A Study on the Influence of Omnidirectional Distortion on CNN-based Stereo Vision

Julian Bruno Seuffert[a], Ana Cecilia Perez Grassi[b], Tobias Scheck[c] and Gangolf Hirtz[d]
Faculty of Electrical Engineering and Information Technology,
Chemnitz University of Technology, Reichenhainer Str. 70, Chemnitz, Germany

Keywords: Omnidirectional, Fish Eye, Indoor, 3D, CNN, Stereo Vision.

Abstract: Stereo vision is one of the most prominent strategies to reconstruct a 3D scene with computer vision techniques. With the advent of Convolutional Neural Networks (CNN), stereo vision has undergone a breakthrough. Always more works attend to recover the depth information from stereo images by using CNNs. However, most of the existing approaches are developed for images captured with perspective cameras. Perspective cameras have a very limited field of view of around 60° and only a small portion of a scene can be reconstructed with a standard binocular stereo system. In the last decades, much effort has been conducted in the research field of omnidirectional stereo vision, which allows an almost complete scene reconstruction if the cameras are mounted at the ceiling. However, as omnidirectional images show strong distortion artifacts, most of the approaches perform an image warping to reduce the reconstruction complexity. In this work, we examine the impact of the omnidirectional image distortion on the learning process of a CNN. We compare the results of a network training with perspective and omnidirectional stereo images. For this work, we use AnyNet and a novel dataset of synthetic omnidirectional and perspective stereo images.

1 INTRODUCTION

Convolutional Neural Networks (CNN) have gained an undisputed protagonist in different fields of computer vision such as detection and classification of objects and semantic segmentation. In recent years, their use has also reached stereo vision.

A typical stereo vision algorithm can be described through four steps: matching cost computation, cost aggregation, optimization and disparity refinement (Scharstein et al., 2001). A disparity map gives the difference between the position of matching pixels on the left and right images of a stereo pair. This values reveal information about the depth of the scene, i.e., the distance of the scene's objects to the camera.

Multiple works have advocated to perform these steps, or a subset of them, using CNNs. Some architectures focus on predicting the matching cost and leave the other steps for a post-processing (Žbontar and LeCun, 2016) and some CNNs include all stereo steps (Mayer et al., 2016).

One of the most important assumption in the existing works is that the matching cost computation is performed on rectified images. The matching cost gives the degree of correspondence between intensity values in the right and left images of the stereo pair. In the case of rectified images, this correspondence occurs only along horizontal epipolar lines (See Sec. 3.1). This latter means that the search for matching pixels is restricted to a single dimension, which significantly simplifies the algorithm. For perspective images, the rectification process is straightforward and it only implies a correction of the images regarding the cameras' arrangement. However, this latter is no longer the case if we consider a different camera model such as the omnidirectional one.

Omnidirectional cameras are gaining attention in computer vision because of their wide field of view (FOV). However, this advantage is accompanied by a high radial distortion of the image. In the case of stereo images, this distortion also affects the epipolar lines, which take the form of curves. As a consequence, now the search for the matching cost must be carried out in two dimensions. To overcome this problem, omnidirectional stereo vision approaches, both standard and CNN-based, pre-process the images by

[a] https://orcid.org/0000-0002-0636-3385
[b] https://orcid.org/0000-0003-1171-903X
[c] https://orcid.org/0000-0002-1829-0996
[d] https://orcid.org/0000-0002-4393-5354

Seuffert, J., Grassi, A., Scheck, T. and Hirtz, G.
A Study on the Influence of Omnidirectional Distortion on CNN-based Stereo Vision.
DOI: 10.5220/0010324808090816
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 809-816
ISBN: 978-989-758-488-6

unwrapping them according to a given model, in order to obtain parallel epipolar lines. This latter allows using the same approaches as for perspective images. However, the unwrapping process introduces errors and loss of information.

In this work, we study how the difference between epipolar lines and curves influences the learning process of a CNN trained to predict the disparity map. Although the first CNN architectures especially developed for omnidirectional stereo images have been recently introduced, they still include image transformation to obtain horizontal epipolar lines (Won et al., 2019b).

A direct comparison of how the geometry of omnidirectional images affects the prediction results is still missing in the literature. This paper investigates how well a CNN architecture, which can successfully learn the disparity maps from perspective images, can also learn the disparity maps from omnidirectional ones. For this purpose, we select the network AnyNet (Wang et al., 2019), which has achieved state-of-the-art results on perspective images.

