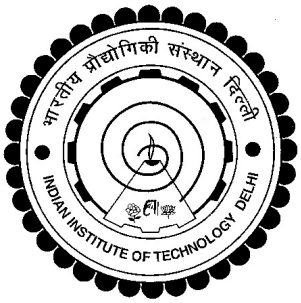
SELF SUPERVISED SCENE DEPTH AND EGO-MOTION ESTIMATION FOR AUTONOMOUS NAVIGATION

VINAY KAUSHIK



DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
OCTOBER 2021

SELF SUPERVISED SCENE DEPTH AND EGO-MOTION ESTIMATION FOR AUTONOMOUS NAVIGATION

by

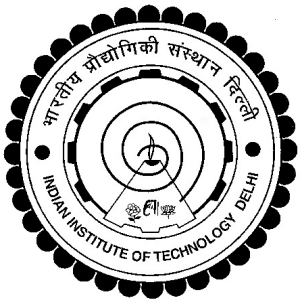
VINAY KAUSHIK

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2021

To my parents

Certificate

This is to certify that the thesis entitled "**SELF SUPERVISED SCENE DEPTH AND EGO-MOTION ESTIMATION FOR AUTONOMOUS NAVIGATION**" being submitted by **Mr. VINAY KAUSHIK** to the Department of Electrical Engineering, Indian Institute of Technology Delhi, for the award of the degree of **Doctor of Philosophy** is the record of the bona-fide research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted either in part or in full to any other university or institute for the award of any degree or diploma.

Prof. Brejesh Lall
Professor
Department of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi - 110016, India.

Date: 28-10-2021
Place: New Delhi

Acknowledgments

I would like to express my sincere gratitude to my advisor **Prof. Brejesh Lall**. The genesis of this thesis would not have been possible without his guidance, critical review and encouragement. His intelligent ideas and comprehensive understanding helped me to establish the overall direction of the research, got me past many significant challenges and always motivated me to strive for excellence.

I thank the members of my thesis committee: Prof. Subhasis Bannerjee, Prof. S.D. Joshi and Dr. Sumantra Dutta Roy for their insightful suggestions which encouraged me to widen my research from various perspectives. I sincerely admire the contribution of all my lab mates, lab incharge, project staff and most importantly, my friends for extending their unstinting support and cooperating attitude during the course of this study.

At the end I acknowledge the most important contribution of my parents. Their infallible love and reassuring rock solid support has always been my strength. They instilled in me the inspiration for excellence and hard work. They remain life-models worthy of emulation.

VINAY KAUSHIK

Abstract

Depth estimation is a significant problem in core computer vision. It's an invaluable fundamental task that has been computationally expensive since it's inception. A task which exemplifies these challenges and is of extreme importance is navigation in the wild. In this work, we attempt to address and propose novel solutions for the same.

A first attempt as part of this effort was complexity reduction of traditional stereo matching. We proposed a pyramidal approach for depth prediction that not only drastically reduced the computational overhead of computing depth, but also fused a threshold based median filter, thus minimizing post-processing overhead.

Although the method was very fast with reasonable accuracy, soon it became apparent [1, 2] that the capabilities of neural networks will drive the future solutions in computer vision including those for solving depth and egomotion. Being motivated by the self-supervised paradigm [2, 3], we propose novel geometric constraints and architectural innovations that improve the accuracy of learning depth from unlabeled stereo data. We propose techniques that improve learning of depth without increasing the complexity of the deep learning model.

Our second work utilises self-supervised learning for training our cascaded model in a two stage process, exploiting sub-pixel convolutions for depth super resolution. The first stage predicts depth, reconstructed images and warp error that is fed to

the second stage for refinement. Our third work further improves the depth accuracy while decreasing the number of training parameters by proposing an architecture that utilises deep feature fusion at multiple scales. We incorporate CoordConv solution for learning depth and propose a refinement module utilising residual fused features for learning super resolved depth.

