Article

Mobile-based 3D modeling: an in-depth evaluation for the application in indoor scenarios

Martin De Pellegrini [1,‡*], Lorenzo Orlandi [1,‡], Daniele Sevegnani[1,] and Nicola Conci [2,*]

[1] ARCODA s.r.l., (Trento, Italy)
[2] University of Trento, Italy
* Correspondence: martin.depellegrini@arcoda.it (M.DP.), nicola.conci@unitn.it (NC)
‡ These authors contributed equally to this work.

Abstract: Indoor environment modeling has become a relevant topic in several applications fields including Augmented, Virtual, and Extended Reality. With the digital transformation, many industries have investigated the possibility to generate detailed models of an indoor environment allowing the viewers to navigate through it, and mapping surfaces so as to insert virtual elements overlaid to the real scene. The scope of the paper is twofold. We first review the existing state-of-the-art (SoA) of learning-based methods for 3D scene reconstruction based on Structure From Motion (SFM) that predict depth maps and camera poses from a video stream. We then present an extensive evaluation using a recent SoA network, with particular attention to the capability of generalizing on new unseen data of indoor environments. The evaluation was conducted based using as a baseline metric the Absolute relative (AbsRel) measure on the depth map prediction.

Keywords: Computer Vision; 3D Reconstruction; Deep Learning; Indoor; Digital Twin; Point Cloud.

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Journal Not Specified* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

The ability of sensing the 3D space using single cameras is a widely investigated topic in image processing and computer vision. Several solutions have been developed over the years to ensure a reliable reconstruction of the observed environment, adopting both traditional image processing [1][2][3], as well as more up-to-date learning approaches [9][34]. In fact, 3D sensing and reconstruction is a necessary building block behind a number of technologies in industry, including robotics, landslide mapping, gaming, mixed reality, archaeology, medicine, to name a few [4][5][6]. Despite the many efforts spent by the research community in providing progressively more accurate models capable and sensing and reconstructing a 3D environment, a number of challenges remains still open. In fact, the acquisition of 3D information can serve multiple purposes, and can be used in real-time in a multi-sensorial context, as seen in robots or, more in general, autonomous systems. This often implies that the visual information is only one among the multiple inputs to a localization and navigation system. In such conditions, the potential errors emerging from inaccuracies and/or wrong reconstruction of portions of the environment are often compensated and mitigated thanks to the presence of additional sensing devices. Vice versa, in a more restrictive context, in which multi-modal equipment is not a viable option, 3D reconstruction is performed using the visual information solely, thus requiring high resolution images for better feature detection, and accurate camera calibration with distortion correction in order to generate the 3D model, consisting of a sparse or dense point cloud.

In this paper, we present an in-depth evaluation of a robust state-of-the-art method for depth estimation, which is used as the core element for 3D reconstruction applications. In particular, we focus our research on the indoor scenario, in which we expect the user to collect data using an arbitrary camera, and following subjective criteria. In other words, the acquisition is not conducted following a rigorous path in scanning the environment, thus not imposing any constraint on the user's side. Such conditions are indeed very common, and cover a wide spectrum of application scenarios, often involving on-the-field workers, which rely on such augmented/extended reality tools for inspection and maintenance operations.

The paper is structured as follows: in Section 2 we present some recent relevant related work; Section 3 discusses the motivation of this work and the main contributions; in Section 4 we focus on the validation pipeline we have envisaged, describing the methodology and the metrics used. In Section 5 the achieved results are presented and discussed. Final remarks and conclusions are drawn in Section 6.

2. Related Work

In the following paragraphs, we report the most relevant works presented in the SoA, starting from the traditional Structure from Motion algorithm and surveying the most recent developments based on deep-learning. Structure from Motion (SfM) [28] allows the estimation of the three-dimensional structure of objects and environments based on the motion parallax that describes the appearance changes of an object when the observer's viewpoint changes. By doing so, it is possible to infer the 3D structure of a target, and retrieve the distance from the camera to generate a 3D representation. Another basic principle of SfM is the stereo photogrammetry triangulation used to calculate the relative position of points from stereo pairs. More in general, SfM is required to solve three main problems. Firstly (i) it is required to find correspondences between the images and measure the distances between the features extracted with respect the two image planes. Typically, SIFT [32] features are used in this phase due to their robustness against change in scale, large variation of view point and challenging conditions such as different illumination and partial occlusions; as a second step, (ii) the camera position associated to each of the images processed is computed, via bundle adjustment (BA), to calculate and optimize 3D structure, camera pose and intrinsic calibration; lastly, (iii) generate a 3D dense point cloud by using the camera parameters to back project the points computed before on the 3D space, also called *multi view stereo matching*.