As there is a lack of omnidirectional stereo datasets for aligned cameras with ground truth depth maps, we present a novel synthetic stereo dataset called *THEOStereo* (available on https://www.tu-chemnitz.de/etit/dst/forschung/comp_vision/theostereo). We train AnyNet with both omnidirectional and perspective images from *THEOStereo* and analyze the accuracy of the predicted disparity maps. Moreover, we study the importance of considering global information on omnidirectional stereo images in contrast with perspective ones.

2 RELATED WORK

With the advent of CNNs, the generation of disparity maps in stereo vision has undergone a breakthrough. Different network architectures have been developed to predict the disparity maps from image pairs. MC-CNN (Matching Cost by using CNN) is an architecture based on siamese networks to predict the matching cost on small image patches (Žbontar and LeCun, 2016). The resulting matching cost is then improved by a series of post-processing steps as cross-based cost aggregation, semiglobal matching (SGM) (Hirschmüller, 2008), a left-right consistency check, subpixel enhancement, a median and a bilateral filter.

In (Luo et al., 2016) a faster siamese matching network is presented. Luo et al. treat the problem as a multi-class classification, where the classes are all possible disparities and compute the inner product between the two representations of the given siamese architecture by using a product layer. DispNet (Mayer et al., 2016) is an end-to-end network that directly predicts disparities for an image pair without post-processing.

GC-Net (Geometry and Context Network) is a deep learning architecture for regressing disparity (Kendall et al., 2017). By using a soft-argmin layer, GC-Net learns disparity as a regression problem, rather than classification, improving the performance and enabling sub-pixel accuracy.

PSMNet (Pyramid Stereo Matching Network) incorporates global context information in stereo matching though a pyramid network (Chang and Chen, 2018).

AnyNet (Wang et al., 2019) predicts a fast initial disparity map and then progressively improves it by predicting residual maps. In this way, the disparity estimation is carried out in stages, being able to extract a prediction from each one at any time. With this architecture, AnyNet can trade off computation effort and accuracy at inference time. We describe AnyNet in Sec. 3.3.

All previously mentioned works assume a rectified perspective image pair as input. Recently the first architectures for omnidirectional stereo images have been introduced. Won et al. present a series of works for omnidirectional depth estimation from a wide-baseline multi-view (four cameras) omnidirectional stereo setup (Won et al., 2019a; Won et al., 2019b; Won et al., 2020).

Their first work presents SweepNet (Won et al., 2019b), a CNN that computes the matching costs of grayscale equirectangular image pairs warped from the omnidirectional images. The resulting cost volume is then refined by applying SGM and the final depth map is estimated. However, SweepNet has problems to manage occlusions, which are typical for the proposed wide-baseline omnidirectional setup. To overcome this problem, Won et al. propose OmniMVS (Won et al., 2019a), an end-to-end deep neural network consisting of three blocks: Feature extractor, spherical sweeping and cost volume computation. In (Won et al., 2020), they extend OmniMVS to consider an entropy boundary loss for learning better regularization in the cost computation block.

All these three works include the warp of the input omnidirectional images (SweepNet) or of their feature maps (OmniMVS) onto concentric global spheres. In our work, we are interested in studying how well a CNN that successfully predicts disparity from perspective stereo images, like AnyNet, is also capable to generate disparity maps from omnidirectional stereo images. This involves learning a different geometry, which requires global information as shown in Sec. 5.

There are different datasets of stereo images. Two of the most used perspective stereo image datasets are Scene Flow (Mayer et al., 2016) and KITTI (Geiger et al., 2012; Menze and Geiger, 2015).

Won et al. present synthetic omnidirectional datasets for indoor and outdoor scenes (Won et al., 2019b; Won et al., 2019a; Won et al., 2020). However, all these datasets are exclusively designed, and therefore limited, for their proposed setup of four cameras with wide-baseline and non-aligned viewing directions.

In our work, we introduce a datatset similar to THEODORE (Scheck et al., 2020), called *THEO-Stereo*. This comprises synthetic omnidirectional stereo images of indoor scenes and their depth maps as ground truth.

3 STEREO VISION

Stereo vision is beside structure from motion, structured light, time of flight and other techniques a common approach to reconstruct scenes in 3D (Giancola et al., 2018). Stereo vision works in a similar manner to the spatial scene perception of humans by retrieving depth information from at least two cameras analogous to two eyes. Those two cameras are located at position $\boldsymbol{C}_0$ and $\boldsymbol{C}_1$ in a camera coordinate system (CCS) with the origin at $\boldsymbol{C}_0$. In this section, we briefly introduce the concepts of perspective and omnidirectional stereo vision. A more complete and detailed explanation of these concepts can be found in (Hartley and Zisserman, 2004) and (Kannala and Brandt, 2006).

