This WACV 2022 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

An experimental comparison of multi-view stereo approaches on satellite images

Alvaro Gómez, Gregory Randall
Facultad de Ingeniería
Universidad de la República, Uruguay
agomez@fing.edu.uy

Gabriele Facciolo, Rafael Grompone von Gioi
Centre Borelli
ENS Paris-Saclay, France

Abstract

Different methods can be applied to satellite images to derive an altitude map from a set of images. In this article we evaluate a set of representative methods from different approaches. We consider true multi-view stereo methods as well as pair-wise ones, classic methods and deep learning based ones, methods already in use on satellite images and others that were originally devised for close range imaging and are adapted to satellite imagery. While deep learning (DL) methods have taken over multi-view stereo reconstruction in the last years, this tendency has not fully reached satellite stereo pipelines that still largely rely on pair-wise classic algorithms. For the comparison, we set-up a framework that allows to interface a DL-based stereo method taken from the computer vision literature with a satellite stereo pipeline. For multi-view stereo algorithms we build on a recently proposed framework originally devised to apply Colmap method to satellite images. Methods are compared on several datasets that include sets of images taken within a few days and sets of images taken months apart. Results show that DL methods have, in general, a good generalization power. In particular, the use of the GANet DL method as the matching step in a pair-wise stereo pipeline is promising as it already performs better than the classic counterpart, even without a specific training.

1. Introduction

Stereo vision is an area that has been extensively researched and multiple algorithms have been proposed along the last decades [30, 15, 20]. Initial methods worked on one stereo pair. Then, Multi-View Stereo (MVS) was first approached as an extension of the stereo algorithms by aggregating the information of multiple stereo pairs. True MVS algorithms considering directly all the images of the scene arrived some time afterwards [5]. True MVS algorithms were mostly devised, by the computer vision community, for the reconstruction of objects, buildings and interiors with images taken with standard pinhole-like cameras at close distance [10, 32]. Deep Learning (DL) MVS methods [20] flourished in the last years and have taken over the top rankings of the main benchmarks of the area [33, 18] but classic methods are still a valid option when dealing with many and/or large size images since DL methods struggle to accommodate large 3D structures in GPU memories.

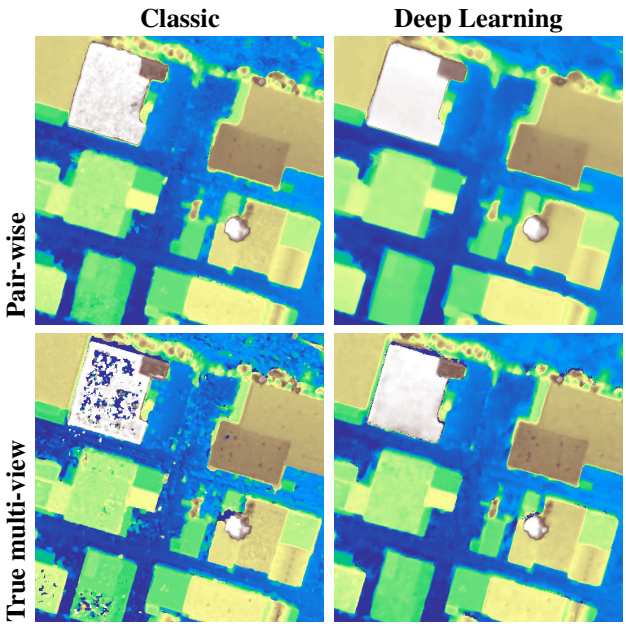

Figure 1. Digital Surface Models (DSM) computed by the methods analyzed in this work on the subregion 156 of JAX_NIT dataset. Methods from top to bottom and left to right: S2P, S2P-GANet, COLMAP, and CasMVSNet. The ground truth altitude for this example can be seen in Figure 5.

In the case of satellite images, MVS has traditionally been performed with pair-wise approaches where the multiple views are treated by pairs doing traditional two-view stereo and then aggregating the pair-wise reconstructions (elevation models or point clouds, for example) to get the final result [6, 26, 19]. Satellite images have specific characteristics that have historically discouraged the use of true MVS methods, for example: (a) the extremely small ratio between the depth range and the distance from the camera

to the scene implies working with a camera model that deviates from the standard pinhole and deals with structures that occupy few pixels in the images; (b) the images for a certain location can only be acquired through several sweeps which may be days, months or, even years apart, introducing variability in illumination, seasonal changes and man-made changes, among others. The variability poses important challenges for the matching of correspondent regions across the images. This variability problem has usually been tackled with a heuristic selection of best pairs that tend to minimize separation in time of the images and prefer the view angles that ensure less error in triangulation [8].

A recent work [37] showed that classic true MVS algorithms used in computer vision could be adapted to satellite images for the benefit of the remote sensing field. We build upon that work and extend the concept to include stereo and MVS methods based on DL.

This work is a concise evaluation of a set of methods which are representative of different approaches that can be applied to satellite images in order to derive an altitude map from a set of images. It is not an extensive benchmark attempt. The selected methods span different interesting aspects: methods already in use on satellite images and others originally devised for close range imaging and adapted to satellite imagery, classic and DL approaches, pair-wise and multi-view reconstruction methods.

A simple and modular satellite image processing pipeline (S2P) [6] is considered as a baseline. Experiments in this article explore if modifications in the pipeline or the use of other methods adapted to satellite imagery can give promising or better results in comparison to the already established pipeline. A framework is built in order to compare different methods on several datasets that include sets of images from different sites and that consider acquisitions over short and long periods of time. The comparison shows that the analyzed methods attain comparable and sometimes better performance than the classical ones. Figure 1 shows typical results, where rows compare pair-wise vs. multi-view approaches and columns compare classic vs. DL methods.

The rest of the article is organized as follows: Section 2 explains the methodology used to compare the methods introduced in Section 3. Section 4 details the datasets on which evaluation was performed. Section 5 presents the experiments. The results are finally analyzed in Section 6.

