Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen
International Center for
Information Technology

Master Thesis Proposal

Efficient Multi-View Stereo Temporal Fusion

Manoj Kolpe Lingappa

Supervised by

Prof. Dr Nico Hochgeschwender
Second Supervisor
MSc. Deebul Sivarajan Nair

Month 2022

1 Introduction

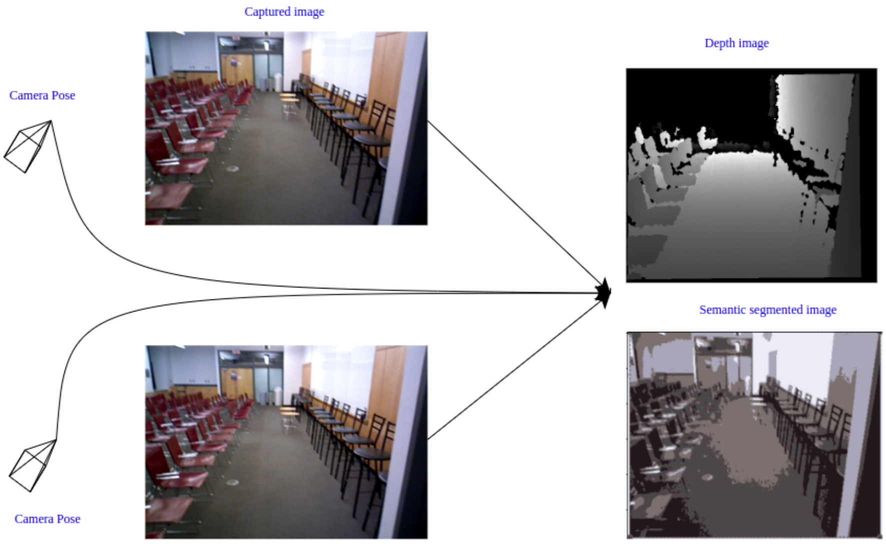


Figure 1: Depth estimation/Semantic segmentation from pair of images and pose. Courtesy of [13]

Computer vision aims to understand the surrounding environment using various mathematical modeling techniques. Reconstruction of three-dimensional views from images is a classic problem in the computer vision domain. Multi-view stereo algorithms can reconstruct the disparity maps or three-dimensional view of an object from the images [4]. It is the process of reproducing the 3D scenes from the multiple images given the camera poses and internal camera matrix. A number of areas take advantage of the reconstruction such as 3D mapping, 3D printing, video games, online shopping in the consumer domain, visual effect industry, digital mapping [11], vehicle tracking, aircraft estimation, and positioning [32], depth estimation [43]. Depth estimation can be performed with the active and passive image-based methods [13].

With the development of 3D reconstruction, active depth sensors are becoming increasingly popular in areas such as self-driving cars. These sensors are used to obtain information about the surrounding environment. However, the acquired depth maps/segmented images from these sensors are sparse in nature due to low computational power resulting in information loss of the captured depth map.

Another approach to reconstructing the 3D scene of an object is with the help of high-quality images captured from the camera where the texture and lighting information is captured [47].

Depth estimation using images is the process of extracting the depth of objects present in the images by capturing and processing multiple images of the object taken from different locations. Semantic segmentation is the process of labeling the each pixel in a image to a class. First-generation depth estimation was based on pixel matching between multiple images taken from calibrated cameras [28]. Images can be obtained from a stereo camera or a monocular camera. This work is based on monocular camera images. Estimation of depth/semantic segmentation from the unconstrained monocular camera images is a challenging task. Most of the multi task state-of-the-art depth estimation/ semantic segmentation algorithms are based on deep learning and compute cost volume according to the hypothesized depths. 3D convolution is applied to this cost volume to regress and predict the depth map [14]. Semantic maps can be estimated from the monocular images using semantic segmentation [16]. The goal of this thesis is to study the functionality testing with different cost volumes and, study the impact of the Gaussian process on depth estimation and deploy depth estimation algorithms in a mobile device, finally, end the research work with a feasibility study of depth estimation architecture to cross-domain application such as segmentation or object detection.

