet and CDNet, learn a mapping function between RGN+NIR to RGB without any pre-processing; these images were collected in RAW format with all of information captured by the single sensor cameras. The proposed architectures might be used to solve similar problems such as deblurring and super-resolution.

5 Edge Detection from a Multispectral Framework

This chapter focuses on the edge detection problem by proposing a novel RGBN multispectral framework that take advantage of the information provided by the NIR channel to improve results. Although the proposed framework is intended to be used with RGBN images, it can be also considered when just RGB images are provided. In the proposed framework two different deep learning based architectures have been developed and evaluated. Although the approach is proposed to tackle the edge detection problem, it can be useful in other related problems such as contour or object's boundary detection. Datasets focused on contour or object's boundary detection, which were not considered during the training process, have been evaluated showing the generalization of proposed approach.

5.1 Introduction

Edge detection is a recurrent task in several image processing and computer vision applications (e.g., segmentation [161], image recognition [124, 155], and in the last boom of Generative Adversarial Network (GAN) in the image-to-image translation [164]). Fields such as: medical image analysis [107], remote sensing [52], oceanography [102], just to mention a few, in most of their heart activities require the usage of edge detection. In spite of the large amount of work on the edge detection problem, it still remains as an open problem with space for new contributions.

It has been almost half a century since Sobel operator [128] has been proposed. Since then, a large number of edge detectors [103] has been presented in the literature and most of the techniques like Canny [18] are still being used in nowadays applications [102]. Recently, in the era of Deep Learning (DL), Convolutional Neural Networks (CNN) models like, DeepEdge [14], HED [150], RCF [76], BDCN [44] among others, have been proposed. These kind of models, which have been reviewed in Section 2.4.4, are capable of producing edge-maps like the ones obtained by Canny [18] or other low level processing edge detectors (e.g., [5]), reaching the state-of-the-art results according to a given edge ground truth. The success of these methods is mainly by the application of numerous CNN filters at different scale levels to a large set

of images, together with regularization techniques and hyperparameters tuning. The main drawback of these techniques lies on the fact that they work just on specific datasets and secondly they are not focused on edge detection as it will be explained below.

Before starting with the proposed approaches, and to avoid misunderstandings, let's review the differences between edge, contour and boundary detection tasks. Edge detection, back in 1980s, was defined as the intensity changes on neighbour pixels of the given image, which is produced by surface discontinuities, reflectance, or illumination boundaries [18, 88]; later in [165], defined in the ambit of computer vision as the process that attempts to detect the significant properties of objects in the scene captured by an image sensor; those properties include discontinuities in the photometrical, geometrical and physical characteristics of objects. Regarding contour and object's boundary detection, they are related tasks, which some times are treated as synonymous of edge detection, but being precise they are representative features of the objects in the image [35] but not necessarily include all the edges from the given image. These features are outlines representing or bounding the shape or form of the objects in the image [94]. On the one hand, boundary refers to the border of an object in the image plane that represents a change in the pixel ownership from one object or surface to another [89]. On the other hand contour refer to the borders of a region of a given object. To summarize, while edge detection is the whole change in the intensity function without taking care of the ownership of the edge and without taking care of defining an open or closed shape, contour and boundary detection tackle the task filtering the edges that not represent a salient feature of the object or shapes from the objects. A detailed description of contour and boundary detection can be found in [38] and [92].

As mentioned above, in this chapter deep learning based approaches are proposed, hence, the dataset used for training become a critical feature for the algorithm success; in spite of the importance of this fact most of the widely used datasets to train and evaluate DL based edge detection algorithms are BSDS300 [89] and BSDS500 [7] (BSDS: Berkeley Segmentation Data Set), which are dataset intended for segmentation—boundary detections. Generally, BSDS300 is used to train and validate during the training stage, while the BSDS500 is just used to test the resulting model. The images from these datasets have at least 5 boundary annotations, in some cases the GT images also contains edges and contours annotations, but these annotations are provided as a plus without additional information. As there are at least five ground truths (GTs), DL based approaches (e.g., [44, 75, 150]) make a GT consensus before the training process. In the consensus stage, a new GT is generated with the provided annotations. A pixel is considered a true annotation in case that three or more annotators have indicated that; otherwise the annotation is discarded. This process helps to filter edges that affect the convergence of the training process [150]. In other words, with this filtering processes, most of the annotations at the edge level are filtered since just few annotators marked them; hence the trained neural network results in a boundary detector. Figure 5.1 shows a couple of illustrations from the BSDS300 and BSDS500 datasets. In these figures, it can be appreciated the different representations obtained if all the annotations are considered or different level of consensus (consensus of two and three are depicted). In general, the consensus of three annotations is considered in most of the approaches (see right column in

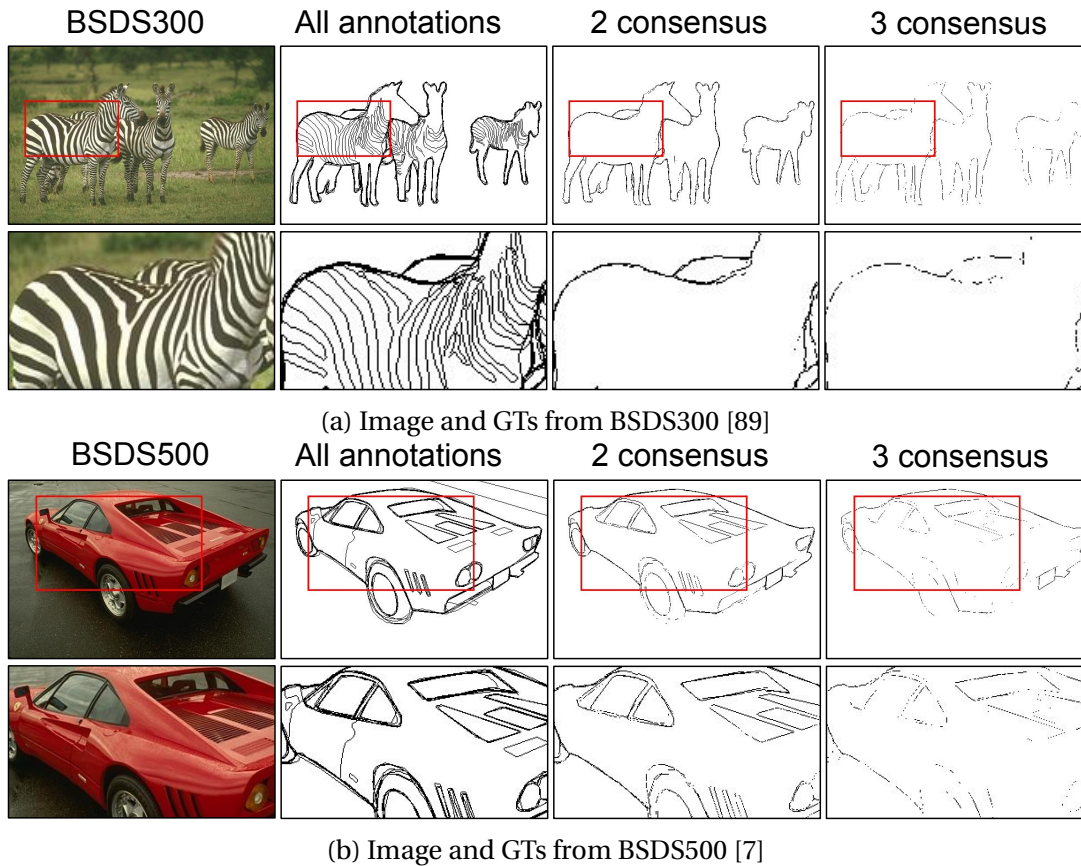


Figure 5.1: BSDS images used for training DL models for boundary detections—the five provided annotations (All) and different level of consensus (2 and 3) are depicted.

Fig. 5.1), which in most of the cases correspond to boundary of the image's content. Hence, DL models trained with such a GT, learn to detect such a boundary as well as to remove other edges from the given image. In other words, CNN based models are very dependent to the type of dataset used during the training process, which do not allow generalization.

Having in mind the aforementioned drawbacks of CNN models, with respect to the non-learning based algorithms, this chapter proposes DL based approaches to the edge detection. As it will be shown, the proposed approaches can be used in any arbitrary image, just like the non-learning based methods do in the edge detection tasks, but by improving their performance. In other words, the proposed networks are trained using just one dataset and then evaluated using all datasets available in the literature of edge detection. In addition, the training of the proposed networks starts from the scratch, without major training settings as usually done in the current DL models for edge, contour and boundary detection tasks. Figure 5.2 shows four end to end predicted edge-maps (actually, CED has been designed for boundary detection) from BSDS500 [7] by the state-of-the-art deep learning models. While

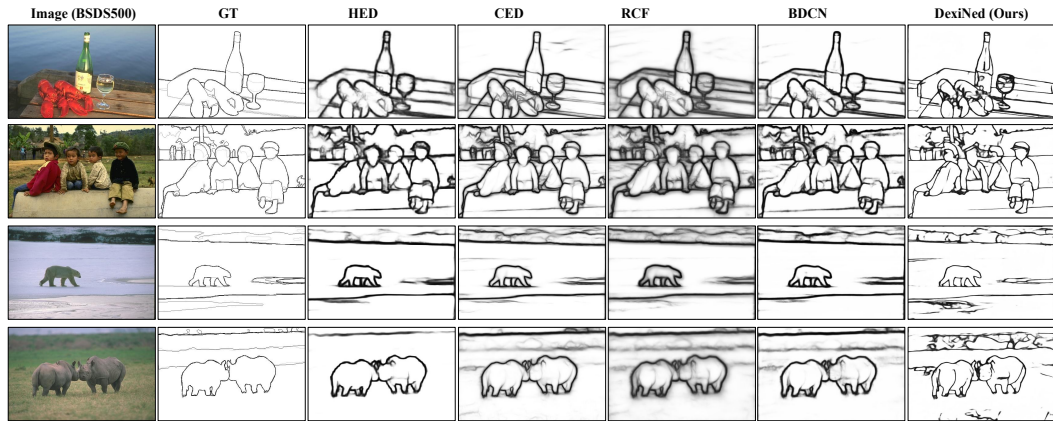


Figure 5.2: Results from the state-of-the-art algorithms and the proposed methods (DexiNed). Note that HED [150], CED [143], RCF [75], and BDCN [44] have been trained with the BSDS500 [7] while DexiNed was trained just in our dataset (BIPED, see Chapter 3).

these models have been trained on the BSDS500 dataset, the proposed method did not use this dataset during the training stage, however, it is able to predict most of the edges with better quality than state-of-the-art approaches (see the last column in Fig. 5.2).

The main contributions of this chapter are the following:

- It presents a multispectral version of HED [150] to detect edges from RGB and NIR images.
- It proposes a robust CNN architecture for edge detection, referred to as DexiNed: Dense Extreme Inception Network. Up to our knowledge this is the first time that a CNN based approach results in such a sharp edge representation, furthermore, outperforming the state-of-the-art results in most of the considered datasets.
- It proposes a RGB-NIR edge detection approach, based on the DexiNed architecture, which can be also used in RGB dataset.
- It presents exhaustive quantitative evaluations on the most used datasets for edge, contour, and boundary detection by using state-of-the-art methods and the proposed approach. These evaluations can be used as a benchmark for further research.

5.2 Proposed Approaches

This section presents visible spectrum and multispectral models developed for the edge detection purpose. Firstly, the multispectral model based on HED [150] is detailed in Sec. 5.2.1; secondly, a new network for visible spectrum images (RGB) is presented in Sec. 5.2.2;

finally, an extension of the model presented in Sec. 5.2.2, to tackle the multispectral case, is detailed in Sec. 5.2.3.

The notation in the different DL based models will be as follow: given a CNN model, $\mathfrak{D}(\cdot)$, the input image for the architecture will be a I_{rgb} (RGB), or a I_{rgbn} (RGBN), image; Y_{rgb} is the corresponding annotated edge-maps (GT) of the given input image; \hat{Y} is the edge-maps predicted from the used DL model $\mathfrak{D}(\cdot)$. Depending on the DL approach, \hat{Y} will contain single or multiple predictions, which will be highlighted in the respective section.

5.2.1 Multispectral Holistically-Nested Edge Detection

Focusing in the multispectral edge detection, the architecture proposed by [150] is redesigned to train images from visible and near infrared spectral bands I_{rgbn} . Figure 5.3 presents an illustration of the modified VGG16 architecture, termed herein as MS-HED, which is described next. Given an I_{rgbn} input image, similar to HED, forward propagates such data throughout five VGG16 [126] convolutional blocks. As illustrated in Fig. 5.3, the first two blocks have two convolutional (conv) layers each, since then, each block has three conv layers. As in HED, also in MS-HED, the feature maps resulting from each block are fed to the layer in the second column (see Fig. 5.3 "Side-output_" into the rectangle). The edge-map prediction in the given image is summarized in the following equation:

$$\hat{Y} = \mathfrak{D}(I_{rgbn}, Y_{rgb}). \quad (5.1)$$

The MS-HED input layer mentioned above is an $I_{rgbn} \in R^{w \times h \times 4}$, where w and h are image spatial dimensions and 4 corresponds to R, G, B, and NIR channels. The kernel size for each VGG conv layer is 3×3 . At the end of the last conv layer in each block there is a maxpooling operator that downsamples the feature maps (the last block, block 5, does not have the maxpooling operator), this operator makes the edge detection process a multi scale task. Therefore, the output from every block is at a different scale, hence a re-sizing of predicted edge-maps is performed in the "Side-output_" rectangle shown in Fig. 5.3. The processes of edge-map prediction re-dimensioning into such rectangle is accomplished by convolutional and transpose convolutional layer. The transpose convolution or deconvolution is added at the end of such "Side-output" to, according to [150], lead the best performance restoring the size by the up-sampling process. This process is like a bilinear interpolation but in a single layer. In addition, another edge-maps prediction is generated in "Fuse-output" rectangle (see Fig. 5.3); this prediction is accomplished by fusing the concatenated outputs from the "Side-outputs" in a convolutional layer. The outputs from "Side-outputs" and "Fuse-output" are deeply supervised by a modified cross-entropy loss function, which is described below. Note that the predictions of all "Side-outputs" and "Fuse-output" has the same size as the GT.

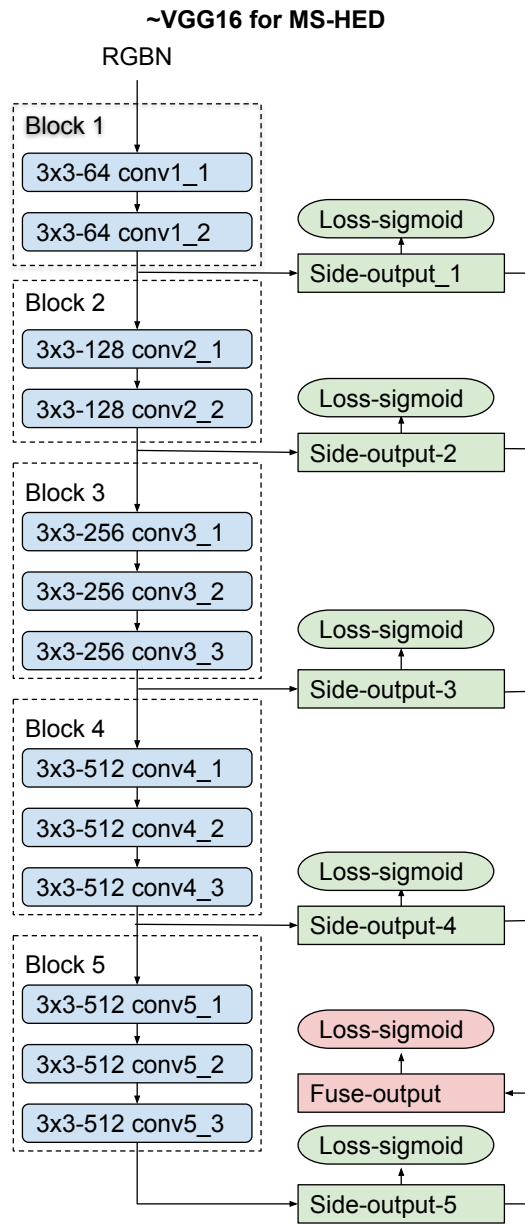


Figure 5.3: VGG16 [126] based multispectral holistically-nested edge detection CNN architecture.

