Towards Deep Multi-View Stereo for 3D Reconstruction from Satellite Imagery

Bradley Collicott
Stanford University, Dept. of Aeronautics & Astronautics
collicott@stanford.edu

Abstract

The problem of recovering 3D geometry from multiple viewpoints is not only fundamental to the field of computer vision, but has also seen recently seen great strides with the use of deep learning. However, the remote sensing community has fallen behind the state-of-the-art in leveraging these techniques. This project aims to implement a state-of-the-art multi-view stereo depth estimation network and investigate the feasibility of using an end-to-end deep learning pipeline for 3D reconstruction from satellite imagery. This report details the implementation of the MVS-Net framework from literature, qualitative and quantitative results from a lightly-trained network, and discussion on future considerations for extension to experiments on satellite imagery.[1]

1. Introduction

Recovering a functional model of the 3D geometry using only satellite or high-altitude imagery is of particular interest to the remote sensing community with regard to both scientific, emergency response, and national defense missions. Recently, there has even been commercial interest in imaging the planet using high-resolution synthetic aperture radar (SAR). Rather than launching a dedicated SAR satellite fleet in a strict formation to produce high-resolution models of the Earth with radar, it would be particularly desirable to leverage existing satellite imagery with advanced computer vision algorithms to reconstruct areas of interest using the large backlog of extremely high resolution Earth images.

With recent advances in the application of deep learning to depth map estimation from unstructured multi-view stereo (MVS) [16] overtaking traditional methods, this appears to be a logical time to close the gap between the computer vision and remote sensing fields. However, to the author's best knowledge, there is no publicly available information on a full end-to-end deep learning approach to 3D reconstruction from satellite imagery. This project, therefore, is the first step towards novel application of a promising technology to a significant remote sensing problem.

[1] https://github.com/bcollico/Deep_Multiview_Depth_Estimation.

2. Related Work

The related work on multi-view stereo reconstruction may be divided into the categories of model-based and learning-based methods. Learning-based methods leverage hand-crafted metrics and traditional feature extraction and matching methods to stitch multiple views together while minimizing some cost of reconstruction, whereas the learning-based methods rely on deep features and learned representations for estimating the depth of an image and therefore permitting a reconstruction.

In the realm of model-based methods, a comprehensive review by Schonberger et. al. of the structure-from-motion (SfM) problem [14] details how the SfM and subsequent reconstruction problem can be broken down into an incremental process consisting of feature extraction, image pairing, feature matching, triangulation, and bundle adjustment. The authors introduce COLMAP, an open-source SfM library that relies on an expensive nonlinear optimization process, extensive outlier rejection, and geometry consistency metrics. Unlike other methods, this SfM pipeline does not require calibrated cameras – i.e. any unstructured set of images may be used. Schonberger published another work [13] that details pixel-wise view selection for unstructured multi-view stereo. This view selection process enables the "incremental" technique utilized in their SfM method; however, it disregards the multi-view nature of the problem by only considering image pairs.

Similarly, work by Furukawa et. al. [9] use a sparse-to-dense feature extraction process and constraints of epipolar geometry to generate dense "patches" along the surface of the target scene. This reconstruction method enforces local photometric constistency as well as global visibility constraints to reject occluded or poor feature matches. In contrast to the SfM approach, this work requires calibrated cameras, such that the full camera matrix is known a priori.

Another quasi-dense approach to surface reconstruction was proposed by Lhuiliier et. al. [12] which leverages

course-to-fine reconstruction by sampling sub-pixel points from the initially generated disparity map. This method does not require calibrated cameras, but still leverages only pairwise image information, leading to an inefficient use of the avaiable information. This method, however, is limited to closed and smooth surfaces, making it unsuitable for outdoor scenes.

In regards to learning-based methods, there have been multiple attempts at extending neural network approaches to stereo and multi-view applications. The focus of this project is on the MVSNet proposed by Yao et. al. [16], which is largely characterized as the first learning-based end-to-end multi-view stereo pipeline. Enabling this advancement was the reduction of the large multi-view reconstruction problem into per-view depth map estimation. This in combination with contributions in other domains, such as successful 2D and 3D encoder-decoder structures and background segmentation approaches, allowed the MVSNet to outperform other methods which were restricted to low resolution, synthetic data and/or small-scale reconstructions.