Traditional 3D reconstruction algorithms require to perform heavy operations and despite the proven effectiveness of these methods, they rely on high quality images as input. This may introduce some or limitations when it comes to process complex geometries, occlusions and low-texture areas. Such issues have been partially tackled replacing traditional feature and geometry-based approaches with deep learning. In particular, some stages of the traditional 3D reconstruction pipeline have been rethought following a deep learning-based formulation. To this aim, we present some of the methods explored for the purpose of our research, which implement the principles of SfM using Convolutional Neural Networks (CNNs). One of the most relevant works exploiting neural networks for depth estimation is DispNet [7]. DispNet is used for single view depth prediction. It is composed by an initial contracting stage, made of convolutional layers, followed by up sampling to perform deconvolutions, convolutions and computation of the loss function. Features from the contracting part are sent to the corresponding layer in the expanding portion. The network operates with a traditional encoder-decoder architecture with skip connections and multi-scale side prediction. The DispNet architecture is reported for convenience in Figure 1.

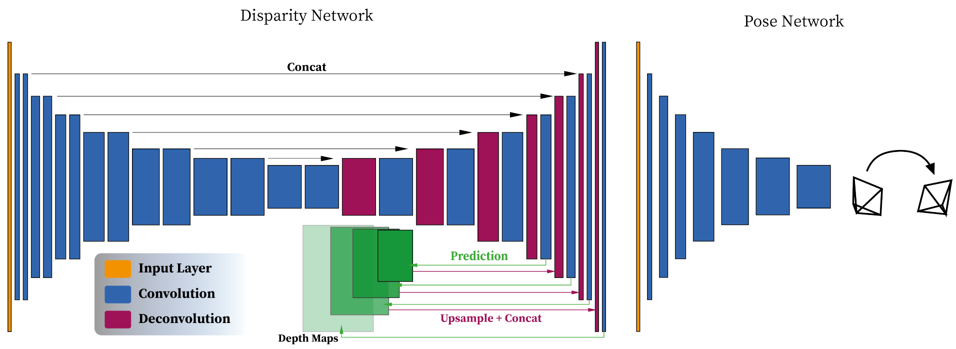

Figure 1. Illustration of the architecture of the Disparity estimation Network (DispNet) with encoder-decoder layout, and Pose estimation Network. Additional details in terms of the size of each layer can be found in the original manuscript.

Many solutions have been developed employing Convolutional Neural Networks (CNNs) for the task of estimating the depth information. Some of them are used for stereo view synthesis such as DeepStereo [8], which learns how to generate new views from single images in order to recreate a synthetic stereoscopic system where the underlying geometry is represented by quantized depth plane. Similarly, Deep3D [9] implements CNNs to convert 2D video into 3D sequences such as Anaglyph for 3D movie or Side-by-Side view for Virtual Reality (VR) applications. In this case the scene geometry is represented by probabilistic disparity maps. As well as Deep3D, other methods are following the recent research in learning three-dimensional structure from single view. Some of them introduced supervision signals such as in the work proposed by Garg *et al.* [33]. The authors propose a supervision consisting of a calibrated stereo twin for single-view depth estimation. The recent trends in depth estimation aim for unsupervised or self-supervised learning from video sequences. These methods work well in the task of inferring the scene geometry and ego-motion (similarly to SfM), but in addition they show great potential for other tasks such as segmentation, object motion mask prediction, tracking and other levels of semantics (please refer to [18][19][20][21][22][23]).

Among the unsupervised/self-supervised methods, three important researches have been conducted by Vijayanarasimhan *et al.* [34], Zhou *et al.* [10] and Bian *et al.* [11]. These approaches implement two sub networks: the first one focuses on single view depth prediction, and the second one is used for camera pose estimation in support to the depth network, so as to replicate a pseudo stereo vision setting. These implementations mostly differ on the loss function, which is applied as supervision signal. In terms of performances the methods achieve state-of-the-art scores on the KITTI [31] and Cityscapes [35] datasets.

Ref.	Method	Indoor	Dataset	Note
[34]	SfM Net	✗	KITTI[31] & Cityscapes[35]	O
[10]	SfM Learner	✗	KITTI[31] & Cityscapes[35]	O
[11]	SC-SfM Learner	✗	KITTI[31] & Cityscapes[35]	O
[12]	**Indoor SC-SfM Learner**	✓	NYUv2[27]	R

Table 1: Methods from literature for depth estimation from video sequences. In the column **Note** symbols (**O**) and (**R**) refer to Original and Rectified Training data.

3. Motivation and Contribution

Despite the proven effectiveness in street mapping contexts, the previous methods do not perform well when it comes to infer the 3D structure of indoor environments, and also by training the network with indoor RGB-D datasets, it does not allow to achieve satisfactory results, as also mentioned in [12]. Indeed, DispNet aims to learn the disparity between frames and due to the nature of hand-recorded sequences, typical of indoor data collection, the spatial relationship between adjacent frames might be of pure rotation, leading to a disparity equal to zero. More in detail, it has been demonstrated that the estimation of the depth map is strictly related to a dominance of translation with respect to rotations in the video sequences acquisition. In fact, previous implementations have been tested on datasets like KITTI [31], where the camera configuration and the forward motion did not give evidence to this issue. A research conducted by Bian *et al.* [12] has proven the existence of such limitation of the DispNet and proposes a weak rectification algorithm to pre-process indoor datasets before training the network. The authors have applied the rectification on the NUYv2 [27][25] dataset used to train the network and tested the generalization capability on the *7Scene* dataset [17]. Since the generalization was evaluated on one dataset only, we aim to provide additional benchmarks evaluating other RGB-D datasets and comment on the network generalization capability.