Loss Function

As described in Eq. 5.1 the model can be summarized as a regression function $\bar{\delta}$. That is, $\hat{Y} = \bar{\delta}(I_{rgb}, Y_{rgb})$, \hat{Y} is a set of predicted edge-maps, $\hat{Y} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_P]$, where \hat{y}_i has the same size as Y , and P is the number of edge-maps predicted by the MS-HED; \hat{y}_P is the result from the last fusion layer f ($\hat{y}_P = \hat{y}_f$). Then, as the model is deep supervised, it uses the same loss as [150] (weighted cross-entropy), which is tackled as follow:

$$\begin{aligned} \mathcal{L}^p(W, w^p) = & -\beta \sum_{j \in Y^+} \log \sigma(y_j = 1 | X; W, w^p) \\ & - (1 - \beta) \sum_{j \in Y^-} \log \sigma(y_j = 0 | X; W, w^p), \end{aligned} \quad (5.2)$$

then,

$$\mathcal{L}(W, w) = \sum_{n=1}^P \delta^n \times \mathcal{L}^n(W, w^n), \quad (5.3)$$

where W is the collection of all network parameters and w is the p corresponding parameter, δ is a weight for each scale level (its setting will be described in the Sec. 5.3.1). $\beta = |Y^-| / (|Y^+ + Y^-|)$ and $(1 - \beta) = |Y^+| / (|Y^+ + Y^-|)$ ($|Y^-|$, $|Y^+|$ denote the edge and non-edge in the ground truth).

5.2.2 Dense Extreme Inception Network for Edge Detection

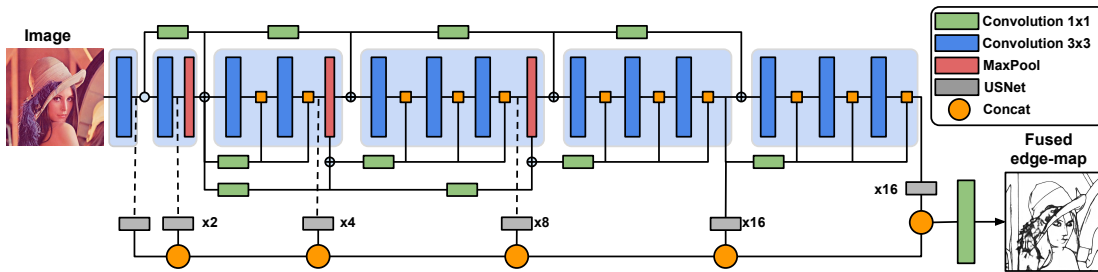


Figure 5.4: Proposed Dense Extreme Inception Network (DexiNed), consists of an encoder composed by six main blocks (showed in light blue). The main blocks are connected between them through 1x1 convolutional blocks. Each of the main blocks is composed by sub-blocks that are densely interconnected by the output of the previous main block. The output from each of the main blocks is fed to an upsampling block that produces an intermediate edge-map in order to build a Scale Space Volume, which is used to compose a final fused edge-map.

After evaluating the performance of an extension to the multispectral domain of the most widely used CNN architecture for edge detection, training from scratch and without pre-trained weights, the main conclusion that has been obtained is that from block 4 the MS-HED

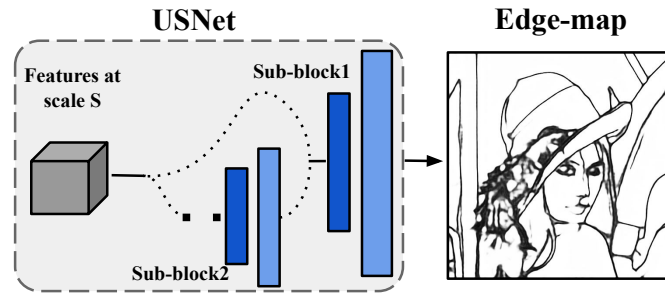


Figure 5.5: Detail of the upsampling block that receives as an input the learned features extracted from each of the main blocks. The features are fed into a stack of learned convolutional and transposed convolutional filters in order to extract an intermediate edge-map.

architecture loses most of the edges detected in the previous blocks, even by using NIR channel. The previous conclusion motivate the design of a new architecture that tackle the drawbacks of MS-HED. Although this thesis is developed in a multispectral framework, where information from one spectral band is used to improve results from another spectral band, the architecture proposed in this section is just intended to work in the visible spectrum domain (RGB). The idea behind this strategy is to develop a new architecture and compare its performance with respect to state-of-the-art approaches. Once the best edge detection architecture is developed, an extension of it is considered to tackle the multispectral case. The proposed new architecture is termed **DexiNed**, which consists of a stack of learned filters that receive as input an image (I_{rgb}) and predict edge-maps (\hat{Y}) with the same resolution as Y . DexiNed can be seen as two sub networks (see Fig. 5.4 and Fig. 5.5): Dense extreme inception network (Dexi) and the upsampling network (USNet). While Dexi is fed with the RGB image, USNet is fed with feature maps from each block of Dexi. The resulting network (DexiNed) generates thin edge-maps, avoiding missed edges in the deep layers. Note that even though without pre-trained data, the edges predicted from DexiNed are in most of the cases better than state-of-the-art results, see illustration in Fig. 5.2.

DexiNed Architecture

The proposed architecture is depicted in Fig. 5.4; it consists of an encoder with 6 main blocks inspired in the xception network [21]. The feature maps resulting as output at each of the main blocks are resized by the Upsampling network (USNet) presented in the next section. All the edge-maps resulting from the USNet are concatenated to feed the stack of learned filters at the very end of the network and produce a fused edge-map. All the six upsampling blocks do not share weights.

The blocks in blue (sub-block) consists of a stack of two convolutional layers with kernel size 3×3 , followed by batch normalization and ReLU as the activation function (just the last convs in the last sub-blocks does not have such activation). The max-pool is set by 3×3 kernel and stride 2. As the architecture follows the multi-scale learning, like in MS-HED, an

upsampling process (horizontal blocks in gray, Fig. 5.5) is followed.

Even though DexiNed is inspired in xception, the similarity is just in the structure of the main blocks and connections. Major differences are detailed below:

- While in xception separable convolutions are used, DexiNed uses standard convolutions.
- As the output is a 2D edge-map, there is "not exit flow", instead, another block at the end of block five has been added. This block has 256 filters and as in block 5 there is not maxpooling operator.
- In block 4 and block 5, instead of 728 filters, 512 filters have been set. The separations of the main blocks are done with the blocks connections (rectangles in green) drawn on the top side of Fig. 5.4.
- Concerning to skip connections, in xception there is one kind of connection, while in DexiNed there are two type of connections, see rectangles in green on the top and bottom of Fig. 5.4.

Whenever the network is enlarged by conv layers (or blocks of conv layers), every deep block losses important edge features and just one main-connection is not sufficient, as highlighted in DeepEdge [14], from the forth convolutional layer the edge feature loss is more chaotic. Therefore, since block 3, the output of each sub-block is averaged with *edge-connection* (orange squares in Fig. 5.4). These processes are inspired in ResNet [45] with the following notes: *i*) as shown in Fig. 5.4, after the max-pooling operation and before summation with the main-connection, the edge-connection is set to average each sub-blocks output (see rectangles in green, bottom side); *ii*) from the max-pool, block 2, edge-connections feed sub-blocks in block 3, 4 and 5, however, the sub-blocks in 6 are feed just from block 5 output.

Upsampling Network (USNet)

DexiNed has been designed to produce thin edges in order to enhance the visualization of predicted edge-maps. One of the key component of DexiNed for the edge thinning is the upsampling network, as appreciated in Fig. 5.4, each output from the Dexi blocks feeds the USNet. The USNet consists of the conditional stacked sub-blocks. Each sub-block has 2 layers, one convolutional and the other deconvolutional; there are two types of sub-blocks. The first sub-block (sub-block1) is feed from Dexi or sub-block2; it is only used when the scale difference between the feature map and the ground truth is equal to 2. The other sub-block (sub-block2), is considered when the difference is greater than 2. This sub-blocks is iterated till the feature map scale reaches 2 with respect to the GT. The sub-block1 is set as follow: kernel size of the conv layer 1×1 ; followed by a ReLU activation function; kernel size of the deconv layer or transpose convolution $s \times s$, where s is the input feature map scale level; both layers return one filter and the last one gives a feature map with the same size as the GT. The last conv layer does not have activation function.

The sub-block2 is set similar to sub-block1 with just one difference in the number of filters, which is 16 instead of 1 in sub-block1. For example, the output feature maps from block 6 in Dexi has the scale of 16, there will be three iterations in the sub-block2 before fed the sub-block1. The upsampling process of the second layer from the sub-blocks can be performed by bi-linear interpolation, sub-pixel convolution and transpose convolution, this study will be tackled in Sec. 5.3.3.

The definition of the loss function for DexiNed is the same as in MS-HED. That is, the Eq. 5.2 is applied to the predicted edge-maps: $\hat{Y} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_6, \hat{y}_f]$, where \hat{y}_i is the edge map obtained by USNet from the feature map predicted at Block i of DexiNed; and \hat{y}_f is the edge map obtained by fusing all USNet results, last conv layer—see rectangle in green of Fig. 5.4 (bottom right side).

5.2.3 Multispectral Dense Extreme Inception Network for Edge Detection

Once an architecture that overcome the drawbacks of MS-HED has been designed, and after deep evaluations that have shown the proposed architecture (DexiNed) reaches the best performance of state of the art, its extension to the multispectral framework is considered. It should be mentioned that the proposed architecture is trained from the scratch and without pre-trained weights. The proposed extension results in an architecture fed by multispectral images; in other words more information will be processed, hence in order to avoid heavy computation, the proposed DexiNed architecture is slightly optimized, by reducing the number of parameters of DexiNed; from the 6 blocks of DexiNed architecture, just the 5 first blocks are considered (see illustration of the reduced architecture in Fig. 5.6). Furthermore, the max-pooling operator from the block 4 is eliminated with the purpose of reducing artifacts in the final predicted edge-maps. The information from the new spectral band is expected to compensate this reduction. The new MultiSpectral Dense eXtreme inception Network for edge detection is termed as MS-DXN.

The input images for the MS-DXN are a RGB image together with its corresponding NIR image (I_{rgbn}). In other words, the dataset used for training the proposed architecture should contain registered visible and near infrared images. However, as most of the edge detection datasets are only in the visible band, a NIR estimator network is developed in order to tackle the edge detection of those datasets—see illustration in the bottom left of Fig. 5.6(into a broken lines rectangle). This NIR estimation network obtains NIR component from the given RGB input; it is referred to as NIR hallucination network (hallu-net). This hallu-net allows MS-DXN to be used in both multispectral datasets (e.g., MBIPED) or classical visible spectrum datasets (e.g., BSDS500, NYUD).

The MS-DXN architecture mainly consists of two parts, like DexiNed: *i*) Dexi network, which is the composition of main blocks and skip connections, light blue rectangles, containing blue and red rectangles, and rectangles in green, respectively (see Fig. 5.6 to appreciate Dexi network's components). *ii*) the second part is USNet (rectangles in gray in the bottom of the Fig. 5.6), which is the conditional iterative network used to generate edge-maps predictions

5.2. Proposed Approaches

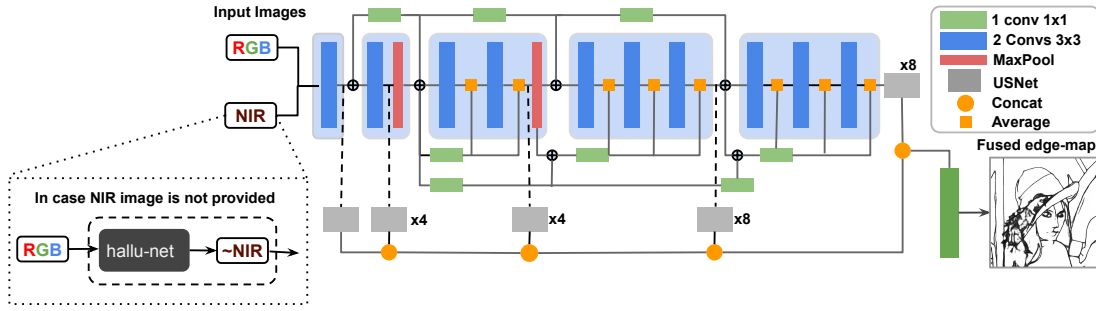


Figure 5.6: Proposed multispectral architecture of DexiNed network (MS-DXN).

from each main block of Dexi network (five predictions). A representative design of USNet is illustrated in Fig. 5.5, its detailed description can be found in Sec. 5.2.2. MS-DXN, as mentioned above, can be fed by a $I_{rgb\hat{n}}$ or $I_{rgb\hat{n}}$. That is, if the dataset for training or testing the MS-DXN, contains the NIR component of the RGB image, it does not use hallu-net and MS-DXN is trained by $I_{rgb\hat{n}}$. The dataset used in this thesis for the MS-DXN training and testing is MBIPED, see dataset descriptions in Chapter 3; for the purposes of evaluation in other datasets, just the RGB images from those datasets are considered and the NIR components are estimated by the hallucination network proposed in the next section. As mentioned above, all the datasets used by the state-of-the-art edge detection models are in the visible domain (I_{rgb}). In these cases the NIR hallucination network is used; hence, before fed Dexi network with input image I_{rgb} the \hat{I}_{nir} is estimated by the hallu-net, turning the input of MS-DXN in $I_{rgb\hat{n}}$.

The settings in the Dexi network and USNet are the same as in DexiNed (i.e., filter size of the conv layers, connection operators, the fusion layer and so on). The feature maps at the end of every block (as in DexiNed), follow two directions, one add with skip connection to fed the next high level block and the other to fed USNet (rectangle in gray in the bottom side of Fig. 5.6) to generate intermediate edge maps from every block. The whole set of predicted edge-maps \hat{Y} are as follow: $\hat{Y} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_5, \hat{y}_f]$; where \hat{y}_i and \hat{y}_f are predictions from the block i and fusion layer, respectively. The loss function to train these networks is the same as the one used in DexiNed, see details of such a loss function in the section above. Before training the MS-DXN, in the case that the NIR images are not provided from the dataset, the hallu-net is prepared to predict such NIR absence. In the next section the hallu-net is described.

5.2.4 NIR Estimation Model

In the last few years, with the purpose to reach better accuracy in the face recognition problem (e.g., [70, 129]), CNN models considering images from different spectral bands have been proposed. These models extract features from each spectral band and thus reduce the error rate the bio-metrical identification. For example, [70] developed a face recognition model for NIR images; it has been trained by using NIR face dataset of CASIA v2 [73]. However, before fed

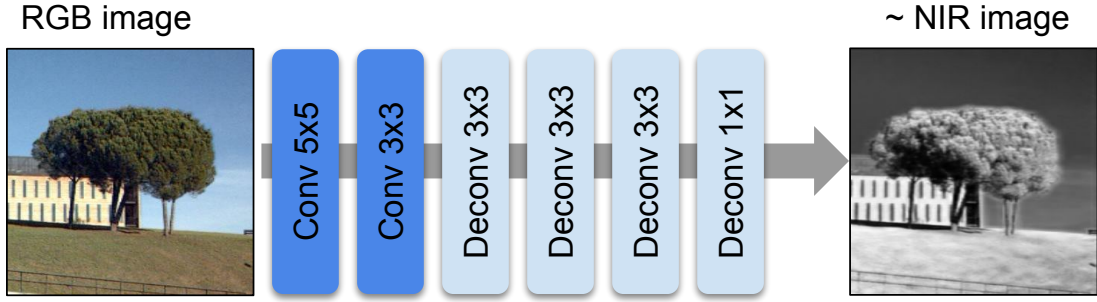


Figure 5.7: Hallu-net architecture, the images used in the figure are from OMSIV dataset ("Conv 5×5 " refers to a convolutional layer with a kernel size of 5×5 , Deconv refers to a deconvolutional layer).

the NIR image to the principal CNN architecture, authors propose another CNN model to map from the NIR image to the RGB domain, then fed the principal network. This cross-spectral proposal improved the accuracy from 75%, when just NIR images were considered, to 80% in the case the hallucinated RGB images were also considered. Similarly to this approach, but on the other direction (from RGB to NIR) MS-DXN proposes to make a NIR hallucination from RGB images of the state-of-the-art datasets for edge detection (which are just in the visible domain). The proposed network architecture is referred to as hallu-net—CNN model for NIR estimation.