Despite these advancements, the remote sensing and satellite imagery literature does not reflect the demonstrated success of learning-based multi-view stereo reconstruction. Even state-of-the-art reconstruction methods from satellite imagery, such as the winner of the Intelligence Advanced Research Projects Activity (IARPA) 2016 3D reconstruction challenge [8], rely on selecting suitable stereo pairs for epipolar rectification. Further, research has only begun moving in the direction of more sophisticated structure from motion (SFM) approaches, exemplified by the authors of [18], who investigate open-source SFM libraries (such as the previously discussed COLMAP) applied to satellite imagery, and the authors of [3], who take a novel approach to designing a satellite-imagery-specific SFM pipeline.

3. Technical Approach

The proposed approach is to re-produce the MVSNet developed in [16] and investigate it's applicability to the problem of 3D reconstruction from satellite imagery. The MVSNet is an end-to-end deep learning pipeline that accepts unstructured multi-view images as an input and outputs the estimated depth map in the target image. The general idea of the MVSNet pipeline is shown in Figure 1. The network was devised to replace traditional dense matching methods, including those that require pre-rectified image pairs, to produce more complete reconstructions while leveraging the the multi-view information present in the problem. The MVSNet introduced several novel contributions to the burgeoning field of deep-learning-based multi-view stereo: (1) A differentiable homography warping that enables the end-to-end training, (2) a cost volume built upon the camera frustrum rather than Euclidean space, and (3) a decoupling of the reconstruction process into the smaller problem of per-view depth map estimation.

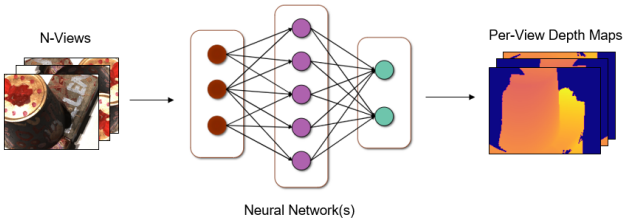


Figure 1. MVSNet General Framework

3.1. MVSNet Architecture

The MVSNet neural network architecture is detailed in the following sections.

3.1.1 Feature Extraction

The MVSNet is comprised of several sub-networks that enable the end-to-end learning process. The first of which is a feature encoder that downsizes the input image while maintaining global image information in each pixel descriptor. The feature extraction network is a 2D convolutional neural network (CNN) with eight layers – the network weights are shared for each view to expedite training. Specifically, the MVSNet uses a 2D U-Net to downsample the input images into 32-channel feature maps. The U-Net encoder is summarized in Fig. 2.

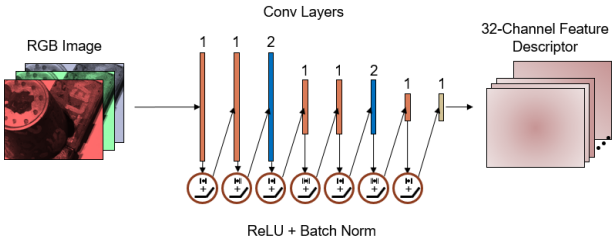


Figure 2. U-Net Feature Encoder

3.1.2 Cost Volume

A 3D cost volume is built from the extracted feature maps upon the camera frustum. The feature maps from each images are warped into planes parallel to the reference image using a differentiable homograpy constructed as

$$\boldsymbol{H}_i(d) = \boldsymbol{K}_i \boldsymbol{R}_i \left(\boldsymbol{I} - \frac{\boldsymbol{R}_i^T \boldsymbol{t}_i - \boldsymbol{R}_1^T \boldsymbol{t}_1}{d} \cdot \boldsymbol{n}_1^T \right) \boldsymbol{R}_1^T \boldsymbol{K}_1^{-1} \tag{1}$$

where the homography between the i^{th} feature map and the reference feature map at depth d is $\boldsymbol{H}_i(d)$, and the i^{th} camera extrinsics and intrinsics are $\{\boldsymbol{K}_i,\ \boldsymbol{R}_i,\ \boldsymbol{t}_i\}$. Care must

be taken to rotate the camera translations to the world reference frame as was shown in Eq. (1). Points of this warped feature map are denoted by the set $\boldsymbol{V}_i(d) = \{x' \mid x \sim \boldsymbol{H}_i(d)x' \ \forall \ x \in \boldsymbol{F}_i\}$ where $\boldsymbol{F}_i$ is the set of encoded features for the i^{th} image and $\sim$ denotes the projective equality. This process is what enables depth estimation downstream in the network - by warping each view to the frustum of the reference camera at varying depths, the network is essentially sampling a pre-defined depth range over which to estimate the depth of individual pixels.