Since monocular cameras is what one finds on most real world platforms such as mobile phone, smart glasses or camcorders. The contributions in this space are incomplete if not adapted for monocular cameras. Estimating camera ego-motion is a crucial component of navigation. Also, further improvement in accuracy can be achieved by simultaneous end-to-end depth and egomotion estimation. Thus our next work learns depth and ego-motion in an end-to-end manner. We propose temporal constraints, novel self driven losses and masking strategies that help in better learning of optimal depth and camera egomotion. Our fourth work introduces novel data augmentation and Minimum 3D consistency constraints along with L1-Regularisation over the auto-mask for learning depth and egomotion from a shared ResNet18 encoder. In, our final work we do an in-depth study of various attention models to learn robust features for improving learnt depth and egomotion. The combination of relational self-attention and contextual attention improves depth accuracy significantly. We also propose a novel architecture for learning egomotion that improves camera pose and learnt depth. We achieve state-of-the art in unsupervised depth and camera egomotion for KITTI 2015 benchmark.

In summary, we have proposed an architecture which is data agnostic and can work in typical monocular camera based platforms. This platform can easily be adapted for application to use cases which require to navigate and avoid obstacles in the wild.

सार

कोर कंप्यूटर विज़न में गहराई का अनुमान एक महत्वपूर्ण समस्या है। यह एक अमूल्य मौलिक कार्य है जो अपनी स्थापना के बाद से कम्प्यूटेशनल रूप से महंगा रहा है। यह कार्य जो इन चुनौतियों का उदाहरण देता है और अत्यधिक महत्व का है वह है जंगल में नेविगेशन। इस काम में, हमारे लिए उपन्यास समाधानों को संबोधित करने और प्रस्तावित करने का वैसा ही प्रयास करते हैं।

इस प्रयास के हिस्से के रूप में पहला प्रयास पारंपरिक स्टीरियो की जटिलता में कमी का मेल मिलाना। हमने गहराई की भविष्यवाणी के लिए एक पिरामिडल दृष्टिकोण प्रस्तावित किया है कि न केवल कंप्यूटिंग गहराई के कम्प्यूटेशनल ओवरहेड को काफी कम कर दिया है, लेकिन थ्रेशोल्ड आधारित माध्य फ़िल्टर भी जुड़ा हुआ है, यह इस प्रकार पोस्ट-प्रोसेसिंग ओवरहेड को कम करता है।

यद्यपि विधि उचित सटीकता के साथ बहुत तेज थी, जल्द ही यह बन गई स्पष्ट [1, 2] कि तंत्रिका नेटवर्क की क्षमताएं भविष्य के समाधानों को आगे बढ़ाएगी कंप्यूटर दृष्टि में गहराई और ईगोमोशन को हल करने के लिए शामिल हैं। प्रेरित होना स्व-पर्यवेक्षित प्रतिमान [2, 3] द्वारा, हम उपन्यास ज्यामितीय बाधाओं का प्रस्ताव करते हैं और वास्तुशिल्प नवाचार जो बिना लेबल वाले सीखने की गहराई की सटीकता में सुधार करते हैं स्टीरियो डेटा। हम ऐसी तकनीकों का प्रस्ताव करते हैं जो बिना बढ़ाए गहराई के सीखने में सुधार करती हैं गहन शिक्षण मॉडल की जटिलता।

हमारा दूसरा काम हमारे कैस्केड मॉडल को प्रशिक्षित करने के लिए स्व-पर्यवेक्षित शिक्षण का उपयोग करता है दो चरणों की प्रक्रिया में, गहराई से सुपर रिज़ॉल्यूशन के लिए उप-पिक्सेल संकल्पों का शोषण करना। पहला चरण गहराई, पुनर्निर्मित छवियों और ताना त्रुटि की भविष्यवाणी करता है जिसे शोधन के लिए दूसरे चरण में फीड किया जाता है। हमारा तीसरा काम गहराई की सटीकता में और सुधार करता है एक वास्तुकला का प्रस्ताव करके प्रशिक्षण मापदंडों की संख्या को कम करते हुए कई पैमानों पर डीप फीचर फ्यूजन का उपयोग करता है। हम इसके लिए CoordCov समाधान शामिल करते हैं सीखने की गहराई और के लिए अवशिष्ट फ़्यूज्ड सुविधाओं का उपयोग करते हुए एक शोधन मॉड्यूल का प्रस्ताव सुपर हल गहराई सीखना।