The NIR estimation model, hallu-net, is the CNN model that estimate a NIR image from a given RGB image. That is, $\hat{I}_{nir} = H(I_{rgb}, I_{nir})$. A modification of the proposed CDNet architecture [131], presented in Section 4.2.2 for color restoration, is considered for NIR estimation (see the architecture in Fig 5.7). The CDNet is a set of two convolutional (conv) and 4 deconvolutional (deconv) layers. The hallu-net modification lies just in the number of filters, the first and the last hidden layers remains as in CDNet, while the rest of layers are set with 64 filters. The kernel sizes are maintained similar to in CDNet; 5×5 and 1×1 for the first and the last layers; in the rest of layers are set by 3×3 kernel size. The loss function used to train hallu-net is L1 defined as $l_1 = \sum_{i=1}^P |I_{nir}(i) - \hat{I}_{nir}(i)|$, P is the total number of pixels in an image and the sub-index i is the pixel index. The \hat{I}_{nir} is the predicted NIR image for the given $I_{rgb} \in \mathbb{R}^{w \times h \times 3}$; the I_{nir} is the target NIR image with the same dimension as I_{rgb} ; w and h are image's width and height respectively. Another loss function, Mean Square Error (MSE), has been considered to train hallu-net, this study is presented in Sec. 5.3.4.

5.3 Experimental Results

Section 5.2 presents three approaches MS-HED, DexiNed and MS-DXN. In this section, firstly results from MS-HED are presented; then, the proposed DexiNed is deeply evaluated, both quantitative and qualitatively, with the BIPED dataset. Finally, MS-DXN, trained in the mul-

tispectral edge detection datasets collected and presented in Chapter 3 is evaluated. Before presenting experimental results, the setup used for training and testing with some implementation notes, are detailed; additionally, assessment metrics used for the quantitative evaluation are described.

5.3.1 Experiment Setup

Dataset used for Training

The models proposed in Sec. 5.2 have been trained only one time by using edge detection based dataset, MBIPED (Multispectral Barcelona Images for Perceptual Edge Detection) and/or BIPED, introduced in Sec. 3.2.2. As highlighted in Chapter 3, under the umbrella of edge detection, most of the DL based models have been trained on dataset for boundary detection or segmentation. Therefore, in this thesis these datasets have been also considered for the evaluations as detailed below.

Dataset used for Testing

The performance of MS-HED is evaluated in MBIPED test dataset (50 images), see Sec. 3.2.2 for more details. The datasets used to evaluate the performance of DexiNed are BIPED, MDBD [92], CID [38], BSDS300 [89], BSDS500 [7], NYUD [125] and PASCAL [99]. Finally, the dataset considered to evaluate MS-DXN are: MBIPED, MDBD [92], CID [38], DCD [71], BSDS300 [89] BSDS500 [7], NYUD [125] and PASCAL [99]. Descriptions of such datasets are presented below.

MDBD: The Multicue Dataset for Boundary Detection has been collected for the purpose of psychophysical studies on object boundary detection in natural complex scenes, considering multiple cues such as luminance, color, motion and binocular disparity [92]. The MDBD dataset consist of short binocular video sequences of natural images (containing 10 frames per scene), there are 100 scenes in high definition (1280×720). Each image in this dataset has been annotated by more than one subjects. Although this dataset is intended for boundary detection, each image has been annotated six times for edge detection and five times for boundary detection. In the current work just the annotations corresponding to edge detection have been considered. Usually, in this dataset 80% of the images are used for training, while the remaining are used for evaluating the learning algorithm. In the current work 20% randomly selected images have been considered for evaluating the performance of DexiNed and MS-DXN.

CID: The Contour Image Dataset is a set of 40 gray-scale images with their respective ground truth, edges manually annotated [38]. The images in this dataset are 512×512 . The main limitation of this dataset is related with the small number of annotated images, as mentioned above just 40 annotated images are provided. Hence, in this case DexiNed and MS-DXN have been evaluated with the whole dataset, as in previous works (e.g., [5, 138]). This dataset, as well as NYUD, are difficult datasets for DL based approaches due to their grayscale

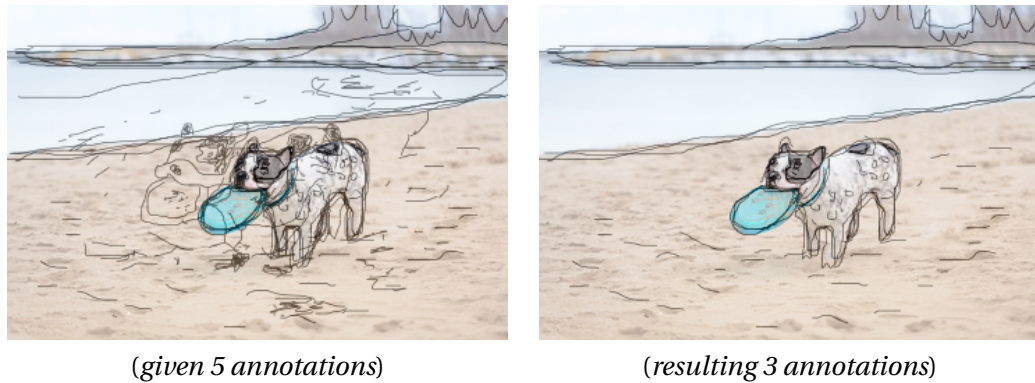


Figure 5.8: (*left*) A DCD test image with the provided annotations [71]. (*right*) Contours after removing wrong annotations.

nature and missed annotations in some of the provided images.

DCD: Dataset of Contour Drawings is collected with the aim to generate contour drawings (boundary-like drawings that capture the outline of the visual scenes) [71]. The DCD contains 1000 images, from that set 100 images are selected for testing. Ground truth annotations were collected by using a game application termed sketch drawing. For each image several annotations were collected; then, after an exhaustive evaluation just five contours per each image were kept as ground truths. However, in the current work, after checking the provided ground truth several wrong annotations have been found; Figure 5.8(*left*) shows an illustration where it can be easily appreciated the wrong contours provided by [71]. Hence, in order to perform a fair evaluation of trained architecture the annotations of the 100 testing images have been carefully checked, one by one, to remove wrong GTs. From this filtering process, several annotations have been removed. Hence, in the resulting set, which still have 100 images, each image contains between two and five annotations. Figure 5.8 (*right side*) presents results after removing wrong annotations. This dataset is considered for evaluating the MS-DXN model.

BSDS500: Berkeley Segmentation Dataset (BSDS), the first version has been published in 2001 [89], which consists of 300 images split up into 200 for training and 100 for validation, termed BSDS300; the last version [7], adds 200 new images for the testing. Every image in BSDS is annotated at least by 6 subjects, the dataset contains images of 481×321 . This dataset is mainly intended for image segmentation and boundary detection, therefore, as it will be illustrated in next sections, for the edge detection purpose some images are not well annotated. Generally, to evaluate the performance of a DL model in BSDS500, the new 200 images are used for testing while the BSDS300 are used for the network training purpose. As DexiNed and MS-DXN did not use BSDS300 for the training (just BIPED or MBIPED are considered) in Sec 5.3.3 and 5.3.4 results from the two datasets (i.e., BSDS300 and BSDS500) are presented. Concerning to BSDS300, for a fair comparison, the quantitative evaluation refers to the test part of that dataset (100 samples).

5.3. Experimental Results

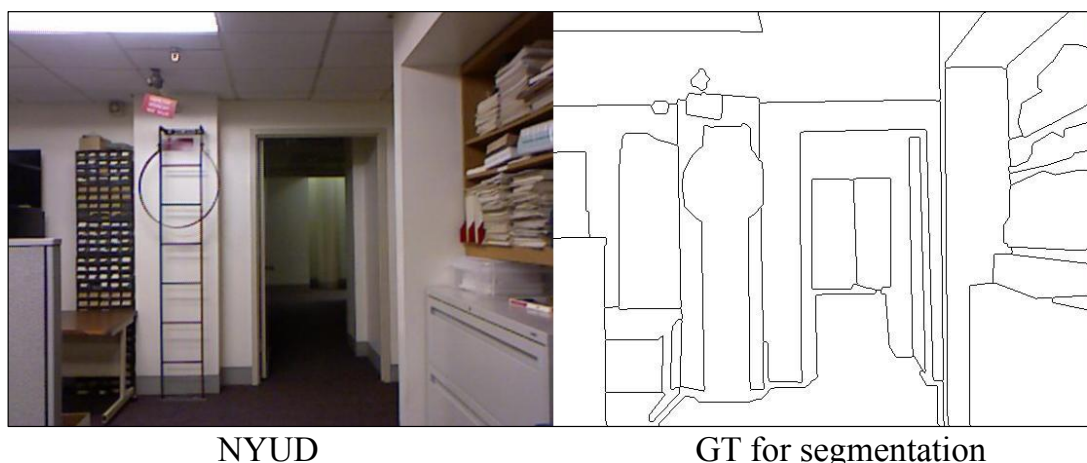


Figure 5.9: A sample of NYUD, note there are a large amount of missed edges in the ground truth.

NYUD: New York University Dataset is a set of 1449 RGBD images that contains 464 indoor scenarios, intended for segmentation purposes. This dataset is split up by [40] into three subsets—i.e., training, validation and testing sets. The testing set contains 654 images, while the remaining images are used for training and validation purposes. In the current work just the testing set has been selected for evaluating the performance of the proposed models. Although most of the images in NYUD are fully annotated for segmentation tasks, there are a few of them with poor annotations. The NYUD has been considered for evaluating edge detection tasks just because most of the last state of the art consider it for edge detection. However, for the edge detection purpose, the NYUD should be considered just for qualitative evaluation. Figure 5.9 shows a RGB image together with its corresponding GT—annotated edges. As it can be appreciated in the GT, there are a lot of missed edges in the provided annotations, such a lack of edges penalize the training and evaluation of a DL based edge detector.

PASCAL: The PASCAL-Context [99], termed in this paper as PASCAL, is a popular dataset used for segmentation with a wide variety of object categories. Currently, most of the major DL methods for edge detection use PASCAL-Context for training and testing edges (e.g., [150], [76]), due to its ground-truths correspond to scenes different to the ones depicted in BSDS dataset. This dataset contains 11530 annotated images, about 5% of randomly selected images (500) have been considered for testing DexiNed and MS-DXN. Although the images in PASCAL have more diverse labeled data, most of the images are annotated only for a couple of objects even though the scene has a vast number of features, in this case vast number of edges (even contours).

Evaluation Metrics

All the edge detector models proposed in this Chapter use the same quantitative evaluation metrics. The assessment metrics used by such approaches are detailed below.

The validation of the proposed algorithms is performed by widely used metrics in the state-of-the-art models for edge, contour and boundary detection [35]. They are the F-measure (F_m) [42, 89] of Optimal Dataset Scale (ODS), Optimal Image Scale (OIS), and the Average Precision (AP) from the Precision (P) over Recall (R) curves: $F_m = \frac{2 \times P \times R}{P + R}$; P is the fraction of edges detected with respect to the GT rather than false positives and R is the fraction of GTs detected by the proposed model rather than missed [89].

Implementation Notes

MS-HED: Since MS-HED implementation does not use VGG16 pre-trained data, the training iterations are 90k with a 0.003 learning rate, 10 mini-batch size and 0.2 fusion weights initialization. With a difference on the number of training iteration, all the settings have the same parameters as in [150] (δ is set by 1.1), which has the best performance in F-Score on the BSDS500 validation set. They experimented using different hyper-parameters, non-linear functions and conclude that the parameters exposed above are the best. Therefore, this work used such values. The training process took about 5 days for both, the HED and MSI-HED models, with the input image size [480, 480, 3] for I_{rgb} and [480, 480, 4] for I_{rgbn} . The training and testing of MS-HED model is done with images from MBIPED dataset. Therefore, the data augmentation process is performed with a small difference with respect to the one presented in Chapter 3. Firstly the given images are split up into two parts; secondly, each of the obtained sub-images is rotated in 15 different angles (cropping the maximum square of a rotated image); and finally, they are horizontally flipped, which augmented the dataset by a factor of 64.

DexiNed: The training of the network is performed from scratch and without pre-trained weights. On average, the model converges after 120k iterations (the quantitative and qualitative results are performed in 150k iterations) with a batch size of 8 using Adam optimizer and learning rate of 10^{-4} . The training process takes about 2 days. After the data augmentation, the images have different size, therefore, in the training process, the images were cropped or resized to 480×480 . 10% of the augmented data were used for the validation during the training process. The weights for fusion layer are initialized as: $\frac{1}{P-1}$ (see Section 5.2.1 for details on P). After a hyperparameter search to reduce the number of parameters, best performance was obtained using kernel sizes of 3×3 , 1×1 and $s \times s$ on its in-block convolutions, connection layers (main and edge) and transpose convolution, respectively.

MS-DXN: The implementation settings were performed similarly to DexiNed, with a little difference highlighted as follow: the quantitative and qualitative results are performed in 23 epochs (around 149k iterations). The training process takes about 2 days with images of size 440×440 as input. The weights for fusion layer are initialized as: $\frac{1}{P}$ (see Section 5.2.1 for details on P). The kernel size of the different convolutional and transpose convolutional layers are set

5.3. Experimental Results

Predicted Edge-maps	HED [150] (I_{rgb})			MS-HED (I_{rgbn})		
	ODS	OIS	AP	ODS	OIS	AP
Fused	0.79	0.80	0.80	0.78	0.80	0.81
Averaged	0.78	0.80	0.82	0.78	0.80	0.83

Table 5.1: Quantitative evaluation conducted on BIPED and MBIPED datasets.

similarly to in DexiNed.

The whole proposed models have been implemented in TensorFlow [1], the training and testing of such models are conducted on a GPU TITAN X.

5.3.2 Results from MS-HED

Quantitative results with the dataset collected for edge detection are depicted in Table 5.1. The evaluations have been performed over two models, HED [150] and MS-HED (introduced in Sec. 5.2.1). Both models have been trained from the scratch, without pre-trained weights and evaluated after 90k iterations. While the "Fused" is the edge prediction from the fusion layer (\hat{y}_f), "Averaged" is the average performed from the whole set of predictions. Whenever OIS is considered, both models reach the same values; according to ODS, \hat{y}_f from HED overcome to the \hat{y}_f from MS-HED; however, in the Average both models reach the same results. The major difference of the models considered for comparison is appreciated whenever the AP values are compared. Both predictions from MS-HED, Fused and Averaged, overcome the results from HED.

The quantitative difference depicted in Table 5.1 shows a slight difference between the two models considered for comparison. Figure 5.10 presents the whole edge-maps predicted from HED and MS-HED for two different samples. Both, RGB and RGBN based predictions, present similar amount of edges like those presented in the GT whenever the rows for FUSED and AVERAGED are considered. Furthermore, both fused and averaged results present plausible edges even in challenging scenes, note that the illustration in the second column of MS-HED in Fig. 5.10 shows how small details like windows in the buildings are detected. However, in the first column, the first light post (left to right) is not properly formed when the edge is predicted from the I_{rgb} images (see in the top, the post through the window in "FUSED" row). Both fused and averaged predictions from the MS images, when the light post is considered, has a presence of edges in the whole of its contour. Output i corresponds to the Side-output- i block in Fig. 5.10.

With the quantitative and qualitative results, many conclusions can highlighted. However, the most important thing is that, the training of a DL model from the scratch without pre-trained weights makes that the predicted edge-maps be generated with less artifacts but also deeper layers will loss high level features (see OUTPUT 5 row in HED, OUTPUT 4 and OUTPUT

Chapter 5. Edge Detection from a Multispectral Framework



Figure 5.10: RGB and NIR images from MBIPED dataset, GT, and all edge-maps predicted from both models considered for evaluation in Table 5.1.

5.3. Experimental Results

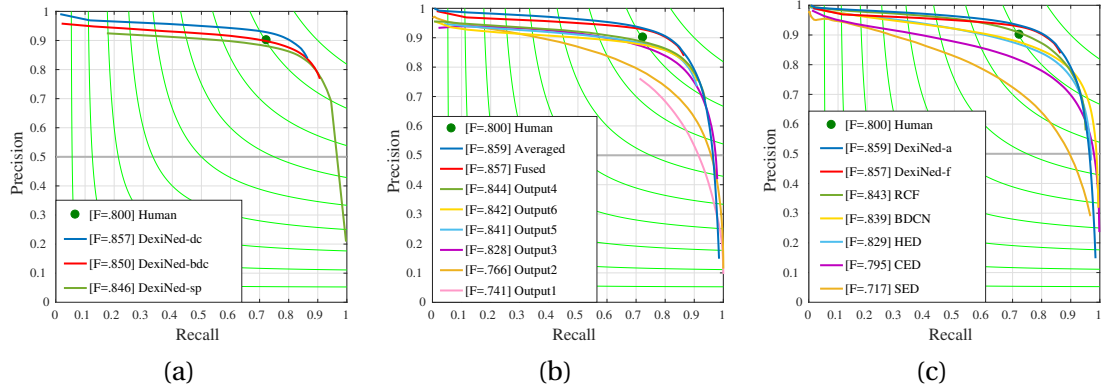


Figure 5.11: Precision/recall curves on BIPED dataset. (a) DexiNed upsampling versions. (b) The outputs of Dexined in testing stage, the 8 outputs are considered. (c) DexiNed comparison with other DL based edge detectors.

