ITC Faculty, University of Twente
Hengelosestraat 99 7514 AE
Enschede
The Netherlands
https://www.itc.nl/EOS

UAV-based multi-sensor datasets for geospatial research - USEGEO

REPORT OF THE SCIENTIFIC INITIATIVE

Principal Investigator:
Dr. **Francesco Nex**, ITC, University of Twente, The Netherlands, f.nex@utwente.nl

Co-Investigator:
Dr. **Fabio Remondino**, 3DOM, FBK, Italy, remondino@fbk.eu
Dr. **Ellie Stathopoulou**, 3DOM, FBK, Italy, estathopoulou@fbk.eu
Dr. **Michael Yang**, University of Twente, The Netherlands, michael.yang@utwente.nl
Dr. **Martin Weinmann**, KIT, Germany, martin.weinmann@kit.edu
Dr. **Boris Jutzi**, KIT, Germany, boris.jutzi@kit.edu

MAIN MOTIVATIONS, OBJECTIVES AND PARTNERSHIPS

The generation of accurate 3D representations of the world using images has been one of the main research topics for the last three decades in photogrammetry as well as in the computer vision community. Various methods with different outputs have been developed toward this scope for various applications such as mapping, autonomous navigation (Geiger et al., 2012), localization, and virtual reality. Depth estimation and reconstruction from image data have led to incredible improvements in recent years, as also witnessed by the release of many commercial solutions that are nowadays common practice among practitioners in different domains. These solutions have mainly been developed following standard Structure from Motion (SfM) and Multiple View Stereo (MVS) pipelines and, for real-world and large-scale applications, are normally based on conventional approaches using hand-crafted features and user-defined parameters. Despite the impressive results of such solutions, delivering precise, complete, and aesthetically pleasing 3D reconstruction results in multi-view scenarios is still an open challenge for the scientific community. Inevitably, acquisition conditions such as image network geometry, illumination conditions, and sensor quality can severely affect the reconstruction results; however, the efficiency of the implemented algorithm is of utmost importance to ensure high-fidelity outcomes (Seitz et al., 2006; Wenzel et al., 2013; Remondino et al., 2014).

Deep neural networks have been recently used in several visual recognition tasks such as image classification (Krizhevsky et al., 2012; He et al., 2016), object detection (Girshick et al., 2014; He et al., 2017), and semantic segmentation (Long et al., 2015; Chen et al., 2017; Badrinarayanan et al., 2017) with great success, mainly due to their capability to consider the global semantic context of the image. For depth and disparity estimation,

convolutional networks have been exploited in the two-view (Zbontar et al., 2016; Kendall et al., 2017; Guo et al., 2019) and multi-view scenarios (Yao et al., 2018; Im et al., 2019, Xu et al., 2021b). Deep learning has also revitalized the interest in monocular depth estimation algorithms (aka SIDE or MDE) (Eigen et al., 2014; Laina et al., 2016; Godard et al., 2019; Hermann et al., 2020; Khan et al., 2020; Madhuanand et al. 2020) that have become popular in many indoor and outdoor applications due to their cost effectiveness and flexibility.

Although the undeniable potential, the applicability of these methods in real-world scenarios is still debatable. Among the biggest concerns is the need for a vast amount of training data, the high memory requirements, and the limited generalization and domain adaptation performance. Several benchmarks have already been released by different communities in the last years to promote the development of efficient and reliable algorithms across different applications (Geiger et al., 2012; Mayer et al., 2016; Schöps et al., 2017; Koch et al., 2019; Welponer et al., 2022). The ground truth in these benchmarks is usually provided by point clouds or surface models obtained by using active sensors or a-priory generated synthetic models of the scene.
In photogrammetry and 3D vision, most of the benchmarks focus on satellite and terrestrial datasets (Welponer et al., 2022). Airborne benchmark data are historically less popular, although their number is progressively increasing: only a few of these benchmarks have been dedicated to UAV datasets (Nex et al., 2015; Lyu et al., 2020; Zhao et al., 2020). These UAV datasets are mostly collections of images for semantic segmentation or aim to assess the image orientation process with ground control points, while the quality of the 3D reconstruction is mainly qualitative. UAV datasets allow ultra-dense 3D reconstructions but their comparison with conventional airborne LiDAR data is normally insufficient to allow a thorough comparison (Nex et al., 2022), given the different point densities of these data.