3.1 Stereo Vision with Perspective Cameras

For performance reasons, perspective cameras are commonly arranged in a so-called canonical camera setup, where image planes are coplanar and their x-axes are collinear. A canonical setup following the perspective camera model is shown in Fig. 1. We define the camera on the left, i.e., at position $\boldsymbol{C}_0$, as the reference camera.

A point in the CCS is denoted as $\boldsymbol{P} = [x_{\mathrm{cam}}, y_{\mathrm{cam}}, z_{\mathrm{cam}}]^T$. For a perspective camera model, the projection of $\boldsymbol{P}$ on the first camera's image plane is given by the point $\boldsymbol{P}^0_{\mathrm{persp}} = [x^0_{\mathrm{img}}, y^0_{\mathrm{img}}]^T$. This image point is defined in an image coordinate system with origin on the image's upper left corner. In order to retrieve the location of the point $\boldsymbol{P}$, the point

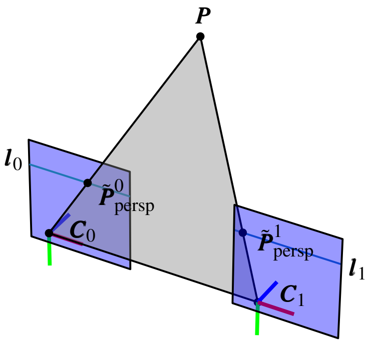


Figure 1: Canonical binocular stereo vision setup following the perspective camera model. The point $\boldsymbol{P}$ is projected onto the image planes at position $\boldsymbol{P}^0_{\mathrm{persp}}$ and $\boldsymbol{P}^1_{\mathrm{persp}}$. Both points have the same y-coordinate as the epipolar lines $\boldsymbol{l}_0$ and $\boldsymbol{l}_1$ are collinear and parallel to the x-axes of the images.

$\boldsymbol{P}^1_{\mathrm{persp}} = [x^1_{\mathrm{img}}, y^1_{\mathrm{img}}]^T$ on the right image, which corresponds to $\boldsymbol{P}^0_{\mathrm{persp}}$ on the left image, must be found. This is done by applying a stereo matching technique that searches for $\boldsymbol{P}^1_{\mathrm{persp}}$ along the corresponding epipolar line $\boldsymbol{l}_1$.

We define the x-distance between each image point $\boldsymbol{P}^0_{\mathrm{perps}}$ and $\boldsymbol{P}^1_{\mathrm{perps}}$ and their corresponding optical axes as $x_l = x^0_{\mathrm{img}} - c_x$ and $x_r = x^1_{\mathrm{img}} - c_x$, respectively, where c_x is the x-coordinate of the image center (See Fig. 2). Applying similar triangles, following relations can be obtained:

$$\frac{x_l}{f} = \frac{x_{\mathrm{cam}}}{z_{\mathrm{cam}}}, \tag{1}$$

$$\frac{x_r}{f} = \frac{x_{\mathrm{cam}} - b}{z_{\mathrm{cam}}}, \tag{2}$$

where f stands for the focal length of the cameras and b for the baseline (distance between cameras). By subtracting Eq. 2 from Eq. 1, the relationship between the depth z_{cam} and the disparity $d = x^0_{\mathrm{img}} - x^1_{\mathrm{img}}$ can be retrieved as follows:

$$z_{\mathrm{cam}} = \frac{f \cdot b}{x_l - x_r} = \frac{f \cdot b}{x^0_{\mathrm{img}} - x^1_{\mathrm{img}}} = \frac{f \cdot b}{d}. \tag{3}$$

The arrangement of disparity and depth values on two-dimensional maps with respect to the image of the reference camera (at $\boldsymbol{C}_0$) are called disparity map and depth map, respectively.