The constructed feature volumes are then aggregated into a global cost volume $\boldsymbol{C}$ using a variance-based metric $\mathcal{M}$.

$$\boldsymbol{C} = \mathcal{M}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_N) = \frac{\sum_{i=1}^{N}(\boldsymbol{V}_i - \bar{\boldsymbol{V}}_i)^2}{N} \tag{2}$$

where $\bar{\boldsymbol{V}}_i$ is the average volume among all feature volumes. All operations are element-wise. The authors note that no preference is given to the reference image in this operation. It is also worth noting that this cost volume is based on the variance of the feature volumes (i.e. the second statistical moment of the Gaussian distribution). Precaution is taken, however, to regularize the cost volume to alleviate noise from occlusions, non-lambertian features, and camera errors. The cost volume is used to generate a probability volume $\boldsymbol{P}$ for depth inference. A four-scale 3D CNN, similar to the 3D U-Net [5] is applied for this step, upsampling and downsampling the 32-channel cost volume before re-combining and decoding it to the corresponding probability volume, applying the softmax operation to normalize the resulting probabilities. The cost regularization network is visualized in Fig. 3, where each box represents a 3D CNN followed by batch normalization and an activation layer.

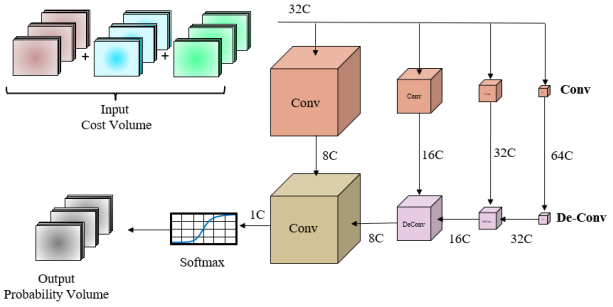


Figure 3. 3D U-Net-like Cost Volume Regularization CNN

3.1.3 Depth Map Inference

To approximate the argmax operation in a differentiable manner, the "soft argmin" operation is applied to retrieve the depth map from the probability volume.

$$\boldsymbol{D} = \sum_{d=d_{min}}^{d_{max}} d \times \boldsymbol{P}(d) \tag{3}$$

This operation amounts to computing the expectation along the depth direction. The resulting depth map is the same size as the 2D image feature maps. To prevent against the influence of small probabilities on the computed depth, only a subset of the depth hypotheses are used to compute the initial depth map. The subset of selected depth probabilities must be re-scaled using the softmax operation to ensure that the probabilities still sum to 1 and the expected depth range is not affected by this operation.

Finally, the reference image is used as a guide for refining the depth map, and a residual learning network is applied. To prevent biasing at a certain depth scale, the initial depth map is normalized to the range $[0, 1]$ using the range of depth hypotheses. The normalized initial depth map and reference image are input to this network, which is a 2D CNN, shown in Fig. 4 that outputs the normalized residual depth of the input image. This residual is added to the normalized initial depth map and re-scaled to produce the refined depth estimate.