5 rows in MS-HED of Fig.5.10). They loss important edges from the previous blocks. More importantly, the edges predicted from block 4 are more coarser in HED, and in MS-HED there are not predicted edges; in the block 5 of both models there are not predicted edge. For the edge detection generalization from a DL model, another CNN architecture is needed, and the dataset used for the training needs more supervision or much edge annotations.

5.3.3 Results from DexiNed

Before tackle the edge detection task from the multispectral domain, the visible band is considered. Therefore the training of DexiNed is preformed in BIPED dataset and evaluated in the datasets presented in Sec. 5.3.1.

Firstly, in order to select the upsampling process that achieves the best result, an empiric evaluation has been performed, see Fig. 5.11a. The evaluation consists in conducting the same experiments by using the three upsampling methods; **DexiNed-bdc** refers to upsampling performed by a transpose convolution initialized with a bi-linear kernel; **DexiNed-dc** uses transpose convolution with trainable kernels; and **DexiNed-sp** uses subpixel convolution. According to F-measure, the three versions of DexiNed get the similar results, however, when analyzing the curves in Fig. 5.11a, it can be appreciated that although all of them have the same behaviour (shape), a small difference in the performance of DexiNed-dc appears. As a conclusion, the DexiNed-dc upsampling strategy is selected; from now on, all the evaluations performed on this section are obtained using a DexiNed-dc upsampling; for simplicity of notation just the term DexiNed is used instead of DexiNed-dc.

Figure 5.11b and Table 5.2a present the quantitative results reached from each DexiNed edge-map prediction. The results from the eight predicted edge-maps are depicted, the best quantitative results, corresponding to the fused (DexiNed-f) and averaged (DexiNed-a) edge-

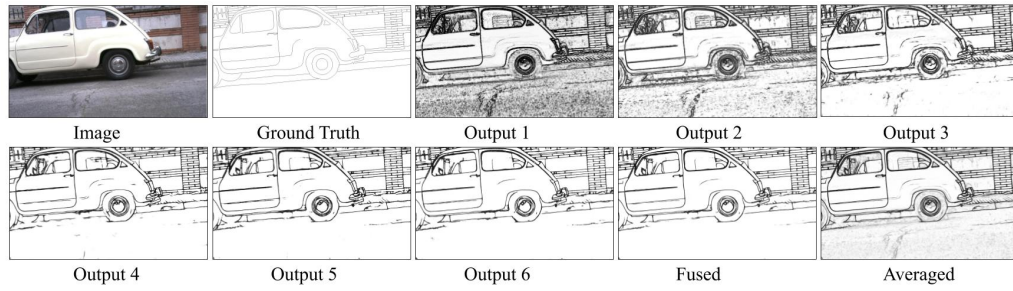


Figure 5.12: DexiNed predictions from BIPED dataset.

Outputs	ODS	OIS	AP	Methods	ODS	OIS	AP
Output 1 (\hat{y}_1)	.741	.760	.162	SED[5]	.717	.731	.756
Output 2 (\hat{y}_2)	.766	.803	.817	HED[150]	.829	.847	.869
Output 3 (\hat{y}_3)	.828	.846	.838	CED[142]	.795	.815	.830
Output 4 (\hat{y}_4)	.844	.858	.843	RCF[75]	.843	.859	.882
Output 5 (\hat{y}_5)	.841	.8530	.776	BDCN[44]	.839	.854	.887
Output 6 (\hat{y}_6)	.842	.852	.805	DexiNed-f	.857	.861	.805
Fused (\hat{y}_f)	.857	.861	.805	DexiNed-a	.859	.867	.905
Averaged	.859	.865	.905				

(a)

(b)

Table 5.2: (a) Quantitative evaluation of the 8 predictions of DexiNed on BIPED test dataset. (b) Comparisons between the state-of-the-art methods trained and evaluated with BIPED.

maps are selected for the comparisons. Similarly to [150] the averaged of all predictions (DexiNed-a) gets the best results in the three evaluation metrics, followed by the prediction generated in the fusion layer. Note that the edge-maps predicted from the block 2 till the 6 get similar results to DexiNed-f, this is due to the fact of the proposed skip-connections. For a qualitative illustration, Fig. 5.12 presents all edge-maps predicted from the proposed architecture. Qualitatively, the result from DexiNed-f is considerably better than the one from DexiNed-a (see illustration in Fig. 5.12). However, according to Table 5.2a, DexiNed-a produces slightly better quantitative results than DexiNed-f. As a conclusion both approaches (fused and averaged) reach similar results; through this manuscript whenever the term DexiNed is used it corresponds to DexiNed-f.

Table 5.2b presents a comparison between the DexiNed and the state-of-the-art techniques on edge and boundary detection. In all the cases BIPED dataset has been considered, both for training and evaluating the DL based models (i.e., HED [150], RCF [76], CED [142]) and BDCN [44], the training process for each model took about two days. As can be appreciated from Table 5.2b, DexiNed-a reaches the best results in all evaluation metrics. Actually both, DexiNed-a and DexiNed-f obtain the best results in almost all evaluation metrics. The F-measure obtained by comparing these approaches is presented in Fig. 5.11c; it can be appreciated how for Recall above 75% DexiNed gets the best results. Illustrations of the edges obtained with DexiNed and the state-of-the-art techniques are depicted in Figure 5.13, just for four images from the

5.3. Experimental Results

BIPED dataset. As it can be appreciated, although RCF and BDCN obtain similar quantitative results than DexiNed, which were the second best ranked algorithms in Table 5.2b, DexiNed predicts qualitative better results. Note that the proposed approach was trained from scratch without pre-trained weights.

The main objective of DexiNed is to get a representation that contains all edges from the given image (RGB or Grayscale) being trained only on BIPED. Therefore, in order to evaluate this capability, all the datasets presented in Sec. 5.3.1 have been considered, split up into two categories for a fair analysis; the first category contains a dataset intended for **edge detection** while the second category contains datasets intended for **contour/boundary detection/segmentation**. Results of edge-maps obtained with state-of-the-art methods are presented in Table 5.3. It should be noted that for each dataset the methods compared with DexiNed have been trained using images from that dataset, while DexiNed is trained just once with BIPED. Values presented in Table 5.3 are provided for comparisons and they come from the corresponding publication. It can be appreciated that DexiNed obtains the best performance in the MDBD dataset (up to our knowledge, the unique dataset for edge detection). It should be noted that DexiNed is evaluated in CID and BSDS300, even though these datasets contain a few images, which are not enough for training other approaches (e.g., HED, RCF, CED). Regarding BSDS500, NYUD and PASCAL, DexiNed does not reach the best results since these datasets have not been intended for edge detection, hence the evaluation metrics penalize edges detected by DexiNed. To highlight this situation, Fig. 5.14 depicts results from Table 5.3. Four samples from each dataset are considered. They are selected according to the best and worst F measure. Therefore, as shown in Fig. 5.14, when the image is fully annotated the score reaches around 100%, otherwise it reaches less than 50%.

As highlighted in previous section, when the deep learning based edge detection approaches are evaluated in datasets intended for objects' boundary detection or objects segmentation, the results will be penalized. To support this claim, we present in Fig. 5.14 four predictions (two best and two worst results according to F-measure) from all datasets used for evaluating the proposed approach (except BIPED that has been used for training). The F-measure obtained in the three most used datasets (i.e., BSDS500, BSDS300 and NYUD) reaches over 80% in those cases where images are fully annotated; otherwise, the F-measure reaches about 30%. However, when the edge dataset (MDBD [92]) is considered the worst F-measure reaches over 75%.

Concerning to **qualitative evaluation**, results obtained with the proposed approach, just trained in the BIPED dataset, when it is evaluated in other publicly available databases. The provided illustrations correspond to the two best and the two worst results in each dataset. In Fig. 5.14, the first row corresponds to the two best F-measure scores for the BSDS500 dataset, while the second row shows the two results with the worst scores. These results are provided just for qualitative evaluations of the performance of the proposed approach. It can be appreciated how details, which are missed in the ground truths, are correctly detected by DexiNed; more details on each dataset are provided follow:

Chapter 5. Edge Detection from a Multispectral Framework



Figure 5.13: Three resultant edge-maps from BIPED test dataset from the approaches presented in Table 5.2b.

5.3. Experimental Results

Dataset	Methods	ODS	OIS	AP
Edge detection dataset				
MDBD[92]	HED[150]	.851	.864	.890
	RCF[76]	.857	.862	-
	DexiNed-f	.837	.837	.751
	DexiNed-a	.859	.864	.917
Contour/boundary detection/segmentation datasets				
CID[38]	SCO[154]	.58	.64	.61
	SED[5]	.65	.69	.68
	DexiNed-f	.65	.67	.59
	DexiNed-a	.65	.69	.71
BSDS300[89]	gPb[7]	.700	.720	.660
	SED[5]	.69	.71	.71
	DexiNed-f	.707	.723	.52
	DexiNed-a	.709	.726	.738
BSDS500[7]	HED[150]	.790	.808	.811
	RCF[76]	.806	.823	-
	CED[142]	.803	.820	.871
	SED[5]	.710	.740	.740
	DexiNed-f	.729	.745	.583
	DexiNed-a	.728	.745	.689
NYUD[125]	gPb[7]	.632	.661	.562
	HED[150]	.720	.761	.786
	RCF[76]	.743	.757	-
	DexiNed-f	.658	.674	.556
	DexiNed-a	.602	.615	.490
PASCAL[99]	CED[142]	.726	.750	.778
	HED[150]	.584	.592	.443
	DexiNed-f	.431	.458	.274
	DexiNed-a	.475	.497	.329

Table 5.3: Comparisons between DexiNed trained on BIPED with respect to the state-of-the-art methods in the literature trained with the corresponding datasets (values from other approaches come from the corresponding publications).

Chapter 5. Edge Detection from a Multispectral Framework

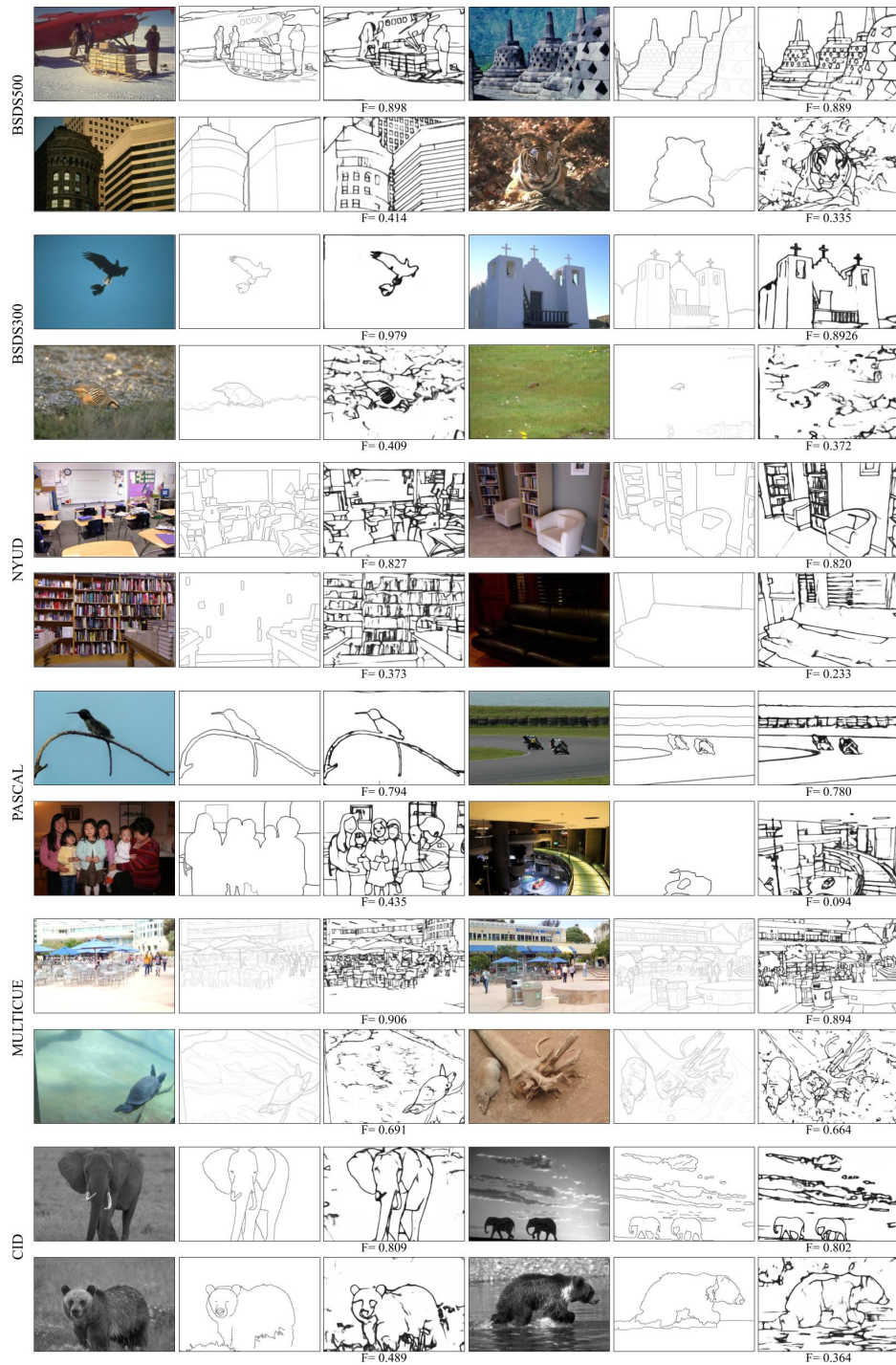


Figure 5.14: Four edge-maps predicted by DexiNed for each tested dataset; they correspond to the best and worst F-measure for the given dataset. The 2nd and 5th columns correspond to the GT of the respective image.

5.3. Experimental Results

- *BSDS500*: The two highest score on this dataset are near to the 90% F-measure score but the worst scores are under the 50%. As depicted in Fig. 5.14 (first two rows), while there is a well annotated GT, the scores are high, otherwise, the scores turn bad. Furthermore, some of the edges in the ground truth do not exist in the given image (for instance, the vertical line on the facade of the building in the right side does not exist in the RGB image). Therefore, even the predicted edges are well localized, the score is still 41.4%.
- *BSDS300*: Results are similarly to the ones obtained with BSDS500. However, in this case a considerably higher performance is reached in the best case. Note that almost a 98% score is obtained, even though these kind of images have not been used for the training stage. The BIPED data set contains scenarios considerable different, see illustrations in Fig. 5.14; in none of them there is a single object.
- *NYUD*: In this dataset the two highest scores overcome the 80% in the F-measure. It should be mentioned that all images in NYUD correspond to indoor scenarios, while in BIPED all the scenes correspond to outdoor scenarios. With these results we could conclude that DexiNed is able to handle different scenarios, even though just outdoor has been used for its training. Finally, the two worst results correspond to scene with poor ground truth annotations. It can be appreciated that DexiNed is able to extract edges in scenes with poor lighting conditions (see illustration with $F=0.233$).
- *PASCAL*: This dataset is challenging due to the purpose of its nature. PASCAL has been conceive for semantic image segmentation. In this case the worst results has a score below 10%, but as mentioned in the previous section, this bad result corresponds to a scene where just a few objects have been annotated, while most of the edges in the scene are neglected. In case images are well annotated, the score overcome 75%.
- *MDBD*: As shown in Fig 5.14, in the GT most of the edges plausible for human eyes are annotated, hence, the worst result gets over 65%, while the best one gets over 90%. As highlighted in the paper, DexiNed overcome state-of-the-art results when GTs are well annotated. It could be also concluded that BIPED dataset helps to generalize the edge detection work.
- *CID*: Since this dataset is complex, due to the nature of its composition (i.e., grayscale domain), the range of results goes from 36.4% till 80.9%. Even though DexiNed did not learn edge detection in grayscale domain, its scores still overcome the state-of-the-art performance.

As a conclusion, it should be stated that edge detection and contour/boundary detection are different problems that need to be tackled separately when a DL based model is considered. For instance, in datasets such as BSDS500, NYUD, and PASCAL, the performance of the proposed models are lower than the results from the state-of-the-art DL models (which are trained in these datasets). These lower quantitative performance can be easily understood by the fact that in these datasets not all the edges are annotated.

5.3.4 Results from MS-DXN

Once the quantitative and qualitative evaluations are performed, DexiNed architecture demonstrated its capabilities in the edge detection problem by training the model from the scratch, even though, without pre-trained weights. Motivated by the obtained results, the DexiNed architecture has been considered to elaborate the MS-DXN architecture as mentioned above. In this new architecture, if the provided dataset contains RGB and NIR images, like MBIPED, during the training process the hallu-net module is not considered. Otherwise, the hallu-net model is used to estimate the NIR component of the provided RGB images. Before presenting the quantitative and qualitative results of MS-DXN in edge detection, the hallu-net experiments are depicted.