3.2 Stereo Vision with Omnidirectional Cameras

The main advantage of using omnidirectional cameras in stereo vision is their large FOV, which allows, e.g., to capture a whole indoor scene when the cameras are

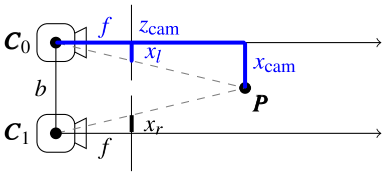


Figure 2: Reconstruction of a point $\boldsymbol{P}$ in a CCS following the perspective camera model. The position of the point $\boldsymbol{P}$ can be retrieved by mean of similar triangles.

mounted on the ceiling. This large FOV is associated with image distortions that increases the complexity of the disparity and depth estimation. Similar to Fig. 1, a stereo setup with omnidirectional cameras is shown in Fig. 3.

For omnidirectional images, the projection of a point $\boldsymbol{P}$ from the camera coordinate system on the images plane can be modeled using the equidistant projection model for dioptric cameras of (Kannala and Brandt, 2006).

In the case of omnidirectional cameras, the point $\boldsymbol{P}$ is projected on a hemisphere. We follow the process in (Findeisen et al., 2014) and project $\boldsymbol{P}$ first on a left and a right virtual hemispherical image. Each point $\tilde{\boldsymbol{P}}^0_{\text{omni}}$ and $\tilde{\boldsymbol{P}}^1_{\text{omni}}$ on these hemispheres can be described in terms of polar coordinates, i.e. through radius r, elevation angle θ and azimuth angle ϕ.

Before the projection from the hemispheres on the image planes is performed, a normalized image plane is defined as follows (Kannala and Brandt, 2006):

$$\begin{bmatrix} x_{\text{norm}} \\ y_{\text{norm}} \end{bmatrix} = \theta \cdot \begin{bmatrix} \cos\phi \\ \sin\phi \end{bmatrix}, \tag{4}$$

where the focal length of the normalized image plane is 1 and the origin of coordinates coincides with the image's center.

Finally, to obtain the coordinates on the image plane, an affine transformation with the so-called calibration matrix $\boldsymbol{K}$ is applied on the normalized image of Eq. 4:

$$\begin{bmatrix} x_{\text{img}} \\ y_{\text{img}} \end{bmatrix} = \underbrace{\begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}}_{\boldsymbol{K}} \cdot \begin{bmatrix} x_{\text{norm}} \\ y_{\text{norm}} \end{bmatrix}, \tag{5}$$

where $f_x = \frac{w}{\pi}$ and $f_y = \frac{h}{\pi}$ are the image focal lengths, w and h are the image width and height, respectively, c_x and c_y are the coordinates of the image's center and s is the screw coefficient between the image's x- and y-axis.

The described omnidirectional model causes that the points $\boldsymbol{P}^0_{\text{omni}}$ and $\boldsymbol{P}^1_{\text{omni}}$ (both projections of $\boldsymbol{P}$) on the left and right images no longer are arranged along a line, as for the perspective camera model, but along a curve as shown in Fig. 4.

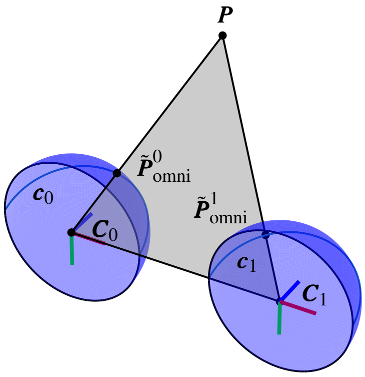
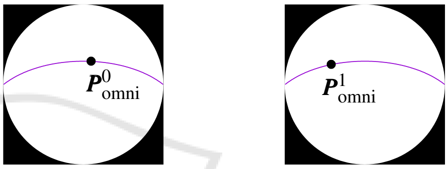


Figure 3: Canonical binocular stereo vision setup following the omnidirectional camera model. The point $\boldsymbol{P}$ is projected on the image hemispheres at position $\tilde{\boldsymbol{P}}^0_{\text{omni}}$ and $\tilde{\boldsymbol{P}}^1_{\text{omni}}$. In contrast to Fig. 1, the points do not have the same y-coordinate as they do not lie on epipolar lines but on epipolar curves $\boldsymbol{c}_0$ and $\boldsymbol{c}_1$.


$\boldsymbol{P}^0_{\text{omni}}$
$\boldsymbol{P}^1_{\text{omni}}$

(a) $\boldsymbol{C}_0$ - omnidirectional (b) $\boldsymbol{C}_1$ - omnidirectional

Figure 4: Stereo image planes of the setup in Fig. 3. The points $\boldsymbol{P}^0_{\text{omni}}$ and $\boldsymbol{P}^1_{\text{omni}}$ are projected from $\boldsymbol{P}$ on the left and right epipolar curves.