In summary the main contributions of the paper are:

- We provide additional benchmarks for the network proposed by Bian *et al.* in order to allow a better understanding on the network generalization performances.
- We analyze the network generalization capability in connection with the statistics of the scene, from which the depth has to be estimated. We compute the depth standard deviation from depth ground truth to describe the amount of depth information that the network has to estimate, and then discuss how the generalization is related to this parameter.

4. Materials and Methods

As anticipated in the previous section, the results and evaluation that are presented in the following paragraphs are based on the work by Bian *et al.* [11] [12]. Here, the network model is pre-trained on ImageNet [13] using ResNet-18 [14] in substitution to the depth and pose sub networks. Next, a fine-tuning on the rectified NYUv2 (Figure 3) [27][25] dataset is applied. Differently from the other architectures, the framework has been developed to overcome the scale ambiguity in [10], but it preserves the capability to test the depth and pose

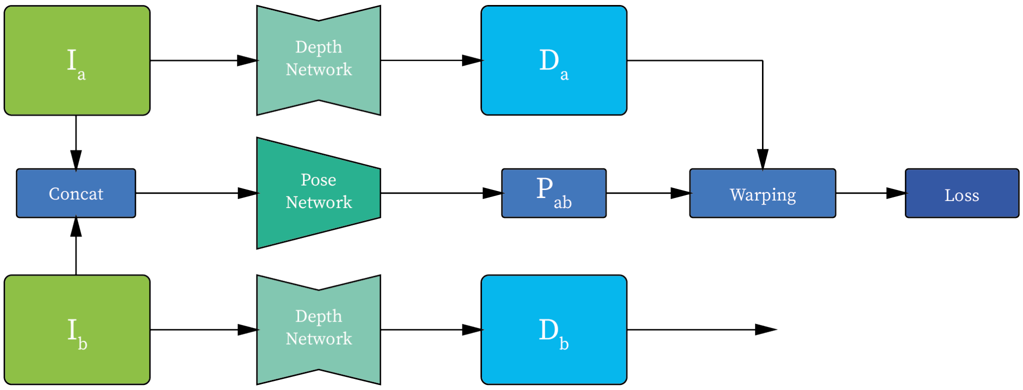

Figure 2. Illustration of the architecture used for the experiments, where I_a, I_b are the input RGB images, D_a, D_b the corresponding estimated depth maps, and P_{ab} is the relative camera position between I_a and I_b.

networks independently. We run our first tests on the depth map prediction on various RGB-D datasets of indoor environments (see Table 2) achieving results comparable to the ground truth (**GT**) except for a scale factor that can be calculated by normalizing the depth map with its median pixel value. The tests are conducted using the pre-trained model publicly available on the authors' GitHub repository [26]. We feed the unseen datasets in input to the model and retrieve the predicted disparity maps. For the evaluation, we adopt the Absolute Relative difference used in literature which is computed as follow:

$$\frac{1}{|V|}\sum_{p\in V}\frac{|d(p)-d^*(p)|}{d^*(p)} \tag{1}$$

where V denotes the set of valid depth pixels, $d(p)$ and $d^*(p)$ are the depth pixel value of the predicted depth map $\boldsymbol{D}$ and the depth ground truth $\boldsymbol{D}^*$, respectively. As mentioned before, the predictions are at different scale with respect to the ground truth. Scaling is then applied via the scaling factor s computed as follows, where $med\{\}$ refers to the median value:

$$s = \frac{med_{p\in V}\{\boldsymbol{D}^*\}}{med\{\boldsymbol{D}\}} \tag{2}$$

Note that, unlike the prediction, the ground truth exhibits some pixels equal to zero or one due to reflective surfaces or distances out of the sensor range. Such non-valid pixels are discarded in the computation above.

4.1. Dataset

The need of virtually reconstructing environments for autonomous navigation and/or extended reality applications has increased the availability of indoor RGB-D data to train more and more data-hungry networks; however, the amount of data is still limited to few common environments. In this section we present a brief overview of the datasets used in our experiments. We tested the network performance on four different datasets containing sequences from several indoor environments. In particular, for the testing purposes we selected the sequences *freiburg_360* and *freiburg_pioneer* from RGB-D TUM Dataset [24], all the sequences from RGB-D 7 Scene [17], the RGB-D Scene dataset from Washington RGB-D