The objective of this benchmark, UseGeo, is to bridge the aforementioned gaps, providing images and ground truth point clouds acquired by UAV platforms for the rigorous assessment of 3D reconstruction algorithms. The benchmark has been supported by ISPRS and aims to foster research on very high-resolution images and deep learning methods, giving a useful training set for both multi-view- and monocular 3D reconstruction algorithms.
Simultaneous acquisitions of images and LiDAR were performed in different urban and peri-urban areas. While LiDAR was acquired (primarily but not necessarily) as a reference (GT), the image data sources were used for the training and testing of MVS and monocular 3D reconstruction algorithms. Different typologies of landscapes have been considered in the acquisition to deliver relatively heterogeneous scenes. The size of the benchmark datasets has been defined in such a way as to allow the training and the testing of both MVS- and monocular algorithms, with a specific focus on deep learning approaches; in this regard, the benchmark has been set to deliver both point clouds and depth maps as ground truth. The data have already been validated using a few meaningful state-of-the-art algorithms to assess their usability for image registration, single- and MVS 3D reconstruction tasks.

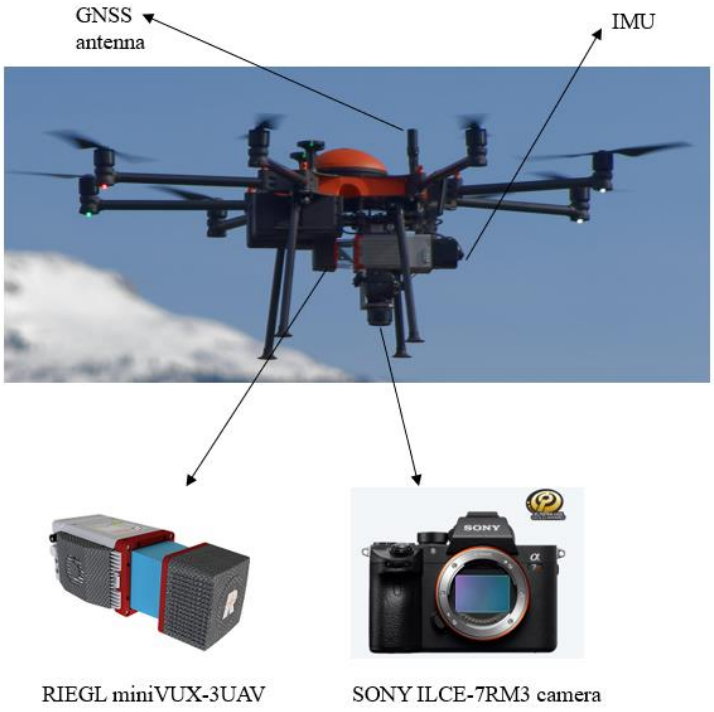
This project has been completed thanks to the collaboration of researchers from three well-known institutions: ITC Faculty (University of Twente, The Netherlands), 3DOM (Fondazione Bruno Kessler, FBK, Italy) and KIT (Kurlsruhe Institute of Technology, Germany). The PI Francesco Nex (University of Twente, The Netherlands since May 2015) has been supported by co-PIs in the development of the different parts of this work (ITC / University of Twente, Netherlands).

BENCHMARK DATA COLLECTION – COMPLETION OF THE WORK

In the following sections, the technical specifications of the used UAV platform and the description of the collected data will be given. The full benchmark includes some parts of the 3D Hesseinheim dataset as detailed in the following: for information on this dataset please refer to (Kölle et al., 2021).

The UAS platform

An Unmanned Aerial System (UAS) was used for data acquisition. The sensor setup encompasses a RIEGL miniVUX-3UAV scanner and a SONY ILCE-7RM3 camera. The average height above ground was 80m with a GSD of 2 cm. The average LiDAR point cloud spacing was 10 cm: in some areas, higher point density was achieved.



