AN EVALUATION OF STEREO AND MULTIVIEW ALGORITHMS FOR 3D RECONSTRUCTION WITH SYNTHETIC DATA

M. Fuentes Reyes[1,*], P. d'Angelo[1], F. Fraundorfer[1, 2]

[1] Department of Photogrammetry and Image Analysis, Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling 82234, Germany - (mario.fuentesreyes, pablo.angelo, friedrich.fraundorfer)@dlr.de
[2] Institute of Computer Graphics and Vision, Graz University of Technology (TU-Graz), Graz 8010, Austria - fraundorfer@icg.tugraz.at

Commission II, WG II/1

KEY WORDS: Stereo matching, multi-view stereo, synthetic datasets, urban reconstruction

ABSTRACT:

The reconstruction of 3D scenes from images has usually been addressed with two different strategies, namely stereo and multi-view. The former requires rectified images and generates a disparity map, while the latter relies on the camera parameters and directly retrieves a depth map. For both cases, deep learning architectures have shown an outstanding performance. However, due to the differences between input and output data, the two strategies are difficult to compare on a common scene. Moreover, for remote sensing applications multi-view data is hard to acquire and the ground truth is either sparse or affected by outliers. Hence, in this article we evaluate the performance of stereo and multi-view architectures trained on synthetic data resembling remote sensing images. The data has been and processed and organized to be compatible with both kind of neural networks. For a fair comparison, training and testing are done only with two views. We focus on the accuracy of the reconstruction, as well as the impact of the depth range and the baseline of the stereo array. Results are presented for deep learning architectures and non-learning algorithms.

1. INTRODUCTION

Within the computer vision community, the research into 3D reconstruction has been a recurrent topic. By having two or more images from the same scene, the task is to reconstruct a 3D representation of such scene based on the matching of corresponding points between the images. This concept is helpful for many fields, as additional sensors to estimate the distance to the objects in the area of interest are not required, just the information captured by the cameras. In the particular case of remote sensing, sensors to retrieve distance such as LiDAR or Synthetic-aperture radar (SAR) are expensive and produce a sparse ground truth with many outliers. Additionally, if the acquisition times of the these sensors differ largely from the one of the cameras, the measurements might not correspond to the content of the images. This discrepancy could happen due to factors such as city growth, seasonal changes, natural disasters, among others. Therefore, an accurate algorithm for 3D reconstruction relying only on optical cameras is a valuable resource for the processing of aerial and satellite images.

Most of the algorithms use either the stereo matching or the multi-view stereo (MVS) approach. For stereo matching, pairs of epipolar rectified images are given as input to compute a disparity map, which will be later converted into a depth map according to the configuration of the stereo array. The stereo algorithms generally follow a pipeline consisting of matching cost computation, cost aggregation, disparity estimation and disparity refinement (Scharstein and Szeliski, 2002). Contrarily, MVS algorithms deal with two or more views and directly work in 3D space. A common strategy for computing the depth is the plane sweep algorithm, where a plane is swung in the 3D space in front of the camera and depth is computed at each location from the different views based on the 2D projections of such plane. A widely known pipeline for MVS based on the plane sweep algorithm is COLMAP (Schönberger et al., 2016).

Lately, deep learning algorithms are leading in terms of accuracy and completeness. However, the stereo matching and MVS architectures have been developed separately due to the nature of the datasets and and the output domain (disparity and depth maps respectively). Even when both would be a viable solution for the reconstruction of the same scene, this has not been explored yet. In addition, learning models require large amounts of data and ground truth, which is hard to acquire and the ground truth is often incomplete. Hence, using synthetic data is an option to evaluate the performance of the networks, as we can generate data in different formats and retrieve all the required parameters, such as camera extrinsics and intrinsics, and the configuration of the stereo rig. Thus, synthetic data can be suitable for both stereo and MVS frameworks.

In this article, we present an evaluation of stereo and MVS deep learning algorithms applied to the same scenes. We train the algorithms in the same datasets to set a fair comparison, for which the datasets have been properly adapted. We utilise the SyntCities dataset from our previous work (Fuentes Reyes et al., 2022), as this resembles remote sensing aerial imagery and provides all necessary input data for the selected algorithms and the SceneFlow (Mayer et al., 2016) dataset, which has been widely used for training. Non-learning algorithms are considered as well for a comparable baseline. As accuracy is a very important factor in remote sensing applications, such as the generation of Digital Surface Models (DSMs), we evaluate the prediction error with a margin of 3 and 1 meters.

Our main contributions are as follow:

- We prepared synthetic data to be compatible with stereo and MVS frameworks, setting similar training conditions.

* Corresponding author

- We trained different models and evaluated the performance in terms of the accuracy for the predicted depth.
- We study the effect of the baseline and occlusions in the depth predictions.

2. RELATED WORK

In this section we describe some of the existing reconstruction algorithms as well as the related datasets, some of which are also used as benchmarks. Detailed differences between stereo and MVS frameworks are also discussed.