2. Framework for the comparison

The different methods are applied to the datasets following the scheme depicted in Figure 2. For pair-wise methods, a Digital Surface Model (DSM) is computed for every possible pair of images of a subregion; these DSMs are then aggregated to get an enhanced multi-pair DSM. On the other hand, true multi-view methods are fed with all the images of a subregion.

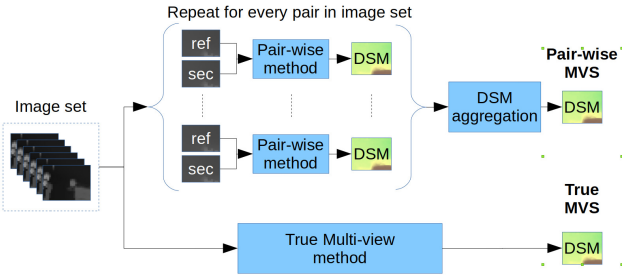

Figure 2. Scheme of multi-pair and true multi-view methods.

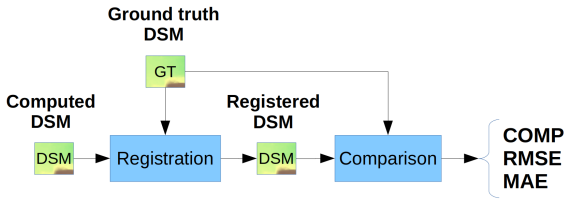

Figure 3. Flow diagram for the comparison of computed DSM with respect to the ground truth altitude.

In order to assess the performance of the different approaches, the computed DSMs are compared against the ground-truth DSM on all the datasets. The comparison is done as shown in Figure 3. First each altitude map is registered to the ground-truth map. A normalized cross correlation approach is used in order to obtain integer X,Y translations that register the DSMs and the altitude is adjusted by the median of the difference to the ground-truth. Once registered, the following metrics [3] are computed:

Completeness (COMP): Proportion of evaluated pixels where the altitude of the computed map differs from the ground-truth less or equal than $z\_tol = 1m$. We consider only pixels with ground-truth information.

RMSE Accuracy: Root Mean Square Error (RMSE) between computed and ground truth maps considering only the pixels with valid information in both maps.

MAE Accuracy: Median Absolute Error (MAE) between computed and ground truth maps considering only the pixels that have valid information in both maps.

3. Methods

Table 1 summarizes the evaluated methods. A brief description of each one and the necessary adaptations to satellite imagery are presented hereafter.

3.1. S2P

Figure 4 shows an overview of the S2P [6] pipeline[1]. The input is a stereo pair of images with their respective camera models expressed by rational polynomial coefficients

[1] https://github.com/centreborelli/s2p

Table 1. Tested methods

Method	Type	DL	Notes
S2P [6]	Pair-wise	No	MGM [7] in disparity computation
S2P-GANet	Pair-wise	Partially	GANet [36] in disparity computation
COLMAP [32, 31]	Multi-view	No	Adapted for satellite images by [37]
CasMVSNet [14]	Multi-view	Yes	Adapted for satellite images in this work

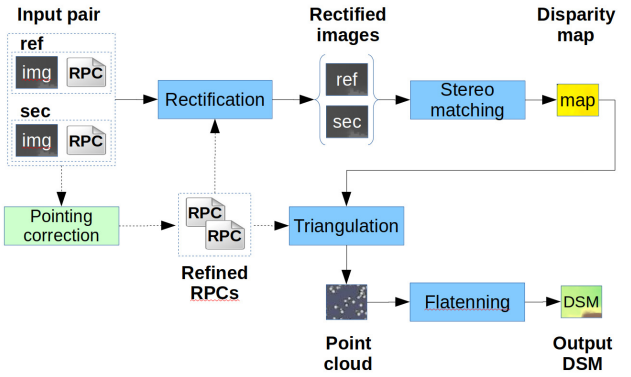

Figure 4. S2P overview. The input is a pair of images with their respective RPC camera models, the output is a DSM given as a georeferenced 3D point cloud and as an altitude image.

(RPC). Each image pair undergoes a pointing correction and is rectified. Disparity is computed on rectified images using MGM stereo matching algorithm [7]. Computed correspondences are then triangulated to produce a geo-referenced 3D point cloud and an altitude map.

3.2. GANet

GANet [36] uses Deep Neural Networks (DNN) to compute a disparity map. As other DNN methods [20], it follows the traditional stereo steps: dense features are extracted for both images, the cost of matching the features at different disparities is organized in a Cost Volume (CV), which is regularized by aggregation and/or filtering and finally a map with minimal cost is derived from the CV.

In most DNN based stereo methods, cost aggregation is done by 3D convolutions, usually in an hourglass configuration [20]. 3D convolutions imply large memory requirements; the computational burden restricts the size of images that can be processed. GANet takes a different approach by introducing a Semi-global Guided Aggregation layer (SGA) which implements a differentiable approximation of Semi-Global Matching (SGM) [16]. SGA is followed by a Local Guided Aggregation layer (LGA) that performs a local filtering. SGA and LGA weights are generated by an auxiliary "guidance subnet" fed with original images and the extracted features.