VUX 3 specifications

Frequency: 200 HZ

Up to 200,000 measurements/sec

Multitarget capacity

SONY ILCE-7RM3

Image: 7952 x 5304 pixels

Pixel size: 4.5 µm

Focal: 21 mm

Figure 1 – The used drone and technical specifications of the payload

The collected data

Using this setup, a total area of 1.1*650 meters was acquired during a campaign over the Italian territory (Sardinia region) in April 2021. Three flights of several strips were performed, acquiring a total of 930 images (7952*5304). In particular, 224, 328 and 277 images were then used for each sub-dataset. An average flight height of about 80 m was adopted during the flights, while 10 cm point cloud spacing resulted from the LiDAR acquisition. RTK-GNSS techniques have been used to determine the position of the UAV during the flights. Trajectories were corrected using riPRECISION as first pre-processing step.

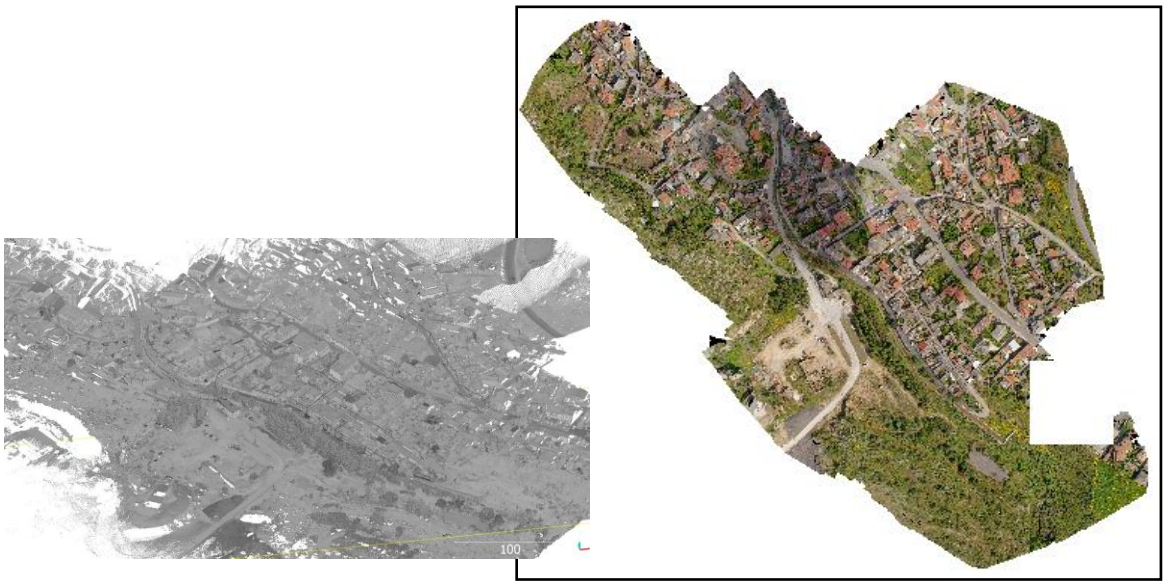


Figure 2 – Example of LiDAR point cloud, and the corresponding true orthophoto in correspondence of Area 1

Thanks to the collaboration with the group led by Prof. Norbert Haala our benchmark was further enriched using some of the data generated for the 3D Hesseinheim dataset. In particular 270 images of this benchmark were added for the Single Depth Estimation task. For further information on the 3D Hesseinheim dataset, please refer to: https://ifpwww.ifp.uni-stuttgart.de/benchmark/hessigheim/default.aspx

DATA PROCESSING

The collected data were processed to deliver reliable input for the research community. These data went through different processing steps in order to deliver some useful outputs to be directly used by the community. In the following sections, a more detailed description of each of these steps is provided. Please note that the data alignment has delivered the final georeferenced LiDAR point clouds and the final orientation of the image blocks, while the generated depth maps have been used as input for the single image depth estimation task. Additionally, we delivered the LiDAR point clouds as a ground truth for the assessment of stereo-matching algorithms.

