QUANTITATIVE COMPARISON BETWEEN NEURAL NETWORK- AND SGM-BASED STEREO MATCHING

A. Frenzel[1], N. Deckers[1], R. Reulke[1]*

[1] Computer Vision, Humboldt-Universität zu Berlin, Germany - reulke@informatik.hu-berlin.de

Commission I, WG I/10

KEY WORDS: Stereo Matching, Deep Learning, ZED Camera

ABSTRACT:

Over the last decades, various methods for three-dimensional detection of the environment have been developed and successfully used. This work considers classical stereo methods, which can determine depth information by the means of correspondence analysis on the basis of two pictures of a scene. Recently, neural networks have been used to solve correspondence analysis. These procedures came first places on corresponding benchmarks and are ahead of many already established solutions. In this work, images captured by the ZED camera are evaluated for accuracy of the depth maps generated by several approaches. This includes modern methods based on neural networks.

1. INTRODUCTION AND MOTIVATION

For the 3D capture of scenes and their semantic interpretation, there are at the moment a number of systems for reconstruction. A standard procedure relies on the use of stereo sensors.

The central point is the correspondence analysis (stereo matching) between the left and right images. Classical methods, such as pixel correlation, result in the use of environmental information to smear edges. Global procedures could help, but computationally expensive. Therefore, the approach of (Hirschmuller, 2005) with the semi-global matching introduced a paradigm shift. This algorithm provides dense disparity maps in image resolution with sharp edges and near real-time. Lately, new algorithms based on neural networks for solving this task have become available. In this paper the different approaches are compared by using real image data. As a reference system, the stereo camera ZED created by stereolabs was used.

2. RELATED WORK

Stereo matching and disparity computation is a way of reconstructing depth from at least two images, captured from two different view points or viewing angles. In photogrammetry the desire to measure depth has evolved ever since the invention of photography and cameras. Early photogrammetric models compute the disparity on a point pair basis.

With the rising available computing power, researchers wanted to compute dense disparity maps from two cameras. This introduces many problems. Since, in the best case, the stereo reconstruction is working unsupervised, distinctive pixel similarity measures have become extremely important, for instance (Birchfield, Tomasi, 1998). However there were still a lot of problems regarding the different view points, for instance different illumination angles, occlusion, spot lights, smooth and texture-free surfaces, etc.

(Scharstein, Szeliski, 2002) have introduced a database of stereo scenarios, called the Middlebury database. They also compared different methods for their accuracy and density .

One of the first really successful methods was presented by Hirschmüller (Hirschmuller, 2005). It works locally as well as semi-globally on the image by sampling from the pixels to different directions and optimising the matching in the neighbourhood of pixels. Researches ever since have worked on different aspects of the stereo matching problem. For instance (Hermann, Vaudrey, 2010) improved the pixel similarity. These methods use classical computer vision methods.

Ever since the rise and success of deep neuronal networks, publications also tried to make use of the computational power for the stereo matching problem. (Pal et al., 2012) used a learning approach utilising conditional random fields. One of the first results on the imagenet database (e.g. (Krizhevsky et al., 2012)) lead to the increased research interest and specialised databases like the KITTI vision benchmark suite of (Geiger et al., 2012). Also the dense stereo problem became more and more interesting for autonomous driving applications to foresee the structure of the street, the vehicles in front of the car and other obstacles like pedestrians.

The concept of the *s*cene flow can be interpreted as a dense stereo matching problem. In the context of autonomous driving, the scene flow and vehicles were estimated jointly by (Menze et al., 2015). Optical flow and hence scene flow can also be estimated by deep neuronal networks, as in (Dosovitskiy et al., 2015). This was implemented using a contracting and an expanding part in the design of the network layers.

Optimisations in the learning speed and better convergence lead to the development of even deeper networks, as in (He et al., 2016).

A data set for scene and optical flow computation comparison was provided by (Mayer et al., 2016). It makes use of synthetic data and enables training of more complex networks as the FlowNet, since the DNNs require a lot of stereo data. Improvements on deep stereo matching were made by (Zbontar et al., 2016) and by (Kendall et al., 2017), the latter by providing a learning method to include estimating the whole process of geometry and context.

*Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-1-W2-17-2019 | © Authors 2019. CC BY 4.0 License.

Table 1. Supported resolutions and refresh rates of the ZED camera.