S2P-GANet: Adaptation to satellite images In this work, GANet is used as an alternative "stereo matching" step in the S2P pipeline, see Figure 4. The stereo matching step receives a rectified stereo pair of images and computes disparity maps in both directions: left-to-right and right-to-left. A consistency check is performed to filter out pixels with non congruent disparities [9, 4]. In order to use GANet in the S2P pipeline, some adaptations have to be considered:

a) Negative disparities: In most stereo algorithms, a CV is computed and regularized, and a disparity map is derived from it. The CV is computed for a certain range of possible disparity values, which must be known *a priori* or estimated. In the S2P pipeline, the disparity range is traditionally estimated by the sparse matching of interest points (e.g. SIFT keypoints [23]), but other strategies are allowed such as specifying a fixed known disparity range or estimating the disparity range from a known altitude range. In several stereo matching algorithms, including [7] used by default in S2P, disparity admits positive and negative values. GANet, however, accept only negative disparities. That is, all pixels in the secondary rectified image must "move" to the left relative to the rectified reference image. In this work, a fixed known altitude range is used, based on the ground truth plus an additional safety guard. The S2P pipeline was adapted to get a rectification compatible with negative disparities.

b) GPU memory restrictions: The size of the rectified stereo images that can be handled by GANet is bounded by the available memory in the GPU. Also, images' width and height must be multiple of 48. A tiling strategy is thus implemented to process large images. The disparity estimation is more error prone at tile borders. So tiles are chosen as large as possible and the overlaps are merged considering the distance to the border as a weight. The experiments reported in this work use tiles of 1872×480 pixels and were run on a Nvidia Tesla P100 GPU with 12Gb of RAM.

3.3. COLMAP

COLMAP [32] is part of a family of methods [13, 10, 11] that focus on large-scale dense reconstruction and fusion. These methods aim to integrate the information of diverse multiple images of a scene such as crowd-sourced image datasets. COLMAP is closely related to [38] and considers, as other MVS methods, a variety of photometric and geometric priors ensuring consistency among different views. The method follows a generalized Expectation-Maximization (EM) scheme with alternating and interleaved estimation of occlusions in the E-step and depths in the M-step. The depth estimation M-step is based on PatchMatch [1] where the tested hypothesis are based on the depths, the normals and their perturbations.

Zhang et al. [37] adapted COLMAP to the peculiarities of satellite images. For that, they work on local scene co-

ordinates, approximate the RPC camera model with a perspective camera and reparameterize the depths as heights over a horizontal reference plane that lies below the scene. In the adaptation, the hypothesis depth planes become the horizontal planes in the scene.

To approximate the camera, the scene volume is sampled in a regular grid and projected by the RPC model. This yields a set of 3D-2D correspondences that are used to fit a 3×4 projection matrix.

Being $(0, 0, 1, -d)$ the coefficients of the horizontal reference plane (plane equation $n^t x - d = 0$ with $n = (0, 0, 1)^t$), the reparameterization of the depths is handled extending the 3×4 projection matrix with a fourth row

$$\begin{pmatrix} u \\ v \\ 1 \\ m \end{pmatrix} = \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{P}_{13} & \mathbf{P}_{14} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \mathbf{P}_{23} & \mathbf{P}_{24} \\ \mathbf{P}_{31} & \mathbf{P}_{32} & \mathbf{P}_{33} & \mathbf{P}_{34} \\ 0 & 0 & \overline{Z} & -\overline{Z}d \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}, \quad (1)$$

where Z is the conventional depth (distance from the camera center) and m the reparameterized depth (the height over the reference plane). $\overline{Z}$ is computed as the average conventional depth of all the sparse scene points in the Structure From Motion (SFM) step. In order to check the photometric consistency between two views at a certain depth, COLMAP, as other MVS methods, computes an homography between the first and second view as:

$$\mathbf{H} = \mathbf{K}_2 \left(\mathbf{R}_{12} - \frac{n^T t_{12}}{f} \right) \mathbf{K}_1^{-1}, \quad (2)$$

where $(n, -f)$ are the coefficients of the hypothesis plane that induces the homography, $\mathbf{R}_{12}$ and t_{12} indicate the pose of camera 2 w.r.t camera 1; and $\mathbf{K}_1$ and $\mathbf{K}_2$ are the intrinsic parameters of the cameras. In the adaptation, Zhang et al. [37] propose a more stable homography computation based on the 4×4 projection matrices instead of operating with the intrinsic and extrinsic parameters.

3.4. CasMVSNet

True multi-view stereo methods based on DNN construct a CV in a similar manner as DNN stereo methods [20]. MVSNet [35] is a well known representative of this category. Given a reference image, fronto-parallel hypothesis planes at different depths are considered. Features are then extracted for each image and differentiable homographies are used to warp the 2D feature maps into the hypothesis planes of the reference camera to form feature volumes (one feature volume per image). The CV is computed as the variance of all the feature volumes. The cost regularization is done by 3D convolutions. A raw depth map is regressed from the regularized CV and finally refined taking into account the information of the reference image.

As mentioned above, aggregation with 3D convolutions struggles with memory and computational costs growing cubically. The Cascade Cost Volume for High-Resolution Multi-View Stereo (CasMVSNet) [14] is a true multi-view stereo method that tries to overcome this issue with a coarse-to-fine approach. Features are extracted at multiple scales. At the early stages of the cascade the CV is built taking into account large scale features and sparse plane hypothesis. This low resolution CV leads to a raw depth that is subsequently used to adjust the depth sampling. Progressing through the stages, features at finer scales are considered but on the other hand the considered depths hypothesis are narrower bands surrounding the previous estimated depths. This keeps bounded the size of the CV along the stages.

Adaptation to satellite images The adaptation of the algorithm in order to use it with satellite images is based on the COLMAP adaptation [37] summarized in Section 3.3. The perspective approximation of the RPC cameras into 4×4 extended projection matrices is borrowed from the output of the SFM step of a run of the adapted COLMAP.

In the code of the adapted CasMVSNet, projections are handled with Equation 1 and homographies are computed as proposed by [37]. In order to preserve the ordering of the planes as trained in CasMVSNet (from near to far relative to the camera), the horizontal planes are traversed in decreasing height (from near to far).

3.5. DSM aggregation criteria

In the case of pair-wise MVS, it is well known that DSM aggregation improves in general the completeness [27, 8]. However, if the DSM computed from a bad pair is included, the result degrades. For this reason it is essential to preselect the pairs to be aggregated. In [8] a very simple heuristic based on two conditions on the metadata of the images was proposed: **a) filtering:** both images in the pair must have an incidence angle smaller than 40º and the angle between views should be in the [5º, 45º] range, preferably around 20º; **b) ordering:** pairs are ordered by the increasing absolute difference between the dates of the images. This is the *a priori* selection criterion used in our experiments. But, this is not enough to filter out all bad matches. Indeed, we observed that strong seasonal changes can lead to very bad results independently of the metadata.