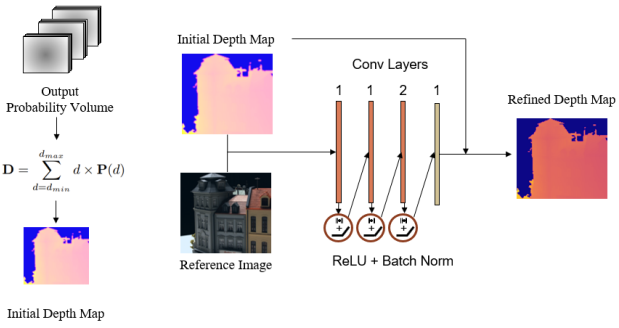


Figure 4. 2D CNN for Depth Map Refinement

3.1.4 Loss Function

Losses are considered for both the initial and refined depth maps. A ground truth depth map is used to compute the mean absolute errors of all pixels with ground truth labels. The loss is

$$\mathcal{L} = \sum_{p \in \boldsymbol{p}_{valid}} \left\| d(p) - \hat{d}_i(p) \right\|_1 + \lambda \left\| d(p) - \hat{d}_r(p) \right\|_1 \tag{4}$$

where $d(p)$, $\hat{d}_i(p)$, and $\hat{d}_i(p)$ are the ground truth depth, initial depth estimate, and refined depth estimate respectively for each image in the batch. The weight parameters, λ, is set to 1.0.

3.1.5 3D Reconstruction via Depth Map Fusion

The output depth maps could then be fused using simple point cloud merging techniques such as a iterative-closest-point (ICP) search. The technique implemented by the original MVSNet authors is to first filter the depth maps and then implement two consistency metrics for fusing the depth maps. (1) pixels with a depth probability less than 0.8 will be excluded, and (2) multi-view geometric consistency is enforced by projecting a pixel through it's depth to another view, and then reprojecting the new pixel back to the reference image using it's depth estimate in the new view. If the pixel reprojected to the reference view is "close" enough to the original, it is included in the depth map. Depth map fusion was not considered in this project.

3.2. Datasets and Evaluation

The neural network was originally planned to be pre-trained on the pre-prepared DTU dataset, provided by [1] and prepared for training the MVSNet by [16], and fine-tuned using portions of the IARPA MVS3DM dataset [2]. Particularly, the modified and reduced dataset introduced in [18] would have been used due to the provided camera intrinsic and extrinsic parameter estimates. Satellite images are generally accompanied by a rational polynomial coefficient (RPC) camera model which uses a ratio of cubic polynomials, defined by 78 coefficients and 10 normalization constants. This camera representation was shown, in [18], to be locally approximatable as a weak perspective pinhole camera with intrinsics familiar to traditional computer vision techniques.

This second phase of this approach, however, did not come to fruition. Due to limitations in the model for learning large-scale scenes, it is not feasible to train or evaluate the MVSNet architecture on satellite imagery. This limitation will be discussed further in the discussion and conclusions.

4. Implementation

The original MVSNet implementation was closely followed for this project with some exceptions. Notable exceptions and implementation details will be explored in this section.

4.1. Neural Network

Some changes were made to the neural network to improve memory efficiency, correct upon inaccuracies, and tune performance to a resource-limited machine. These changes include:

1. Bias parameters were excluded from all convolutional layers. The batch normalization after each layer nullifies any bias effect, and excluding this parameter provides GPU memory savings.

2. The original homography warping included the term $\boldsymbol{K}_1^T$ rather than $\boldsymbol{K}_1^{-1}$ as was shown above. The inverse must be used to preserve the generality of the transformation, i.e. the warping of an image to itself is the identity matrix. Additionally, the original equation did not make mention of rotating the translation vectors to a common reference frame, which is also necessary.

3. It should be noted that the original implementation used the Tensorflow machine learning framework, whereas this work makes use of PyTorch.

A summary of the trainable parameters in the neural networks is shown in the following table. Notice that the number of parameters in the 3D Cost Volume Regularization network is approximately an order of magnitude greater than the 2D CNNs.

Table 1. Trainable Model Parameters by component

Feature Encoder	40088
Cost Volume Reg	321864
Depth Refine	20064
Total	382016

The Adam optimizer with a variable learning rate was used in training the model. The learning rate scheduler was set to decrease the learning rate by 20% if the validation loss did not decrease after 3 epochs, waiting 5 epochs between successive changes in the learning rate. The learning rate was assigned an initial value of 0.005 and a minimum value of 0.0001.

4.2. Training/Validation/Testing

The data splits were selected to be similar to the original MVSNet convention. The dataset, however, was expansive, so only one lighting condition was selected for each camera view for training. A subset of the original validation and evaluation sets were also selected and used to compute the accuracy metrics. A summary of the data splits is shown in the table below. The neural network was trained for 14 epochs on the training split with a batch size of 5 samples (each containing 3 images), with the learning rate scheduled according to performance on the validation split. The loss, initial depth map accuracy, and refined depth map accuracy across all training samples is plotted in Fig. 5.