चूंकि मोनोकुलर कैमरे वही हैं जो वास्तविक दुनिया के अधिकांश प्लेटफॉर्म पर मिलते हैं जैसे मोबाइल फोन, स्मार्ट ग्लास या कैमकोर्डर के रूप में। इस स्थान में योगदान हैं एककोशिकीय कैमरों के लिए अनुकूलित नहीं होने पर अधूरा। कैमरा ईगोमोशन-गति का आकलन नेविगेशन का एक महत्वपूर्ण घटक है। साथ ही, सटीकता में और सुधार किया जा सकता है
एक साथ एंड-टू-एंड गहराई और ईगोमोशन अनुमान द्वारा प्राप्त किया गया। इस प्रकार हमारा अगला कार्य गहराई और ईगोमोशन को अंत से अंत तक सीखता है। हम अस्थायी प्रस्तावित करते हैं बाधाओं, उपन्यास स्व-चालित नुकसान और बेहतर सीखने में मदद करने वाली मास्किंग रणनीतियाँ इष्टतम गहराई और कैमरा ईगोमोशन की। हमारा चौथा काम उपन्यास डेटा पेश करता है एल 1-नियमन के साथ वृद्धि और न्यूनतम 3डी स्थिरता बाधाएं साझा ResNet18 एन्कोडर से गहराई और ईगोमोशन सीखने के लिए ऑटो-मास्क पर। हमारे अंतिम कार्य में, हम मजबूत सीखने के लिए विभिन्न ध्यान मॉडल का गहन अध्ययन करते हैं सीखी गई गहराई और ईगोमोशन में सुधार के लिए सुविधाएँ।
संबंधपरक का संयोजन आत्म-ध्यान और प्रासंगिक ध्यान गहराई सटीकता में काफी सुधार करता है। हम ईगोमोशन सीखने के लिए एक नई वास्तुकला का भी प्रस्ताव है जो कैमरा मुद्रा में सुधार करता है और गहराई सीखी। हम पर्यवेक्षित गहराई और कैमरे में अत्याधुनिक हासिल करते हैं KITTI 2015 बेंचमार्क के लिए इगोमोशन।

संक्षेप में, हमने एक आर्किटेक्चर प्रस्तावित किया है जो डेटा अज्ञेयवादी है और काम कर सकता है ठेठ एककोशिकीय कैमरा आधारित प्लेटफार्मों में। इस मंच को आसानी से अनुकूलित किया जा सकता है उन मामलों का उपयोग करने के लिए आवेदन के लिए जिन्हें नेविगेट करने और जंगली में बाधाओं से बचने की आवश्यकता होती है।

Contents

Certificate **i**

Acknowledgments **ii**

Abstract **iii**

List of Figures **xii**

List of Tables **xvi**

List of Abbreviations **xviii**

1 Introduction **1**

1.1 Overview 1

1.2 Motivation and Scope 5

1.3 Challenges & Objectives 8

1.4 Major Contributions of the Thesis 11

1.5 Layout of the Thesis 13

2 Literature Review **16**

2.1 Depth from Stereo Matching 18

2.2 Depth from Stereo Matching 19

2.2.1 Problems In Stereo Block-Matching 20
2.3 Supervised Depth Estimation 21
2.4 Self-supervised Depth Estimation 22
2.4.1 Self-supervised Learning of Depth from Stereo images 22
2.4.2 Self-supervised Learning of Depth from Monocular images . . 24
2.5 Self Supervised Egomotion Estimation 25

3 UnDispNet: A cascaded framework for self-supervised learning of depth from stereo images 29
3.1 Introduction 29
3.2 Proposed work 30
3.3 Methodology 31
3.3.1 Depth Prediction as Image Reconstruction 32
3.3.2 Network Architecture 32
3.3.3 Training Loss 36
3.3.4 Occlusion Regularization 37
3.4 Experiments and Results 39
3.4.1 Implementation Details 39
3.4.2 KITTI Split 41
3.4.3 Eigen Split 42
3.4.4 Make3D 42
3.4.5 Generalizing to Our dataset 44
3.4.6 Limitations 46
3.5 Summary 46