2.1 Stereo networks

Before deep learning frameworks, most of the algorithms followed the pipeline mentioned above with matching cost computation, cost aggregation, disparity estimation and disparity refinement. In these cases, a cost volume is created for the disparity candidates and those disparities with the smallest cost are selected and refined for the final disparity map. A well known algorithm derived from this principle is Semi-Global Matching (SGM) (Hirschmuller, 2008) thanks to its trade-off between accuracy and computational cost. SGM computes the cost along different paths and penalizes large disparity changes. Similarly, More Global Matching (MGM) (Facciolo et al., 2015) takes into account more than one direction for the cost computation and achieves higher performance than SGM, with slightly more computational resources.

MC-CNN (Zbontar and LeCun, 2016) was the first deep learning architecture used in the stereo matching and conceived only to replace the cost volume generation part, while the refinement was still conducted with no-learning algorithms such as SGM. After that, some end-to-end networks were designed to encompass the whole stereo pipeline and generate directly the disparity map as output. DispNet (Mayer et al., 2016) utilized an encoder-decoder architecture, whereas GC-Net (Kendall et al., 2017) included 3D convolutions to enhance the contextual information. PSMNet (Chang and Chen, 2018) additionally introduced a spatial pyramid pooling module to collect information from different scales. GA-Net (Zhang et al., 2019) incorporated layers which are a differentiable form of SGM and led to a performance boost in accuracy, smooth results and improved the estimation on textureless areas. AANet (Xu and Zhang, 2020) replaced 3D convolutions with intra- and cross-scale cost aggregation layers, reducing significantly the computational costs, inference times and with little impact on the accuracy. RAFT-Stereo (Lipson et al., 2021) combined gated recurrent units (GRUs) and a correlation pyramid, showing a robust result for textureless and overexposed areas. Newer architectures such as STTR (Li et al., 2021) contain Transformers and offer a good generalization across domains.

For our experiments we selected GA-Net and AANet due to its accuracy and reduced computational cost respectively. They are also a common framework to compare with new architectures and both are based on a cost-volume network.

2.2 Multi-view networks

Multi-view stereo algorithms take two or more views into account while estimating the depth of the objects in the scene. The input images do not need to be stereo-rectified and can be taken from different points of view, but extrinsic and intrinsic parameters are required to know how the cameras relate to one another. Normally, the views are sorted according to the camera position and orientation, so views close together are preferred as input for the algorithm. For a reference image, n-additional views are selected to estimate the depth map of the reference image.

COLMAP (Schönberger et al., 2016) selects the views according to the geometric and photogrammetric information, and then computes the depth estimation through multi-view geometric consistency and further refinement. GIPUMA (Galliani et al., 2015) generates random 3D planes in space and the most suitable ones are iteratively propagated to get an accurate depth map, which is estimated efficiently thanks to its GPU implementation.

In a similar way to stereo matching, deep learning has also achieved an outstanding performance for MVS. MVSNet (Yao et al., 2018) proposed to create a depth volume approach based on the plane sweep algorithm, where the best candidate of the volume for each pixel is selected as the depth value. MVSNet is also the base for the development of newer architectures. CasMVSNet (Gu et al., 2020) improved the efficiency in terms of computational costs by using a coarse to fine scheme. Here, the predictions at the coarse volume are taken as a starting point to build a small volume around the estimated value for the next stages, reducing the number of simultaneous candidates. In VisMVSNet (Zhang et al., 2020) an additional uncertainty estimation is computed for the visibility of each pixel, including in that way the information related to the occlusions. Such occluded pixels are then avoided in the fusion process and generated a more robust result. Another case is UniMVSNet (Peng et al., 2022), where a coarse to fine scheme similar to CasMVSNet is enhanced by a unified representation that deals with the prediction as a regression and a classification task simultaneously. UniMVSNet did not only show a very good performance, but can handle the computational resources efficiently.

Another two important cases are R-MVSNet (Yao et al., 2019) and PatchMatchnet (Wang et al., 2021), although these two are not based on a depth-volume strategy as the previous cases. R-MVSNet applies a regularization through a GRU network sequentially, reducing the memory requirements with a higher performance than MVSNet. PatchMatchNet follows an idea based on PatchMatch (Barnes et al., 2009) similar to GIPUMA, leading to both good performance and efficient memory. In our analysis we decided to use UniMVSNet because of its accuracy and memory efficiency. Besides, it is based on a cost volume strategy as GANet and AANet.

2.3 Datasets

Deep learning strategies are not only known for their performance, but also for being data demanding. In the autonomous driving field for example, the KITTI 2012 (Geiger et al., 2012) and KITTI (Menze and Geiger, 2015) datasets are regularly not enough to train a neural network model because of their size and the incomplete ground truth. To help overcome this, synthetic data can be generated with thousands of samples and accurate ground truth, as the geometric details of the 3D models can be retrieved by the rendering software. Hence, it is a common strategy to pre-train the model in a extensive synthetic dataset and later apply a fine-tuning stage to compensate for the domain gap. A notable example of synthetic data is the SceneFlow dataset, the main reference to train stereo networks. The

dataset comprises more than 35k stereo pairs with corresponding ground truth and a large variety of shapes and textures.