Video mode	max. frames per second	image resolution
2.2K	15	2208×1242
1080p	30	1920×1080
720p	60	1280×720
WVGA	100	672×376

The focus of this article is the evaluation of the dense and deep stereo matching methods presented in (Dosovitskiy et al., 2015, Mayer et al., 2016, Kendall et al., 2017). These networks are called DispNet, DispNetC and GCNet.

3. PRE-PROCESSING AND STEREO SENSOR SDK

This article assumes a calibrated stereo camera system. The model description is based on the pin hole camera model, the distortion is described according to the Brown model (Duane, 1971). The object is located in the stereo area and is completely imaged by both cameras (hardly any occlusions). Before matching, the image data of both cameras are rectified into an epipolar geometry.

By introducing the ZED camera, the company stereolabs launched a stereo camera on the market in 2015 (Stereolabs, 2015). There are two identical (synchronised) RGB cameras with a 1/3 inch sensor, each pixel is square and has the size of 2μm. The aspect ratio of the sensor is $16 : 9$. The optics are specified with a field of view of 110 degrees (F-Number = 2). The camera supports various resolutions and refresh rates, which are listed in Table 1. The maximum possible resolution of a camera is 2208×1242; the minimum resolution 672×376. The image repetition rate is expected to be dependent on the resolution. Both cameras are arranged in parallel (stereo normal case) and stand at a distance of about 12cm to each other. There is a SDK available, which requires a CUDA capable graphics card. In addition to camera control, the SDK provides the following functions:

- Stereo Capture
- Depth Perception
- Positional Tracking and
- Spatial Mapping.

The stereo-matching and the Depth Perception methods can be used both in- and outdoor in a range of $0.5m$ to $20m$. The results are created in real time. The disparity map has the same resolution as the captured image. The quality and frequency of the output disparity maps can be controlled by using three predefined modes (Performance, Medium, Quality).

4. MATCHING APPROACHES

We limited ourselves to various derivations of semi-global (block) matching and to methods with neural networks (Hirschmuller, 2005). The choice fell on the networks DispNetC and DispNet, which were in 16th place on the KITTI benchmark at this time (Geiger et al., 2012). The decisive criterion was the availability of the source code and a ready-made model for a comparison.

4.1 Semi-Global Matching

This method minimises a global cost function. In order to reduce the computational effort, different paths leading from the edge of the image to each pixel are considered and the required costs for each step are determined by the sum of two terms. The first term describes the direct matching cost, while the second term penalises different disparities of the neighbour pixel in that direction. In doing so, it differentiates between the same or similar disparities and disparity jumps, whereby the path with the lowest cost is selected from this set. In this work we consider the implementation provided by OpenCV with 5 or 8 paths (Itseez, 2015).

4.2 DispNetC

The neural network DispNetC is a network which is trained end-to-end and consists of two parts: A Contracting Part and an Expanding Part (Dosovitskiy et al., 2015, Mayer et al., 2016). An integral part of the Expanding Part is the Correlation Layer introduced by the authors. This merges the feature vectors of both images and passes the result to the Expanding Part. As input, the network receives a rectified and normalised stereo image pair. The output is the disparity map.

4.3 DispNet

DispNet is a simplified form of DispNetC. Instead of an individual treatment of the right and left image, the two images are superimposed and edited together by the network. Compared to DispNetC, the correlation layer and the independent processing of the stereo images are abandoned. The number of convolutional layers or convolutional transposed layers in the contracting and expanding part will remain identical. As a result of this change the resulting network no longer specialises in the task of correspondence analysis. Since it does not receive any specifications, the network must learn independently how to extract features from the images, identify correspondences and generate a disparity map from them.

4.4 Implementation and training of the networks

The original authors have implemented the two networks in a modified version of the framework Caffe (Jia et al., 2014) and trained them with the SceneFlow data set FlyingThings3D (Mayer et al., 2016). Note that we did not add own data.

5. METHOD FOR COMPARISON OF STEREO SYSTEMS

In order to test the methods for typical problems of correspondence analysis, scenarios for the determination of depth estimation, interpretation of homogeneous surfaces, edge transitions and round surfaces have been created. With the ZED camera, stereo images were taken of these scenes, which were then examined using the described methods.

5.1 Experimental setup

The ZED camera equipped with a laser rangefinder was mounted to a tripod and aligned with a wall (see Fig. 1a). To make the measurements comparable, a rail was attached to the floor orthogonal to the wall. Along this trail, the tripod and the ZED camera were led along, in order to allow recordings from different distances of up to 4m. The schematic structure is shown in Fig. 1b.