4 Architectural improvements for learning better depth by self-supervision 48
4.1 Introduction 48

4.2 Proposed Approach 49
4.3 Methodology 51
4.3.1 Depth Prediction as View Reconstruction 51
4.4 Experiments 54
4.4.1 KITTI split 55
4.4.2 Eigen Split 56
4.5 Summary 56

5 Structural constraints and data augmentation for predicting depth and camera ego-motion from a single camera 57

5.1 Introduction 57
5.2 Proposed Work 59
5.3 Methodology 60
5.3.1 Preliminaries 62
5.3.2 Training Model Architecture 64
5.3.3 Constraints for depth prediction 64
5.4 Experiments and Results 70
5.4.1 Dataset 70
5.4.2 Parameter Settings 70
5.5 Main Results 71
5.5.1 Comparison of Pose Encoder 72
5.5.2 Ablation Study of Losses 72
5.5.3 Progressive Training Strategy 74
5.6 Summary 74

6 Deep Self-Attention for Self-Supervised Monocular Depth and Camera Egomotion Estimation 76

6.1 Introduction . . . 76
6.1.1 Self-Attention in Deep learning . . . 77
6.2 Proposed Work . . . 78
6.3 Methodology . . . 80
6.3.1 Training Model Architecture . . . 81
6.3.2 Constraints for depth prediction . . . 82
6.4 Experiments and Results . . . 86
6.4.1 Dataset . . . 87
6.4.2 Technical Specifications and Parameter Settings . . . 88
6.4.3 Main Results . . . 89
6.4.4 Qualitative Analysis . . . 90
6.4.5 Make3D . . . 90
6.4.6 Ablation Study of Losses . . . 91
6.5 Visual Odometry Results . . . 100
6.5.1 Dataset and Metrics . . . 100
6.5.2 Detailed Analysis . . . 101
6.6 Summary . . . 102

7 A pyramidal hierarchical depth propagation approach for generating depth maps from stereo data 103

7.1 Introduction . . . 103
7.2 Proposed Work . . . 104
7.3 Methodology . . . 105
7.3.1 Selecting Disparity for highest Gaussian level stereo pair . . . 106
7.3.2 Disparity Refinement . . . 107
7.3.3 Hierarchical Disparity Computation . . . 108
7.3.4 Selective Median Filtering . . . 109

7.4 Results 110

8 Conclusions and Future work **114**

8.1 Conclusion of this research 114

8.2 Future work 117

Bibliography **118**

List of publications **137**

Technical Biography of the Author **138**

List of Figures

1.1 Challenges associated with depth estimation 3
1.2 Our five works are related by the central theme of depth and ego-motion estimation utilising different geometric constraints for learning the same. The work also incorporates strategies optimising architecture for learning richer features for developing smaller robust network architectures. 14
1.3 The structure of this thesis 15

2.1 Overview of self-supervised learning for Single Image Depth Prediction (SIDE). 23

3.1 Architecture for UnDispNet 32
3.2 Qualitative comparison performance of our method on specular surfaces such as car's windows. 33
3.3 Qualitative comparison of our self-supervised cascaded super-resolved monocular depth estimation method with previous state-of-the-art methods. Results on KITTI dataset [46] using Eigen split [18]. We predict fine-grained depth across boundaries and reduce shadow artifacts in occluded and far away regions 38

3.4 Qualitative results on Make3D [47]. Despite training on Cityscapes dataset, we achieve visually accurate results. 43
3.5 Qualitative results on our own dataset captured using GoPro Hero 6. 45

4.1 Qualitative results KITTI diving dataset 2015 [46] 50
4.2 Architecture for deep feature fusion encoder-decoder with depth refinement for predicting self-supervised depth. 51

5.1 Depth predicted from our network 59
5.2 Architecture diagram 60
5.3 Overview of the depth prediction and egomotion estimation framework. 61
5.4 Qualitative comparison of our self-supervised monocular depth estimation method with other contemporary methods. Results on KITTI dataset [46] using Eigen split [18]. We predict sharp accurate depth across boundaries and reduce artifacts in moving regions 63