In parallel, datasets have also been developed for the MVS architectures. The DTU dataset (Aanæs et al., 2016) made use of a robot arm to take pictures of small objects from different directions. Various lighting conditions are also included to enhance the color distribution of the images. Another remarkable case is the Tanks and Temples (T&T) dataset (Knapitsch et al., 2017) with images taken from real indoors and outdoors environments, making the 3D reconstruction a challenging task. Both DTU and T&T are a common benchmark to evaluate the performance of MVS architectures. However, the ground truth is not accurate for all the pixels due to the sensor and scene properties. Same as for stereo matching, the synthetic data also represent a solution to train or at least pre-train the models. In this context, BlendedMVS (Yao et al., 2020) is a computer generated dataset with a large variety of textures, shapes and points of view that is compatible with MVS frameworks, being a common reference for training as SceneFlow is for stereo frameworks.

Still, available large datasets have a format not compatible for the two studied frameworks. Stereo datasets would require additional views from the same scene and the respective camera parameters to be used in a MVS algorithm. Contrariwise, MVS datasets would require epipolar rectification to be applied to a stereo algorithm, which might affect the quality of the ground truth due to the rectification process. Given this situation, we refer to the SyntCities dataset, as the stereo pairs also include the camera parameters, facilitating both stereo and MVS applications.

3. METHODOLOGY

In the present section we describe how the datasets have been processed to be compatible with the selected neural networks, as well as the series of experiments and considerations aiming to carry a fair comparison of the algorithms.

3.1 Data preparation

As discussed above, available datasets in their current formats cannot be directly implemented in both stereo and MVS architectures. Therefore, we have selected only two cases, SceneFlow and SyntCities to be processed in a compatible format.

3.1.1 SceneFlow preparation The images included in the SceneFlow dataset are already paired and fulfill the epipolar constraints. To apply them for a MVS algorithm we require to include the camera parameters, which can be derived from the information provided by the authors. Focal length, as well as the principal points and the baseline (defined as 1 in Blender units) are provided for all the images, which helps to create the intrinsic matrices. For the extrinsic matrices, we simplify the parameters to a basic position and rotation. Since the images are taken originally form a video sequence, two pairs of images do not show the exact same scene. Thus, the camera translation between frames is not relevant, as a full reconstruction from the scene is not even possible. For the rotation part, both left and right views can be assumed to come from a camera that has no rotations, as the camera planes are co-planar. Therefore, we can use as extrinsic matrices:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{1}$$

for the left and right images, respectively. To generate the depth ground truth, we compute the depth maps from the provided disparities with the formula:

$$z = f * b/d, \tag{2}$$

where z = depth, f = focal length, b = baseline and d = disparity. MVS approaches make use of a pre-defined depth range for each image, which is usually given by the sensor and acquisition conditions. For SceneFlow, we take the depth map values of each image and we set the depth range to 2th percentile as minimum and $\mu + \sigma$ as maximum, being μ and σ the mean and standard deviation respectively. This also helps to focus on objects that are close to the camera.

3.1.2 SyntCities preparation SyntCities is a dataset to train stereo matching networks with patches resembling remote sensing scenes and under controlled simulated conditions. Three 3D city models are used to render the dataset: Paris, Venice and New York. The samples are given for ground sample distances (GSD) of 10cm, 30cm and 100cm and provided with training and testing subsets. Although not originally designed to work in a MVS framework, the camera parameters are available and samples along the same epipolar line can be used as the additional views. For the current article, we do not use the additional views simultaneously for the reconstruction, but we use the views to create diverse stereo pairs and study the effect of the baseline, which also implies differences in terms of occlusion.

In Figure 1 we can observe how the samples are selected for both the stereo and MVS implementation. Within the SyntCities dataset, many samples are rendered with the same conditions but different base height ratios (bhr) for the stereo rig, which helps us to study the effects of the baseline. By default, SyntCities images are given in pairs, which are represented for simplicity by the legends Baseline 1, Baseline 2 and Baseline 3 in Figure 1. From there, we take the left sample from the largest baseline as a reference (R) and use the other images as additional views (V) for stereo pairing. The base height ratio determines the baseline b from the height h as $bhr = b/h$, where h =2000m for all cases. Bhr values are 0.1, 0.2, 0.3, 0.4 and 0.5 (with baselines of 200m, 400m, 600m, 800m and 1000m respectively) for the Paris and Venice models. For a $bhr = 0.5$, the simulated camera resembles an acquisition field of view (FOV) around 28°. Images from the New York samples were not used, as these have a smaller baseline. In Figure 2, examples for Paris are given for a reference image with its respective 5 additional views. As expected, bigger changes in the images occur at larger baselines, which also implies larger occluded areas.