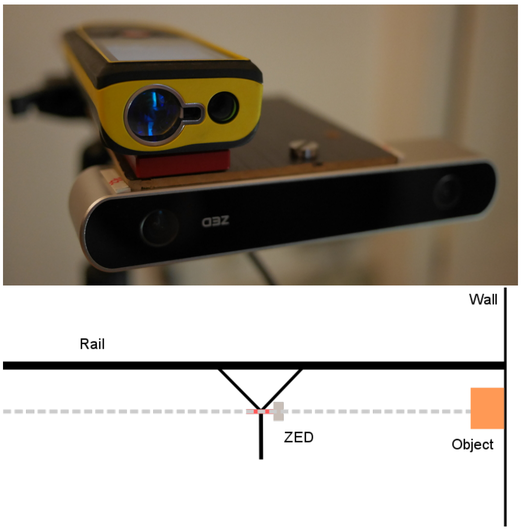

Figure 1. Experimental setup: (Upper image) ZED camera and laser distance meter. (Lower image) The ZED camera is guided along a rail perpendicular to the wall. The camera is aligned with the wall or an object. Images of the scene are taken in equidistant steps

Depending on the scene, an additional object was placed in front of the wall. This then counted as a reference point for the distance measurement and reduced the maximum possible distance. Image frames were taken at intervals of $0.5m$ starting at $1.0m$, thus, depending on the scene 6 to 7 images were taken. For the evaluation the disparity maps were generated with all procedures using the rectified images of the ZED camera and then the associated point clouds were generated.

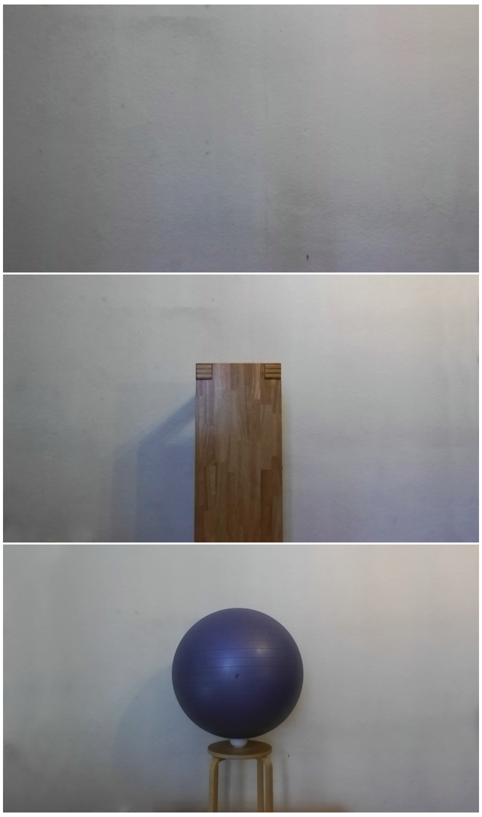

Figure 2. Images of the scenes from the perspective of the left camera. (a) white wall, (b) edge and (c) ball scene.

10 disparity maps and point clouds were evaluated for three selected scenes each (Fig. 2):

- 6 for the processes of stereolabs,
- one for each for DispNet and DispNetC and
- two for the variants of SGBM (OpenCV).

All pictures were taken with a resolution of 1280 × 720. This was chosen because it was closest to the resolution of the set used for training the SceneFlow data set (resolution: 960 × 540).

5.2 Scenarios

In Fig. 2, the individual scenarios from the perspective of the left camera at 1.5m distance are shown.

Scene 1: White area. In this simple case we wanted to test how the procedures could deal with flat surfaces. The camera itself is aimed at a white surface with a slight texture (rough-grained wallpaper).

Scene 2: Edge. Disparity leaps always present matching methods with challenges. To put these to a test, a wooden box is placed parallel to the wall. The wooden box is 0.35m wide and the distance from the surface to the wall is 0.49m. Because the wooden box is made of wood, it has a pronounced texture. So one can now evaluate the influence of the texture on the accuracy of the derived 3D structure.

Scene 3: Ball. Finally, the reconstruction of curved surfaces should be checked. The object used is an exercise ball.

6. EVALUATION AND RESULTS

In this chapter, the methods to be investigated (methods of stereolabs, SGBM, DispNet, DispNetC) are analysed on the basis of the described scenarios. This includes the density of the disparity map itself, the distance estimation, reconstruction of a sharp edge and a round object.