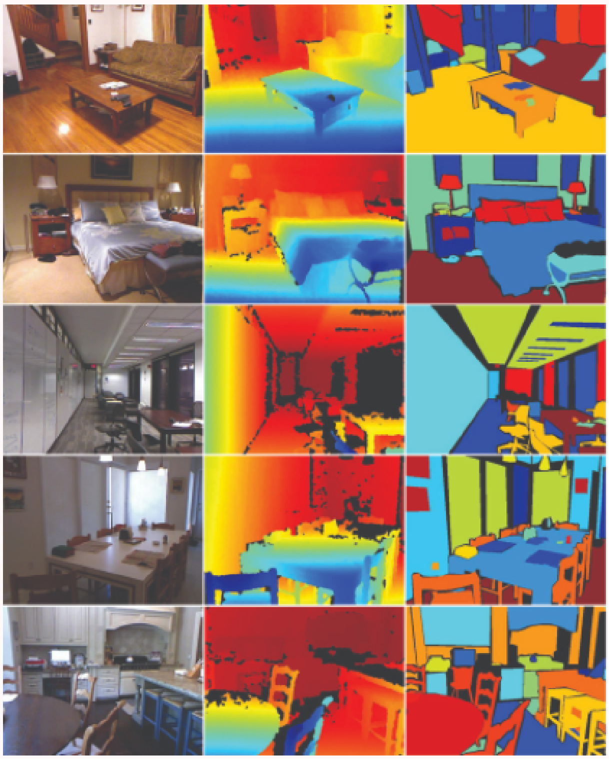

Figure 3. NYU dataset [25]

Object Dataset [36] and the SUN RGB-D Dataset [37]. Details about the number of samples and resolution are reported in Table 2.

- **RGB-D TUM Dataset**: the sequence *freiburg1_360* contains a 360 degree acquisition in a typical office environment; the *freiburg_pioneer* sequence shows a quite open indoor environment captured by a robot with depth sensor attached on top of it (Figure 4). The dataset is provided with depth ground truth acquired by the Kinect sensor, and camera pose ground truth as rotation and translation are acquired with an external motion capture system, it is typically used for SLAM systems. For additional details we refer to the dataset website[15] and to the original paper[24]. Among the available sequences we decided to choose two of them (*freiburg1_360* and *freiburg_pioneer*) since they represent distinct environments with interesting characteristics useful to test the generalization of the network. In particular, in *freiburg_360* there are many complex geometry defined by the office furniture; *freiburg_pioneer* is instead characterized by wide spaces, usually implying more homogeneous depth maps but larger depth range.

Figure 4. RGB-D TUM Dataset, frame taken from the two sequences.

- **RGB-D Microsoft Dataset**: this dataset [17] consists in sequences of tracked RGB-D frames of various indoor environments, and it is provided with the corresponding depth ground truth (Figure 5). This dataset is the one used by the authors in [12] to test the generalization capability of the network. Accordingly, we decided to re-run the tests as well, to ensure the replicability of the paper results.

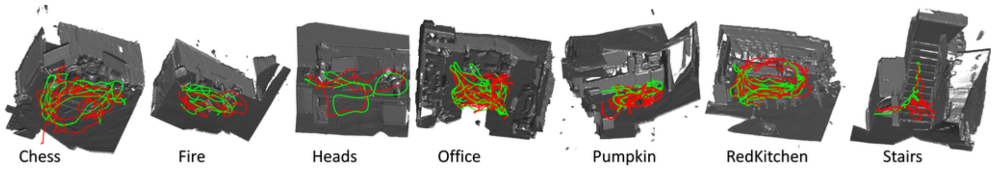

Figure 5. 7 Scene dataset [17]

- **Washington RGB-D Object Dataset**: the dataset [36] was created with the purpose of providing structure data of real objects. Aside the isolated objects, the dataset provides 22 annotated sequences of various indoor environment with depth ground truth. Also in this case, RGB-D data are collected using Micorsoft Kinect using aligned 640x480 RGB and depth images (Figure 6).

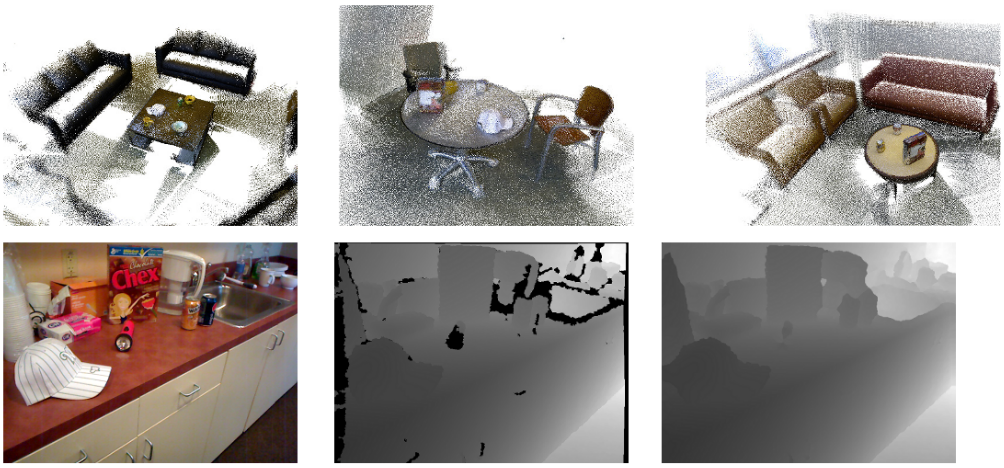

Figure 6. Washington RGB-D Object Dataset [36]

- **SUN RGB-D Dataset**: the dataset [37] is a collection of several common indoor environments from different datasets; it contains RGB-D images from NYUv2 [27], Berkeley B3DO [38] and SUN3D [39]. The dataset has in total 10335 RGB-D images. In order to make the experiments comparable, we have selected only the samples acquired using Kinect (Figure 7).