As a consequence, in omnidirectional stereo vision, the matching point $\boldsymbol{P}^1_{\text{omni}}$ for a given point $\boldsymbol{P}^0_{\text{omni}}$ must be searched along a 2D epipolar curve on the right image. To avoid this problem, some approaches transform the hemispherical image to a half-cylindrical image and then project the resulting epipolar curves as lines onto the image plane (Li, 2006; Findeisen and Hirtz, 2014; Findeisen et al., 2014). Although, this allows to reduce the search domain for point matching to one dimension, the re-projection step on a half-cylindrical image introduces interpolation artifacts that affect the disparity map generation.

CNNs appear to have the potential to be able to learn the geometry of epipolar curves and therefore to generate disparity maps directly from untransformed omnidirectional images. In the next section, we describe AnyNet (Wang et al., 2019), a state-of-the-art network to generate disparity maps from perspective images and in Sec. 5 we analyze if its architecture is also able to learn epipolar curves.

3.3 Stereo Vision with AnyNet

To study the impact of omnidirectional distortion on the generation of disparity maps through CNNs, we have selected AnyNet (Wang et al., 2019).

AnyNet is a recently developed network that has reached state-of-the-art results on perspective images. This network has the particularity of predicting successive disparity maps while increasing their quality with each step. In this way, AnyNet achieves a compromise between computing time and precision.

AnyNet's architecture consists of four stages, each of which offers a disparity map as output and feeds the next stage with information. Stage one to three predict the disparity map in different resolutions from coarse to fine. The forth stage sharpens the disparity map of the third stage with an SPNet (Liu et al., 2017).

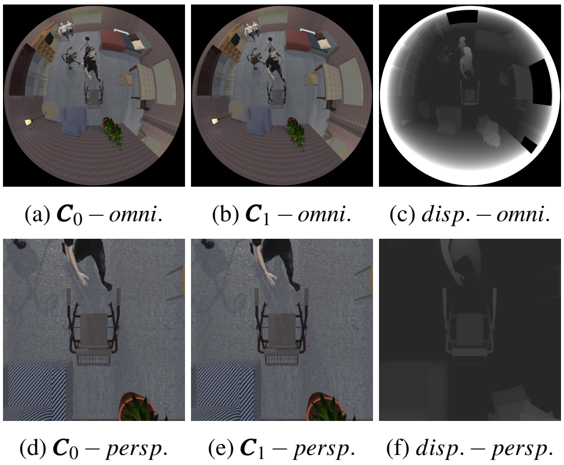
The stages one to three take feature vectors as an input, which are calculated from the stereo pair by using a U-Net (Ronneberger et al., 2015). SPNet in stage four has, on the other hand, the stereo pair and the disparity map of stage three as input.

In addition to the feature vectors, the disparity networks in stage two and three also take as input the disparity map from the previous stage and calculate a residual map that is added to the current disparity map. In (Wang et al., 2019), AnyNet is trained with the synthetic dataset Scene Flow (Mayer et al., 2016) and later fine-tuned with KITTI (Geiger et al., 2012; Menze and Geiger, 2015). The end-to-end training is performed on patches of 512 × 256 pixels randomly cropped from the original images, while the evaluation is done on KITTI's full size images (1242 × 375 pixels).

4 EXPERIMENTS

We present a dataset similar to THEODORE (Scheck et al., 2020), called *THEOStereo*. THEODORE is a synthetic dataset of images captured from omnidirectional top-view cameras with 180° FOV. This dataset comes along with segmentation masks but does not provide depth maps as ground truth. We incorporate a pair of virtual omnidirectional cameras in THEODORE's environment and generate $31,250$ new stereo image pairs of indoor scenes with their corresponding depth maps.

The two virtual cameras are arranged in a canonical stereo camera setup using an omnidirectional (equidistant) camera model, i.e., the viewing directions are aligned and their *x*-*y*-planes are coplanar. The baseline b was 0.3 m. Both virtual cameras are perfect (perfect hemispherical images) dioptric cameras, which can be described using Kannala's model (Kannala and Brandt, 2006) with distortion paramters $k_1 = 1$ and $k_i = 0$ for $i \in \{2,3,4,5\}$. The stereo camera height varies between 4.73 m and 5.6 m.