First, we will consider the dot density of the generated disparity maps. Then the distance values recorded for scenes 1 and 2 are being compared with the measured values. Finally, the reconstruction of the sphere is being considered.

6.1 Density of the depth map

Procedure	Coverage
SGBM	71%
SGBM (8 paths)	74%
Disp Net	100%
Disp NetC	100%
ZED	95%
ZED (FILL)	100%

Table 2. Mean coverage of disparity maps. First line - Procedure, second line - Coverage

In the best case scenario, the coverage with 3D points is equivalent to the full number of image pixels. In general, the DispNet, DispNetC and ZED camera algorithms achieve 100% coverage in FILL mode (hole fill filter) (see Table 2). All other methods have holes or undefined areas. Fig. 3 shows that the algorithms have difficulties at the edges.

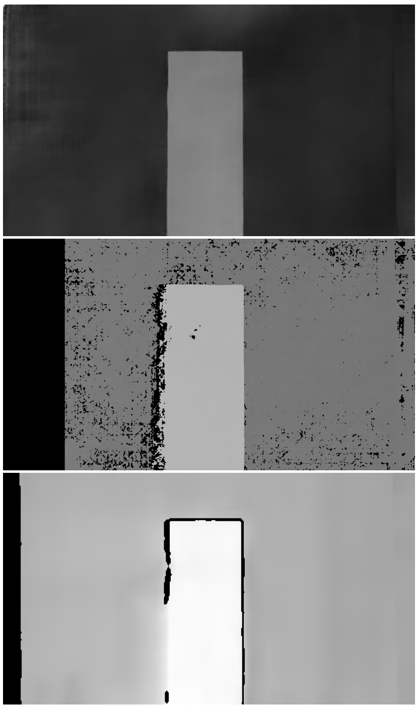

Figure 3. Depth maps of edge scene using (a) DispNet, (b) SGBM and (c) stereolabs.

6.2 Distance measurements (scene 1 and 2)

Basically, one can say that none of the methods was able to determine the distances correctly and that the error increases (squared) with the distance. As shown in Figure 4, all processes are subject to variations, and the processes of stereolabs are much more reliable. When the surface is well textured, the variations in all processes will decrease. It is noteworthy that the processes of stereolabs then fluctuate only for a few centimetres around the mean (1-2cm, see Fig. 4b), whereas all other processes are subject to greater fluctuations from a distance of 2.5m (up to 7cm).

6.3 Edge reconstruction

The following section will show how well the procedures are able to handle disparity leaps. Fig. 5b shows a highlight on the surface, which causes problems for some reconstruction methods. A good result is generated by DispNet and DispNetC. The edge is clearly visible in Fig. 5c and 5d, the surface also appears homogeneous and is barely affected by the highlight. In addition, the hidden area behind the edge is correctly assigned to the background.

The procedure of stereolabs produces good results (see Figs. 5e to 5g). The highlight has, depending on the mode, only little or no effect on the result. However, if one needs a complete disparity map and thus activates the FILL mode, an error occurs, which is produced by the highlight (see Figure 5h to 5j). It should be noted that the filters used by stereolabs reduce the error with increasing quality of the depth map.

6.4 Reconstruction of a sphere

The evaluation of the recordings showed that none of the algorithms could reconstruct the round ball (see Figure 6). The surface of the ball was interpreted rather as a plane. This applies