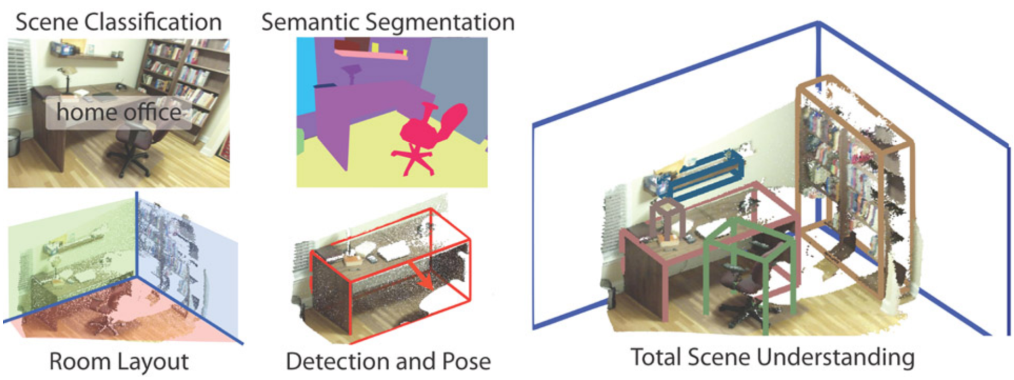

Figure 7. SUN RGB-D Dataset [37]

As reported above, in all selected datasets, the RGB-D data is acquired with Microsoft Kinect version 1. The device is equipped with an RGB camera and a structured light sensor working on the near infrared light spectrum, where a known infrared pattern is projected onto the scene and the depth is computed after distortion correction. For additional information about the sensor and the related performances, please refer to the study by Wasenmüller *et al.* [40]. In terms of accuracy, the sensor exhibits an exponentially increasing offset going from 10mm at 0.5m, of up to 40mm at distance of 1.8m. Although the performances of the sensor are not as accurate as other more recent devices made available on the market, most benchmark datasets in the literature still have the Kinect depth map as ground truth.

Name	#Images	Img. Size	Ref.
freiburg_360 (TUM RGB-D)	756	640x480	[24]
freiburg_pioneer (TUM RGB-D)	1225	640x480	[24]
7Scene	29000	640x480	[17]
Washington	11440	640x480	[36]
SUN	10335	640x480	[37]

Table 2: Details of the three dataset used in the testing phase.

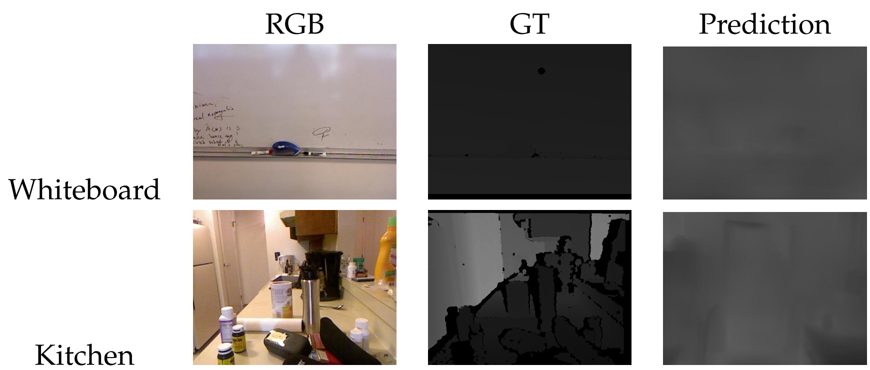

Figure 8. Example of depth map prediction with different depth standard deviation.

5. Results

In this section we present the results we obtained in our simulations. Since the author of [12] already compared the network performances with previous state-of-the-art unsupervised methods, and in particular with [11] and [41] showing an improvement in terms of Absolute Relative error after training data rectification, we focus on enriching the benchmark by testing the network on different unseen data. We evaluate the datasets described in the previous section by feeding frame sequences to the network and computing the Absolute Relative difference (AbsRel) for each prediction-ground truth pair every **5** frames. The results are reported in Table 3. We notice that the network generalization performance highly depends on the images depth range that has to be estimated. As an example, environments containing various structural features are more likely to result in a higher error, and frames depicting an homogeneous scenario with lower depth variation result in a lower error.