2 Problem Statement

Multi-view stereo is one of the fields of computer vision that targets to construct the most likely 3D model of an object using images. Reconstruction of the true 3D geometry is an ill-posed problem. Over the past years, a large number of algorithms and architectures have been proposed to find the 3D geometry of the object. However, a lack of datasets taken at varying environmental conditions made it difficult to compare the performance of the algorithms [36]. It takes a lot of time to process large images and with the low textured images a bad reconstruction is observed[21], [36], [38]. Most the state-of-the-art depth estimation algorithm is computationally heavy and cannot be deployed on the edge device. A lightweight architecture with reasonable performance needs to be developed to deploy in low computational power devices. Conventional approaches use two-view stereo rigs for reconstruction. However, estimation of depth from unconstrained monocular camera images is a challenging task. There are advantages to using a moving camera. Firstly with a larger baseline, the accuracy of the distant object can be improved. Secondly, multiple varying point images are able to fuse all the information for robust and stable depth estimation [17]. This work concentrate on the depth estimation from unconstrained monocular camera images, deployment on the edge device, and extension of the disparity map estimation architecture to the semantic segmentation.

3 Related Work

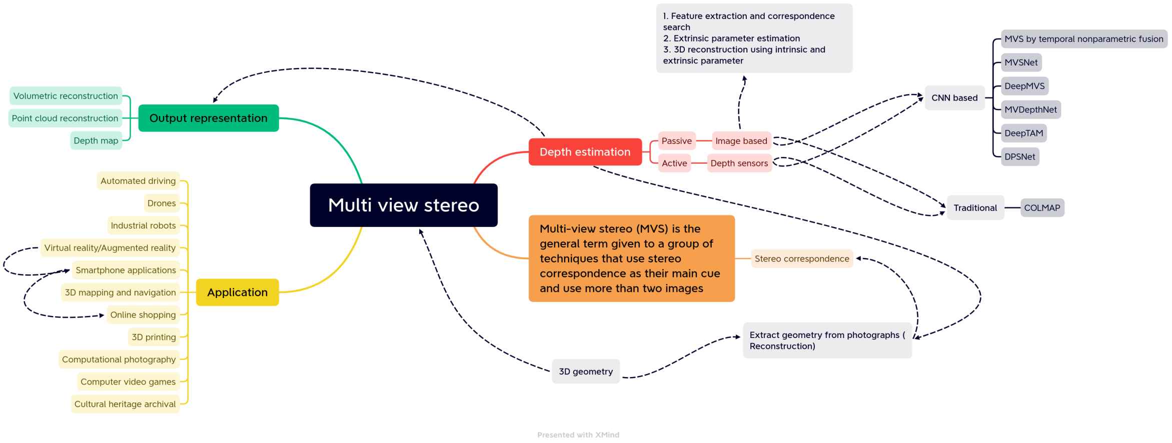

Multi view stereo
Output representation
Volumetric reconstruction
Point cloud reconstruction
Depth map
Application
Automated driving
Drones
Industrial robots
Virtual reality/Augmented reality
Smartphone applications
3D mapping and navigation
Online shopping
3D printing
Computational photography
Computer video games
Cultural heritage archival
Depth estimation
Passive
Image based
Active
Depth sensors
1. Feature extraction and correspondence search
2. Extrinsic parameter estimation
3. 3D reconstruction using intrinsic and extrinsic parameter
CNN based
MVS by temporal nonparametric fusion
MVSNet
DeepMVS
MVDepthNet
DeepTAM
DPSNet
Traditional
COLMAP
Multi-view stereo (MVS) is the general term given to a group of techniques that use stereo correspondence as their main cue and use more than two images
Stereo correspondence
Extract geometry from photographs (Reconstruction)
3D geometry
Presented with XMind