Table 2. DTU Dataset Split

Training Samples	3871
Validation Samples	196
Evaluation Samples	1078
Total Samples	5145

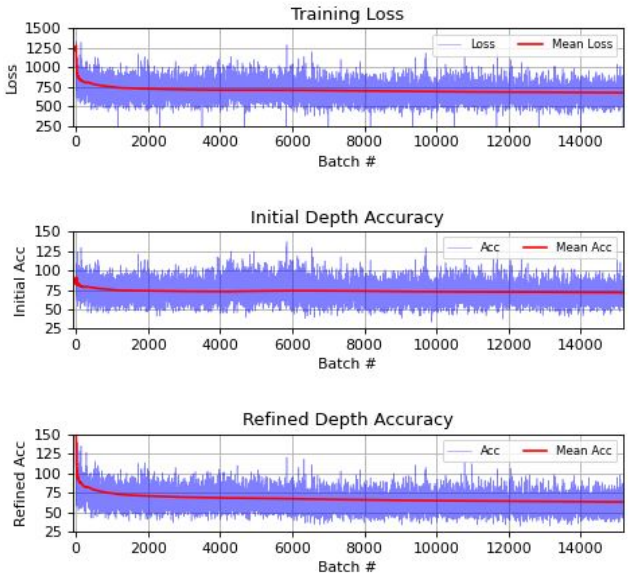


Figure 5. Training Statistics

5. Results

The network was trained for 14 epochs with a batch size of 5 samples, resulting in a total training time of approx. 40 hours on an Nvidia K80 GPU with 12GB vRAM hosted on a Google Cloud virtual machine. The testing results for mean pixel-wise accuracy compared to the ground truth depth maps is shown in Tab. 3. The depth accuracy values here represent the mean difference in the depth estimation of valid pixels, where the depth of the scene is on the order of 400-900 units. The table, therefore, shows that the average pixel accuracy of the depth map is approximately 12% of the depth range of the image. This value may have a skew induced by large outliers in the scene.

Table 3. MVSNet Depth Map Accuracy Results

Initial Depth Map	Refined Depth Map
65.17	59.34

Estimated depth maps are shown for several views. The original image and ground truth depth maps are shown for qualitative comparison – note that the depth maps were rescaled to the range [0, 255] for visualization.

6. Discussion

This section will discuss the presented results as well as limitations and extensions of the MVSNet and future directions for multi-view stereo in satellite imagery.

6.1. Estimated Depth Maps

Despite only being trained for 14 epochs, the estimated depth maps are produced as expected. The initial depth map is a rough approximation of the depth in the scene, whereas the refined depth map uses the original image to capture edges and features that were not observed in the initial approximation. Across all depth maps, we see that the network has difficulty in assigning values to the far-field depth points, i.e. those beyond the scene of interest. This is due to the training process of only comparing "valid" depth points in the ground truth. This leads one to believe that some segmentation or other attempt at background removal should be implemented to alleviate the network's burden of estimating the depth of background points.

In reference to specific depth images, Fig. 6a showcases the ability of the depth refinement network to capture edges and visual features from the image starting with a blurry and inaccurate initial depth map. In Fig. 6b, we show an example where the network successfully estimated the depth of points that were not fully incorporated in the ground truth depth map. This generalization capability is a strong indication that the network has learned the intended depth estimation behavior. Finally in Fig. 6c, we see an example where the network only partially captures the details of the image, struggling with resolving subtle differences in depth, such as the difference between the figurine's feet and the blocks upon which it stands.