3.2 Conducted experiments

We utilized few well-known algorithms to test stereo pairs from SyntCities with different baselines. Both learn-based and traditional algorithms were considered. For the traditional part we selected SMG and MGM, as these are a common reference to compare other algorithms. The used SGM implementation is the one in the CATENA pipeline (Krauß, 2014) and for

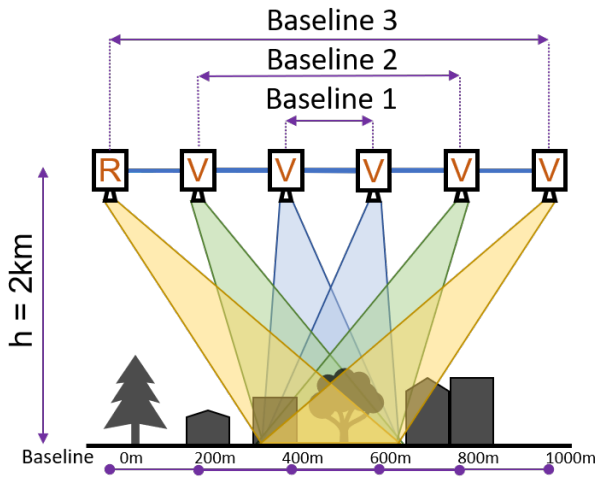


Figure 1. Selected geometry for SyntCities samples. All images lie on the same epipolar line with different baselines. For a reference view (R), 5 additional views (V) are available. Baseline distances are given for each view.

MGM we utilized the one provided by the author [1]. We used $P1 = 400$ and $P2 = 800$ with 16 directions and a Census-cost (Zabih and Woodfill, 1994) for SGM. In the case of MGM, we used $P1 = 8$ and $P2 = 32$. Both SGM and MGM were given [-10, 192] as disparity range. Disparity maps are computed before and after applying the left-right consistency (LRC) check. Similar to the neural network results, the case before LRC check produces values for most of the pixels, so we used these results for the comparison. The results after LRC are also relevant, as these show the refinement effect.

We trained all the selected networks (GANet, AANet and UniMVSNet) on the SceneFlow dataset, as this is a common practice for stereo algorithms and it has a large pool of images. Testing, on the other hand, was done for SyntCities images. By avoiding training and testing on the same domain, we do not give additional advantage to the learning algorithms. We trained UniMVSNet with 2 views, so all models are based on the same training dataset with the paired images. GANet was trained for 27 epochs with a disparity range of [0,192] in 4 x GeForce RTX 2080 GPU. AANet was trained with the same conditions but 350 epochs, having a similar training time. UniMVSNet was trained for 16 epochs with 192 depth planes in 1x GeForce RTX 2080 GPU.

An important point to note here is to differentiate between the disparity and depth ranges. From the equation 2, we can see that disparity and depth are inversely related. The deep learning MVS frameworks already perform in the 3D space based on the plane sweep algorithm, where the planes hypotheses are uniformly distributed in the space of the camera. Contrarily, the stereo networks search for the disparity candidates in a uniform sampling, which is later non-uniform when the disparities are converted into depth values. This relation also discussed in detail in the CIDER (Xu and Tao, 2020) network. Because of this non-linear relationship, stereo and MVS algorithms are affected by the distribution of the depth values in space. In the figure 3, such relationship is displayed for an image of the SceneFlow dataset with $f = 450$ and $b = 1$ for the disparity range [0, 192]. As we can see, the depth values are sparsely sampled for the low disparities and densely sampled for high disparities in stereo algorithms. We have adapted the depth ranges of the images to cover most of the content and alleviate the problem given by the depth - disparity range inconsistencies.

[1] https://github.com/gfacciol/mgm

4. EVALUATION

To asses the results in terms of accuracy, completeness and effect of the baseline, we tested the algorithms:

- SGM: SGM result before LRC check.
- SGM w/LRC: SGM result after LRC check.
- MGM: MGM result before LRC check.
- MGM w/LRC: MGM result after LRC check.
- AANet: result of AANet converted to depth.
- GANet: result of GANet converted to depth.
- UniMVSNet: UniMVSNet result directly as depth map.

The first metric used to analyze the results is the Median Absolute Deviation (MAD), as this is a robust metric for skew distributions (Höhle and Höhle, 2009). This is computed as:

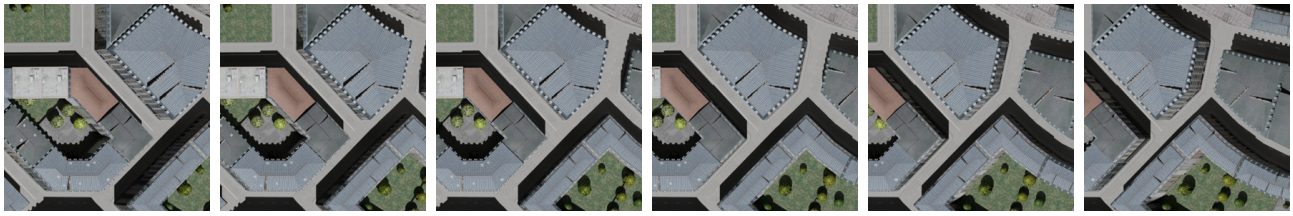
$$\mathrm{MAD}_{\mathrm{diff}} = \mathrm{median}(\mid X_{\mathrm{diff}} - \widetilde{X}_{\mathrm{diff}} \mid), \tag{3}$$

where $\widetilde{X}_{\mathrm{diff}} = \mathrm{median}(X_{\mathrm{diff}})$, and $X_{\mathrm{diff}} = X - \bar{X}$, being X the ground truth, $\bar{X}$ the generated result and X_{diff} the difference between both.