The **hallu-net** presented in Sec. 5.2.4 has been evaluated by training it with L_1 , MSE and L_1 +MSE loss functions. Additionally, two learning rates (lr) have been considered in each training. Each setting has been trained during 3000 epochs; PSNR and SSIM are considered for the quantitative evaluations of the estimated NIR images when the OMSIV dataset has been considered. Results from these evaluations are presented in Table 5.4. Even though, the best score is for L_1 with $lr = 1e - 2$ considering PSNR; while considering SSIM the best scores are in the first and the last rows. In summary, there is not a clear configuration that gets the best results in both evaluation metrics. Looking at the results in Table 5.4, the second best score in PSNR and SSIM corresponds to the same setup, which has been finally selected. The training and evaluations have been conducted in OMSIV dataset by using 400 images for training and validation and 100 for the testing. The input color image size to train hallu-net was 164×164 .

For the quantitative and qualitative evaluations of MS-DXN, the HED [150], CED [142], RCF [75] and BDCN [44] datasets have been considered. Currently, SED [5] is the state-of-the-art method for edge detection, which is based on low level features (non-learning algorithm), this method will be considered as the baseline, see details of such models in Sec. 2.4. For a fair comparison, all the DL based models have been trained using the same dataset, which is RGB images from MBIPED. The training of HED and CED took about 3 days (20 epochs), RCF has been trained on about 5 days (till 20 epochs), BDCN has been trained till 6k iterations (about four days), and DexiNed has been trained till 150K iterations for about 2 days. The proposed MS-DXN model has been trained about 2.5 days, 23 epochs (around 149K iterations) twice, firstly with MBIPED to evaluate the performance when real NIR images are considered, and then with RGB from MBIPED and NIR images estimated by hallu-net; in the last case the hallu-net has been previously trained with OMSIV. The setting of the DL based models maintains the same configuration as described in their own manuscript.

The comparisons are performed according to the type of dataset used for the evaluation; the dataset classification as edge, contour, boundary, and segmentation is performed according to the statement given in each manuscript where the approaches have been proposed. A deep study is conducted to the dataset intended for edge detection tasks (MBIPED and MDBD); more precisely, MBIPED is also evaluated in the multispectral domain by MS-DXN, since the corresponding multispectral images (RGB and NIR) are provided (I_{rgbn}). Table 5.5 and Fig.

5.3. Experimental Results

Loss functions	lr	Epochs	PSNR	SSIM
L_1	$1e-1$	3000	20.148	0.751
L_1	$1e-2$	3000	20.258	0.748
MSE	$1e-1$	3000	20.182	0.745
MSE	$1e-2$	3000	20.214	0.750
$L_1 + \text{MSE}$	$1e-1$	3000	20.164	0.745
$L_1 + \text{MSE}$	$1e-2$	3000	20.153	0.751

Table 5.4: Quantitative evaluation of the different version of hallu-net. lr is the learning rate used on each experiment.

5.15 present comparisons from edge detection based datasets, and Tables 5.6, 5.7 and 5.8 present results from datasets intended for contour, boundary, detection/object segmentation, respectively. With respect to the qualitative results, whenever the proposed models do not use sub-index at the end like MS-DXN-a or MS-DXN-f, the edge-maps prediction corresponds to MS-DXN-f, even though if in the figure is termed just as MS-DXN without "-f". The choice is based on the suggestion of DexiNed, after quantitatively evaluating each output, the fused (DexiNed-f) and the averaged (DexiNed-a) predicted edge-maps scored the second and the first best marks, respectively. While the second one shows the best edge-map contrast, the first one reaches the best quantitative result. Therefore, for the qualitative comparisons DexiNed-f and MS-DXN-f are considered, termed in the illustrations as DexiNed and MS-DXN, respectively.

Figure 5.15a shows precision/recall curve for all the approaches when MBIPED is used. The graph shows an improvement of MS-DXN-a over DexiNed-a; although MS-DXN has fewer parameters, it still reaches the state-of-the-art results, which means that considering ODS the two best marks are for the two proposed models, MS-DXN and DexiNed. A more detailed quantitative results, considering ODS, OIS, and AP, is presented in Table 5.5. Considering two assessment metrics, ODS and AP, the MS-DXN-a (a model trained with visible and NIR bands images I_{rgb}) reaches the best result, whenever OIS is considered the best results is from BDCN; actually, the difference with the second best result (from MS-DXN-a) is just 0.1%.

As previously highlighted, all learning algorithms for edge detection are trained using the RGB images from MBIPED. So now, other datasets, which have not been previously used in the training process, are considered to validate the edge detection generalization. Figure 5.15b presents results from MDBD dataset, note that the same 20 images are considered to compare the performance of all models; according to F_m in Fig. 5.15b, the best result correspond to DexiNed-a followed by BDCN; SED reaches the third position, which even though is not a learning based algorithm its results are in the state of the art. As a conclusion from the MDBD evaluation we could state that the sought generalization of a DL based model for edge detection has been reached, the proposed DL model can be used as a low level based edge detector or biological visual perception based method (see Sec. 2.4 for details); it should be mentioned that although the proposed approach has not been trained on such a dataset

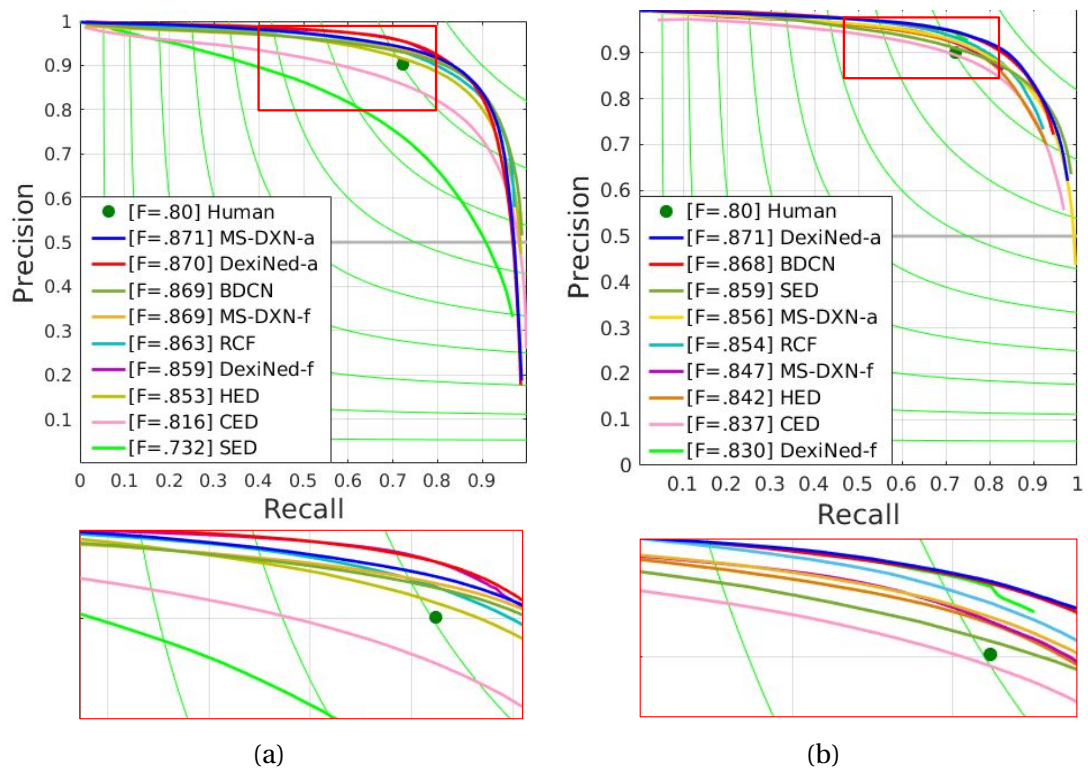


Figure 5.15: (a) Precision/recall curves on MBIPED dataset, results from all methods considered for comparison (the multispectral images from MBIPED have been used just for MS-DXN, while the RGB images to all other approaches). (b) Precision/recall curves on MDBD dataset, the results is from the 20 test images. Note that the same images used for the proposed model are also evaluated in the other models.

5.3. Experimental Results

Dataset	Methods	ODS	OIS	AP
Edge detection dataset				
MBIPED[130]	SED[5]	.732	.745	.786
	HED[150]	.853	.863	.911
	CED[142]	.816	.830	.871
	RCF[75]	.850	.863	.910
	BDCN[44]	.869	.878	.918
	DexiNed-f[130]	.859	.860	.782
	DexiNed-a[130]	.870	.870	.913
	MS-DXN-f	.869	.874	.873
	MS-DXN-a	.871	.877	.921
MDBD[92]	SED[5]	.859	.867	.913
	HED[150]	.842	.856	.866
	CED[142]	.837	.849	.8370
	RCF[75]	.854	.865	.877
	BDCN[44]	.868	.883	.898
	DexiNed-f[130]	.830	.830	.731
	DexiNed-a[130]	.871	.877	.919
	MS-DXN-f	.847	.850	.795
	MS-DXN-a	.856	.870	.918

Table 5.5: Comparisons between results obtained with DexiNed and MS-DXN with respect to the state-of-the-art methods in the literature. All methods were trained on MBIPED—in this case the multispectral images from MBIPED have been used just for MS-DXN, while the RGB images to all other approaches.

(MDBD), its results score over the 80%, which is more than human level (the human level is 75%). Table 5.5 (section MDBD) shows results from three metrics considered to compare the performance of the proposed approaches with the state of the art. According to ODS and AP, DexiNed-a gets the best results, nevertheless, whenever OIS is considered BDCN reaches the best results followed by DexiNed-a, and MS-DXN-a.

For a qualitative comparison, Fig. 5.16 presents edge-maps predictions from the whole DL based models presented in Table 5.5. There are two samples for each edge based dataset. Although most of the models considered for evaluation predicted most of the edges present in the GT, from the seventh row (DexiNed), the predicted edge-maps are cleaner in both dataset MBIPED and MDBD. In the last column of MBIPED, there are edges predicted from the multispectral edge detector (MS-DXN) that are not present in DexiNed results, neither in HED, RCF and BDCN (see details on the bricks on the image in the second column of Fig. 5.16); MS-DXN detects more edges than the other approaches (AP=0.921).

Since most of state-of-the-art learning based algorithms for edge detection use datasets like BSDS500, NYUD, and even PASCAL context [99], which are not specifically intended for edge detection, this Chapter also includes quantitative and qualitative results in these datasets. Furthermore, CID, BSDS300, and DCD datasets, which have been recently presented in the

Chapter 5. Edge Detection from a Multispectral Framework

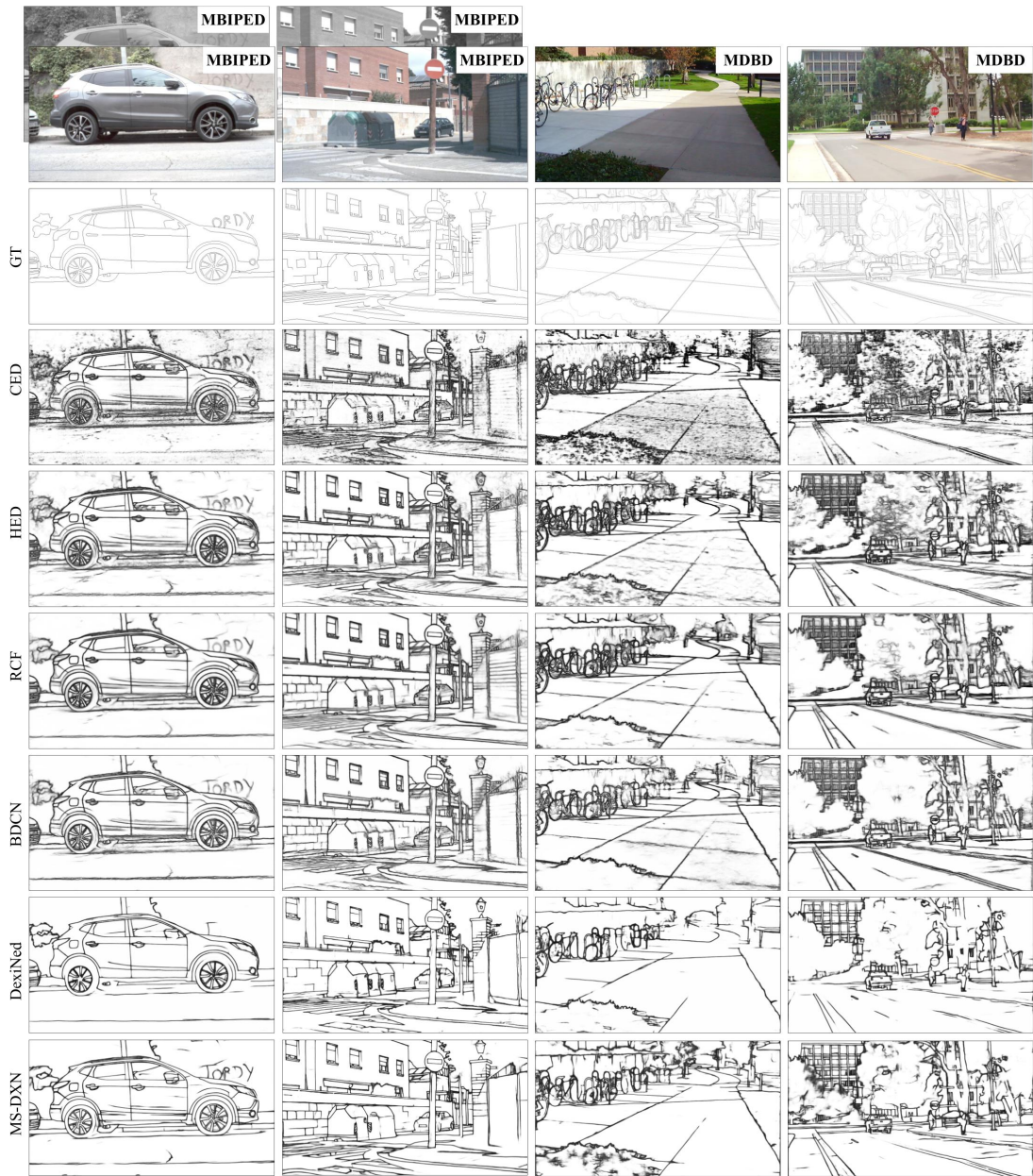


Figure 5.16: Four samples from the results obtained with the models depicted in Table 5.5. The label in the images (1st row) corresponds to the datasets where they come from.

literature (2019), are also considered to evaluate the performance of the proposed models.

Table 5.6 presents quantitative results from Contour Image Data base (CID) and Dataset of Contour Drawing (DCD); it should be noted these datasets have been proposed for contour detection. In the CID dataset, the baseline is SED [5] and since the 2003 (year of the dataset publication) none of the models proposed in the literature have reached the 70% in ODS, OIS or AP. After the DL based models (depicted in Table 5.6) are trained on RGB images of MBIPED, the RCF and BDCN overcome the 70% on each evaluations metrics. Concerning to the proposed approaches, DexiNed-a and MS-DXN-a still reach the best score ever published. As highlighted above, in the visual appreciation of predicted edge-maps from the proposed models (DexiNed-f, and MS-DXN-f), they are thin and clearer edge-maps than those from the best scored in the Table 5.6. Figure 5.17 presents four challenging scenarios where it can be appreciated the quality of edge-maps obtained whit the proposed approaches. According to the values presented in Table 5.6 in the case of DCD dataset, the CED (in OIS), BDCN (in ODS), and DexiNed-a (in AP) are the best ranked models to detect edges in such dataset. Note that the DCD was proposed for photo-sketching [71], and this is the first time that DCD is evaluated on the metrics proposed in Sec. 5.3.1. Although most of the results scored over the 80% (even 90% in CED), this dataset cannot be used for training due to some inconsistency in the contour-maps GT (see details in DCD presentation of Sec. 5.3.1). In the case of CID, since it is a gray-scale image dataset, the MS-DXN estimates the NIR component by using the provided graysacle image as RGB components, in other words the provided gray-sacle image is used three times.

Figure 5.17 presents two images from each dataset (CID and DCD) for qualitative evaluations. Although quantitatively, RCF and BDCN scored with the best performance, visually, results from the models proposed in this Chapter (DexiNed and MS-DXN) show more contrast edge-map predictions. See from the seventh row (DexiNed) how all four images are clearer and the edge-maps are more sharp. As mentioned above the images from CID are in grayscale, and none of the DL based models were trained on images in such a domain, even though, the edge-map predictions are slightly similar to the ground truth edge-maps (second row).