Data alignment

The initial dataset was LiDAR strips (.rdbx) in Scanner Coordinate System (SCOS), camera images with their orientations from GNSS/IMU, and the initial trajectory of the drone. For the alignment of the two datasets, the LiDAR and camera datasets first needed to be in the same reference system with a rough alignment. The hybrid adjustment (Pfeifer et al., 2014) approach was used as an efficient approach to align camera and LiDAR datasets. The hybrid adjustment approach simultaneously optimizes the orientation of LiDAR and camera data for their optimal alignment with minimal errors. The camera images were pre-processed to obtain exterior image orientations, image point observations, and tie points for the hybrid adjustment. In the hybrid adjustment, correspondences were established and selected between image pairs (IMG-IMG), overlapping LiDAR strips (STR-STR), image tie points, and LiDAR strips (IMG-STR) with a modified ICP algorithm. The false correspondences and outliers were rejected based on several threshold criteria. The subsets of the points are selected from the correspondences with a uniform sampling technique to make the adjustment process computationally efficient, which selects the points from both the datasets in the object space as consistently as possible. The uniform sampling technique ensures that the uniform distribution of points in the correspondences and equal-area regions are weighted equally within the hybrid adjustment.

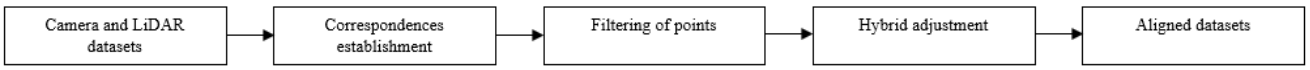


Figure 3: LiDAR and camera data alignment process

The main iteration loop in the hybrid adjustment starts with the direct georeferencing of the LiDAR strips with the initial parameters in the first loop and estimated parameters from the hybrid adjustment in the subsequent loops. The potential correspondences are matched, i.e., the nearest neighbour of a query point in the overlapping point cloud. The false correspondences are rejected and removed in the subsequent step based on roughness criteria, the distance between the corresponding points, and the threshold angle between the normals of the corresponding points. The correspondences that remained after the rejection step are weighted based on their surface roughness and angle between respective surface normals. It is worth mentioning that the correspondences were newly established in each iteration of hybrid adjustment. After a given number of iterations are completed, the LiDAR strips were georeferenced with the estimated parameters in the final iteration loop of the hybrid adjustment. As a final product of the hybrid adjustment, adjusted LiDAR strips, adjusted image orientations, and the adjusted trajectory of the UAS were obtained.
After the hybrid adjustment implementation, the camera images with adjusted orientations were processed in Pix4DMapper software. The primary quality analysis for camera alignment was carried out in CloudCompare software with the computation of the mean cloud-to-cloud (C2C) distances between LiDAR and camera point clouds. The mean C2C distances were chosen as lower C2C distances as a measure of better alignment of point clouds. The mean C2C distances after implementation are summarized in table 2.

Table 1: Mean C2C distances between LiDAR and camera point clouds after hybrid adjustment

Dataset	Mean C2C distances between LiDAR and camera point clouds after hybrid adjustment (cm)
Dataset_A	8.8
Dataset_B	8.5
Dataset_C	6.7

After implementing the hybrid adjustment, the alignment between LiDAR strips and camera data was obtained up to a cm level of accuracy without using any ground truth inputs in the form of Ground Control Points (GCPs) or Control Point Clouds (CPCs). Accordingly, the quality of results after hybrid adjustment is promising, especially if we expect to increase the accuracy with the use of the reference data in the adjustment when available.

Depth map generation

After the camera orientation step, for each image, the corresponding projection matrix P is obtained, encapsulating the extrinsic and intrinsic parameters. For this task, it was decided to use the undistorted imaged generated by Pix4Dmapper instead of the original images. Given this information along with the GT point cloud acquired with the laser scanner, we generated a GT depth map for each image by projecting the 3D points into the image plane. Occluded areas are taken into consideration and potential gaps are treated using nearest neighbour interpolation.