(a) $\boldsymbol{C}_0 - omni.$ (b) $\boldsymbol{C}_1 - omni.$ (c) $disp. - omni.$

(d) $\boldsymbol{C}_0 - persp.$ (e) $\boldsymbol{C}_1 - persp.$ (f) $disp. - persp.$

Figure 5: Sample images of the dataset *THEOStereo* and their calculated disparity maps. Their corresponding disparity maps are stored in 16-bit images. For the sake of visibility, the intensity range of (c) and (f) is clipped at 255.

To train AnyNet, we need disparity maps as ground truth, however *THEOStereo* provides depth maps. The values of a depth map depend only on the height of the scene's objects and the position of the camera. Therefore, they are the same for equally positioned perspective and omnidirectional cameras. However, as described before, this is not the case for disparity maps. In the case of perspective stereo, each value of the disparity map describes a horizontal translation to the right, whereas for omnidirectional disparity two values are necessary in order to describe a translation along an epipolar curve, which would result in a two-channel disparity map.

In order not to change the architecture of AnyNet, we generate single-channel disparity maps from the dataset's depth maps. *THEOStereo*'s depth maps present the same omnidirectional distortion as the input images. In this way, by applying Eq. 3 on these omnidirectional depth maps, we obtain a hybrid disparity map with values that correspond to a perspective geometry, but conserving the omnidirectional distortion on their arrangement (See first row Fig. 5).

We generate a second dataset from *THEOStereo* by transforming its images to a perspective view with a typical FOV of 60°. Fig. 5 shows an example of a *THEOStereo* stereo image pair and its perspective version, as well as their disparity maps. This second dataset is used to compare the performance of AnyNet for perspective and omnidirectional im-

Table 1: Results of AnyNet on *THEOStereo*. The error refers to the testing subset of *THEOStereo* after training 200 epochs.

Exp.	Camera model	Input size in pixels	MAE in pixels	$\delta > 1$	$\delta > 2$	$\delta > 4$	$\delta > 3$ & $\varepsilon > 5$ %
A	perps.	512×256	0.33	9.6 %	4.9 %	2.2 %	2.9 %
B	omni.	512×256	6.36	75.0 %	57.0 %	35.9 %	39.3 %
C	omni.	1024×1024	3.94	68.8 %	46.8 %	23.4 %	26.4 %

ages. Both datasets are split into three subsets: training (80%), validation (10%) and testing (10%).

We perform three experiments:

- Exp. A: In this experiment, we use the pre-trained model of AnyNet with Scene Flow and fine-tuned it with our perspective dataset. The network is trained on randomly selected image patches of 512×256 pixels (See Sec. 3.3).
- Exp. B: The pre-trained AnyNet is fine-tuned with randomly selected omnidirectional image patches of *THEOStereo* as in Exp. A.
- Exp. C: We repeat experiment B by fine-tuning on the full omnidirectional images. This experiment considers global information during the training. In this way, with each image pair, the network learns the complete epipolar curves and not a section of them as in Exp. B.

For all experiments the evaluation is performed on the testing dataset considering full images. AnyNet is trained for all experiments during 200 epochs using Adam (Kingma and Ba, 2015) with an initial learning rate of $5 \cdot 10^{-4}$. We applied cosine annealing (Loshchilov and Hutter, 2017) to successively reduce the learning rate to zero until the end of the training. We chose a batch size of 48 for all experiments and activated the SPNet in the second epoch.

5 RESULTS

To evaluate the performance of the network in all three experiments, five error metrics are considered: Mean Absolute Error, the ratio $\delta > i$ of pixels with errors greater than i, with $i \in \{1,2,4\}$ and the 3-Pixel-Error.

The Mean Absolute Error (MAE) is calculated by taking the mean value of the absolute difference per pixel between the predicted disparity map and the ground truth (L_1 loss).

The metric $\delta > i$ indicates the percentage of pixels in the predicted disparity map that have an error bigger than i pixels with respect to the ground truth. The metrics $\delta > 1$, $\delta > 2$ and $\delta > 4$ are the so-called bad-1, bad-2 and bad-4 error (Scharstein et al., 2014).