Similarly to disparity maps evaluations, we also compared the error rate of the prediction but in this case oriented to the depth values. We computed the error rate 3 meters (ER-3), which is the percentage of pixels where the prediction error is larger than 3 meters. Similarly, we compute the error rate 1 meter (ER-1) following the same principle. The latter is critical for remote sensing, where accuracy within 1 meter is expected for applications such as DSM generation. The thresholds are based also on the influence of the disparity - depth relationship. We took an image with 600m baseline, its respective camera parameters and $d = 1, 2$. The corresponding depth values were $z = 1999.01, 1998.01$, having a difference of 1m. In any case, considering that the objects are located at 2000m from the camera, 1m error is a strict margin, so we also evaluate for 3m.

Completeness is also a desired feature for the reconstruction algorithms. Non-learning based approaches like SGM or MGM routinely refine the predicted disparity map with LRC to retrieve only the pixels where the disparities are more reliable and thus shortening the presence of outliers. However, this refinement reduces sometimes significantly the density of the result, creating a lot of no defined regions in the disparity maps. Neural networks on the other hand generate a prediction for each pixel in the image, but this allows the outliers to remain in the predicted disparity map. Hence, we do also report the percentage of pixels that were used for the metrics.

We also study the performance with and without occluded areas. As we have a dense ground truth for disparities, we also created LRC masks from them to identify the occluded areas. Such masks apply to pixels that are only visible in one of the images. While it is expected that the algorithms cannot estimate the correct depth value in such areas, the prediction can still be satisfactory due to the neighbouring context. For instance, deep learning approaches gather contextual information to smoothly interpolate on the occluded areas. Besides, we want to observe how large is the error in the non-occluded areas, where the error is assumed to be low.



(a) Reference Image (b) bhr=0.1 (c) bhr=0.2 (d) bhr=0.3 (e) bhr=0.4 (f) bhr=0.5

Figure 2. Examples of paired images from SyntCities along a common epipolar line. For the reference image (a), images with 5 different base height ratios (b-f) are given.

		With occluded pixels				Non-occluded pixels			
Algorithm	**Baseline(m)**	**ER-1(↓)**	**ER-3(↓)**	**MAD(↓)**	**Val.pix.(↑)**	**ER-1(↓)**	**ER-3(↓)**	**MAD(↓)**	**Val.pix.(↑)**
SGM	200	24.86	11.58	0.50	99.53	23.26	10.04	0.48	97.08
	400	22.18	14.33	0.28	98.63	15.79	9.25	0.25	89.31
	600	25.91	19.38	0.24	97.53	14.71	9.85	0.18	82.29
	800	30.77	24.76	**0.23**	96.76	15.19	10.52	0.16	76.39
	1000	35.52	29.62	0.27	95.83	16.14	11.40	**0.14**	70.82
MGM	200	23.67	**9.03**	0.49	99.76	21.86	7.20	0.48	97.05
	400	**21.90**	13.45	0.30	99.71	15.00	7.60	0.26	89.31
	600	28.23	21.53	0.27	99.67	14.97	9.06	0.20	82.31
	800	34.54	28.50	0.29	99.67	16.27	10.64	0.17	76.41
	1000	40.53	34.57	0.39	99.67	18.47	12.45	0.16	70.85
AANet	200	32.43	12.71	0.54	100.00	31.37	11.95	0.52	97.11
	400	30.93	14.88	0.43	100.00	25.57	10.28	0.37	89.36
	600	32.41	17.30	0.43	100.00	23.37	9.25	0.33	82.35
	800	34.27	20.25	0.45	100.00	22.57	10.08	0.31	76.45
	1000	38.18	23.62	0.55	100.00	23.65	10.67	0.31	70.88
GANet	200	36.04	12.32	0.68	100.00	34.80	11.22	0.66	97.11
	400	24.78	13.02	0.41	100.00	18.25	7.42	0.36	89.36
	600	24.95	15.87	0.36	100.00	13.81	**6.37**	0.28	82.35
	800	27.16	18.91	0.36	100.00	13.11	7.06	0.25	76.45
	1000	29.90	21.75	0.36	100.00	**13.07**	7.54	0.23	70.88
UniMVSNet	200	26.94	12.00	0.42	100.00	25.50	10.95	0.40	97.11
	400	26.52	14.21	0.35	100.00	20.09	9.14	0.31	89.36
	600	29.83	17.66	0.37	100.00	19.01	8.69	0.28	82.35
	800	34.52	21.87	0.43	100.00	20.36	9.72	0.29	76.45
	1000	39.52	26.80	0.55	100.00	21.87	10.82	0.29	70.88
SGM w/LRC	200	20.81	7.86	0.46	93.35	20.15	7.39	0.46	92.24
	400	12.86	6.51	0.23	85.52	10.81	5.37	0.23	82.45
	600	11.69	7.11	0.17	78.05	8.21	4.57	0.16	73.83
	800	11.64	7.65	0.14	71.16	6.73	3.59	0.13	66.12
	1000	12.31	8.73	0.13	65.00	6.18	3.43	0.11	59.42
MGM w/LRC	200	19.47	5.32	0.45	92.54	18.95	4.95	0.45	91.63
	400	10.82	4.06	0.24	82.81	9.52	3.41	0.23	80.70
	600	9.80	4.76	0.17	74.60	7.71	3.18	0.17	72.13
	800	11.02	6.65	0.15	67.85	7.33	3.40	0.14	64.39
	1000	13.08	8.82	0.13	61.63	7.63	3.72	0.12	57.22

Table 1. Experiments results for Paris and Venice images. MAD represents the Median Absolute Deviation, ER-3 the 3 meters error rate, ER-1 the 1 meter error rate and Val. pix. the percentage of pixels with a valid value generated by the algorithm. Underlined bold numbers show the best result (cases w/LRC excluded) for MAD, ER-3 and ER-1.