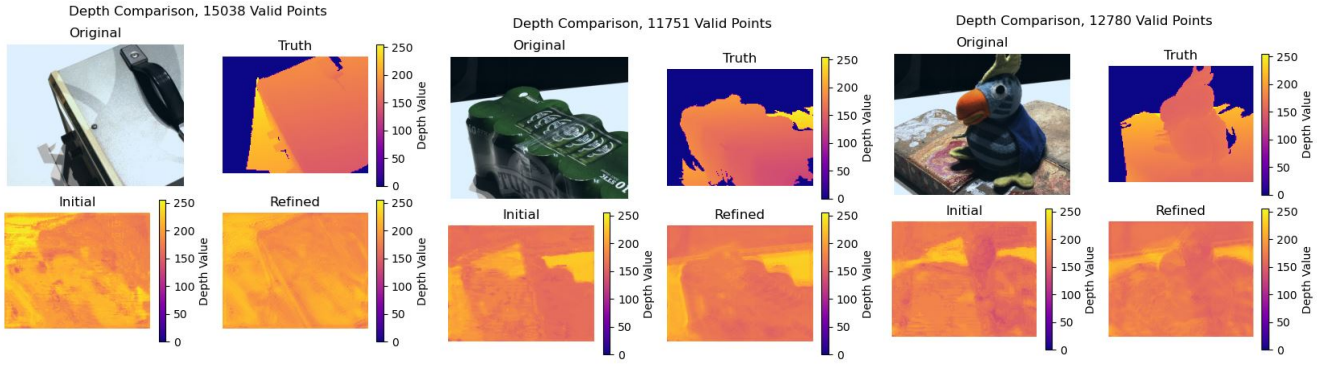
6.2. Limitations

The methods implemented here have several severe limitations that make it difficult to use for general purpose depth estimation and high resolution or large scenes.

1. The algorithm requires known camera intrinsic and extrinsic parameters. If the intended training dataset does not contain these truth labels, they will have to be estimated a la structure from motion. This estimation process is prone to error and may introduce untenable noise into the training process.

2. The algorithm requires an initial estimate of the range of depths in the scene. For completely new scenes, a poor estimate of the scene depth will result in nonsensical results from the network. In the training process, the minimum scene depth was provided for determining this range.

3. Another imposed constraint on the training dataset is the existence of reliable, dense ground-truth depth maps for all scenes. In the context of satellite imagery, this is particularly difficult to obtain and process considering the large scale of environment. Several works in literature have attempted to remedy this issue with self-supervised or unsupervised training processes that



(a) Scan 9 (b) Scan 10 (c) Scan 12

Figure 6. Qualitative Comparison of Estimated Depth Maps

learn depth while enforcing consistency between the views. In [6], the unsupervised process is implemented with cross-view consistency metrics to produce depths maps that comply with the underlying geometry. Other methods such as [4] rebuke the notion of enforcing photometric consistency, and propose a neural rendering technique to handle non-Lambertian surfaces and object occlusions.

4. Another limiting factor for extending the traditional MVSNet to large scenes is the high memory consumption of the cost volume regularization network. Using the MVSNet for large scenes would require prohibitively large amounts of dedicated GPU memory. To this end, several papers have addressed the issue of high-resolution and large-scale environments: Yao et. al. proposed the Recurrent MVSNet (R-MVSNet) [17] which makes use of a Gated Recurrent Unit (GRU) to sequentially regularize the 2D depth maps rather than the entire 3D cost volume, drastically reducing memory consumption; Weilharter et. al. propose the High-Res MVSNet [15] and employ a course-to-fine depth search across images to reduce memory consumption while still achieving depth quality comparable to other end-to-end deep MVS methods.

6.3. Future Work

This project serves as a starting point for investigating learning-based methods for 3D reconstruction from satellite imagery. There is much left too be discovered in this field due to the lack of available literature of neural-network-based approaches in this domain. To this end, the author has recommendations for future research endeavors on this topic.

1. Investigation into methods for high-resolution multi-view reconstruction using self-supervised or unsupervised learning. There is no definitive solution for removing the ground truth depth labels from the MVS learning process, and much less a solution for doing so with large-scale scenes. One may draw inspiration from the methods mentioned above as well as monocular depth estimation techniques such as [10, 11].

2. Methods for directly integrating the RPC camera model and into an end-to-end deep learning pipeline for MVS reconstruction. As mention, the camera information directly available from satellite imagery currently must be converted to the more familiar pinhole model for processing. Removing this process would reduce approximation errors and lessen the computational and logistical burden associated with undertaking this problem.