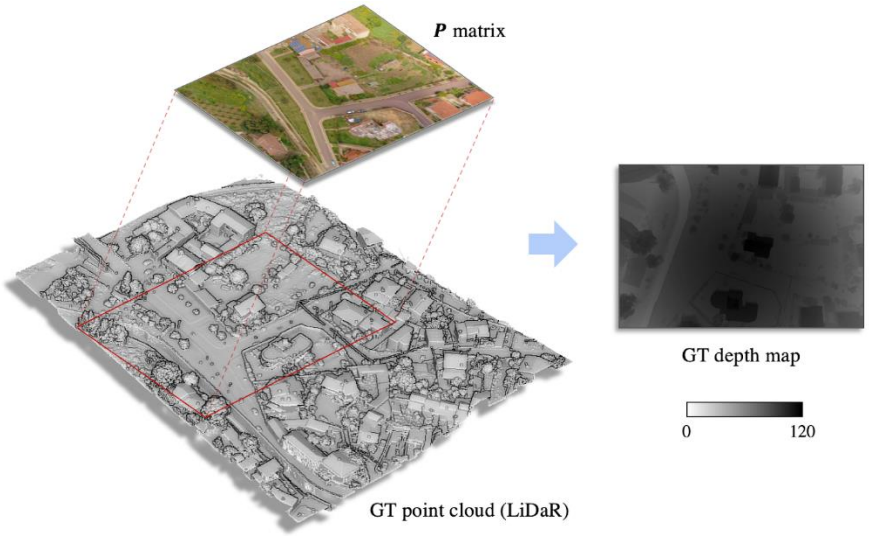


Figure 4. For each image of known orientation parameters, a GT depth map is generated by projecting the LiDaR 3D points

3D point clouds

As already mentioned, the LiDAR point clouds have been added to the downloadable material of the benchmark in order to allow their use for image matching algorithm assessment or any other further use from the scientific community.

EVALUATION CRITERIA AND PROCEDURES

The benchmark focus is on two main goals: a) **Single image depth estimation algorithm assessment**, b) **stereo matching 3D point cloud assessment**. In the following, we report the processing of few preliminary results achieved using a SIDE state of the art algorithm (Madhuanand et al., 2021). Details on the training strategy and the performed tests are reported too.

The input images for training the model (Madhuanand et al., 2021) are a combination of nadir images from Hessigheim and Zeche Zollern (Nex et al., 2015) datasets. The images from Zeche Zollern and Hessigheim datasets are of size 6132*8176 and 1989*1320, respectively, with an overlap of 80% with consecutive images. While both datasets comprise features like rooftops, vegetation, roads, and barren land, the Zeche Zollern dataset appears to be more of an urban setting than the Hessigheim dataset. From both datasets, a total number of 1036 images were used for training, 136 images for validation, and 88 images for testing.
The model (Madhuanand et al., 2021) was implemented using PyTorch framework (Paszke et al., 2017) and trained using resized input images of resolution 640×352 pixels. Learning rate was set to 10^{-5} and the Adam optimizer (Kigma et al., 2014) was considered. The training lasted 40 epochs with a batch size of 12. The weights between different loss terms are taken from Madhuanand et al., (2021). We used a single Nvidia Titan Xp GPU with 16 GB memory for the computation which took ca. 11 hours. The trained model is tested on the 88 images from the Hessigheim dataset. To assess the performance, various pixel-wise metrics are calculated between the model depths and reference depths as given in the Table.

Table 2: Preliminary results of the used SIDE algorithm (Madhuanand et al., 2021)

Method	Training dataset	Testing dataset	Abs Rel	Sq Rel	RMSE	δ1.25 (Higher is better)	δ1.15 (Higher is better)	δ1.05 (Higher is better)
Madhuanand et al., (2021)	Zeche Zollern + Heissinheim	Heissinheim	0.085	1.020	10.013	0.979	0.827	0.319

From the model depths it can be observed that the edges of buildings, roads, and other structures are reconstructed sufficiently in comparison to the trees and objects that are closer to ground level. Also, the objects near the margins are slightly distorted.