Table 5.7 presents quantitative results on the datasets intended for boundary detection. The baseline (BDCN) on ODS, OIS and AP are 0.828, 0.844 and 0.890, respectively [44] (2019); however, as highlighted in Sec. 5.1, even though with such quantitative results, the edge-maps predicted from such model, trained in BSDS300+PASCAL datasets, does not detect perceptual edges from the test dataset of BSDS500 (see this comparison in the Fig. 5.2). The results from the proposed models presented in Table 5.7, although does not overcome the baseline, (due to that not all GT are annotated in the edge level), most of the proposed approaches still overcome the state-of-the-art non-learning algorithm for edge detection, SED.

Figure 5.18 shows two images from BSDS300 and two from BSDS500. The annotation-maps in the second row of the respective datasets are from GTs. Two challenging and textured images are presented in the second columns of each dataset; the GTs for these two images, are badly annotated. As it can be appreciated, the edge maps presented in last two rows are clearer and

Dataset	Methods	ODS	OIS	AP
Contour detection datasets				
CID[38]	SED[5]	.650	.690	.680
	HED[150]	.698	.731	.741
	CED[142]	.662	.698	.700
	RCF[75]	.721	.743	.738
	BDCN[44]	.720	.742	.741
	DexiNed-f[130]	.609	.610	.442
	DexiNed-a[130]	.675	.719	.730
	MS-DXN-f	.660	.667	.541
	MS-DXN-a	.680	.716	.724
DCD [71]	SED[5]	.801	.842	.856
	HED[150]	.803	.866	.823
	CED[142]	.853	.915	.632
	RCF[75]	.809	.857	.760
	BDCN[44]	.869	.899	.849
	DexiNed-f[130]	.784	.815	.660
	DexiNed-a[130]	.807	.842	.860
	MS-DXN-f	.805	.826	.791
	MS-DXN-a	.815	.844	.850

Table 5.6: Comparisons of the results obtained with the proposed approaches (DexiNed and MS-DXN) and the by using RGB component of MBIPED, in the case of MS-DXN the NIR component has been obtained from hallu-net. In addition, SED (a non-learning based algorithm based edge detector) is considered for evaluation.

5.3. Experimental Results

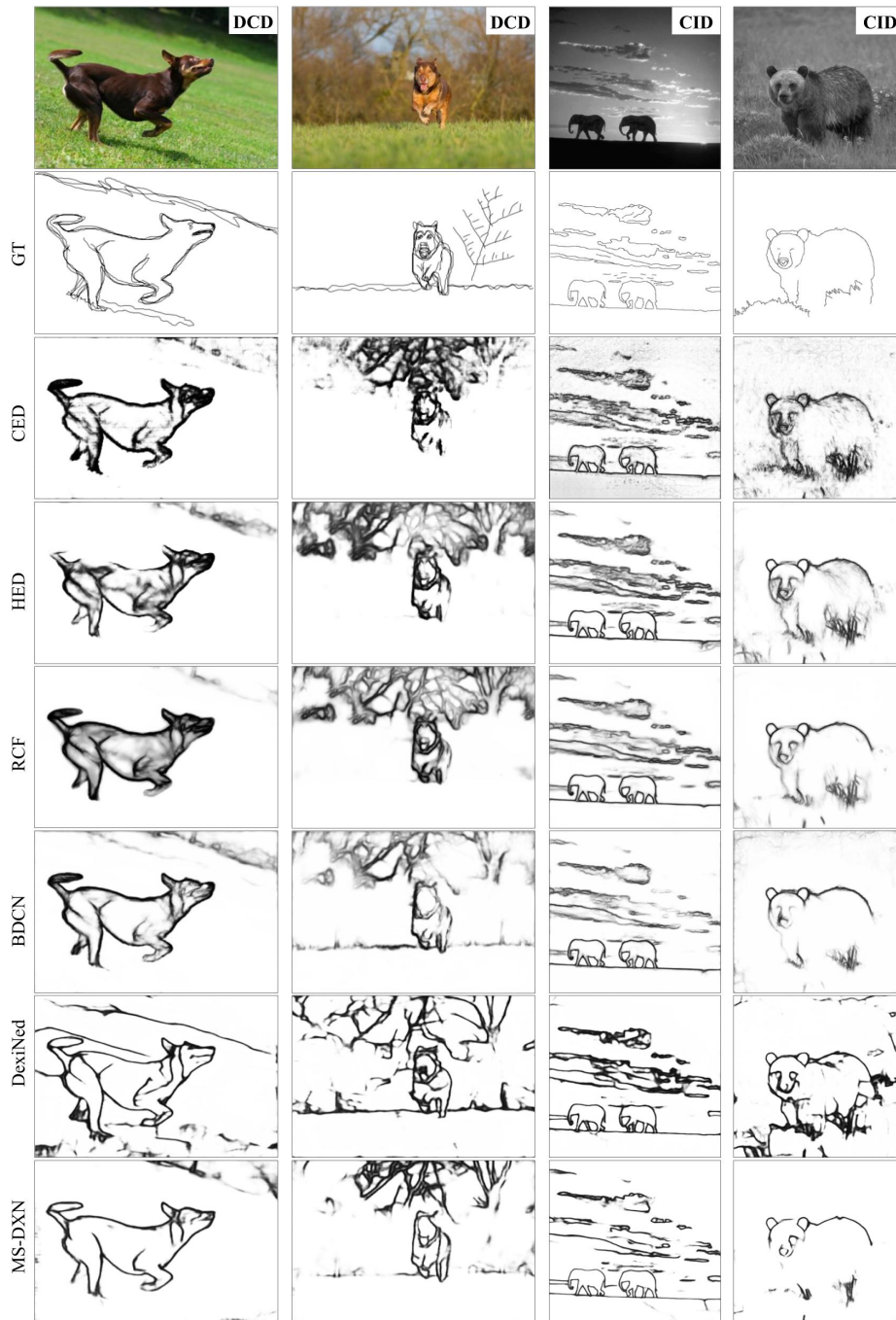


Figure 5.17: Four samples from the result obtained with the models depicted in Table 5.6. The label in the images (1st row) corresponds to the datasets where they are from.

Dataset	Methods	ODS	OIS	AP
Boundary detection dataset				
BSDS300 [89]	SED[5]	.700	.720	.660
	HED[150]	.624	.671	.585
	CED[142]	.618	.648	.506
	RCF[75]	.639	.680	.595
	BDCN[44]	.712	.740	.698
	DexiNed-f[130]	.707	.723	.520
	DexiNed-a[130]	.709	.726	.738
	MS-DXN-f	.651	.667	.554
	MS-DXN-a	.654	.678	.625
BSDS500 [7]	SED[5]	.710	.740	.740
	HED[150]	.646	.688	.594
	CED[142]	.635	.658	.496
	RCF[75]	.663	.701	.634
	BDCN[44]	.690	.714	.686
	DexiNed-f[130]	.729	.745	.583
	DexiNed-a[130]	.728	.745	.689
	MS-DXN-f	.677	.698	.597
	MS-DXN-a	.680	.703	.655

Table 5.7: Comparisons of the results obtained in boundary datasets with the proposed approaches (DexiNed and MS-DXN) and the state-of-the-art methods proposed in the literature. All methods were trained by using the RGB component of MBIPED, in the case of MS-DXN the NIR component has been obtained from hallu-net. In addition, SED (a non-learning based algorithm based edge detector) is considered for evaluation.

5.3. Experimental Results

Dataset	Methods	ODS	OIS	AP
Segmentation datasets				
NYUD [125]	SED[5]	.548	.565	.502
	HED[150]	.575	.590	.446
	CED[142]	.547	.562	.398
	RCF[75]	.587	.605	.463
	BDCN[44]	.594	.609	.481
	DexiNed-f[130]	.658	.674	.556
	DexiNed-a[130]	.602	.615	.490
	MS-DXN-f	.571	.587	.411
	MS-DXN-a	.573	.590	.482
PASCAL [99]	SED[5]	.436	.466	.321
	HED[150]	.428	.452	.265
	CED[142]	.420	.443	.259
	RCF[75]	.440	.465	.275
	BDCN[44]	.450	.476	.294
	DexiNed-f[130]	.475	.497	.329
	DexiNed-a[130]	.423	.443	.274
	MS-DXN-f	.419	.435	.228
	MS-DXN-a	.420	.441	.272

Table 5.8: Comparisons of the results obtained in the scene or object segmentation datasets with the proposed approaches and the state-of-the-art methods proposed in the literature. All methods were trained with the RGB images from MBIPED. In addition, SED (a non-learning based algorithm based edge detector) is considered for evaluation.

sharp. The deep learning models generalization are more visible in the edge prediction from the BSDS500 image (third column), almost all models are able to detect edges, even in the complex zones like in the center of the image (shadows).

Finally, the last datasets considered to evaluate the performance of the deep learning edge detector are datasets from scene or object segmentation, NYUD [125] and PASCAL [99]. Table 5.8 presents a comparative results on those datasets and Fig 5.19 shows edge-maps predicted from each dataset, two illustrations per dataset are depicted. Quantitatively, Dexined-f reaches the best score in three evaluation metrics (ODS, OIS and AP) on NYUD and also in PASCAL dataset. Qualitatively, all models considered for qualitative comparisons present thin edge-maps; however, edges presented in the last two rows are clearer. Regarding NYUD dataset, MS-DXN reaches the best perceptual edge-map predictions (clean and sharp); furthermore, in PASCAL, those models give edge-maps with the same quality as in NYUD—note that MS-DXN is able to capture more edges from the given image, see dog’s eyes in third column from PASCAL dataset.

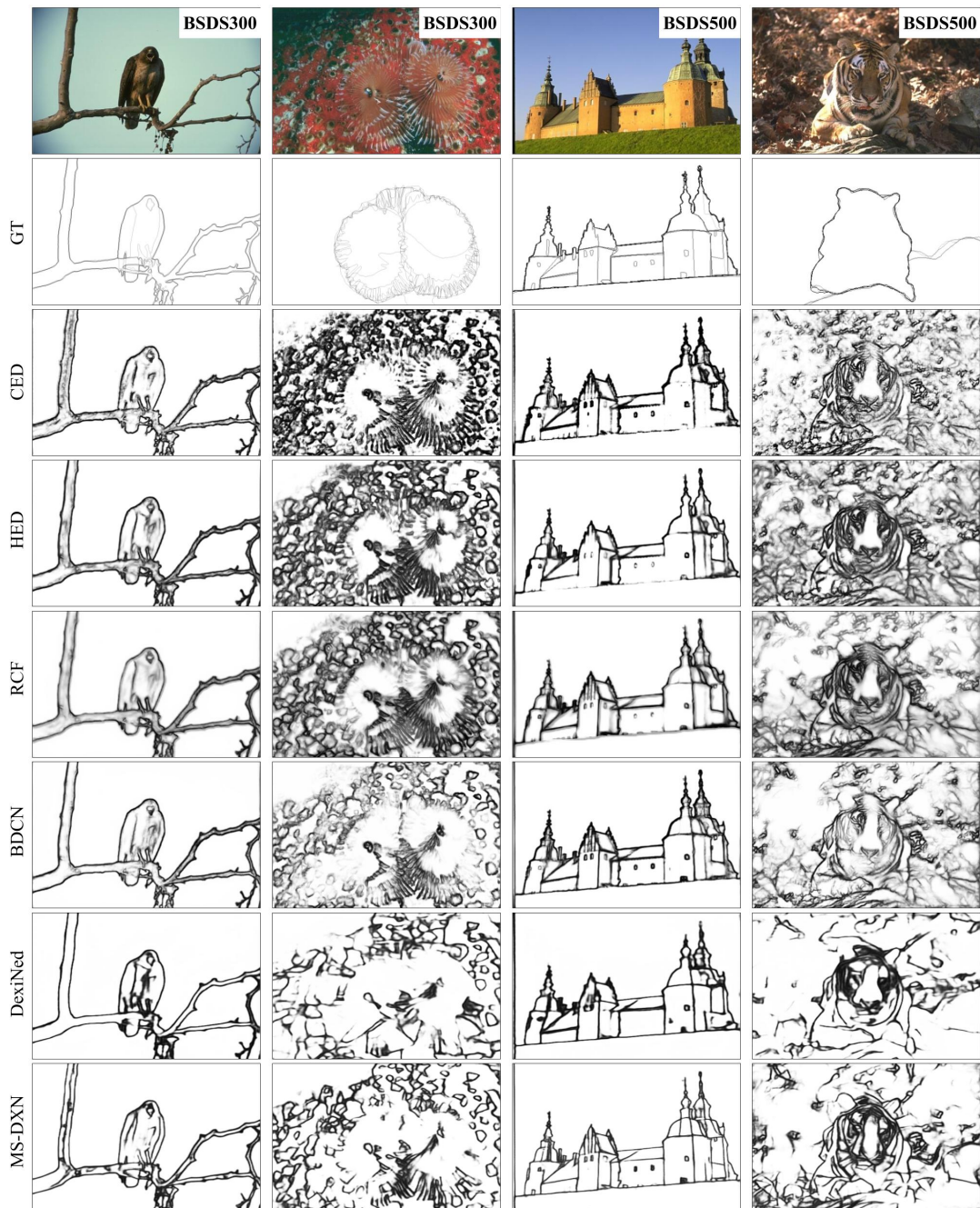


Figure 5.18: Four samples from the result obtained with the models depicted in Table 5.7. The label in the images (1st row) corresponds to the datasets where they are from.

5.3. Experimental Results

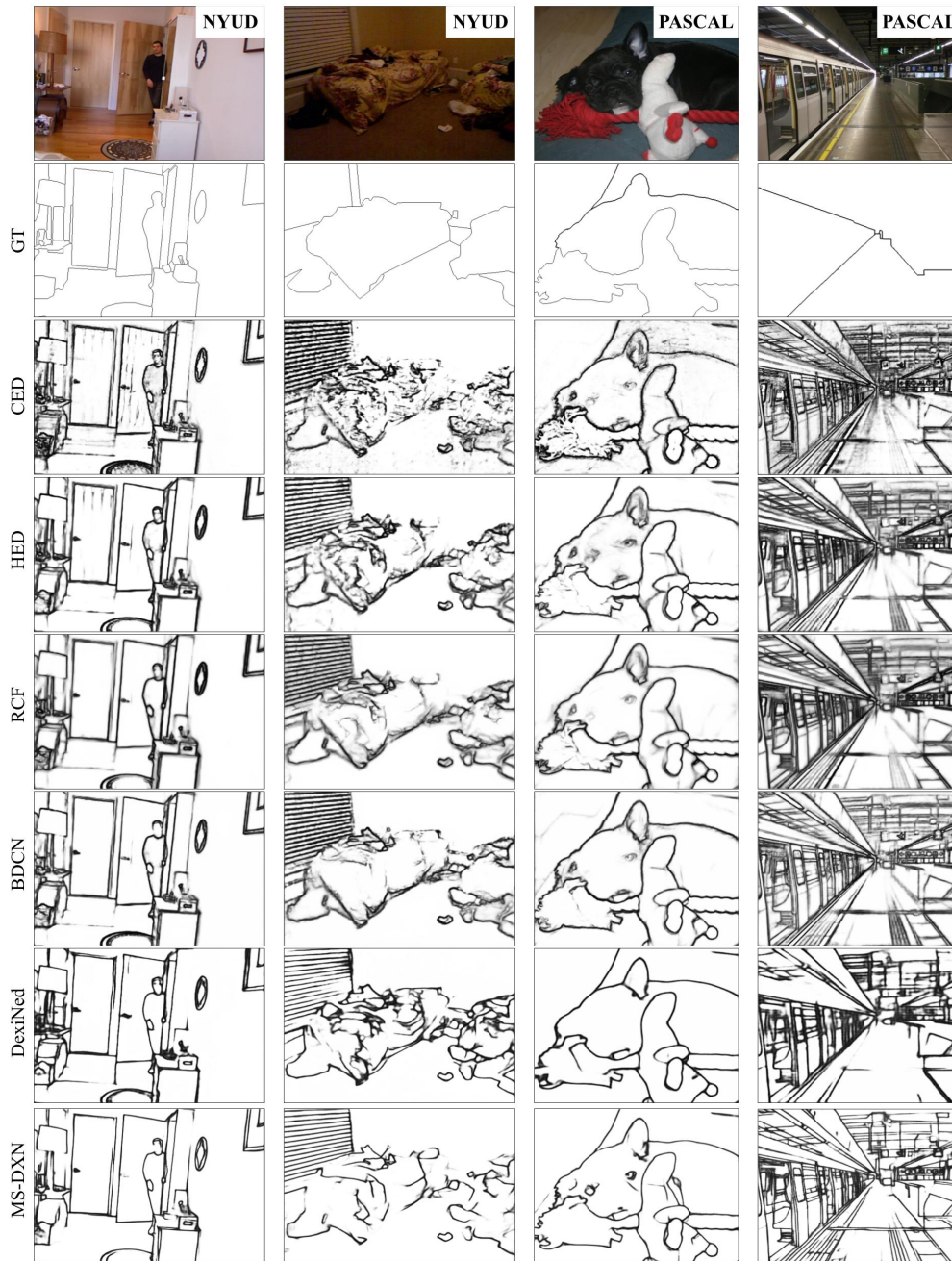


Figure 5.19: Four samples from the results obtained with the models depicted in Table 5.8. The label in the images (1st row) correspond to the datasets where they are from.