6.1 Depth predicted from our network 77
6.2 Architecture diagram 79
6.3 Qualitative results on the KITTI Eigen split [19] test set. Our models perform better on thinner objects such as trees, signs and bollards, as well as being better at delineating difficult object boundaries. The depth of far objects including sky is further improved. 80
6.4 Overview of the depth prediction and egomotion estimation framework. 81
6.5 Contextual Attention Module from Yu [120] 85

6.6 Qualitative comparison of our proposed approach with other state-of-the-arts on Make3D [47] dataset. Despite training on KITTI dataset, we achieve visually accurate results than comparable methods. Here, A) Monodepth2 [22]; B) Monodepth2 dh [83]; C) MLDA-NET [124]. The depth is shown in color mode, and from yellow to purple means depth transition from 0 to respectively 90

6.7 Qualitative comparison of pose trajectory of various state of art methods on 09 sequence of KITTI 2015 dataset. 96

6.8 *Qualitative Ablation study of Trajectories estimated by our pose network on sequence 09 of the KITTI Odometry dataset. Here,* ***Aug*** *represents Augmentation Loss,* ***Sep Pose*** *is the separate pose framework module,* ***Att*** *is relational self-attention module,* ***MinSCL*** *is the minimum structural consistency loss.* 97

7.1 Block Diagram of Pyramidal Disparity Map Computation Method . . 106

7.2 a) Cost match for a 5x5 patch b) Disparity Map for a 5x5 patch. Median filtering is performed on selected green colour disparity values. 109

7.3 1st and 2nd Row: Original Image, 3rd and 4th Row: Disparity Map . 111

List of Tables

3.1 Results on the KITTI 2015 Stereo 200 training set disparity images [46].For training, K is the KITTI dataset [46] and CS is the cityscapes dataset [9]. pp is the disparity post processing specified by [3]. 35

3.2 Self-Supervised depth estimation results on the KITTI dataset [46] using the Eigen Split [18] for depths at cap 80m, as described in [18]. 40

3.3 Results on the Make3D dataset [47]. Methods marked with an * use ground truth depth data of Make3D training dataset for supervision to predict depth. Using the standard C1 metric, we use the standard 70m cap for computing all evaluation metrics. 44

4.1 Results on the KITTI 2015 Stereo 200 training set disparity images [46].For training, K is the KITTI dataset [46]. pp is the disparity post processing specified by [3]. 54

4.2 Self-Supervised depth estimation results on the KITTI dataset [46] using the Eigen Split [18] for depths at cap 80m, as described in [18]. 55

5.1 Self-supervised depth prediction results on KITTI dataset [46] trained at 640x192 resolution. Results on Eigen split [18] for depths at cap 80m, as described in [18]. 69

5.2 Effect of various losses on depth prediction at 512x192 image resolution on Eigen split [18]. We observe that the combination of Augmentation Loss, L1 regularisation in mask and Minimum Structure Consistency Loss with automasking (A) gives us the best depth results. 71
5.3 Comparing our shared encoder with different pose networks of Monodepth2 [22] for resolution 640x192 images. 73
5.4 Comparing results of progressive learning. Small model is trained at 512x192 resolution while the big model is trained at 640x192 resolution. We show that optimal training at smaller resolution further boosts depth prediction at Bigger resolution. Here Ft. implies finetuned model. 74

6.1 Self-supervised depth prediction results on KITTI dataset [46] trained at 1024 x 384 resolution. Results on Eigen split [18] for depths at cap 80m, as described in [18]. 87
6.2 Ablation study for depth prediction at 640x192 image resolution using ResNet18 Encoder on Eigen split [18]. We observe that the combination of L1 Regularisation(L1R), Augmentation Loss and our Attention framework gives us the best depth results. 88
6.3 Comparing our method at 640x192 resolution with other methods utilising same ResNet18 network backbone. 88
6.4 Make3D [47] results. All self-supervised monocular (M) methods use median scaling. 89
6.5 Comparing our method at 640x192 resolution with multiple variations of self-attention features. Augmentation loss is applied during training. Feature is the ResNet18 encoder's output feature. SS and MS mean single scale and multi-scale respectively. 91