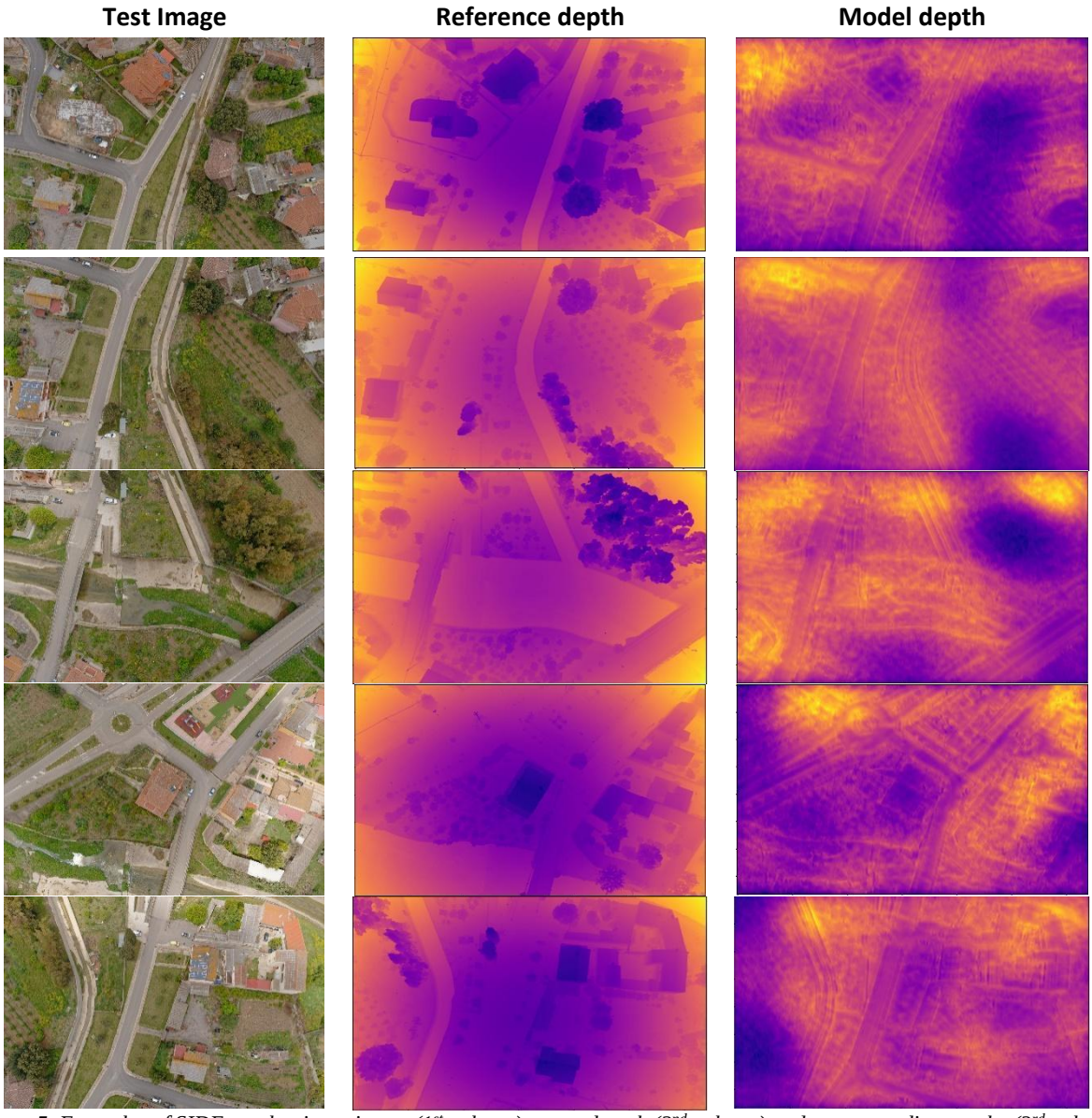


Figure 5. Examples of SIDE results: input image (1st column), ground truth (2nd column) and corresponding results (3rd column)

DATA DELIVERY

The dedicated webpage on the FBK's website has been implemented: (https://usegeo.fbk.eu/home). Any interested researcher can learn more about the dataset from this website. The data are freely available at the following link: https://drive.google.com/drive/u/0/folders/1csxyFxDDF3x3MwbmMdC1XwkeEPElKMb1
The folders are structured in an intuitive way. As the amount of data is very high, it was decided to downsample undistorted images and corresponding depth maps in the final release of the data. This makes the amount of data more manageable (within a few hours) for the download.

The Team of this SI will communicate with Markus Englich (ISPRS webmaster) to include the relevant information about this benchmark in the ISPRS website too, guaranteeing the best possible visibility to this initiative.

DISSEMINATION

The Scientific Initiative has been largely advertised during the ISPRS Congress 2022 in Nice with a dedicated presentation given by the PI Francesco Nex. It is still under development a scientific paper that will be submitted to the ISPRS Open Journal in the next months. This paper will give more details on the collected data, the pre-processing and the final delivery for the community.