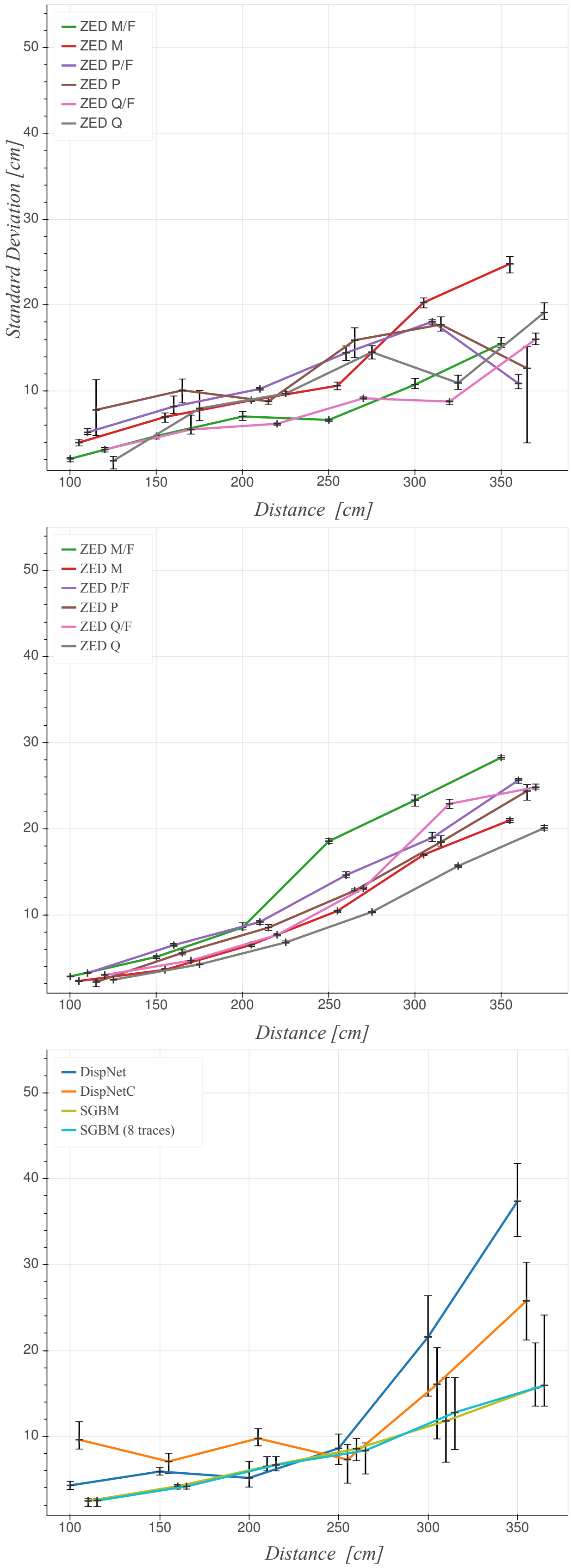

Figure 4. Deviation between the measured and the determined distance for 10 measurements. (a) Homogeneous area using stereolabs; (b) Textured area using stereolabs; (c) Textured area using DispNet and SGBM.

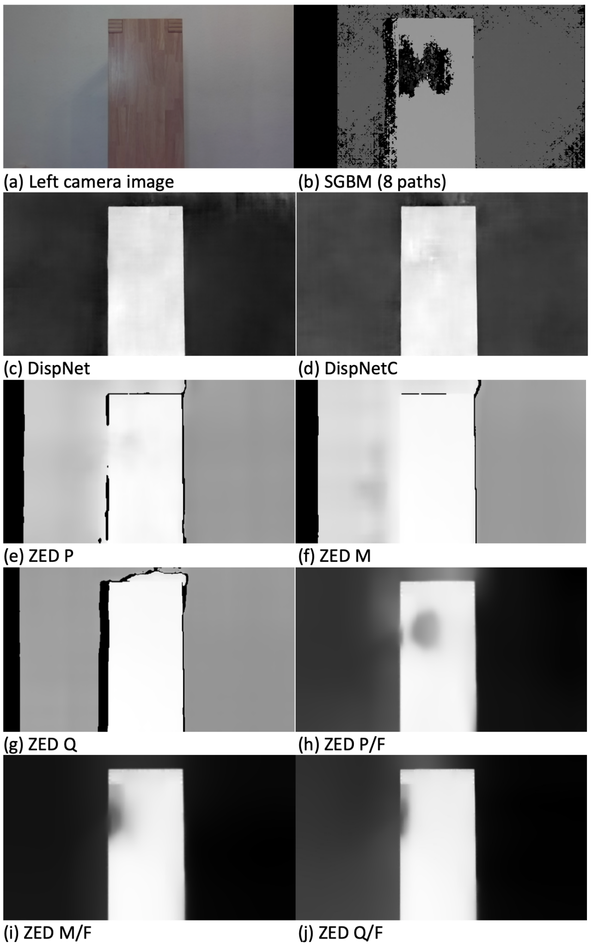
(a) Left camera image
(b) SGBM (8 paths)
(c) DispNet
(d) DispNetC
(e) ZED P
(f) ZED M
(g) ZED Q
(h) ZED P/F
(i) ZED M/F
(j) ZED Q/F

Figure 5. Edge at 0.7m distance. The ”P” ”M” and ”Q” options are quality options of the ZED system, with increasing reconstruction quality. The ”F” option is for filling gaps.

to all procedures and distancing. The best visual reconstruction was done by DispNet and ZED QUALITY (the stereolabs process using the highest quality level) at a distance of 1.5m (see Figures 7).