3. Leveraging multi-spectral information in the reconstruction process. As discussed before, interest is growing in the collection of data using satellite as synthetic apertures, e.g. SAR Radar image collection. It was demonstrated in [7] that pan-chromatic images could be enhanced by incorporating multi-spectral information, so perhaps further gains could be made in the domain of multi-spectral fusion for 3D reconstruction.

7. Conclusions

This project implemented the MVSNet multi-view depth inference network in the PyTorch machine learning framework. The network was shown to qualitatively produce depth maps that capture the overall shape and depth of the target image, and to quantitatively differ from the ground truth on the order of 10-15% of the depth range of the scene. Despite the large magnitude of this mean error, the results are promising considering that the network was training for only 14 epochs and did not include any fine-tuning. The network was shown to generalize to unseen evaluation data and perform at the same level as was demonstrated during

training. Future work on this model include further training on the existing data and extension of the training data to include the full DTU dataset.

In the context of extending this MVS solution to satellite imagery, discussion was provided on why this MVS-Net would not be suitable for large-scale or high-resolution depth estimation. Discussion was presented on possible extensions to unsupervised training to alleviate the difficulty of collecting ground truth depth information for satellite images. Future directions were proposed to continue investigating this problem, including unsupervised high-resolution reconstruction, integration of the complex RPC camera model into the neural network framework, and usage of multi-spectral information to improve depth estimation accuracy and robustness.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-Scale Data for Multiple-View Stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. 4

[2] Marc Bosch, Zachary Kurtz, Shea Hagstrom, and Myron Brown. A multiple view stereo benchmark for satellite imagery. *Proceedings - Applied Imagery Pattern Recognition Workshop*, 2017. 4

[3] Sebastian Bullinger, Christoph Bodensteiner, and Michael Arens. 3D surface reconstruction from multi-date satellite images. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 43(B2-2021):313–320, 2021. 2

[4] Di Chang, Aljaz Bozic, Tong Zhang, Qingsong Yan, Yingcong Chen, Sabine Susstrunk, and Matthias Niessner. Rc-mvsnet: Unsupervised multi-view stereo with neural rendering. 2022. 6

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-net: Learning dense volumetric segmentation from sparse annotation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9901 LNCS:424–432, 2016. 3

[6] Yuchao Dai, Zhidong Zhu, Zhibo Rao, and Bo Li. Mvs2: Deep unsupervised multi-view stereo with multi-view symmetry. In *2019 International Conference on 3D Vision (3DV)*, pages 1–8, 2019. 6

[7] Sajjad Eghbalian and Hassan Ghassemian. Multi spectral image fusion by deep convolutional neural network and new spectral loss function. *International Journal of Remote Sensing*, 39:3983–4002, 06 2018. 6

[8] Gabriele Facciolo, Carlo De Franchis, and Enric Meinhardt-Llopis. Automatic 3D Reconstruction from Multi-date Satellite Images. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017-July:1542–1551, 2017. 2

[9] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010. 1

[10] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October(1):3827–3837, 2019. 6

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:6602–6611, 2017. 6

[12] Maxime Lhuillier, Long Quan, Maxime Lhuillier, Long Quan, A Quasi-dense Approach, and Surface Reconstruction. A Quasi-Dense Approach to Surface Reconstruction from Uncalibrated Iages To cite this version : HAL Id : hal-00091032 A Quasi-Dense Approach to Surface Reconstruction from Uncalibrated Images. 2006. 1

[13] Johannes L. Schonberger and Jan Michael Frahm. Structure-from-Motion Revisited. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:4104–4113, 2016. 1

[14] Johannes L. Schönberger, Enliang Zheng, Jan Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9907 LNCS:501–518, 2016. 1

[15] Rafael Weilharter and Friedrich Fraundorfer. Highres-mvsnet: A fast multi-view stereo network for dense 3d reconstruction from high-resolution images. *IEEE Access*, PP:1–1, 01 2021. 6

[16] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11212 LNCS:785–801, 2018. 1, 2, 4

[17] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent MVSnet for high-resolution multi-view stereo depth inference. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:5520–5529, 2019. 6

[18] Kai Zhang, Noah Snavely, and Jin Sun. Leveraging vision reconstruction pipelines for satellite imagery. *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*, pages 2139–2148, 2019. 2, 4