Scenes	AbsRel	StdDev (σ^2)
freiburg_360 (TUM RGB-D) [24]	0.16	5056.86
freiburg_pioneer (TUM RGB-D)[24]	**0.28**	**11370.31**
Chess (7Scene) [17]	0.19	5800.00
Fire (7Scene) [17]	**0.15**	**4418.00**
Office (7Scene) [17]	0.16	4438.00
Pumpkin (7Scene) [17]	**0.13**	**3435.00**
RedKitchen (7Scene) [17]	0.20	5700.00
Stairs (7Scene) [17]	0.17	5341.00
Washington [36]	**0.30**	**9656.00**
B3DO (SUN RGB-D) [37]	0.18	6886.21

Table 3: Single-view depth estimation results on selected Datasets

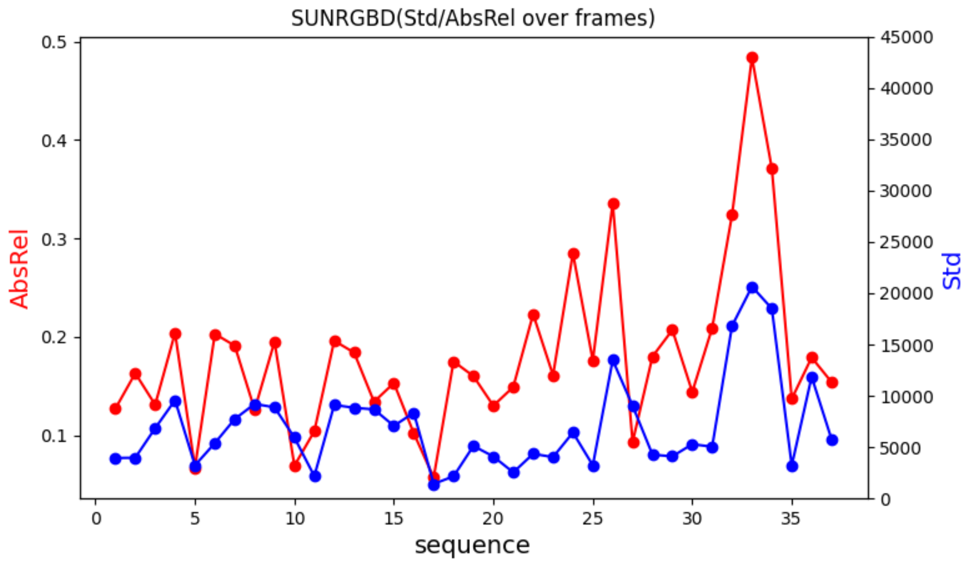

Figure 9. The plot shows the behaviour of Absolute Relative error (AbsRel) and Depth Standard Deviation (Std) over the *B3DO* sequence from SUN RGB-D Dataset.

In addition to the Absolute Relative error, we then analysed the Standard Deviation (σ^2) of depth ground truth images, which gives an insight of how challenging an environment is from the learning perspective. The depth standard deviation shows great potential in understanding the overall structure of the environment, thus it can be employed in further improvements of the network depth prediction. As for the AbsRel, the tests were performed computing σ^2 along with the error for each frame pairs every 5 frames. Figure 8 shows an example of borderline situations taken from SUN RGB-D [37], where in the case of the whiteboard, the measured AbsRel is particularly low, equal to 0.05 and $\sigma^2 = 1416.48$, on the other hand, in the kitchen image the depth range is larger with $\sigma^2 = 20639.78$, and the resulting absolute error is equal to 0.48. By comparing the two examples we can see that frames with a smaller σ^2 consist of relatively simple tasks that the network can easily manage; at the same time they often turn to be *false positives*. This situation is frequent because of the required normalization procedure, which is applied to the predicted depth in order to compare it with the GT. Indeed, for homogeneous surfaces that appear to be orthogonal to the optical axis, the predicted depth map results in an almost flat gray level image, leading, after the normalization, to an apparently optimal prediction, no matter if the scale is consistent or not along the entire sequence. On the other hand, the higher the variation in the depth range, the harder is for the network to predict consistent disparity maps. This behaviour is shown in the plot reported in Figure 9, where the test are conducted on the *B3DO* sequence from SUN RGB-D. Unlike the other sequences, *B3DO* is composed by random frames from different environment, thus it is a good challenge for the generalization capability of the network.. As next step we performed the same test on the remaining (Table 2) to find the contexts in which the network works well and in which ones it is harder for the network to predict the disparity. Figure 10 presents the Absolute Relative error for each considered sequences in relation to the depth Standard Deviation both computed as the mean over the entire sequence. It is arguable from the plot that the Absolute Relative error is directly proportional to the amount of depth information (given by the standard deviation) that the network has to estimate. More precisely, it is noticeable that for datasets such as **7Scene**, **SUN RGB-D** and the sequence *freiburg_360*, where the space is limited and so the overall depth standard deviation, the network tends to remain consistent and more accurate in the prediction, resulting in a lower absolute error. On the other hand, the prediction accuracy decreases when it comes to process wider and more complex environments as the ones belonging to the **Washington** dataset and the sequence *freiburg_pioneer*, and this is due to the higher variation in the environmental depth as it can be seen in Figure 10.