Figure 2: Multi view stereo mind map

Multi view stereo (MVS) is a general term given to the group of techniques that uses stereo correspondence to find the geometry of a object using the images captured from different viewpoint. 3D reconstruction of the target object can be done with classical [5], [8], [27], [26], [9], [13] or modern deep learning based approaches [44], [20], [45], [6]. Goal of the image based 3D reconstruction algorithm can be defined as “given a set of photographs of an object or a scene, estimate the most likely 3D shape that explains those photographs, under the assumptions of known materials, viewpoints, and lighting conditions” [11]. With a set of assumptions state of the art architecture can produce highly detailed reconstructions from large set of images.

Multi-view stereo has a wide range of applications ranging from automated driving, augmented reality, drones, Automated industrial robots, Online shopping, 3D printing, 3D mapping, and navigation. With the increased research on the depth sensors like LIDARs are becoming increasingly cheap and widely used in robotics, and cellphones. However, there is a loss of information with these sensors, and this promoted the development of the image-based depth estimation [40].

MVS is classified as follows (a) volumetric reconstruction method [24] (b) point cloud reconstruction [12] and (c) depth map based method [43]. In a pairwise stereo method image, rectification is performed to limit the correspondences found in the horizontal epipolar lines. This problem is addressed by the volumetric representation of the view [25], [19], [10], [42], [29]. Due to high memory load, the volumetric approach is not suitable for large scenes but it gives good performance for small objects. A lightweight architecture is proposed by Wang et al. A matching cost volume is computed using the plane sweeping approach from the nearby images and then regard depth estimation as a regression problem which is found using the deep neural network [39]. The plane sweep volume method does not require any rectified image. However, the approach requires intrinsic and extrinsic camera parameters in advance or can be computed using structure from motion [19].

Depth estimation can be performed with active and passive image-based methods. Learning-based depth reconstruction can be described as finding a predictor f_θ that can find the depth maps $\hat{D}$ from the set of images I, which are close to the unknown depth map D. Mathematically, we are trying to find a function f_θ such that the loss function $L(I) = d(f_\theta(I), D)$ is minimized. Where θ is the learnable parameter, and $d(.)$ is the measure of distance between the real depth D and the predicted depth $\hat{D}$ [28]. There is two class of depth estimation methods. The first class of the method involves traditional stereo matching approaches to find the correspondences which in turn help to find the disparity map. A depth map can be found from this disparity map [35]. There are three stages for the prediction of function f, first is the feature extraction, feature matching, and cost aggregation, and finally, depth estimation [28]. The second class of method involves an end-to-end trainable network [46]. Training requires a large amount of data and these approaches are similar to the traditional stereo matching algorithm by breaking the problem into small chunks and computing the result[28]. Early multi-view stereo methods work on finding the correspondences between multiple image patches[15]. Most of the state-of-the-art disparity map estimation architecture requires high computational power thereby limiting the deployment in the low computational devices such as mobile phones, and tablets. Work by [17] deploys the depth estimation architecture on an IOS device, current work aims to deploy the architecture in an android device and also find the feasibility of the temporal depth estimation architecture on to the semantic segmentation task. Semantic segmentation deals with the idea of correctly identifying the objects in a image and localizing the object by labeling the pixels. Classical segmentation task is done with the help of decision trees [37] or Markov random field [41]. A deep learning approach was proposed to perform semantic segmentation [7]. A efficient Fully connected network (FCN) [31] was developed with computation shared between overlapping region. A encoder-decoder U-net architecture used to perform the semantic segmentation task especially in the context of medical imaging [33]. Following years a similar architecture to U-Net known as SegNet [1] was introduced. And it does not entirely transfer feature map from encoder to the decoder, rather transfer only the max pooling indices [3].

4 Research questions

Three research questions are defined for the master thesis

RQ1 What are the works on state-of-the-art temporal fusion?