After the quantitative and qualitative evaluations on these datasets, the main observation is that, although the images from NYUD and PASCAL datasets have complex and diverse scenes (better than BSDS500) to evaluate the performance of the models for edge, contour or boundary detection, the ground truths (GT) are also diverse because of such images were generated for segmentation tasks. Therefore, most of the GTs are not annotated to edge, nor contour, nor boundary. The training of DL based approaches on such datasets do not guarantee the edge, contour or boundary detection.

5.4 Conclusion

This Chapter presents two approaches (DexiNed and MS-DXN), which are based on deep learning for edge detection task; these approaches are developed from scratch. The MS-HED, which is also proposed, loses most of the edges predicted by intermediate blocks; hence, it loses important high level edge features while it is used to train from the scratch and without pretrained weights. With a carefully designed CNN architecture and a dataset annotated for a respective task (i.e., BIPED for perceptual edge detection), the DL model can generalize the edge detection in any arbitrary image (grayscale or color) considered to evaluate the performance. The skip connections from DexiNed make that in the deeper layer, it does not lose edge features like MS-HED.

The adaptation of DexiNed to the multispectral domain (MS-DXN) has shown appealing results; furthermore it has been shown that this architecture is able to be used in case that just one spectral band is provided. The proposed multispectral model is trained by using images from two spectral bands, visible and near infrared; with this setting MS-DXN reaches 0.921 in average precision (see Table 5.5, this is the highest score reached in MBIPED dataset. Definitely, the usage of the NIR information helps to detect more edges than those predicted by using just RGB images; furthermore, by using two bands the edge predictions are similar to BDCN but it should be mentioned that MS-DXN has been trained without previous pre-training. The MS-DXN gives the best qualitatively and quantitatively results.

6 Conclusions and Future Work

6.1 Conclusions

This dissertation presents works focused on the processing of images acquired by single sensor cameras. These cameras are capable of capturing visible and a part of near infrared bands in a single shot. Whenever the image is acquired by such cameras in sufficient sunlight conditions, the visible bands (RGB channels) also have near infrared information, hence, the color of captured images looks desaturated. This phenomenon is termed in the literature as crosstalk between RGB and NIR channels. To face this problem, two approaches have been proposed in this thesis; the first one is based on multilayer perceptron (artificial neural network), while the second approach presents variants of convolutional neural network (CNN). Although in the state-of-the-art methods for color correction, the usage of a set of processes are proposed to face the RGB crosstalk with NIR, the proposals in this manuscript restore the RGB images in a single process. The result from these approaches, is an image with a quality where the desaturation problem is almost imperceptible for human eyes. Two datasets have been collected and shared to the community, SSOMSI and OMSIV, for the color restoration. Up to our knowledge, just these two datasets have images from outdoor with sufficient sunlight and with different material where the NIR absorption/reflectance is different than visible band.

Chapter 5 tackle the edge detection task for images captured by the single sensor camera, but generalizing the contribution so that the state-of-the-art dataset used for edge detection can be also processed. Before face the edge detection, two dataset have been collected and carefully annotated; these dataset are referred to as BIPED and MBIPED. Even though the images are in different dataset, those images have been collected at the same time and later on correctly aligned (registered) BIPED to MBIPED. The edge level annotation have been accomplished in two stages; firstly, the BIPED dataset has been annotated, later, MBIPED dataset. The annotation process in the MBIPED images is performed through image fused from RGB+NIR and NIR, with the aim to label as much information as possible from the two bands. The edge detection problem has been firstly tackled in the visible domain (BIPED) by a Dense Extreme Inception Network; secondly, with the images from MBIPED the multispectral edge detection has been considered. With the different methods proposed in this thesis,

edge detection task has been generalized from a Deep Learning model. In other words, the proposed models have been trained in a single dataset but in spite of that they reach the state-of-the-art results in the edge detection dataset MDBD (the sub set of edge detection from the Multicue Dataset for Boundary Detection). Concerning the qualitative results from the proposed approaches, the predicted edge-maps are thin and clearer than the prediction from the state-of-the-art approaches. Actually, the predicted edge-maps from BSDS300, BSDS500, NYUD (the most-used datasets for edge detection) are thin and clearer than never before.

6.1.1 List of Contributions

This thesis covers the following publications, in chronological order:

- Aguilera, C., **Soria, X.**, Sappa, A.D. and Toledo, R., 2017, June. RGBN multispectral Images: A Novel Color Restoration Approach. In International Conference on Practical Applications of Agents and Multi-Agent Systems (pp. 155-163). Springer, Cham.
- **Soria, X.**, Sappa, A.D. and Akbarinia, A., 2017, November. Multispectral Single-Sensor RGB-NIR imaging: New Challenges and Opportunities. In 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA) (pp. 1-6). IEEE.
- **Soria, X.**, Sappa, A. and Hammoud, R., 2018. Wide-Band Color Imagery Restoration for RGB-NIR Single Sensor Images. *Sensors*, 18(7), p.2059. **Journal.**
- **Soria, X.** and Sappa, A.D., 2018, November. Improving Edge Detection in RGB Images by Adding NIR Channel. In 2018 14th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS) (pp. 266-273). IEEE.
- **Soria, X.** and Riba, E., and Sappa, A.D., 2019. Dense Extreme Inception Network: Towards a Robust CNN Model for Edge Detection. In 2020 Winter Conference on Applications of Computer Vision (WACV'20). IEEE. (*under review*)

6.2 Future Work

Behind the methods proposed in this thesis there is a formal invitation to the usage of single sensor cameras in image processing and computer vision fields. In addition to the image sensors, we encourage the usage of deep learning in-single sensors camera operations just like major smartphones nowadays. In the dissertation, the RGB+NIR images have not been deeply studied, hence, the incursion of such type of images in the scene understanding is needed.

Although a large collection of models for RGB+NIR image restoration are proposed in the literature, it still remains as a open problem. As demonstrated in Chapter 4, with the CCN models the color correction of the RGB+NIR images, even in the challenging scenes, are accomplished till reach almost imperceptible difference in the human eyes; therefore, new color restoration techniques should consider the usage of DL. The CCN based approaches

proposed in this dissertation did not use NIR channel to subtract the percentage of NIR infected in the RGB images, hence, the NIR subtraction by using the NIR component of the RGB image should be a major analysis. Even though, OMSIV dataset has different materials to generalize the training of DL models, more images with several material and different environmental condition will be desirable. For example, there are hundred of colors of plants and the color restoration on these images will be a challenging task.

Edge detection, since the 1960s, is an active task that has been tackled from different models, which are based in low level features and learning algorithms. Concerning to the learning based algorithms, new optimized convolutional neural network have been designed, which overcome the result of ResNet or Inception v3 in the classification tasks; this new architectures (i.e., efficientNets) should be considered for edge detection. In Chapter 5 avoiding contour and boundary detection, edge detection is tackled, and as highlighted, the edge detection task should be separately study to the contour or boundary detection; if the contour/boundary detection problems were tackled, new datasets with the corresponding annotations will be generated. Furthermore, as demonstrated with multispectral dense extreme inception network (MS-DXN), adding NIR channel to the RGB image for edge detection, improve qualitatively and quantitatively the final performance; therefore, new multispectral CNN architectures should be considered also with the special attention in the NIR channel.

Bibliography

- [1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In OSDI, volume 16, pages 265–283, 2016.
- [2] Daron Acemoglu and Pascual Restrepo. Artificial antelligence, automation and work. Technical report, National Bureau of Economic Research, 2018.
- [3] James E Adams. Design of practical color filter array interpolation algorithms for digital cameras. 2. In Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269), volume 1, pages 488–492. IEEE, 1998.
- [4] James E Adams Jr. Design of practical color filter array interpolation algorithms for digital cameras. In Real-Time Imaging II, volume 3028, pages 117–125. International Society for Optics and Photonics, 1997.
- [5] Arash Akbarinia and C. Alejandro Parraga. Feedback and surround modulated boundary detection. International Journal of Computer Vision, 126(12):1367–1380, Dec 2018.
- [6] Arash Akbarinia, C Alejandro Parraga, et al. Biologically-inspired edge detection through surround modulation. In Proceedings of the British Machine Vision Conference, pages 1–13, 2016.
- [7] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 33(5):898–916, May 2011.
- [8] Kobus Barnard, Vlad Cardei, and Brian Funt. A comparison of computational color constancy algorithms. i: Methodology and experiments with synthesized data. IEEE Transactions on Image Processing, pages 972–984, 2002.
- [9] Imad A Basheer and Maha Hajmeer. Artificial neural networks: Fundamentals, computing, design, and application. Journal of Microbiological Methods, 43(1):3–31, 2000.
- [10] M. Basu. Gaussian-based edge-detection methods-a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 32(3):252–260, Aug 2002.

Bibliography

- [11] Mitra Basu. Gaussian derivative model for edge enhancement. Pattern Recognition, 27(11):1451–1461, 1994.
- [12] Durga Prasad Bavirisetti and Ravindra Dhuli. Two-scale image fusion of visible and infrared images using saliency detection. Infrared Physics & Technology, 76:52–64, 2016.
- [13] Bryce E Bayer. Color imaging array, jul 1976. US Patent 3,971,065.
- [14] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4380–4389, 2015.
- [15] Matthew Brown and Sabine Süsstrunk. Multi-spectral sift for scene category recognition. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 177–184. IEEE, 2011.
- [16] Harold C Burger, Christian J Schuler, and Stefan Harmeling. Image denoising: Can plain neural networks compete with bm3d? In IEEE Conference on Computer Vision and Pattern Recognition, pages 2392–2399. IEEE, 2012.
- [17] Peter Burt and Edward Adelson. The laplacian pyramid as a compact image code. IEEE Transactions on Communications, 31(4):532–540, 1983.
- [18] John Canny. A computational approach to edge detection. In Readings in computer vision, pages 184–203. Elsevier, 1987.
- [19] Zhenyue Chen, Xia Wang, and Rongguang Liang. Rgb-nir multispectral camera. Optics express, 22(5):4985–4994, 2014.
- [20] Zhenyue Chen, Nan Zhu, Shaun Pacheco, Xia Wang, and Rongguang Liang. Single camera imaging system for color and near-infrared fluorescence image guided surgery. Biomed. Opt. Express, 5(8):2791–2797, Aug 2014.
- [21] François Chollet. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1251–1258, 2017.
- [22] Ryan Dahl, Mohammad Norouzi, and Jonathon Shlens. Pixel recursive super resolution. arXiv preprint arXiv:1702.00783, 2017.
- [23] Jacobs David, Garg Rhul, P. KNAAN Yael, Wadhwa Neal, and Levoy Marc. Estimating depth using a single camera, apr 2019. "WorldWide" Patent: WO2019070299A1.
- [24] Li Deng, Dong Yu, et al. Deep learning: Methods and applications. Foundations and Trends® in Signal Processing, 7(3–4):197–387, 2014.
- [25] Piotr Dollár and C Lawrence Zitnick. Fast edge eetection using structured forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(8):1558–1570, 2015.

- [26] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2):295–307, 2015.
- [27] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2):295–307, 2016.
- [28] Weisheng Dong, Peiyao Wang, Wotao Yin, Guangming Shi, Fangfang Wu, and Xiaotong Lu. Denoising prior driven deep neural network for image restoration. arXiv preprint arXiv:1801.06756, 2018.
- [29] econ-systems company. Single sensor rgb-nir camera, 2015. [Accessed: 03-feb-2018].
- [30] Glenn Elert. The electromagnetic spectrum, the physics hypertextbook. Hypertextbook.com, 1998.
- [31] G. Evangelidis. Iat: A matlab toolbox for image alignment, 2013.
- [32] Hany Farid. Blind inverse gamma correction. IEEE Transactions on Image Processing, 10(10):1428–1433, 2001.
- [33] John Fengler, Paul Westwick, Arthur E Bailey, and Paul Cottle. Imaging system for combined full-color reflectance and near-infrared imaging, Nov 2015. US Patent 9,173,554.
- [34] Yaroslav Ganin and Victor Lempitsky. n^4 fields: Neural network nearest neighbor fields for image transforms. In Asian Conference on Computer Vision, pages 536–551. Springer, 2014.
- [35] Xin-Yi Gong, Hu Su, De Xu, Zheng-Tao Zhang, Fei Shen, and Hua-Bin Yang. An overview of contour detection approaches. International Journal of Automation and Computing, Jun 2018.
- [36] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT press, 2016.
- [37] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pages 2672–2680, 2014.
- [38] Cosmin Grigorescu, Nicolai Petkov, and Michel A Westenberg. Contour detection based on nonclassical receptive field inhibition. IEEE Transactions on Image Processing, 12(7):729–739, 2003.
- [39] Bahadir K Gunturk, John Glotzbach, Yucel Altunbasak, Ronald W Schafer, and Russel M Mersereau. Demosaicking: Color filter array interpolation. IEEE Signal processing magazine, 22(1):44–54, 2005.

Bibliography

- [40] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2013.
- [41] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. Information fusion, 14(2):127–135, 2013.
- [42] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhansu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In 2011 International Conference on Computer Vision, pages 991–998. IEEE, 2011.
- [43] Khizar Hayat. Super-resolution via deep learning. arXiv preprint arXiv:1706.09077, 2017.
- [44] Jianzhong He, Shiliang Zhang, Ming Yang, Yanhu Shan, and Tiejun Huang. Bi-directional cascade network for perceptual edge detection. arXiv preprint arXiv:1902.10903, 2019.
- [45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.
- [46] Robert Hecht-Nielsen. Neurocomputing. Reading Addison-Wesley, 1990.
- [47] Brittany Hillen. Cipas february 2019 report shows huge drop in global digital camera shipments, April 2019. [Accessed: 06-Jun-2019].
- [48] Guowei Hong, M Ronnier Luo, and Peter A Rhodes. A study of digital camera colorimetric characterisation based on polynomial modelling. Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur, 2001.
- [49] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In 20th International Conference on Pattern Recognition, pages 2366–2369. IEEE, 2010.
- [50] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.
- [51] X. Hu, F. Heide, Q. Dai, and G. Wetzstein. Convolutional sparse coding for rgb+nir imaging. IEEE Transactions on Image Processing, 27(4):1611–1625, April 2018.
- [52] Furkan Isikdogan, Alan Bovik, and Paola Passalacqua. Rivamap: An automated river analysis and mapping engine. Remote Sensing of Environment, 202:88–97, 2017.
- [53] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5967–5976. IEEE, 2017.