For this reason we apply an additional selection criterion on the pairs, which determines *a posteriori*, after computing the DSM, if it should be aggregated. A DSM is aggregated if it has more that a certain number of valid (not undefined) pixels. In our experiments a minimum of 70% of valid pixels was considered.

4. Datasets

The methods described above were tested on three datasets, consisting on stereo satellite images from the

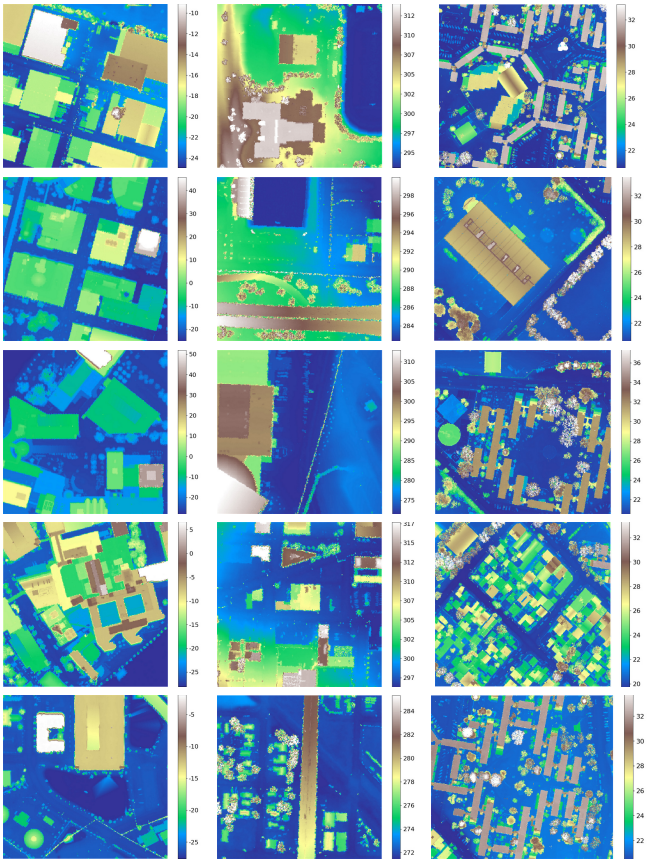

Figure 5. Ground-truth altitude maps from the datasets used in this study. Left: JAX dataset (subregions 156, 165, 214, 251, 264). Center: OMA dataset (subregions 203, 247, 251, 287, 353). Right: Site 1 of MVS3D dataset (subregions 001, 002, 003, 004, 005). Altitudes are in meters.

Multiple View Stereo Benchmark for Satellite Imagery (MVS3D) [3] and the US3D dataset [2].

The MVS3D is a set of 47 satellite images of a neighborhood of the city of Buenos Aires, Argentina. Besides, an airborne Lidar of the same region acquired for the MVS3D challenge [3] (in a different date than the satellite images), is considered the ground truth for the altitudes.

The US3D dataset consists of 26 WorldView-3 target-mode panchromatic, visible, and near infrared (VNIR) images collected between 2014 and 2016 over Jacksonville (JAX), Florida and 43 WorldView-3 target-mode panchromatic and VNIR images collected between 2014 and 2015 over Omaha (OMA), Nebraska. Semantic labels and an airborne Lidar are also available. The Lidar, acquired at a different date than the satellite images by the USGS, is considered the ground-truth for the altitudes.

For our evaluation, 5 subregions from each of the datasets are considered. The selection is representative of the datasets but arbitrary in any other aspect, see Figure 5. In each subregion, a bounded set of images is considered in order to allow a tractable pairwise analysis: 6 images acquired in a small time interval (same day or some days apart) and 6 images acquired in a longer time interval (spanning months) are considered. These sets of images are given the suffixes NIT and FIT which stand for near-in-time and far-in-time, respectively. Table 2 summarizes the set of images used for the experiments.

Table 2. Datasets used. For each location, five subregions are considered. For each subregion, six images acquired in a small time interval (same day or some days apart) and six acquired in a longer time interval (months apart) are considered.

Dataset	Subregions	Time span	Alias
MVS3D	MVS_ {001, 002, 003, 004, 005}	1 d	MVS_NIT
US3D_JAX	JAX_ {151, 165, 214, 251, 264}	44 d	JAX_NIT
US3D_OMA	OMA_ {203, 247, 251, 287, 353}	24 d	OMA_NIT
US3D_JAX	JAX_ {151, 165, 214, 251, 264}	386 d	JAX_FIT
US3D_OMA	OMA_ {203, 247, 251, 287, 353}	405 d	OMA_FIT

5. Experiments

5.1. Methods on stereo pairs

The S2P, S2P-GANet and COLMAP methods were tested on all the possible pairs for each subregion. With 6 images per subregion, there are 30 possible pairs considering the order of the images. Methods based on the S2P pipeline already work on stereo pairs, while COLMAP is set to work on the minimal set of two images. Table 3 presents the results, averaging only the DSMs that have at least 70% of valid pixels.

As expected and consistently with what was reported in [37], S2P based methods outperform the adapted COLMAP which is not intended to work just on image pairs. Among the two variants of S2P, results are quite similar in completeness with S2P-GANet achieving better values of RMSE accuracy, which denotes a lower dispersion in the altitude values of the DSMs. Figure 6 compares the two variants of S2P. S2P-GANet variant achieves comparable or better results than the S2P pipeline, even without having been trained on satellite images, as illustrated in Table 3.