6. Conclusions

The goal of our paper was to test the generalization performance of the architecture proposed in [11], providing additional benchmark evaluations. The evaluation has been conducted using the Absolute Relative error as a standard metric. In addition we aim at providing the reader with some hints to interpret the reasons behind some of the results we achieved, so as to draw more detailed conclusions. We noticed that the network ability to estimate the structure of an indoor environment is related to the amount of information that has to be learnt, as it can be evinced from the plots reported above. In particular, the data from Washington Dataset shows the worst results and this is mostly due to the larger standard deviation on the depth range. We understand that this parameter can be considered as a valuable parameter to describe the network generalization capability on various environment. According to our experience, we believe the employment of the depth standard deviation as a weighting parameter in the learning stage is a useful parameter to better stimulate the network in predicting consistent disparity maps from large and more complex indoor environments.

7. Future Works

We tried to extend the evaluation of DispNet in a diversified set of scenarios, with the purpose of testing the depth extraction accuracy in monocular video, using (SoA) CNN. It is needless to say how such an approach can be revolutionary when deployed in real and unconstrained scenarios, and can be proved to be valuable for the companies engaged in the collection of digital twin, as well as

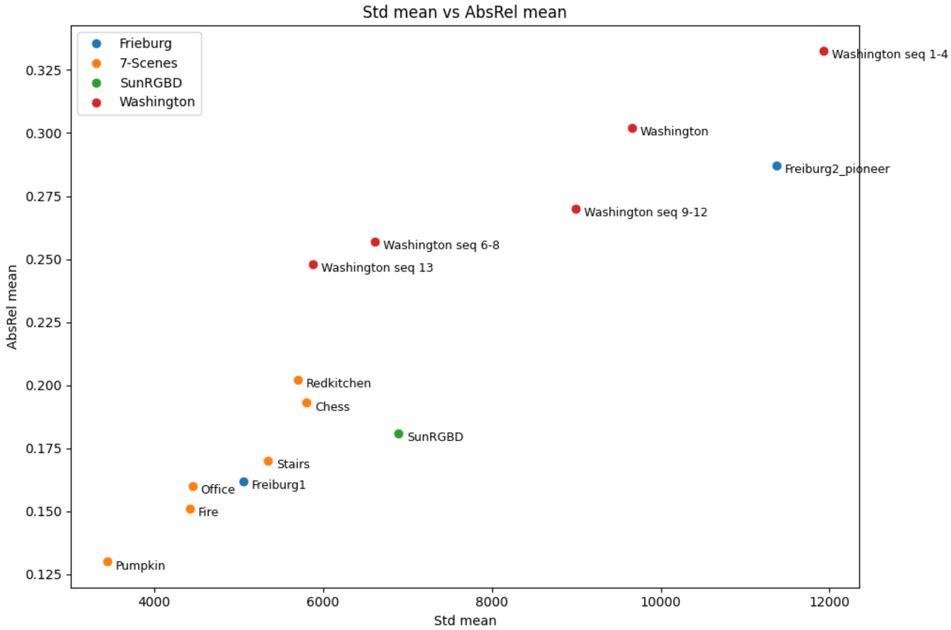

Figure 10. Mean Standard Deviation σ^2 vs. Mean Absolute Relative error of all datasets.

for the ones involved in Mixed and Augmented reality developments. Our aim and recommendation for future studies include:

- the adoption of other SoA architectures for richer comparisons;
- the adoption of a novel metric that considers the depth standard deviation for performance evaluation and in the training stage;
- the extension of the study to additional datasets, where the ground truth is collected with more up-to-date and accurate depth sensors.

Abbreviations

The following abbreviations are used in this manuscript:

SoA	State-of-the-art
SfM	Structure from Motion
SIFT	Scale Invariant Feature Transform
BA	Bundle Adjustemnt
CNN	Convolutional Neural Network
DispNet	Disparity Network
RGB	Red, Green, Blue
RGB-D	Red, Green, Blue and Depth
GT	Ground Truth
AbsRel	Absolute Relative error
StdDev	Standard Deviation

References

1. Fazakas, Tamas, and Róbert T. Fekete. "3D reconstruction system for autonomous robot navigation." 2010 11th International Symposium on Computational Intelligence and Informatics (CINTI). IEEE, 2010.
2. Gupta, Sharad Kumar, and Dericks P. Shukla. "Application of drone for landslide mapping, dimension estimation and its 3D reconstruction." Journal of the Indian Society of Remote Sensing 46.6 (2018): 903-914.