RQ2 How are the results from RQ1 compared with each other to perform temporal fusion?

RQ2.1 What are the results in comparison with different error metrics?

RQ3 How to cross-transfer the temporal nonparametric fusion to the other tasks, such as object detection or semantic segmentation?

RQ3.1 How different loss criteria impact the performance of the semantic segmentation?

RQ3.2 What is the performance of semantic segmentation with respect to different Gaussian kernels?

5 Project Plan

The following sections explain work packages, milestones and project schedules, and deliverables.

5.1 Work Packages

The bare minimum will include the following packages:

WP1 Literature Search
This section aims to extensively search for references to papers that are related to multi-view stereo.

T1.1 Literature review
In this task collection of literature related to multi-view stereo is done and conceptual understanding of the 3D geometry from images.

WP2 Data aggregation and preprocessing
This section explains the data collection and data preprocessing.

T2.1 Data collection
In this section, data is collected from multiple sources, and the nature of the data is examined and analyzed using visualization tools or statistical methods. An analysis is carried out to ensure data is diverse, unbiased, and abundant in nature.

T2.2 Data preprocessing
Preprocessing of data is carried out based on the input requirement of the model. The preprocessing step converts the raw sourced data into a format that enables successful training of the model.

WP3 Model implementation

This section explains the development and implementation of the model.

T3.1 Evaluation of the model

This task aims to reproduce the Multi-view Stereo by Temporal Nonparametric Fusion architecture results.

T3.2 Cross application of the MVS temporal fusion to the segmentation

In this section, the extension of multi-view stereo disparity estimation architecture to the other application areas of computer vision is carried out.

WP4 Evaluation

This package aims to evaluate the results based on the different metrics.

T4.1 Results reporting

In this task, the output of the evaluation is reported.

WP5 Project Report

This work package involves writing the project report. It is done in parallel with all previous work packages.

T5.1 Documentation of reviewed literature

In this task, a detailed analysis of the state of the art is done and all the findings are documented in the project report.

T5.2 Documentation of baseline results

In this task, the implementation result of Multi-view Stereo by Temporal Nonparametric Fusion baseline is done.

T5.3 Documentation on the results for the different temporal fusion This task documents the result of different temporal fusion architecture with different error metric is found.

T5.4 Documentation of cross-domain application of MVS approach
In this task, result of the cross-domain application of the MVS approach is performed.