5.2. Aggregated pair-wise DSM

Pair-wise DSMs computed for S2P and S2P-GANet methods are aggregated after being selected and ordered with the criteria explained in Section 3.5. Among the criteria used, the *a posteriori* filtering becomes relevant in the case of sets of images with important seasonal changes as seen in Figure 7. Note that in OMA_FIT dataset, less than half of pair-wise DSMs pass the criterion and are considered for the aggregation (see column #DSMs in Table 3), however the heuristics with *a posteriori* filtering achieves similar results as the oracle guided integration.

For the aggregation, the selected DSMs are registered to the ground truth DSM and then reduced by the median.

Table 3. Results on stereo pairs for S2P, S2P-GANet and COLMAP methods. The results are the average of the metrics on each of the datasets. Only DSMs with at least 70% of valid pixels are considered. The best metrics on each row are depicted in bold.

	S2P				S2P-GANet				COLMAP			
	# DSMs	COMP	RMSE	MAE	# DSMs	COMP	RMSE	MAE	DSMs	COMP	RMSE	MAE
JAX_NIT	139	**0.61**	4.25	**0.42**	127	**0.61**	**3.22**	0.48	119	0.53	18.03	0.58
MVS_NIT	146	**0.62**	2.89	0.51	131	**0.62**	**2.57**	**0.47**	111	0.56	7.41	0.55
OMA_NIT	136	**0.76**	2.53	**0.27**	143	**0.76**	**2.16**	0.28	123	0.57	17.18	0.53
JAX_FIT	114	0.57	4.85	0.35	90	**0.62**	**2.80**	**0.34**	99	0.38	30.37	1.25
OMA_FIT	71	0.54	4.94	0.50	58	**0.60**	**2.54**	**0.47**	63	0.16	39.25	8.48

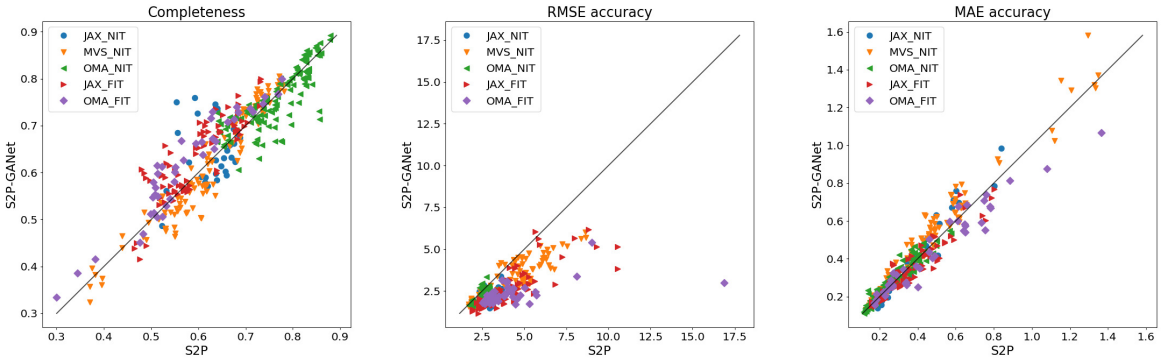

Figure 6. Results on stereo pairs for S2P and S2P-GANet.

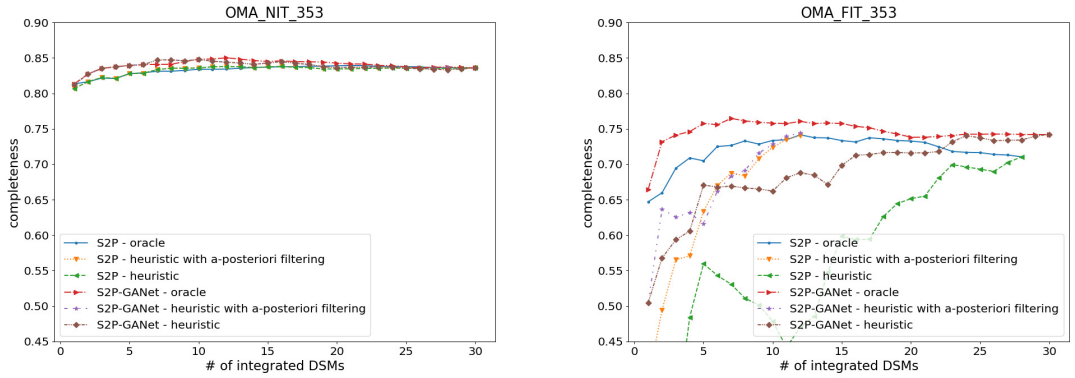

Figure 7. Evaluation of the *a posteriori* pair selection criterium. Comparison of the progressive integration of pair-wise DSMs using the completeness *oracle* (DSM are ordered by completeness) and the heuristic rules based on metadata with and without the *a posteriori* filtering. Left: subregion 353 of OMA_NIT. Right: subregion 353 of OMA_FIT. The use of the *a posteriori* criterium becomes relevant in the case of the OMA_FIT dataset which presents important seasonal changes between images.

Note that this procedure is used for comparing the methods but cannot be used in real cases when ground truth is not available. In a realistic setting a surrogate DSM must be selected as reference for the registration [8].

Figure 8 illustrates the progression of DSM integration for S2P and S2P-GANet on two subregions. Second and fourth rows show the error reduction induced by DSM aggregation. Note the better performance of S2P-GANet, even with less DSMs.

5.3. Pair-wise vs true multi-view methods

Table 4 presents the results for aggregated pair-wise and true multi-view methods. As can be seen, pair-wise methods overcomes true multi-view methods in almost all datasets and metrics.