We selected 20 images for our study from our two virtual cities: 15 from Paris and 5 from Venice. For all the test images, we selected 30cm as GSD and 5 additional views with different baselines Since the images of Paris and Venice have the same baselines, these are averaged for the metrics.

5. RESULTS

From the set of experiments and metrics described above, we present all our results in Table 1. We split the SGM and MGM results depending on whether we used the LRC refinement or not, having the former at the end of the table. In all the remaining cases, the generated results covers most of the pixels. In the figure 4 we have visual results for the ER-3 metric in all tested algorithms for one of the tested images, which corresponds to a 200m baseline case.

We analyse first the results considering occluded areas. In terms of density, all cases (excepting those w/LRC refinement) achieve almost a complete depth map having at least 98% coverage. Looking at MAD, SGM and MGM have the best performance, although it is important to remember that some pixels have no defined values. For ER-3 the traditional methods perform better than the learning ones in the small baselines such as 200m, similar for 400m and worse for 600m, 800m and 1000m. Hence, non-learning algorithms still outperform neural networks but only for small baselines. Nonetheless, non-learning algorithms have the advantage that no training time is required, just fine-tuning of the parameters to enhance the res-

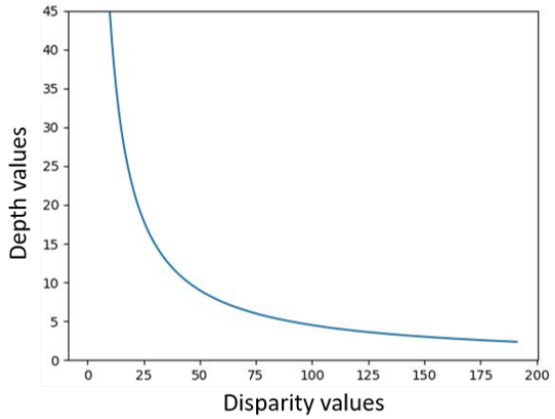


Figure 3. Non-linear relationship between disparity and depth values for an image of the SceneFlow dataset. The disparity range was set to [0, 192], which is common for many implementations.

ult. If we compare ER-1 the trend is similar, where good results for large baselines are given for learning algorithms. Interestingly, GANet is the approach with the best performance in this metric for baselines of 600m and above.

Taking into account the results w/LRC refienement (which applies only to SGM and MGM) we notice much better values in ER-3, ER-1 and MAD. However, this also has a costly price as large sections of the images become undefined. For real applications on the other hand, it is helpful to have an algorithm that delivers only those areas where the estimation offers a good quality, so SGM and MGM are a valuable resource.

For the non-occluded results, in all cases the density is below 100% as expected and for SGM and MGM even lower, as some additional areas are discarded. MAD is better for SGM and MGM, while for neural networks GANet and AANet perform best and worst respectively. Considering ER-3, MGM is the best for a small 200m baseline but GANet outperforms in all the other cases. AANet and UniMVSNet behave similarly. For the strict ER-1, MGM achieves the best result for the small 200m and 400m baselines and GANet for the rest of the cases. In this part, we notice that UniMVSNet overcomes AANet for larger baselines.

Having analysed the dense results, we focus on the SGM w/LRC and MGM w/LRC cases. Accuracy is very high as can be seen from ER-3 and ER-1 values being mostly below 10%. This may be misleading as accuracy has increased while density has decreased. In fact, for the large baselines the depth maps cover less than 60% of the image.

If we look at the LRC check effect more in detail, we can notice that the removed pixels between the before and after results belong mostly to the occluded areas, thus dismissing efficiently the unreliable predictions. For instance, if we compare the case for the 1000m baseline, we notice that ER-1 for SGM goes from 16.14% to 6.18%, while the percentage of valid pixels goes from 70.82% to 59.42%, close to 10% for both values. A similar trend is observed for MGM, where ER-1 values for the 1000m baseline are reduced from 18.47% to 7.63% and the percentage of valid pixels from 70.85 to 57.22, having differences of 10.84% and 13.63% respectively. In figure 4, we can easily notice how most of the pixels with an error larger than 3m are removed by the LRC check, although these regions become undefined outputs.

Comparing only SGM and MGM we notice a similar performance, being MGM slightly improved for small baselines and SGM for the large ones. Cases with or without occlusions, as well as with or without LRC check show a similar behaviour between these two methods. Fine-tuning of the penalty parameters P_1 and P_2 might lead to a better performance, but these have to be set empirically.