- [54] Viren Jain and Sebastian Seung. Natural image denoising with convolutional networks. In Advances in Neural Information Processing Systems, pages 769–776, 2009.
- [55] Robert Keys. Cubic convolution interpolation for digital image processing. IEEE Transactions on Acoustics, Speech, and Signal Processing, 29(6):1153–1160, 1981.
- [56] Haris Ahmad Khan, Jean-Baptiste Thomas, Jon Yngve Hardeberg, and Olivier Laligant. Illuminant estimation in multispectral imaging. JOSA A, 34(7):1085–1098, 2017.
- [57] Daisuke Kiku, Yusuke Monno, Masayuki Tanaka, and Masatoshi Okutomi. Residual interpolation for color image demosaicking. In 2013 IEEE International Conference on Image Processing, pages 2304–2308. IEEE, 2013.
- [58] Daisuke Kiku, Yusuke Monno, Masayuki Tanaka, and Masatoshi Okutomi. Minimized-laplacian residual interpolation for color image demosaicking. In Digital Photography X, volume 9023, page 90230L. International Society for Optics and Photonics, 2014.
- [59] Daisuke Kiku, Yusuke Monno, Masayuki Tanaka, and Masatoshi Okutomi. Simultaneous capturing of rgb and additional band images using hybrid color filter array. In Digital Photography X, volume 9023, page 90230V. International Society for Optics and Photonics, 2014.
- [60] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
- [61] Klaus Kohlmann. Corner detection in natural images based on the 2-d hilbert transform. Signal Processing, 48(3):225–234, 1996.
- [62] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, pages 1097–1105, 2012.
- [63] Rehm Lars. 2017, the year of the new mobile protocol, December 2017. [Accessed: 04-Jun-2019].
- [64] Rehm Lars. Camera expert reviews, 2019. [Accessed: 01-Jul-2019].
- [65] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10):1995, 1995.
- [66] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature, 521(7553):436, 2015.
- [67] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4):541–551, 1989.
- [68] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

Bibliography

- [69] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint, 2016.
- [70] José Lezama, Qiang Qiu, and Guillermo Sapiro. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6628–6637, 2017.
- [71] Mengtian Li, Zhe Lin, Radomír Měch, , Ersin Yumer, and Deva Ramanan. Photo-sketching: Inferring contour drawings from images. WACV, 2019.
- [72] Shutao Li, Xudong Kang, and Jianwen Hu. Image fusion with guided filtering. IEEE Transactions on Image Processing, 22(7):2864–2875, 2013.
- [73] Stan Li, Dong Yi, Zhen Lei, and Shengcai Liao. The casia nir-vis 2.0 face database. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 348–353, 2013.
- [74] Xin Li, Bahadır Gunturk, and Lei Zhang. Image demosaicing: A systematic survey. In Visual Communications and Image Processing 2008, volume 6822, page 68221J. International Society for Optics and Photonics, 2008.
- [75] Y. Liu, M. Cheng, X. Hu, J. Bian, L. Zhang, X. Bai, and J. Tang. Richer convolutional features for edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–1, 2019.
- [76] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Kai Wang, and Xiang Bai. Richer convolutional features for edge detection. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 5872–5881. IEEE, 2017.
- [77] Manolis Loukidakis, José Cano, and Michael O’Boyle. Accelerating deep neural networks on low power heterogeneous architectures. eprints, 2018.
- [78] Y. M. Lu, C. Fredembach, M. Vetterli, and S. Süsstrunk. Designing color filter arrays for the joint capture of visible and near-infrared images. In 2009 16th IEEE International Conference on Image Processing (ICIP), pages 3797–3800, Nov 2009.
- [79] Rastislav Lukac. Single-Sensor Imaging: Methods and Applications for Digital Cameras. CRC Press, 2008.
- [80] Rastislav Lukac. Single-Sensor Imaging: Methods and Applications for Digital Cameras. CRC Press, 2008.
- [81] Rastislav Lukac and Konstantinos N Plataniotis. Color filter arrays: Design and performance analysis. IEEE Transactions on Consumer Electronics, 51(4):1260–1267, 2005.

- [82] M Ronnier Luo, Guihua Cui, and B Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. Color Research & Application, 26(5):340–350, 2001.
- [83] Julia Lüthen, Julian Wörmann, Martin Kleinsteuber, and Johannes Steurer. A rgb/nir data set for evaluating dehazing algorithms. Electronic Imaging, 2017(12):79–87, 2017.
- [84] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. Information Fusion, 31:100–109, 2016.
- [85] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. Information Fusion, 45:153–178, 2019.
- [86] Michael Maire, Pablo Arbeláez, Charless Fowlkes, and Jitendra Malik. Using contours to detect and localize junctions in natural images. In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2008.
- [87] Henrique S Malvar, Li-wei He, and Ross Cutler. High-quality linear interpolation for demosaicing of bayer-patterned color images. In 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 3, pages iii–485. IEEE, 2004.
- [88] David Marr and Ellen Hildreth. Theory of edge detection. Proceedings of the Royal Society of London. Series B. Biological Sciences, 207(1167):187–217, 1980.
- [89] David R Martin, Charless C Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(5):530–549, 2004.
- [90] M. Martinello, A. Wajs, S. Quan, H. Lee, C. Lim, T. Woo, W. Lee, S. S. Kim, and D. Lee. Dual aperture photography: Image and depth from a mobile camera. In IEEE International Conference on Computational Photography, pages 1–10, April 2015.
- [91] Manuel Martinello, Andrew Wajs, Shuxue Quan, Hank Lee, Chien Lim, Taekun Woo, Wonho Lee, Sang-Sik Kim, and David Lee. Dual aperture photography: Image and depth from a mobile camera. In 2015 IEEE International Conference on Computational Photography (ICCP), pages 1–10. IEEE, 2015.
- [92] David A Mély, Junkyung Kim, Mason McGill, Yuliang Guo, and Thomas Serre. A systematic comparison between visual cues for boundary detection. Vision research, 120:93–107, 2016.
- [93] Daniele Menon and Giancarlo Calvagno. Color image demosaicking: An overview. Signal Processing: Image Communication, 26(8):518 – 533, 2011.
- [94] Yansheng Ming, Hongdong Li, and Xuming He. Contour completion without region segmentation. IEEE Transactions on Image Processing, 25(8):3597–3611, 2016.

Bibliography

- [95] Yusuke Monno, Daisuke Kiku, Masayuki Tanaka, and Masatoshi Okutomi. Adaptive residual interpolation for color image demosaicking. In 2015 IEEE International Conference on Image Processing (ICIP), pages 3861–3865. IEEE, 2015.
- [96] Yusuke Monno, Masayuki Tanaka, and Masatoshi Okutomi. N-to-srgb mapping for single-sensor multispectral imaging. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 33–40, 2015.
- [97] Yusuke Monno, Hayato Teranaka, Kazunori Yoshizaki, Masayuki Tanaka, and Masatoshi Okutomi. Single-sensor rgb-nir imaging: High-quality system design and prototype implementation. IEEE Sensors Journal, 19(2):497–507, 2018.
- [98] Yusuke Monno, Hayato Teranaka, Kazunori Yoshizaki, Masayuki Tanaka, and Masatoshi Okutomi. Single-sensor rgb-nir imaging: High-quality system design and prototype implementation. IEEE Sensors Journal, 19(2):497–507, 2018.
- [99] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 891–898, 2014.
- [100] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In 27th International Conference on Machine Learning (ICML), pages 807–814, 2010.
- [101] Michael Negnevitsky. Artificial Intelligence: a Guide to Intelligent Systems. Pearson Education, 2005.
- [102] Mark D Ohman, Russ E Davis, Jeffrey T Sherman, Kyle R Grindley, Benjamin M Whitmore, Catherine F Nickels, and Jeffrey S Ellen. Zooglider: An autonomous vehicle for optical and acoustic sensing of zooplankton. Limnology and Oceanography: Methods, 17(1):69–86, 2019.
- [103] Mohammadreza Asghari Oskoei and Huosheng Hu. A survey on edge detection methods. University of Essex, UK, 33, 2010.
- [104] Chulhee Park and Moon Gi Kang. Color restoration of rgbn multispectral filter array sensor images based on spectral decomposition. Sensors, 16(5):719, 2016.
- [105] J Anthony Parker, Robert V Kenyon, and Donald E Troxel. Comparison of interpolating methods for image resampling. IEEE Transactions on medical imaging, 2(1):31–39, 1983.
- [106] Pietro Perona, Jitendra Malik, et al. Detecting and localizing edges composed of steps, peaks and roofs. MIT, 1991.
- [107] Reza Pourreza, Ying Zhuge, Holly Ning, and Robert Miller. Brain tumor segmentation in mri scans using deeply-supervised neural networks. In International MICCAI Brainlesion Workshop, pages 320–331. Springer, 2017.

- [108] Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. Electronics letters, 38(7):313–315, 2002.
- [109] Rajeev Ramanath, Wesley E Snyder, Youngjun Yoo, and Mark S Drew. Color image processing pipeline. IEEE Signal Processing Magazine, 22(1):34–43, 2005.
- [110] Xiaofeng Ren, Charless C Fowlkes, and Jitendra Malik. Scale-invariant contour completion using conditional random fields. In Tenth IEEE International Conference on Computer Vision (ICCV), volume 2, pages 1214–1221. Citeseer, 2005.
- [111] J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. Journal of Applied Remote Sensing, 2(1):023522, 2008.
- [112] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 234–241. Springer, 2015.
- [113] Azriel Rosenfeld and Avinash C. Kak. Digital Picture Processing: Volume 1. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2 edition, 1982.
- [114] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. Nature, 5(3):1, 1986.
- [115] Zahra Sadeghipoor Kermani. Joint acquisition of color and near-infrared images on a single sensor. Technical report, EPFL, 2015.
- [116] Neda Salamati, Clément Fredembach, and Sabine Süsstrunk. Material classification using color and nir images. In Color and Imaging Conference, number 1 in 2009, pages 216–222. Society for Imaging Science and Technology, 2009.
- [117] Angel Sappa, Juan Carvajal, Cristhian Aguilera, Miguel Oliveira, Dennis Romero, and Boris Vintimilla. Wavelet-based visible and infrared image fusion: A comparative study. Sensors, 16(6):861, 2016.
- [118] Lex Schaul, Clément Fredembach, and Sabine Süsstrunk. Color image dehazing using the near-infrared. In 2009 16th IEEE International Conference on Image Processing (ICIP), pages 1629–1632. IEEE, 2009.
- [119] Brian G Schunck. Edge detection with gaussian filters at multiple scales. In Proceedings of a Workshop on Computer Vision, Published by IEEE Computer Society Press, Washington, DC, pages 208–210, 1987.
- [120] Jun Shen and Serge Castan. An optimal linear operator for step edge detection. CVGIP: Graphical Models and Image Processing, 54(2):112–133, 1992.
- [121] Navid Shiee and Aseem Agarwala. Take your best selfie automatically, with photobooth on pixel 3, Apr 2019. [Accessed: 03-Jun-2019].

Bibliography

- [122] Ming-Yu Shih and Din-Chang Tseng. A wavelet-based multiresolution edge detection and tracking. Image and Vision Computing, 23(4):441 – 451, 2005.
- [123] Ivana Shopovska, Ljubomir Jovanov, and Wilfried Philips. Rgb-nir demosaicing using deep residual u-net. In 2018 26th Telecommunications Forum (TELFOR), pages 1–4. IEEE, 2018.
- [124] Jamie Shotton, Andrew Blake, and Roberto Cipolla. Multiscale categorical object recognition using contour fragments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(7):1270–1281, 2008.
- [125] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In European Conference on Computer Vision, pages 746–760. Springer, 2012.
- [126] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
- [127] DH Sliney. What is light? the visible spectrum and beyond. Eye, 30(2):222, 2016.
- [128] Irwin Sobel. Camera models and machine perception. Technical report, Computer Science Department, Technion, 1972.
- [129] Lingxiao Song, Man Zhang, Xiang Wu, and Ran He. Adversarial discriminative heterogeneous face recognition. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
- [130] Xavier Soria, Edgar Riba, and D. Angel Sappa. Dense extreme inception network: Towards a robust cnn model for edge detection. arXiv preprint arXiv:1902.10903, 2019.
- [131] Xavier Soria, Angel Sappa, and Riad Hammoud. Wide-band color imagery restoration for rgb-nir single sensor images. Sensors, 18(7):2059, 2018.
- [132] Xavier Soria, Angel D Sappa, and Arash Akbarinia. Multispectral single-sensor rgb-nir imaging: New challenges and opportunities. In 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), pages 1–6. IEEE, 2017.
- [133] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2818–2826, 2016.
- [134] Mariarosaria Taddeo and Luciano Floridi. How ai can be a force for good. Science, 361(6404):751–752, 2018.
- [135] Daniel Stanley Tan, Wei-Yang Chen, and Kai-Lung Hua. Deepdemosaicking: Adaptive image demosaicking via multiple deep fully convolutional networks. IEEE Transactions on Image Processing, 27(5):2408–2419, 2018.

- [136] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946, 2019. ICRL2019.
- [137] Runjie Tan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Color image demosaicking via deep residual learning. In 2017 IEEE International Conference on Multimedia and Expo (ICME), pages 793–798. IEEE, 2017.
- [138] Qiling Tang, Nong Sang, and Tianxu Zhang. Extraction of salient contours from cluttered scenes. Pattern Recognition, 40(11):3100–3109, 2007.
- [139] Hayato Teranaka, Yusuke Monno, Masayuki Tanaka, and Masatoshi Ok. Single-sensor rgb and nir image acquisition: Toward optimal performance by taking account of cfa pattern, demosaicking, and color correction. Electronic Imaging, 2016(18):1–6, 2016.
- [140] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 11(Dec):3371–3408, 2010.
- [141] X. Wang. Laplacian operator-based edge detectors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(5):886–890, May 2007.
- [142] Yupei Wang, Xin Zhao, and Kaiqi Huang. Deep crisp boundaries. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3892–3900, 2017.
- [143] Yupei Wang, Xin Zhao, Yin Li, and Kaiqi Huang. Deep crisp boundaries: From boundaries to higher-level tasks. IEEE Transactions on Image Processing, 28(3):1285–1298, 2018.
- [144] Yupei Wang, Xin Zhao, Yin Li, and Kaiqi Huang. Deep crisp boundaries: From boundaries to higher-level tasks. IEEE Transactions on Image Processing, 28(3):1285–1298, 2019.
- [145] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4):600–612, 2004.
- [146] Ching-Chih Weng, Homer Chen, and Chiou-Shann Fuh. A novel automatic white balance method for digital still cameras. In 2005 IEEE International Symposium on Circuits and Systems, pages 3801–3804. IEEE, 2005.
- [147] Nicolas Widynski and Max Mignotte. A multiscale particle filter framework for contour detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(10):1922–1935, 2014.
- [148] D Wu, L Feng, C Zhang, and Y He. Early detection of botrytis cinerea on eggplant leaves based on visible and near-infrared spectroscopy. Transactions of the ASABE, 51(3):1133–1139, 2008.

Bibliography

- [149] Ren Xiaofeng and Liefeng Bo. Discriminatively trained sparse code gradients for contour detection. In Advances in Neural Information Processing Systems, pages 584–592, 2012.
- [150] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. International Journal of Computer Vision, 125(1-3):3–18, 2017.
- [151] Jinru Xue and Baofeng Su. Significant remote sensing vegetation indices: a review of developments and applications. Journal of Sensors, 2017, 2017.
- [152] CS Xydeas, , and Vladimir Petrovic. Objective image fusion performance measure. Electronics letters, 36(4):308–309, 2000.
- [153] Hongju Yang, Yao Li, Xuefeng Yan, and Fuyuan Cao. Contourgan: Image contour detection with generative adversarial network. Knowledge-Based Systems, 164:21–28, 2019.
- [154] Kai-Fu Yang, Shao-Bing Gao, Ce-Feng Guo, Chao-Yi Li, and Yong-Jie Li. Boundary detection using double-opponency and spatial sparseness constraint. IEEE Transactions on Image Processing, 24(8):2565–2578, 2015.
- [155] Ming-Hsuan Yang, David J Kriegman, and Narendra Ahuja. Detecting faces in images: A survey. IEEE Transactions on pattern analysis and machine intelligence, 24(1):34–58, 2002.
- [156] Wei Ye and Kai-Kuang Ma. Color image demosaicing using iterative residual interpolation. IEEE Transactions on Image Processing, 24(12):5879–5891, 2015.
- [157] RICHARDA YOUNG. The gaussian derivative model for spatial vision. i- retinal mechanisms. Spatial vision, 2(4):273–293, 1987.
- [158] Carman K. M. Yuk, Oscar C. Au, Richard Li, and Sui-Yuk Lam. Color demosaicking using direction similarity in color difference spaces. 2007 IEEE International Symposium on Circuits and Systems, pages 1281–1284, 2007.
- [159] Zhiliang Zeng, Ying Kin Yu, and Kin Hong Wong. Adversarial network for edge detection. In 2018 Joint 7th International Conference on Informatics, Electronics Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision Pattern Recognition (icIVPR), 2018.
- [160] F. Zhang, X. Wu, X. Yang, W. Zhang, and L. Zhang. Robust color demosaicking with adaptation to varying spectral correlations. IEEE Transactions on Image Processing, 18(12):2706–2717, Dec 2009.
- [161] Kaihua Zhang, Lei Zhang, Kin-Man Lam, and David Zhang. A level set approach to image segmentation with intensity inhomogeneity. IEEE Transactions on Cybernetics, 46(2):546–557, 2016.

- [162] Yu Zhang, Lijia Zhang, Xiangzhi Bai, and Li Zhang. Infrared and visual image fusion through infrared feature extraction and visual information preservation. Infrared Physics & Technology, 83:227–237, 2017.
- [163] Zhiqiang Zhou, Bo Wang, Sun Li, and Mingjie Dong. Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with gaussian and bilateral filters. Information Fusion, 30:15–26, 2016.
- [164] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, pages 2223–2232, 2017.
- [165] Djemel Ziou, Salvatore Tabbone, et al. Edge detection techniques-an overview. Pattern Recognition and Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii, 8:537–559, 1998.