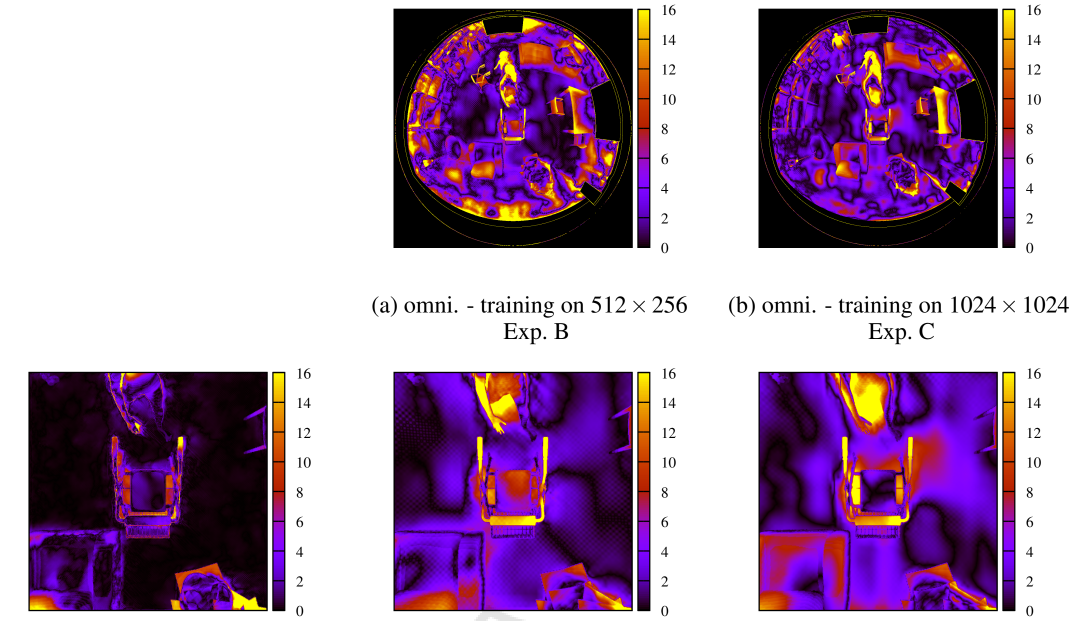
The 3-Pixel-Error (Menze and Geiger, 2015) is defined by $\delta > 3$ & $\varepsilon > 5\%$, where $\delta > 3$ indicates the ratio of pixels with an error bigger than three pixels and $\varepsilon > 5\%$ denotes that such errors are also bigger than 5% of the ground truth disparity value.

Each considered metric was first averaged batch-wise for a batch size of 48 and finally averaged over all batches. Tab. 1 documents the error measurements for each experiment. As expected, AnyNet hat the best performance for the case of perspective images (Exp. A). Experiments B and C present much higher errors, showing that AnyNet has much more difficulty in learning from omnidirectional images. However, as shown later in Fig. 6, it is still able to generate an omnidirectional disparity map.

The difference between Exp. B and Exp. C shows that the network can reach better results, when the complete image is used to train. Learning from the full images helps the network to understand the omnidirectional geometry better.

To visualize the error, Fig. 6 shows for all experiments the Absolute Error Heat Map (AE-HM) for an image of the testing dataset. In the first row, the resulting AE-HM of Exp. B and C are presented. We can see that, while the error in the center of both maps is similar, the error in Exp. B increases with the distance to the center. This latter indicates that by learning from patches, the network losses information about the global omnidirectional geometry.

The second row in Fig. 6 presents the AE-HM for Exp. A on the same image. By considering a perspective camera, the FOV is reduced by two thirds and the absolute error also reduces considerably. In order to facilitate the comparison, the AE-HMs of Exp. B and C have been cropped and converted to perspective in such a way that they coincide with the AE-HM of Exp. A. Areas that correspond in the ground truth disparity map to a larger disparity value than AnyNet's default maximum disparity (192) are excluded in Fig. 6a to Fig. 6e (set to black). This comparison shows that for omnidirectional images, AnyNet does not accurately retrieve the depth of planes such as the floor, but is almost error-free in the case of perspective images.



(a) omni. - training on 512 × 256
Exp. B

(b) omni. - training on 1024 × 1024
Exp. C

(c) persp. - training on 512 × 256
Exp. A

(d) omni. - training on 512 × 256
(undistorted)

(e) omni. - training on 1024 × 1024
(undistorted)

Figure 6: Absolute Error Heat Maps of a sample testing image. In the top row, the training was performed on cropped and full size omnidirectional images. The AE-HM generated by training the network on the perspective image patches is shown in (c). The Figs. (d) and (e) show the same heat maps of (a) and (b) but converted to perspective with the same FOV as (c). The color range visualizes a mean absolute error between 0 and 16 pixels.