Finally, one has to conclude that it is not possible to correctly determine the point correspondences by using these methods. The cause is the largely texture-free and shiny surface, which contains many similar points.

Figure 6. Left and right image of the sphere.

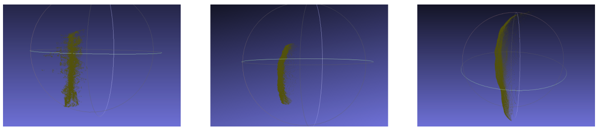

Figure 7. Results of sphere reconstruction. (left-SGBM, middle-DispNet, right-ZED Q)

7. SUMMARY AND OUTLOOK

In this work different stereo-matching methods were considered. These range from classic methods to proprietary ones, as well as to those based on neural networks. Special attention was paid to the stereo camera ZED implemented by stereolabs. On the other hand, methods based on neural networks (DispNet, DispNetC) were established and compared with the method implemented in OpenCV (Semi-Global block matcher). All procedures have been evaluated in problem specific scenarios. It can be concluded from the analysis that all procedures are affected by the typical problems of the correspondence analysis. This concerns in particular the problems in the reconstruction of homogeneous surfaces, highlights and occlusions. In general it can be stated that good texturing is beneficial and that the quality decreases with an increasing distance. The optimal working range for the ZED stereo camera seems to be in the near range (<2m). The best way to do it, is using stereolabs. Due to the adjustable quality, depending on the application, the result can be adjusted and the error reduced. However, this method has problems with edges (soft transition) and highlights. The result of the neural networks does not satisfy what the ranking in the benchmarks suggests. The DispNet and DispNetC networks deliver solid near-field results, but fall behind the stereolabs approach at a distance of 2.5m. It can be assumed that the result depends heavily on the training data set and the network has been optimised for the benchmark, which is the reason, why a direct transfer to other applications does not produce the desired result. Nevertheless, the nets are able to better deal with problem areas such as edges and highlights. The Semi-Global block matcher (SGBM) had major problems with the recordings made. The results are characterised by large fluctuations and irregular areas. Using texturing, the process can again deliver solid results. All this, however, is a very narrow view on the whole problem. Different methods could behave differently in other situations. But for now, the method implemented by sterolabs works best, which is most likely also due to best knowledge of the sensor and device properties.

The investigated methods should be examined in the future in regards to speed, memory consumption and required hardware. For example, a high-performance graphics card is required to train the neural networks. The execution itself can take place with most frameworks but also on the CPU. It would be interesting how big the performance differences between the GPU and the CPU are depending on the resolution. Furthermore, it is obvious that the result of the neural networks depends on the training data. Convolutional networks offer the possibility to work on different resolutions. It would be interesting to take into consideration to what extent the resolution of the training data has an influence on the result and how the result can be improved with specific adaptation of the training data. The introduction of noise, de-calibration in the y-direction, brightness differences in both images, etc. is conceivable.

REFERENCES

Birchfield, S., Tomasi, C., 1998. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(4), 401–406.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T., 2015. Flownet: Learning optical flow with convolutional networks.

Proceedings of the IEEE international conference on computer vision, 2758–2766.

Duane, C. B., 1971. Close-range camera calibration. *Photogramm. Eng*, 37(8), 855–866.

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

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hermann, S., Vaudrey, T., 2010. The gradient-a powerful and robust cost function for stereo matching. *2010 25th International Conference of Image and Vision Computing New Zealand*, IEEE, 1–8.

Hirschmuller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2, IEEE, 807–814.

Itseez, 2015. Open source computer vision library.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*.

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.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097–1105.

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., Heipke, C., Geiger, A., 2015. Joint 3D estimation of vehicles and scene flow. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2.

Pal, C. J., Weinman, J. J., Tran, L. C., Scharstein, D., 2012. On learning conditional random fields for stereo. *International Journal of Computer Vision*, 99(3), 319–337.

Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3), 7–42.

Stereolabs, 2015. ZED camera.

Zbontar, J., LeCun, Y. et al., 2016. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research*, 17(1-32), 2.

https://doi.org/10.5194/isprs-archives-XLII-1-W2-17-2019 | © Authors 2019. CC BY 4.0 License.