5.2 Milestones

M1 Literature search

M2 Data collection and preprocessing

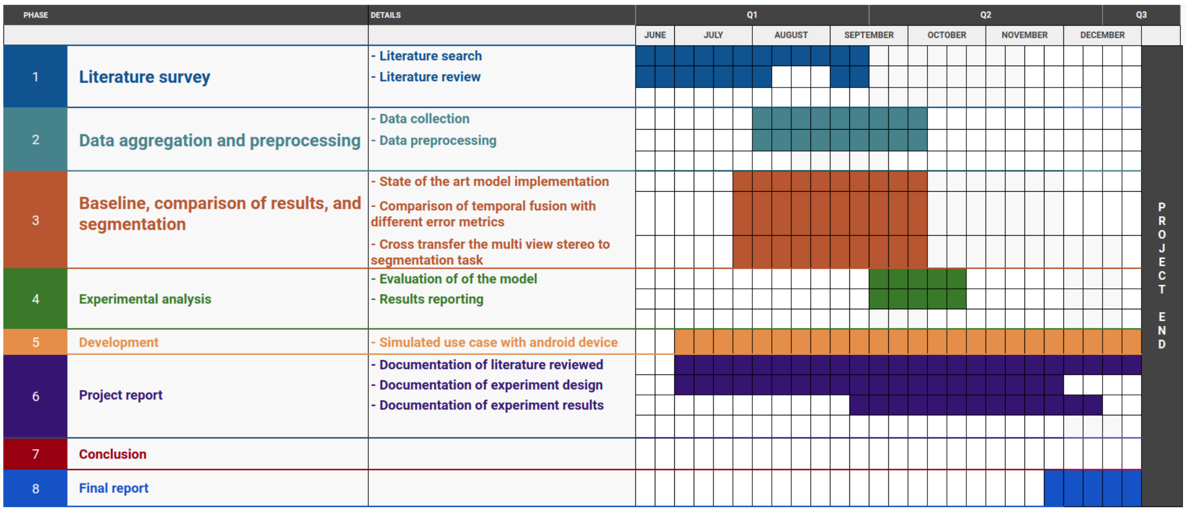
M3 Building a baseline

M4 Experimental Analysis

M5 Development

M6 Report submission

5.3 Project Schedule



PHASE		DETAILS	Q1												Q2												Q3		
			JUNE		JULY				AUGUST				SEPTEMBER				OCTOBER				NOVEMBER				DECEMBER				
1	Literature survey	- Literature search	■	■	■	■	■	■	■	■	■	■	■	■															PROJECT END
		- Literature review	■	■	■	■	■	■	■				■	■															
2	Data aggregation and preprocessing	- Data collection							■	■	■	■	■	■	■	■	■												
		- Data preprocessing							■	■	■	■	■	■	■	■	■												
3	Baseline, comparison of results, and segmentation	- State of the art model implementation						■	■	■	■	■	■	■	■	■	■												
		- Comparison of temporal fusion with different error metrics						■	■	■	■	■	■	■	■	■	■												
		- Cross transfer the multi view stereo to segmentation task						■	■	■	■	■	■	■	■	■	■												
4	Experimental analysis	- Evaluation of of the model													■	■	■	■	■										
		- Results reporting													■	■	■	■	■										
5	Development	- Simulated use case with android device			■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	
6	Project report	- Documentation of literature reviewed			■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	
		- Documentation of experiment design			■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■					
		- Documentation of experiment results											■	■	■	■	■	■	■	■	■	■	■	■	■	■			
7	Conclusion																												
8	Final report																								■	■	■	■	■	■	

Figure 3: Timeline of the project

5.4 Deliverables

Minimum viable

- Literature review on multi view stereo temporal fusion in the context of depth estimation and semantic segmentation
- Analysis of the state of the art temporal fusion architectures
- Create a baseline of multi view stereo temporal fusion with images from monocular camera

Expected

- Compare performances of state of the art multi view stereo temporal fusion techniques with different error metrics
- Simple simulated use case of temporal fusion on the android device

Maximum

- Cross transfer the multi view stereo temporal fusion architecture to the segmentation task
- Evaluation of the temporal segmentation method with different loss criteria
- Improved monocular image multi view temporal fusion technique
- Performance of multi view stereo temporal fusion with respect to different Gaussian kernels

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[2] Gianfranco Bianco, Alessandro Gallo, Fabio Bruno, and Maurizio Muzzupappa. A comparative analysis between active and passive techniques for underwater 3d reconstruction of close-range objects. *Sensors*, 13(8):11007–11031, 2013.

[3] Albert Bou. Deep learning models for semantic segmentation of mammography screenings, 2019.

[4] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021.

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021.

[6] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020.

[7] Dan Ciresan, Alessandro Giusti, Luca Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in neural information processing systems*, 25, 2012.

[8] Jeremy S De Bonet and Paul Viola. Poxels: Probabilistic voxelized volume reconstruction. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 418–425. Citeseer, 1999.

[9] Carlos Hernández Esteban and Francis Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3): 367–392, 2004.

[10] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 5515–5524, 2016.

[11] Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

[12] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.

[13] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.

[14] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020.

[15] Wilfried Hartmann, Silvano Galliani, Michal Havlena, Luc Van Gool, and Konrad Schindler. Learned multi-patch similarity. In *Proceedings of the IEEE international conference on computer vision*, pages 1586–1594, 2017.