6 CONCLUSION AND FUTURE WORK

In this work, we compare the quality of disparity maps predicted from perspective and omnidirectional stereo images. For this study, we train AnyNet on our dataset *THEOStereo*. We demonstrate that it is possible to learn disparity maps from omnidirectional images, which allow wide angle 3D scene reconstructions. Furthermore, we prove that by learning from omnidirectional images, global information is essential as it can significantly reduce the MAE by around 35 %. A comparison of our approach with approaches like OmniMVS (Won et al., 2019a) are planned. Future experiments will also investigate other loss functions, metrics and networks. Instead of single channel disparity maps, the utilization of multichannel maps encoding an n-dimensional disparity metric or feature vector should be investigated. For that purpose, the architecture of the disparity networks must be adapted. With our dataset, we pave the way for further research on omnidirectional stereo vision with aligned cameras.

ACKNOWLEDGEMENTS

This work is funded by the European Regional Development Fund (ERDF) as well as the Free State of Saxony under the grant number 100-241-945.

REFERENCES

Chang, J.-R. and Chen, Y.-S. (2018). Pyramid Stereo Matching Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, Salt Lake City, UT. IEEE.

Findeisen, M. and Hirtz, G. (2014). Trinocular Spherical Stereo Vision for Indoor Surveillance. In *2014 Canadian Conference on Computer and Robot Vision*, pages 364–370, Montreal, QC, Canada. IEEE.

Findeisen, M., Meinel, L., and Hirtz, G. (2014). A trinocular omnidirectional stereo vision system for high-precision RGB-D acquisition. In *Proceedings ELMAR-2014*, pages 1–4, Zadar, Croatia. IEEE.

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

Giancola, S., Valenti, M., and Sala, R. (2018). *A Survey on 3D Cameras: Metrological Comparison of Time-of-Flight, Structured-Light and Active Stereoscopy Technologies*. SpringerBriefs in Computer Science. Springer International Publishing, Cham, 1 edition.

Hartley, R. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK, 2 edition. https://doi.org/10.1017.

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

Kannala, J. and Brandt, S. S. (2006). A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1335–1340. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., and Bry, A. (2017). End-to-End Learning of Geometry and Context for Deep Stereo Regression. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 66–75, Venice. IEEE.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Li, S. (2006). Trinocular Spherical Stereo. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4786–4791, Beijing, China.

Liu, S., De Mello, S., Gu, J., Zhong, G., Yang, M.-H., and Kautz, J. (2017). Learning Affinity via Spatial Propagation Networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 1520–1530. Curran Associates, Inc.

Loshchilov, I. and Hutter, F. (2017). SGDR: Stochastic Gradient Descent with Warm Restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Luo, W., Schwing, A. G., and Urtasun, R. (2016). Efficient Deep Learning for Stereo Matching. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5695–5703, Las Vegas, NV, USA. IEEE.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., and Brox, T. (2016). A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, Las Vegas, NV, USA. IEEE.

Menze, M. and Geiger, A. (2015). Object scene flow for autonomous vehicles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, Boston, MA, USA. IEEE.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. pages 234–241. Springer, Cham.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., and Westling, P. (2014). High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. volume 8753, pages 31–42, Cham. Springer. Series Title: Lecture Notes in Computer Science.

Scharstein, D., Szeliski, R., and Zabih, R. (2001). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, pages 131–140, Kauai, HI, USA. IEEE Comput. Soc.

Scheck, T., Seidel, R., and Hirtz, G. (2020). Learning from theodore: A synthetic omnidirectional top-view indoor dataset for deep transfer learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Žbontar, J. and LeCun, Y. (2016). Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(1):2287–2318.

Wang, Y., Lai, Z., Huang, G., Wang, B. H., van der Maaten, L., Campbell, M., and Weinberger, K. Q. (2019). Anytime Stereo Image Depth Estimation on Mobile Devices. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5893–5900, Montreal, QC, Canada. IEEE.

Won, C., Ryu, J., and Lim, J. (2019a). OmniMVS: End-to-End Learning for Omnidirectional Stereo Matching. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8986–8995, Seoul, Korea (South). IEEE.

Won, C., Ryu, J., and Lim, J. (2019b). SweepNet: Wide-baseline Omnidirectional Depth Estimation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6073–6079, Montreal, QC, Canada. IEEE.

Won, C., Ryu, J., and Lim, J. (2020). End-to-End Learning for Omnidirectional Stereo Matching with Uncertainty Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Early Access.