References

Badrinarayanan, V., Kendall, A., and Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12):2481–2495, 2017.
Bleyer, M., Rhemann, C., and Rother, C. Patchmatch stereo-stereo matching with slanted support windows. In BMVC, volume 11, pages 1–11, 2011.
Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4):834–848, 2017.
Dai, Y., Zhu, Z., Rao, Z., and Li, B. Mvs2: Deep unsupervised multi-view stereo with multi-view symmetry. In 2019 International Conference on 3D Vision (3DV), pages 1–8. IEEE, 2019.
Eigen, D., Puhrsch, C., and Fergus, R. Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems, 27, 2014.
Eigen, D. and Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE international conference on computer vision, pages 2650–2658, 2015.
Faugeras, O. and Keriven, R. Complete dense stereovision using level set methods. In European conference on computer vision, pages 379–393. Springer, 1998.
Fu, H., Gong, M., Wang, C., Batmanghelich, K., and Tao, D. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2002–2011, 2018.
Furukawa, Y., Curless, B., Seitz, S. M., and Szeliski, R. Towards internet-scale multi-view stereo. In 2010 IEEE computer society conference on computer vision and pattern recognition, pages 1434–1441. IEEE, 2010.
Galliani, S., Lasinger, K., and Schindler, K. Massively parallel multiview stereopsis by surface normal diffusion. In Proceedings of the IEEE International Conference on Computer Vision, pages 873–881, 2015.
Gallup, D., Frahm, J.-M., Mordohai, P., Yang, Q., and Pollefeys, M. Real-time plane-sweeping stereo with multiple sweeping directions. In 2007 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2007.
Garg, R., Bg, V. K., Carneiro, G., and Reid, I. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In European conference on computer vision, pages 740–756. Springer, 2016.
Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 3354–3361. IEEE, 2012.
Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014.
Godard, C., Mac Aodha, O., and Brostow, G. J. Unsupervised monocular depth estimation with left-right consistency. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 270–279, 2017.
Guo, X., Yang, K., Yang, W., Wang, X., and Li, H. Group-wise correlation stereo network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3273–3282, 2019.
He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask R-CNN. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.