5.4. Fine-tuning

Although there are multiple datasets for training stereo and multi-view stereo algorithms, these datasets are mostly comprised of close range scenes [30, 18, 34]. In the last years some datasets were deployed that allow training in

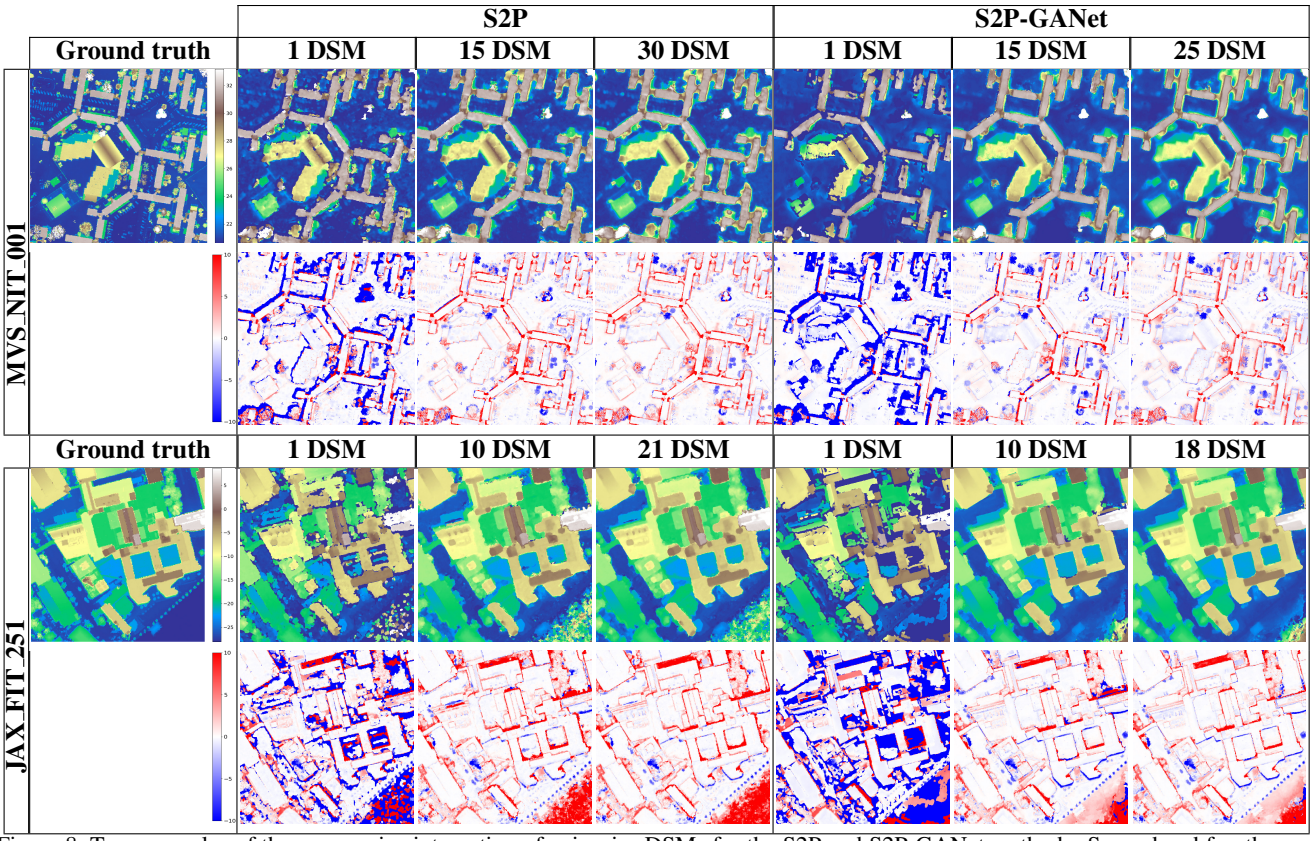

Figure 8. Two examples of the progressive integration of pair-wise DSMs for the S2P and S2P-GANet methods. Second and fourth rows show the difference with the respective ground truth DSM. Colorbars are in meters

long range scenes such as the WHU MVS/Stereo dataset (WHU dataset) [22] and the RVL Purdue SatStereo (RVL dataset) [28]. WHU is a synthetic aerial dataset produced out of thousands of real aerial images covering an area over the Guizhou Province in China. RVL Purdue SatStereo dataset provides a set of stereo-rectified images and associated ground truth disparities for areas of interest drawn from two sources: IARPA's MVS Challenge dataset [3] and the CORE3D-Public dataset [2].

GANet developers made available three pretrained models. The results reported in previous sections use the very basic model trained on the Sceneflow [24] dataset for only 10 epochs. This model is intended, in principle, as a starting point for fine tuning, but it was used as-is in our work. Tests were conducted on the other two available models which are fine-tuned for the Kitti2012 and Kitti2015 [12, 25] benchmarks. Best results were attained with the basic model.

A fine-tuning of GANet, starting from the basic model and using the WHU and RVL datasets was performed. For the training on WHU, over 8000 crops were used and results are reported after 9 epochs. For the RVL dataset, 440 crops were used from the MP1, MP2, MS1 and MS2 subsets of the dataset and results are reported after 65 epochs. The results on stereo pairs for the basic model and the two fine-tuned models are presented in Table 5. In general similar or better results are obtained in completeness and in the accuracy with the fine-tuned models.

6. Discussion

Satellite images have specific characteristics that hinder the adaptation of well established methods used on close range images. Most satellite pipelines in use are based on pair-wise approaches with classic methods that are known to achieve accurate results. This study confirms this fact showing that it is hard to beat the baseline pipeline. On the other hand, results also expose that other valuable methods from the computer vision field can be adapted to work on satellite images since they get results comparable and in some cases slightly better than the baseline even if they have not been trained on satellite images.

Stereo methods can be adapted to work as a stage of an existing satellite stereo pipeline. In this work we tested the GANet method as an alternative "stereo matching" step in the S2P pipeline. An interesting finding is that the results for the S2P-GANet variant were similar and in some cases better than the S2P baseline pipeline without an specific training. The fine-tuning of GANet in more appropriate datasets, such as WHU and RVL, showed a slight

Table 4. Results of the tested pair-wise and true multi-view methods on each subregion of the datasets.