6.6 Parameter tuning for Min-SCL loss at 640x192 resolution. Augmentation loss is applied during training. 92
6.7 Effect on taking per pixel minimum across multiple frames in SCL Loss. The Model is trained at 640x192 resolution. Augmentation loss is applied during training. 93
6.8 Effect on different separate pose frameworks on learning of monocular depth. The Model is trained at 640x192 resolution. Augmentation loss applied during training. 93
6.9 Ablation of adding contextual attention block in our depth decoder architecture. The block is only applied at single scale (encoder's output resolution). Augmentation loss is applied during training. Feature is the ResNet18 encoder's output feature. 94
6.10 Ablation of Augmentation loss hyper-parameter as well as model training parameters (Batch size and Learning Rate) on learning depth prediction. The encoder used is vanilla ResNet18 with no self-attention module as in Monodepth2 [21] for fair evaluation. The network is trained using images at 640x192 resolution. 95
6.11 Effect of fine-tuning the depth decoder with Frozen Encoder and Pose Framework at 640x192 resolution. 95
6.12 Progressive training at 1024x384 resolution. 98
6.13 Ablation study of various types of augmentation done and it's effect on depth prediction. 98
6.14 *Egomotion estimation results on sequence 09 of the KITTI Odometry dataset [46]. **M.** and **S.** stand for the "monocular sequences" and "stereo image pairs" respectively.* 99

6.15 *Ablation study on Egomotion estimation results on sequence 09 of the KITTI Odometry dataset. Here,* ***Aug*** *represents Augmentation Loss,* ***Sep Pose*** *is the separate pose framework module,* ***Att*** *is relational self-attention module,* ***MinSCL*** *is the minimum structural consistency loss.* 100

7.1 Average runtime execution 110

7.2 Average Disparity Error 112

List of Abbreviations

AR Augmented Reality. 10, 25

BM Block Matching. 20, 110

CNN Convolutional Neural Network(s). 19, 73, 85

CS CityScapes (Dataset). xiii, 35, 40

CT Census Transform. 20

DoF Degree(s) of Freedom. 2, 3, 16, 17, 61, 62, 72, 81, 117

DSI Depth (Disparity) Search Image. 20, 105, 107

ELU Exponential Linear Unit. 64, 82

FPGA Field-Programmable Gate Array. 20

FPN Feature Pyramid Network. 49, 50

FPS Frames Per Second. 5, 18, 46

GAN Generative Adversarial Network(s). 56, 77, 118

GPU Graphics Processing Unit. 5, 39, 55, 70, 73, 88

ICP Iterative Closest Point. 3

LIDAR Laser Imaging Detection And Ranging. 1, 6, 16, 19, 22, 26

NCC Normalized Cross Correlation. 3, 17, 20, 21, 103

PnP Perspective N Point. 2, 17

RADAR Radio Detection And Ranging. 1, 6, 16, 26

ResNet Residual Network(s). iv, xiv, xv, 6, 49, 55, 59, 72, 73, 78, 80, 81, 84, 88, 90, 91, 94, 116

RGB Red Green Blue. 1, 8, 13, 16, 25, 26, 27, 61, 64, 81, 98, 114

RT Rank Transform. 20

SAD Sum of Absolute Difference. 3, 17, 20, 21, 103

SCL Structural Consistency Loss. xii, xv, 71, 72, 73, 77, 92, 93, 94, 97

SGM Semi Global Matching. 20, 110

SLAM Simultaneous Localization and Mapping. 2, 3, 16, 17, 25, 26, 138

SSD Sum of Squared Difference. 3, 5, 17, 18, 21, 103

SSIM Structural SIMilarity. 5, 18, 36, 65, 83

VO Visual Odometry. 2, 25

VR Virtual Reality. 25, 117

VRAM Virtual Random Access Memory. 88

ZNCC Zero-mean Normalized Cross Correlation. 20, 21, 103, 108