Hirschmuller, H. Stereo processing by semiglobal matching and mutual information. IEEE Transactions on pattern analysis and machine intelligence, 30(2):328–341, 2008.
Hosni, A., Bleyer, M., Rhemann, C., Gelautz, M., and Rother, C. Real-time local stereo matching using guided image filtering. In 2011 IEEE International Conference on Multimedia and Expo, pages 1–6. IEEE, 2011.
Huang, P.-H., Matzen, K., Kopf, J., Ahuja, N., and Huang, J.-B. Deepmvs: Learning multi-view stereopsis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2821–2830, 2018.
Hu, J., Ozay, M., Zhang, Y. and Okatani, T.. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. Proc. WACV, pp. 1043-1051, 2019.
Im, S., Jeon, H.-G., Lin, S., and Kweon, I.-S. Dpsnet: End-to-end deep plane sweep stereo. In 7th International Conference on Learning Representations, 2019.
Ji, M., Gall, J., Zheng, H., Liu, Y., and Fang, L. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In Proceedings of the IEEE International Conference on Computer Vision, pages 2307–2315, 2017.
Kar, A., Häne, C., and Malik, J. Learning a multi-view stereo machine. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 364–375, 2017.
Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., and Bry, A. End-to-end learning of geometry and context for deep stereo regression. In Proceedings of the IEEE International Conference on Computer Vision, pages 66–75, 2017.
Khot, T., Agrawal, S., Tulsiani, S., Mertz, C., Lucey, S., and Hebert, M. Learning unsupervised multi-view stereopsis via robust photometric consistency. arXiv preprint arXiv:1905.02706, 2019.
Kölle, M., Laupheimer, D., Schmohl, S., Haala, N., Rottensteiner, F., Wegner, J. D., and Ledoux, H. The hessigheim 3d (h3d) benchmark on semantic segmentation of high-resolution 3d point clouds and textured meshes from uav lidar and multi-view-stereo. ISPRS Open Journal of Photogrammetry and Remote Sensing, 1:100001, 2021.
Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25: 1097–1105, 2012.
Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., and Navab, N. Deeper depth prediction with fully convolutional residual networks. In 2016 Fourth international conference on 3D vision (3DV), pages 239–248. IEEE, 2016.
Lee, J. H., Han, M.-K., Ko, D. W., and Suh, I. H. From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326, 2019.
Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015.
Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., and Brox, T. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4040–4048, 2016.
Merrell, P., Akbarzadeh, A., Wang, L., Mordohai, P., Frahm, J.-M., Yang, R., Nistér, D., and Pollefeys, M. Real-time visibility-based fusion of depth maps. In 2007 IEEE 11th International Conference on Computer Vision, pages 1–8. IEEE, 2007.
Paschalidou, D., Ulusoy, O., Schmitt, C., Van Gool, L., and Geiger, A. Raynet: Learning volumetric 3d reconstruction with ray potentials. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3897–3906, 2018.
Saxena, A., Sun, M., and Ng, A. Y. Make3d: Learning 3d scene structure from a single still image. IEEE transactions on pattern analysis and machine intelligence, 31(5):824–840, 2008.
Scharstein, D. Matching images by comparing their gradient fields. In Proceedings of 12th International Conference on Pattern Recognition, volume 1, pages 572–575. IEEE, 1994.
Schönberger, J. L., Zheng, E., Frahm, J.-M., and Pollefeys, M. Pixelwise view selection for unstructured multi-view stereo. In European Conference on Computer Vision, pages 501–518. Springer, 2016.
Schöps, T., Schonberger, J. L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., and Geiger, A. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3260–3269, 2017.
Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. A comparison and evaluation of multi-view stereo reconstruction algorithms. In IEEE computer society conference on computer vision and pattern recognition, volume 1, pages 519–528. IEEE, 2006.
Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In European conference on computer vision, pages 746–760. Springer, 2012.
Strecha, C., Fransens, R., and Van Gool, L. Wide-baseline stereo from multiple views: a probabilistic account. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 1, pages I–I. IEEE, 2004.
Strecha, C., Fransens, R., and Van Gool, L. Combined depth and outlier estimation in multi-view stereo. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2, pages 2394–2401. IEEE, 2006.
Teed, Z. and Deng, J.. DeepV2D: Video to Depth with Differentiable Structure from Motion. In International Conference on Learning Representations, 2019.
Tosi, F., Aleotti, F., Poggi, M., and Mattoccia, S. Learning monocular depth estimation infusing traditional stereo knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9799–9809, 2019.
Watson, J., Firman, M., Brostow, G. J., and Turmukhambetov, D. Self-supervised monocular depth hints. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2162–2171, 2019.
Welponer, M., Stathopoulou, E.-K., and Remondino, F. Monocular depth prediction in photogrammetric applications. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 43:469–476, 2022.
Wenzel, K., Rothermel, M., Haala, N., and Fritsch, D. Sure–the ifp software for dense image matching. In Photogrammetric week, volume 13, pages 59–70, 2013.
Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., and Ricci, E. Structured attention guided convolutional neural fields for monocular depth estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3917–3925, 2018.
Xu, Q. and Tao, W. Learning inverse depth regression for multi-view stereo with correlation cost volume. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 12508–12515, 2020a.
Xu, Q., Oswald, M. R., Tao, W., Pollefeys, M., and Cui, Z. Non-local recurrent regularization networks for multi-view stereo. arXiv preprint arXiv:2110.06436, 2021b.

Xu, Q. and Tao, W. Multi-scale geometric consistency guided multi-view stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5483–5492, 2019.

Yang, J., Mao, W., Alvarez, J. M., and Liu, M. Cost volume pyramid based depth inference for multi-view stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4877–4886, 2020.

Yao, Y., Luo, Z., Li, S., Fang, T., and Quan, L. Mvsnet: Depth inference for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision (ECCV), pages 767–783, 2018.

Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., and Quan, L. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5525–5534, 2019.

Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., and Quan, L. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1790–1799, 2020.

Yin, W., Liu, Y., Shen, C., and Yan, Y. Enforcing geometric constraints of virtual normal for depth prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5684–5693, 2019.

Yin, W., Wang, X., Shen, C., Liu, Y., Tian, Z., Xu, S., Sun, C., and Renyin, D. Diversedepth: Affine-invariant depth prediction using diverse data. arXiv preprint arXiv:2002.00569, 2020.

Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., and Shen, C. Learning to recover 3d scene shape from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 204–213, 2021.

Zbontar, J. and LeCun, Y. Computing the stereo matching cost with a convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1592–1599, 2015.

Zhou, T., Brown, M., Snavely, N. and Lowe, D.G.. Unsupervised learning of depth and ego-motion from video. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1851-1858, 2017.