		Pair-wise								True multi-view					
		S2P				S2P-GANet				COLMAP			CasMVSNet		
		# DSMs	COMP	RMSE	MAE	# DSMs	COMP	RMSE	MAE	COMP	RMSE	MAE	COMP	RMSE	MAE
JAX_NIT	**JAX_156**	30	0.822	1.803	**0.123**	30	0.813	1.864	0.144	0.763	3.322	0.175	**0.830**	**1.423**	0.147
	JAX_165	25	**0.690**	6.042	**0.311**	20	0.676	4.260	0.431	0.570	6.645	0.515	0.624	**3.741**	0.381
	JAX_214	24	**0.699**	5.868	**0.271**	22	0.671	4.847	0.417	0.593	8.480	0.455	0.612	**4.331**	0.485
	JAX_251	30	**0.675**	4.747	**0.313**	25	0.661	3.946	0.396	0.561	12.990	0.621	0.654	**2.968**	0.351
	JAX_264	30	0.792	2.562	**0.160**	30	**0.833**	**2.147**	0.183	0.697	7.341	0.210	0.755	2.231	0.214
MVS_NIT	**MVS_001**	30	**0.744**	2.570	**0.282**	25	0.730	**2.388**	0.367	0.675	2.666	0.389	0.682	2.917	0.329
	MVS_002	30	**0.828**	2.024	**0.158**	30	0.823	**1.945**	0.180	0.787	1.978	0.223	0.827	2.587	0.164
	MVS_003	28	**0.688**	3.873	**0.338**	23	0.667	**3.851**	0.415	0.645	3.915	0.413	0.642	3.913	0.359
	MVS_004	30	**0.707**	**2.218**	**0.358**	27	0.654	2.263	0.544	0.681	2.324	0.392	0.665	2.461	0.412
	MVS_005	28	**0.670**	2.983	0.440	26	0.663	**2.683**	0.433	0.632	2.867	0.488	0.633	2.760	**0.407**
OMA_NIT	**OMA_203**	29	0.867	1.728	**0.105**	30	**0.871**	**1.678**	0.127	0.817	1.779	0.151	0.800	1.923	0.236
	OMA_247	28	0.841	2.458	**0.146**	27	**0.845**	**2.345**	0.149	0.778	2.973	0.217	0.803	2.544	0.221
	OMA_251	25	0.903	2.769	**0.095**	30	**0.915**	**2.240**	0.124	0.856	3.000	0.167	0.835	2.360	0.219
	OMA_287	24	0.832	2.561	0.155	28	**0.853**	**2.025**	**0.138**	0.743	3.801	0.220	0.763	2.858	0.285
	OMA_353	30	0.836	2.455	**0.126**	28	**0.837**	**2.387**	0.138	0.796	2.420	0.192	0.792	2.738	0.193
JAX_FIT	**JAX_156**	29	0.791	2.031	**0.123**	30	**0.828**	1.739	0.124	0.693	7.715	0.188	0.800	**1.579**	0.171
	JAX_165	15	0.692	5.985	**0.274**	6	**0.706**	**4.195**	0.310	0.579	10.845	0.384	0.529	5.267	0.465
	JAX_214	19	0.678	6.615	**0.244**	12	**0.701**	**3.951**	0.285	0.586	12.693	0.334	0.444	4.257	0.655
	JAX_251	21	0.681	4.461	0.285	18	**0.703**	**3.624**	**0.284**	0.611	15.230	0.370	0.610	4.447	0.371
	JAX_264	30	0.728	3.074	0.187	24	**0.794**	**2.407**	**0.181**	0.621	8.564	0.263	0.687	2.807	0.286
OMA_FIT	**OMA_203**	17	**0.729**	2.650	**0.229**	7	0.707	2.381	0.392	0.500	20.750	0.537	0.602	**2.336**	0.504
	OMA_247	13	0.792	2.796	0.229	10	**0.825**	2.501	**0.198**	0.574	13.987	0.383	0.725	**2.419**	0.305
	OMA_251	14	0.859	3.224	**0.156**	17	**0.891**	**2.318**	0.208	0.623	17.756	0.312	0.679	3.049	0.388
	OMA_287	15	0.772	3.853	0.243	12	**0.822**	**2.557**	**0.211**	0.537	16.973	0.534	0.610	4.245	0.462
	OMA_353	12	0.740	2.652	**0.315**	12	**0.744**	**2.511**	0.375	0.528	16.477	0.651	0.693	2.694	0.459

Table 5. Results of S2P-GANet on the stereo pairs trained with SceneFlow, WHU and RVL datasets. Results are the average of the metrics on each of the datasets. Only DSMs with at least 70% of valid pixels are considered. Best metrics on each row are in bold.

	SceneFlow (basic model)				WHU (9 epochs)				RVL (65 epochs)			
	# DSMs	COMP	RMSE	MAE	# DSMs	COMP	RMSE	MAE	DSMs	COMP	RMSE	MAE
JAX_NIT	127	0.61	3.22	0.48	127	**0.62**	3.37	**0.43**	116	0.60	**2.80**	0.55
MVS_NIT	131	0.62	2.57	0.47	134	**0.64**	2.55	**0.42**	140	0.62	**2.36**	0.47
OMA_NIT	143	**0.76**	2.16	0.28	149	**0.76**	2.20	**0.24**	138	**0.76**	**2.14**	0.28
JAX_FIT	90	0.62	2.80	0.34	70	**0.64**	2.32	**0.27**	92	**0.64**	**2.35**	0.35
OMA_FIT	58	0.60	2.54	0.47	26	0.63	2.24	**0.34**	66	**0.64**	**2.03**	0.38

but consistent improvement in mean on all datasets used in the experiments. The improvement are more notorious in the (tougher) far in time acquired images (JAX_FIT, OMA_FIT).