Between the two learning stereo methods, namely AANet and GANet, we notice how GANet has the best metrics for all cases except for the 200m baseline, where AANet leads for the ER-1 metric. In general, both show a competitive reconstruction result. In addition, the conversion from disparity to depth, which would represent sparsity in the depth space does not have a strong effect when compared to the UniMVSNet results, being even similar.

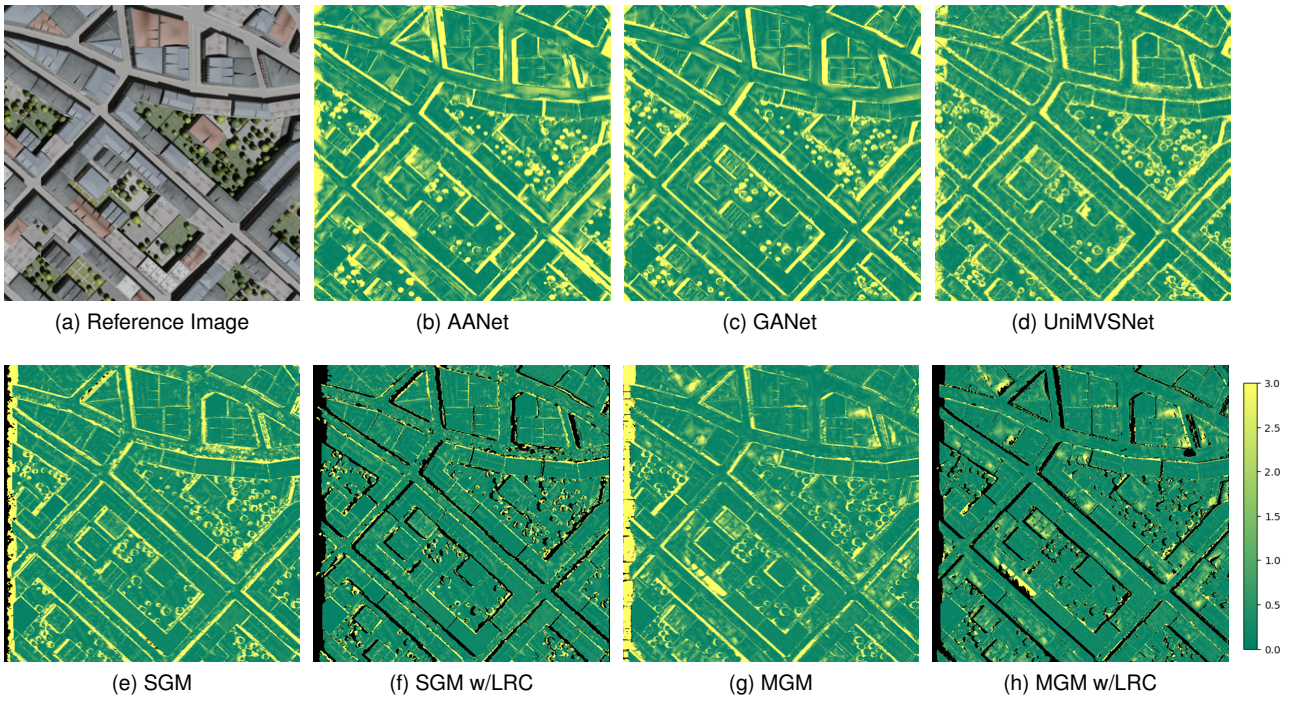
With regard to the main objective of this paper, we also study the performance differences between the stereo (GANet and AANet) and MVS (UniMVSNet) frameworks. The obtained results show that:

- In general, all cases have a comparable performance and are suitable for 3D reconstruction, as ER-3 considering occlusions are between 12% and 27% depending on the baseline
- Overall, for ER-3 the performance degrades when the baseline increases if occluded areas are also counted.
- If we focus only on the non-occluded areas, algorithms tend to perform best for ER-3 in intermediate baselines, while for ER-1 all except the smallest case have a similar performance.
- UniMVSNet is the best for all cases where the baseline is 200m, highlighting its focus on close range views.
- GANet is the best for ER-3 and ER-1 in all baselines except 200m, which shows the best accuracy and is particularly good for ER-1 in the non-occluded regions having a significantly difference with respect to the other algorithms. The matching itself of visible pixels performs the best in this case.
- The prediction for occluded areas in all learning approaches yields better results than the non-learning cases, which shows good capabilities to interpolate from the reliable pixels. Predicted depth maps tend to include smooth regions with sharp boundaries, specially if the baselines are not that large. Such interpolation effect is superior in the stereo networks as the ER-3 scores are lower.

Last but not least, we note the domain gap effect of training and testing in the different datasets. Due to such gap, the performance of the networks is not as high as it can be when it is fine-tuned in the same domain. It is of interest that the non-learning algorithms have a similar performance to the learning ones when the domain gap is present. Thus, for unseen data both options are a valuable resource.