3. Alexiadis, D. S., D. Zarpalas, and P. Daras. "Real-time, realistic, full 3-D reconstruction of moving humans from multiple Kinect streams." IEEE Trans. in Multimedia (2013).
4. Khilar, Rashmita, S. Chitrakala, and SurenderNath SelvamParvathy. "3D image reconstruction: Techniques, applications and challenges." 2013 International Conference on Optical Imaging Sensor and Security (ICOSS). IEEE, 2013.
5. Hosseinian, S., and H. Arefi. "3D RECONSTRUCTION FROM MULTI-VIEW MEDICAL X-RAY IMAGES–REVIEW AND EVALUATION OF EXISTING METHODS." International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences 40 (2015).
6. Bresnan, Joan, and Sam A. Mchombo. "The lexical integrity principle: Evidence from Bantu." Natural Language & Linguistic Theory 13.2 (1995): 181-254.
7. Mayer, Nikolaus, et al. "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. 2016.
8. Flynn, John, et al. "Deepstereo: learning to predict new views from real world imagery." U.S. Patent No. 9,916,679. 13 Mar. 2018.
9. Xie, Junyuan, Ross Girshick, and Ali Farhadi. "Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks." European Conference on Computer Vision. Springer, Cham, 2016.
10. Zhou, Tinghui, et al. "Unsupervised learning of depth and ego-motion from video." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
11. Bian, Jia-Wang, et al. "Unsupervised scale-consistent depth and ego-motion learning from monocular video." arXiv preprint arXiv:1908.10553 (2019).
12. Bian, Jia-Wang, et al. "Unsupervised depth learning in challenging indoor video: Weak rectification to rescue." arXiv preprint arXiv:2006.02708 (2020).
13. Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.
14. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
15. Computer Vision Group TUM Department of Informatics Technical University of Munich, RGB-D SLAM Dataset https://vision.in.tum.de/data/datasets/rgbd-dataset/download(accessed on May 2021)
16. RGB-D Dataset 7-Scene, https://www.microsoft.com/en-us/research/project/rgb-d-dataset-7-scenes/(accessed on April 2021)
17. Glocker, Ben, et al. "Real-time RGB-D camera relocalization." 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, 2013.
18. Agrawal, Pulkit, Joao Carreira, and Jitendra Malik. "Learning to see by moving." Proceedings of the IEEE international conference on computer vision. 2015.
19. Jayaraman, Dinesh, and Kristen Grauman. "Learning image representations equivariant to ego-motion." Proc. ICCV. 2015.
20. Goroshin, Ross, et al. "Unsupervised learning of spatiotemporally coherent metrics." Proceedings of the IEEE international conference on computer vision. 2015.
21. Misra, Ishan, C. Lawrence Zitnick, and Martial Hebert. "Shuffle and learn: unsupervised learning using temporal order verification." European Conference on Computer Vision. Springer, Cham, 2016.
22. Pathak, Deepak, et al. "Learning features by watching objects move." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
23. Wang, Xiaolong, and Abhinav Gupta. "Unsupervised learning of visual representations using videos." Proceedings of the IEEE international conference on computer vision. 2015.
24. Sturm, Jürgen, et al. "A benchmark for the evaluation of RGB-D SLAM systems." 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2012.
25. NYU depth datadet version 2. https://cs.nyu.edu/ silberman/datasets/nyu_depth_v2.html (accessed on May 2021)
26. https://github.com/JiawangBian/Unsupervised-Indoor-Depth (accessed on May 2021)
27. Silberman, Nathan, et al. "Indoor segmentation and support inference from rgbd images." European conference on computer vision. Springer, Berlin, Heidelberg, 2012.
28. Szeliski, Richard. Computer vision: algorithms and applications. Springer Science & Business Media, 2010.
29. Wu, Changchang. "Towards linear-time incremental structure from motion." 2013 International Conference on 3D Vision-3DV 2013. IEEE, 2013.
30. Wu, Changchang. "VisualSFM: A visual structure from motion system." (2011).
31. Geiger, Andreas, et al. "Vision meets robotics: The kitti dataset." The International Journal of Robotics Research 32.11 (2013): 1231-1237.
32. Lowe, David G. "Object recognition from local scale-invariant features." Proceedings of the seventh IEEE international conference on computer vision. Vol. 2. Ieee, 1999.
33. Garg, Ravi, et al. "Unsupervised cnn for single view depth estimation: Geometry to the rescue." European conference on computer vision. Springer, Cham, 2016.
34. Vijayanarasimhan, Sudheendra, et al. "Sfm-net: Learning of structure and motion from video." arXiv preprint arXiv:1704.07804 (2017).
35. Cordts, Marius, et al. "The cityscapes dataset." CVPR Workshop on the Future of Datasets in Vision. Vol. 2. 2015.

36. Lai, Kevin, et al. "A large-scale hierarchical multi-view rgb-d object dataset." 2011 IEEE international conference on robotics and automation. IEEE, 2011.
37. Song, Shuran, Samuel P. Lichtenberg, and Jianxiong Xiao. "Sun rgb-d: A rgb-d scene understanding benchmark suite." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
38. Janoch, Allison, et al. "A category-level 3d object dataset: Putting the kinect to work." Consumer depth cameras for computer vision. Springer, London, 2013. 141-165.
39. Xiao, Jianxiong, Andrew Owens, and Antonio Torralba. "Sun3d: A database of big spaces reconstructed using sfm and object labels." Proceedings of the IEEE international conference on computer vision. 2013.
40. Wasenmüller, Oliver, and Didier Stricker. "Comparison of kinect v1 and v2 depth images in terms of accuracy and precision." Asian Conference on Computer Vision. Springer, Cham, 2016.
41. Cl´ement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. "Digging into self-supervised monocular depth estimation." In International Conference on Computer Vision (ICCV), pages 3828–3838, 2019.