Regarding true MVS methods, Zhang et al. proposed in [37], an adaptation strategy and implemented it for classic methods such as COLMAP and Planesweep. The approach, as reported in their article, does not achieve better results than a pair-wise MVS based on S2P. Nevertheless, the approach is very interesting and can be applied to other methods, not intended initially for satellite imagery. We use that strategy to adapt the CasMVSNet method in this work. Table 4 shows that, although not outstanding, the metrics are close to the ones of pair-wise MVS. Results for CasMVSNet were obtained using a model pretrained on the DTU [17] dataset comprised of close range scenes.

Although the popularity of DL methods, they are still not the preferred option in satellite stereo pipelines. The results obtained with DL methods in this study show the potential of using this kind of algorithms on satellite images as a step in a classic pipeline or as an end-to-end MSV solution. Both tested DL methods exhibited a great generalization power in particular GANet. It is interesting to note that part of the internal structure of the GANet mimics SGM [16] which has been extensively used as the main aggregation strategy in several classic satellite stereo pipelines [26, 29, 21]. This fact, along with its generalization ability, points to this method as a really attractive option to include it in an existing satellite pipeline. The adaptation and enhancement of this and other methods to satellite images depends largely on the existence of aerial and satellite training datasets that are still scarce. The availability of more datasets such as WHU and RVL will surely trigger the adaptation of exiting methods and the development of new methods for the benefit of the remote sensing community.

References

[1] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *British Machine Vision Conference*, volume 11, pages 1–11, 2011.

[2] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D Hager, and Myron Brown. Semantic stereo for incidental satellite images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1524–1532. IEEE, 2019.

[3] Marc Bosch, Zachary Kurtz, Shea Hagstrom, and Myron Brown. A multiple view stereo benchmark for satellite imagery. In *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–9. IEEE, 2016.

[4] Steven D. Cochran and Gérard Medioni. 3-d surface description from binocular stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(10):981–994, oct 1992.

[5] Robert T Collins. A space-sweep approach to true multi-image matching. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 358–363, June 1996. ISSN: 1063-6919.

[6] Carlo de Franchis, Enric Meinhardt-Llopis, Julien Michel, Jean-Michel Morel, and Gabriele Facciolo. An automatic and modular stereo pipeline for pushbroom images. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3:49–56, 2014.

[7] Gabriele Facciolo, Carlo de Franchis, and Enric Meinhardt. MGM: A significantly more global matching for stereovision. In *Procedings of the British Machine Vision Conference 2015*, pages 90.1–90.12. British Machine Vision Association, 2015.

[8] Gabriele Facciolo, Carlo De Franchis, and Enric Meinhardt-Llopis. Automatic 3D Reconstruction from Multi-date Satellite Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1542–1551, Honolulu, HI, July 2017. IEEE.

[9] Pascal Fua. A parallel stereo algorithm that produces dense depth maps and preserves image features. *Machine vision and applications*, 6(1):35–49, 1993.

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

[11] David Gallup, Marc Pollefeys, and Jan-Michael Frahm. 3d reconstruction using an n-layer heightmap. In *Joint Pattern Recognition Symposium*, pages 1–10. Springer, 2010.

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[13] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M Seitz. Multi-view stereo for community photo collections. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[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. page 10.

[15] Rostam Affendi Hamzah and Haidi Ibrahim. Literature survey on stereo vision disparity map algorithms. *Journal of Sensors*, 2016.

[16] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008.

[17] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanaes. Large Scale Multi-view Stereopsis Evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413, Columbus, OH, USA, June 2014. IEEE.

[18] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

[19] Thomas Krauß, Pablo d'Angelo, Mathias Schneider, and Veronika Gstaiger. The fully automatic optical processing system catena at DLR. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 177–183, 2013.

[20] 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.

[21] Matthew J. Leotta, Chengjiang Long, Bastien Jacquet, Matthieu Zins, Dan Lipsa, Jie Shan, Bo Xu, Zhixin Li, Xu Zhang, Shih-Fu Chang, Matthew Purri, Jia Xue, and Kristin Dana. Urban semantic 3d reconstruction from multiview satellite imagery. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[22] Jin Liu and Shunping Ji. A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6050–6059, 2020.

[23] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[24] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134.

[25] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. In *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015.

[26] Zachary M Moratto, Michael J Broxton, Ross A Beyer, Mike Lundy, and Kyle Husmann. Ames stereo pipeline, NASA's open source automated stereogrammetry software. *LPI*, (1533):2364, 2010.

[27] Ozge C Ozcanli, Yi Dong, Joseph L Mundy, Helen Webb, Riad Hammoud, and Victor Tom. A comparison of stereo and multiview 3-D reconstruction using cross-sensor satellite imagery. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 17–25, June 2015.

[28] Sonali Patil, Bharath Comandur, Tanmay Prakash, and Avinash C Kak. A new stereo benchmarking dataset for satellite images. *arXiv preprint arXiv:1907.04404*, 2019.

[29] Ewelina Rupnik, Mehdi Daakir, and Marc Pierrot Deseilligny. Micmac–a free, open-source solution for photogrammetry. *Open Geospatial Data, Software and Standards*, 2(1):1–9, 2017.

[30] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.

[31] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, Las Vegas, NV, USA, June 2016. IEEE.

[32] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016.

[33] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[34] Thomas Schops, Johannes L. Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2538–2547, Honolulu, HI, July 2017. IEEE.

[35] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth Inference for Unstructured Multi-view Stereo. *arXiv:1804.02505 [cs]*, July 2018. arXiv: 1804.02505.

[36] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip H.S. Torr. GA-Net: Guided Aggregation Net for End-To-End Stereo Matching. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 185–194, Long Beach, CA, USA, June 2019. IEEE.

[37] Kai Zhang, Noah Snavely, and Jin Sun. Leveraging vision reconstruction pipelines for satellite imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[38] Enliang Zheng, Enrique Dunn, Vladimir Jojic, and Jan-Michael Frahm. PatchMatch Based Joint View Selection and Depthmap Estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1517, Columbus, OH, USA, June 2014. IEEE.