6. CONCLUSIONS

In the present article we conducted a lot of experiments to compare the performance of learning-based stereo and multi-view approaches on a similar setting. We noticed that stereo networks lead to a better reconstruction, especially GANet. Des-



(a) Reference Image (b) AANet (c) GANet (d) UniMVSNet

(e) SGM (f) SGM w/LRC (g) MGM (h) MGM w/LRC

Figure 4. Error maps for a Paris sample. For the reference image (a), we show the error maps for the algorithms AANet (b), GANet (c), UniMVSNet (d), SGM (e), SGM w/LRC (f), MGM(g) and MGM w/LRC (h). Scale bar for the errors given as a reference. Errors are clipped to a maximum of 3m. Regions in black correspond to undefined pixels by the algorithms.

pite a slightly lower performance, MVS networks are also competitive and are even better for small baselines than stereo networks, but the accuracy drops for the large baselines.

We evaluated first considering also occluded areas in the stereo pairs to observe the robustness of the methods in this challenging regions, observing that the interpolation capabilities to predict these values is working reasonably well and is marginally better for the stereo networks. In non-occluded areas we noticed a good performance for most of the cases, which shows that the matching itself is not a problem. Besides, we also included non-learning algorithms in our comparison, which also yielded good results but reduced the number of valid pixels in the predicted depth maps. For future work, we plan to further evaluate the fusion of multiple stereo pairs against the direct MVS result, and add confidence measures to enhance the fusion process.

ACKNOWLEDGEMENTS

Mario Fuentes Reyes is currently funded by a DLR-DAAD Research Fellowship (No. 57478193) to pursue his PhD studies.

REFERENCES

Aanæs, H., Jensen, R. R., Vogiatzis, G., Tola, E., Dahl, A. B., 2016. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120, 153–168.

Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D. B., 2009. PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3).

Chang, J.-R., Chen, Y.-S., 2018. Pyramid stereo matching network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5410–5418.

Facciolo, G., de Franchis, C., Meinhardt, E., 2015. Mgm: A significantly more global matching for stereovision. *Proceedings of the British Machine Vision Conference (BMVC)*, 90.1–90.12.

Fuentes Reyes, M., D'Angelo, P., Fraundorfer, F., 2022. SyntCities: A Large Synthetic Remote Sensing Dataset for Disparity Estimation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 10087-10098.

Galliani, S., Lasinger, K., Schindler, K., 2015. Massively parallel multiview stereopsis by surface normal diffusion. *2015 IEEE International Conference on Computer Vision (ICCV)*, 873–881.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. *Conference on Computer Vision and Pattern Recognition*, 3354–3361.

Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P., 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2495–2504.

Hirschmuller, H., 2008. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2), 328-341.

Höhle, J., Höhle, M., 2009. Accuracy assessment of digital elevation models by means of robust statistical methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(4), 398–406.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A., 2017. End-to-end learning of geometry and context for deep stereo regression. *Proceedings of the IEEE International Conference on Computer Vision*, 66–75.

Knapitsch, A., Park, J., Zhou, Q.-Y., Koltun, V., 2017. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics*, 36(4).

Krauß, T., 2014. Six years operational processing of satellite data using CATENA at DLR: Experiences and recommendations. *KN-Journal of Cartography and Geographic Information*, 64(2), 74–80.

Li, Z., Liu, X., Drenkow, N., Ding, A., Creighton, F. X., Taylor, R. H., Unberath, M., 2021. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6177–6186.

Lipson, L., Teed, Z., Deng, J., 2021. RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. *2021 International Conference on 3D Vision*, 218–227.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T., 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4040–4048.

Menze, M., Geiger, A., 2015. Object scene flow for autonomous vehicles. *IEEE Conference on Computer Vision and Pattern Recognition 2015*, 3061–3070.

Peng, R., Wang, R., Wang, Z., Lai, Y., Wang, R., 2022. Rethinking depth estimation for multi-view stereo: A unified representation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1), 7–42.

Schönberger, J. L., Zheng, E., Pollefeys, M., Frahm, J.-M., 2016. Pixelwise view selection for unstructured multi-view stereo. *European Conference on Computer Vision (ECCV)*.

Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M., 2021. Patchmatchnet: Learned multi-view patchmatch stereo. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14189–14198.

Xu, H., Zhang, J., 2020. AANet: Adaptive aggregation network for efficient stereo matching. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1959–1968.

Xu, Q., Tao, W., 2020. Learning inverse depth regression for multi-view stereo with correlation cost volume. *AAAI Conference on Artificial Intelligence*, 34, 12508–12515.

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. MVSNet: Depth Inference for Unstructured Multi-view Stereo. *European Conference on Computer Vision (ECCV)*.

Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L., 2019. Recurrent MVSNet for High-resolution Multi-view Stereo Depth Inference. *Computer Vision and Pattern Recognition (CVPR)*.

Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L., 2020. BlendedMVS: A Large-scale Dataset for Generalized Multi-view Stereo Networks. *Computer Vision and Pattern Recognition (CVPR)*.

Zabih, R., Woodfill, J., 1994. Non-parametric local transforms for computing visual correspondence. *European Conference on Computer Vision*, Springer, 151–158.

Zbontar, J., LeCun, Y., 2016. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17, 1–32.

Zhang, F., Prisacariu, V., Yang, R., Torr, P. H., 2019. GA-Net: Guided aggregation net for end-to-end stereo matching. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 185–194.

Zhang, J., Yao, Y., Li, S., Luo, Z., Fang, T., 2020. Visibility-aware Multi-view Stereo Network. *British Machine Vision Conference (BMVC)*.