[16] Lei He, Jiwen Lu, Guanghui Wang, Shiyu Song, and Jie Zhou. Sosd-net: Joint semantic object segmentation and depth estimation from monocular images. *Neurocomputing*, 440:251–263, 2021.

[17] Yuxin Hou, Juho Kannala, and Arno Solin. Multi-view stereo by temporal non-parametric fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2660, 2019.

[18] Yuxin Hou, Arno Solin, and Juho Kannala. Unstructured multi-view depth estimation using mask-based multiplane representation. In *Scandinavian Conference on Image Analysis*, pages 54–66. Springer, 2019.

[19] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018.

[20] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. *arXiv preprint arXiv:1905.00538*, 2019.

[21] Michal Jancosek and Tomás Pajdla. *Segmentation based multi-view stereo.* Citeseer, 2009.

[22] Ray A Jarvis. A perspective on range finding techniques for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):122–139, 1983.

[23] Ray A Jarvis. A perspective on range finding techniques for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):122–139, 1983.

[24] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. *Advances in neural information processing systems*, 30, 2017.

[25] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. *Advances in neural information processing systems*, 30, 2017.

[26] Vladimir Kolmogorov and Ramin Zabih. Multi-camera scene reconstruction via graph cuts. In *European conference on computer vision*, pages 82–96. Springer, 2002.

[27] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 307–314. IEEE, 1999.

[28] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[29] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Shape reconstruction using volume sweeping and learned photoconsistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 781–796, 2018.

[30] Marc Levoy, Kari Pulli, Brian Curless, Szymon Rusinkiewicz, David Koller, Lucas Pereira, Matt Ginzton, Sean Anderson, James Davis, Jeremy Ginsberg, et al. The digital michelangelo project: 3d scanning of large statues. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 131–144, 2000.

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[32] Mendoza Guzmán Vıctor Manuel, Mejıa Munoz José Manuel, Moreno Márquez Nayeli Edith, Rodrıguez Azar Paula Ivone, and SRE Ramirez. Disparity map estimation with deep learning in stereo vision. *Proceedings of the Regional Consortium for Foundations, Research and Spread of Emerging Technologies in Computing Sciences (RCCS+ SPIDTEC2)*, pages 27–40, 2018.

[33] 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.

[34] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002.

[35] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002.

[36] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, 1:519–528, 2006. ISSN 0302-2345.

[37] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[38] Christoph Strecha, Wolfgang Von Hansen, Luc Van Gool, Pascal Fua, and Ulrich Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008.

[39] Kaixuan Wang and Shaojie Shen. Mvdepthnet: Real-time multiview depth estimation neural network. In *2018 International conference on 3d vision (3DV)*, pages 248–257. IEEE, 2018.

[40] Xiang Wang, Chen Wang, Bing Liu, Xiaoqing Zhou, Liang Zhang, Jin Zheng, and Xiao Bai. Multi-view stereo in the deep learning era: A comprehensive revfiew. *Displays*, 70:102102, 2021.

[41] Jianxiong Xiao and Long Quan. Multiple view semantic segmentation for street view images. In *2009 IEEE 12th international conference on computer vision*, pages 686–693. IEEE, 2009.

[42] Youze Xue, Jiansheng Chen, Weitao Wan, Yiqing Huang, Cheng Yu, Tianpeng Li, and Jiayu Bao. Mvscrf: Learning multi-view stereo with conditional random fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4312–4321, 2019.

[43] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[44] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[45] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5525–5534, 2019.

[46] Chaoqiang Zhao, Qiyu Sun, Chongzhen Zhang, Yang Tang, and Feng Qian. Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences*, 63(9):1612–1627, 2020.

[47] Qingtian Zhu, Chen Min, Zizhuang Wei, Yisong Chen, and Guoping Wang. Deep learning for multi-view stereo via plane sweep: A survey. *arXiv preprint arXiv:2106.15328*, 2021.