

# **Learning Based Depth Map Estimation: Considering Noise and Scene Categories**

*A THESIS*

*submitted by*

**Seema Kumari**  
**(PTS1501)**

*for the award of the degree*

*of*

**Master of Science**  
**(by Research)**



**SCHOOL OF COMPUTING AND ELECTRICAL ENGINEERING**  
**INDIAN INSTITUTE OF TECHNOLOGY MANDI**

**2019**



*“Difficulties in your life do not come to destroy you, but to help you realize your hidden potential and power; let difficulties know that you too are difficult” – Dr. A. P. J. Abdul Kalam*

*To My Parents*

**Smt. Santosh Devi and Sri. Dinesh Kumar**

*My Brothers*

**Dr. Rajeev Kumar, Mr. Vinay Kumar and Late Mohit Kumar**

*My Husband*

**Dr. Srimanta Mandal**

## **DECLARATION**

I hereby declare that the entire work embodied in this thesis is the result of the investigations carried out by me in the **School of Computing and Electrical Engineering, Indian Institute of Technology Mandi**, under the supervision of **Dr. Arnav Bhavsar**. This work has not been submitted elsewhere for any degree or diploma. In keeping with the general practice, due acknowledgments have been made wherever the work described is based on a finding of other investigators. In addition, I certify that no part of this work will, in future, be used for submission in my name, for the award of any other degree at any university.

Kamand, 175 005

Date:

**Seema Kumari**

## THESIS CERTIFICATE

This is to certify that the thesis titled **Learning Based Depth Map Estimation: Considering Noise and Scene Categories**, submitted by **Seema Kumari**, to the Indian Institute of Technology, Mandi, for the award of the degree of **Master of Science** (by Research), is a bonafide record of the research work done by her under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other institute or university for the award of any degree or diploma.

Kamand, 175 005

Date:

**Dr. Arnav Bhavsar**  
(Guide)

# Acknowledgments

The success and final outcome of this research work required a lot of guidance and assistance from my supervisor and I am extremely privileged to give my deepest thanks to my guide Dr. Arnav Bhavsar for his expert advice and encouragement throughout my research work. He has always been there to motivate and encourage me in the right direction. Also, I would like to thank all the members of annual progress committee that consists of Dr. Bharat Singh Rajpurohit, Dr. Anil K. Sao, Dr. Aditya Nigam and Dr. Syed Abbas for their valuable inputs in my research work.

I really want to thank my job supervisors Dr. Bharat Singh Rajpurohit and Dr. Satya-jitsinh A. Thakor for their constant support. I would like to also thank my colleagues Mr. Shivam Prajapati, Mr. Arun Kumar and Mr. Tarun Verma for all their support.

I would like to thank all MANAS Lab group members for all their support in all forms. Especially, I would like to extend my sincere gratitude to Dr. Renu M. Rameshan for her advice for research work as well as life. I also want to thank Sujeet Kumar, Krishan Sharma, Ranjeet R. Jha, Krati Gupta, Shikha Gupta and Arshdeep S. Boparai for their support in all forms.

I would like thank to my Pooja *bhabi*, Deepa *bhabi* and Puspa *bhabi* for taking care of me. I owe my gratitude to my Maa, Papa, Brother and other family members for their endless support, care, blessings in my entire journey. I also want to thank my brother Dr. Rajeev Kumar for his all support and care.

Importantly, I love to thank my dear husband Srimanta Mandal without his care, help, motivation, and support I could not have completed my research work smoothly. Last but not least, there are persons not mentioned here but, deserve my token of appreciation, I want to thank them all.

# ABSTRACT

**Keywords:** *Depth map from single intensity image estimation, CNN, Residual connection, Encoder-decoder, Hourglass, Perceptual loss, AWGN, Image denoising, Sparsity, Non-local grouping, Dictionary, Edge preservation, Scene classification*

3D scene analysis can play a crucial role in different 3D vision-related applications, where depth information is pivotal. However, accurate dense depth sensing through active depth sensors (e.g. Laser depth scanner) is costly. An alternative is to employ low-cost depth sensors, which yields noisy depth information. Another common alternative which is still prevalent is that of depth estimation from intensity images using stereo. However, in this case, establishing correspondences among the multiple viewpoints is often not accurate due to various issues such as illumination, occlusion and so on. Thus, in recent years learning based depth estimation from single intensity images has been explored. However, intensity images can be noisy due to sensor characteristics.

On these lines, we propose an approach to estimate depth from a single intensity image using a learning-based strategy. Here, we have developed a novel convolutional neural network (CNN) encoder-decoder architecture, which learns the depth information using example pairs of color images and their corresponding depth maps. The proposed model is based on an integration of residual connections within pooling (down-sampling) and up-sampling layers, and hourglass module which operates on the encoded features, thus processing these at various scales. Furthermore, the model is optimized under the constraints of perceptual loss as well as the mean squared error loss. The perceptual loss considers the high-level features, thus operating at a different scale of abstraction, which is complementary to the mean squared error loss that considers a pixel-to-pixel error.

Considering that the training and testing dataset can be noisy, the estimated depth may not be accurate. Although our depth estimation framework can handle low-level noise in the intensity test image, a higher level of noise can distract the estimated depth map. For this scenario, we propose a denoising algorithm for both intensity images and depth maps that can address higher levels of noise. It has been shown that for denoising, non-local similar patches play an important role. Nevertheless, noise may create ambiguity in finding similar



patches, hence it may degrade the results. However, most of the non-local similarity-based approaches do not consider the issue of noisy patch grouping. Hence, we propose to denoise an image by mitigating the issue of grouping non-local similar patches in the presence of noise in the transform domain using sparsity and edge-preserving constraints. The effectiveness of the transform domain grouping of patches is utilized for learning dictionaries and is further extended for achieving an initial approximation of sparse coefficient vector for the clean image patches. The results are further improved by employing edge preserving constraints and processing at coarser scales. Our technique is useful to preserve the surface discontinuities and prominent details in depth and intensity images while suppressing noise, and we demonstrate clear benefits of denoising.

Another aspect that is considered in this work, is whether an apriori knowledge of scene type can benefit in depth estimation. We demonstrate the improvement in estimating the depth map by classifying different indoor scenes and building different depth estimation models for scene types. Such an approach may be useful in an application involving a small and fixed number of scenes. In order to build a classifier, we have used a smaller version of Residual Convolutional Neural Network (ResNet-18) that discriminates between different indoor scenes (e.g. bookstore, dining, bathroom, classroom, and kitchen, etc.) even in presence of noise in testing images. Here, our denoising method can help in accurate estimation of the depth map. Such an approach can not only serve as an initial step of depth estimation but it can also be useful in scene classification/retrieval application.

# Contents

<b>Abstract</b>	<b>ii</b>
<b>List of Tables</b>	<b>vii</b>
<b>List of Figures</b>	<b>vii</b>
<b>Abbreviations</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Depth Sensing/Estimation Techniques . . . . .	2
1.1.1 Active Sensors . . . . .	2
1.1.2 Passive Techniques from Multiple Images . . . . .	3
1.1.3 Monocular Depth Estimation . . . . .	4
1.2 Objective and Scope of the Work . . . . .	5
1.2.1 Convolutional Neural Network for Depth Map Estimation from a Single Intensity Image . . . . .	6
1.2.2 Image Denoising . . . . .	7
1.2.3 Scene Classification . . . . .	9
1.3 Contributions of the Thesis . . . . .	10
1.4 Organization of the Thesis . . . . .	11
<b>2 Background</b>	<b>13</b>
2.1 Sparse Representation . . . . .	13
2.1.1 Linear Systems of Equations . . . . .	13
2.1.2 Regularization . . . . .	15
2.1.3 Sparse Modelling . . . . .	16

2.1.4	The Dictionary . . . . .	18
2.2	Convolutional Neural Network (CNN) . . . . .	18
2.2.1	Shallow vs Deep Convolutional Neural Network . . . . .	19
2.2.2	Residual Connections . . . . .	20
2.2.3	Basic Autoencoder Framework . . . . .	21
2.2.4	Autoencoder with Skip Connections . . . . .	23
2.3	Summary . . . . .	24
<b>3</b>	<b>Single Image Based Depth Map Estimation via Residual CNN Encoder-Decoder and Stacked Hourglass</b>	<b>25</b>
3.1	Introduction . . . . .	25
3.2	Related Work . . . . .	26
3.3	Proposed Approach for Single Image Depth Estimation with CNN . . . . .	29
3.3.1	Network Model . . . . .	29
3.3.2	Implementation Details . . . . .	31
3.3.3	Loss Function . . . . .	32
3.4	Experimentation . . . . .	33
3.4.1	Dataset . . . . .	33
3.4.2	Assessment metrics . . . . .	34
3.4.3	Results and Comparisons . . . . .	35
3.4.4	Ablation Study . . . . .	38
3.5	Summary . . . . .	41
<b>4</b>	<b>Sparse Representation Based Single Image Denoising with Transform Domain Patch Similarity</b>	<b>44</b>
4.1	Introduction . . . . .	44
4.2	Grouping Similar Patches: Noiseless vs Noisy . . . . .	47
4.3	Patch Similarity in Transform Domain . . . . .	49
4.3.1	Patch Grouping Using Transform Domain Similarity . . . . .	49
4.3.1.1	Learning Dictionaries . . . . .	50
4.3.1.2	Estimation of Sparse Coefficient Vector . . . . .	51
4.4	Edge Preservation . . . . .	51
4.4.1	Preserving edges at coarser scales . . . . .	51

4.4.1.1	Multi-scale edge preservation . . . . .	53
4.4.2	Preventing transitional smoothing . . . . .	54
4.5	Experimentation . . . . .	55
4.5.1	Comparison . . . . .	55
4.5.2	Denoising in Depth Map Estimation . . . . .	58
4.5.3	Comparisons and analysis among orthogonal transformations and filtering technique . . . . .	59
4.6	Summary . . . . .	61
<b>5</b>	<b>Scene Classification for Depth Map Estimation</b>	<b>64</b>
5.1	Introduction . . . . .	64
5.2	Related Work . . . . .	66
5.3	Proposed Framework . . . . .	68
5.3.1	Residual Learning . . . . .	68
5.3.2	Network Architecture . . . . .	68
5.3.3	Implementation Detail . . . . .	70
5.3.4	Depth Estimation with Scene Classification . . . . .	70
5.4	Experimentation . . . . .	70
5.4.1	Results for Classification . . . . .	71
5.4.2	Effect of Scene Classification in Depth Map Estimation . . . . .	73
5.5	Summary . . . . .	76
<b>6</b>	<b>Summary and Future Work</b>	<b>78</b>
6.1	Summary . . . . .	78
6.2	Future Work . . . . .	80
	<b>References</b>	<b>81</b>
	<b>List of Publications</b>	<b>90</b>