Accuracy Evaluation and Improvement of the Calibration of Stereo Vision Datasets

Kai Cordes and Hellward Broszio

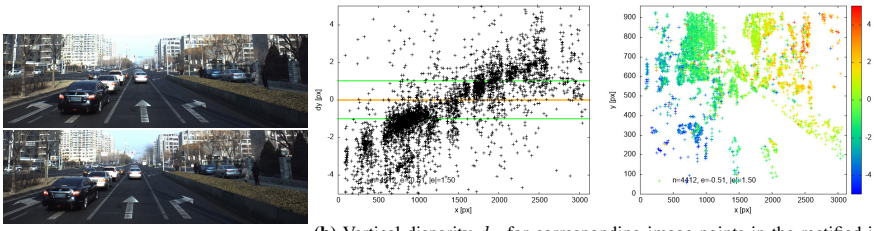
VISCODA GmbH
{cordes,broszio}@viscoda.com
https://www.viscoda.com

Abstract. In automated driving systems, stereo cameras gain in importance for highly accurate perception using classical depth map estimation and machine learning based perception tasks. For this, high-impact datasets were published, e.g., KITTI, Cityscapes, ApolloScape, or Argoverse. These datasets are used for the evaluation of classical stereo vision approaches as well as for the learning of machine learning models. The stereo camera configuration has significant influence on the inferred models which provide object detection, object tracking, and 3D scene reconstruction. Thus, the accuracy of the camera calibration is of high importance, especially when safety critical functions are addressed.
We propose a simple but effective method for the accuracy evaluation of stereo camera calibration and provide a comparison for current highly influential stereo image datasets. The developed metric is then used as cost function to realize the optimization of the given camera parameters.
The evaluations show strongly varying accuracies for different datasets and varying accuracies within a dataset. Datasets with frequent on-site recalibration generally provide higher accuracies while others show suboptimal results. We can infer that mechanical instabilities spoil the usability of these datasets. To counteract this, the proposed optimization minimizes the proposed error metric leading to accurately rectified stereo images.

Keywords: Accuracy · Calibration · Stereo · Camera · Optimization · Dataset

1 Introduction

Depth estimation from stereo cameras is a fundamental *Computer Vision* task [13, 34] with applications in many fields, such as autonomous driving, robotics, scene understanding, and medical diagnosis. Especially for automated driving, stereo cameras have recently gained much importance for challenges such as object detection, tracking, and 3D scene reconstruction [4, 30]. Since these applications are designed for safety critical function, i.e., to prevent accidents, the accuracy of the depth estimation is important. For the accurate mapping from 2D image content to 3D coordinates the calibration of the sensors is required. Thus, stereo vision datasets include intrinsic and extrinsic sensor calibration parameters. Camera calibration is a time-consuming, complicated, and semi-automatic task and there are several approaches for the optimization of the calibration parameters. Calibration procedures employ 2D patterns [17], structured light [26],



(a) Rectified stereo image

(b) Vertical disparity d_y for corresponding image points in the rectified image pair in Fig. 1a. We show d_y with respect to the x-coordinate (left) and color-coded with respect to the (x/y) position in the left stereo image (right).

Fig. 1: Visualization of the vertical disparity d_y in pixels. For rectified images computed from accurate stereo calibration, the vertical disparity should be zero and have green color in the right visualization of Fig. 1b. In this example, a significant error, positive (red) and negative (blue), is visible. The error is dependent on the image position and results from suboptimal calibration.

and calibration tools [14,22]. Nevertheless, the accuracies of the calibrations are usually not provided by the authors of a dataset.

Stereo Vision has a long history in the field of *Computer Vision*. Traditional approaches [19,21,26] use rectified images to estimate the depth of corresponding points from their positions in the images. The rectification eases the disparity estimation since the search space for the correspondence of a point in the left image is limited to a small stripe in the right image. If the calibration is optimal, both points have the same y-coordinate. For erroneous calibration, vertical distances occur, and traditional methods may not be able to establish enough corresponding points. Recent methods employ *Neural Networks* for the end-to-end learning of depth maps [5,6,18,23,28,29,31,32,35] and dominate current performance leaderboards. The learned models implicitly incorporate the camera configuration during training and avoid the explicit search for corresponding image points. Nevertheless, rectified images are usually used for training and evaluation since these images are included in most datasets. Additionally, the standardized camera configuration should increase the transferability of the model from one dataset to another. Finally, rectified images ease the performance evaluation. A rectified stereo image pair is computed based on the calibration of the stereo camera system. Thus, the calibration accuracy is important since errors in the calibration will propagate to the results. A few papers address the misalignment of rectified stereo images. In [9], the sensitivity of 3D reconstruction to erroneous camera calibration is derived. Errors in different parameters of the stereo camera system lead to different effects in the reconstruction. In [2], synthetic data is used to validate the resulting depth error.

Numerous datasets were published for the evaluation of *Stereo Vision* approaches [4,8,15,16,20,26,33]. Often, several recording days are employed to provide a significant variability in the dataset, e.g., different lighting conditions [15] or varying weather [11]. For the usage in the dataset, the authors provide rectified images. They are computed based on the camera calibration resulting from the selected calibration procedure. As variances in the dataset are achieved by data capturing on multiple days, weeks, or even seasons, the camera system should be frequently recalibrated since small differences in the extrinsic parameters are likely to occur.

To quantify the calibration accuracy of stereo camera systems, we propose an evaluation which estimates the deviation of the provided image content from the expected rectified stereo configuration. For a stereo camera with correct calibration parameters, corresponding points in rectified images have the same y-coordinate. Thus, a distance d_y between the y-coordinates indicates the calibration error. The distance d_y is visualized in Fig. 1 for an example stereo image of the *ApolloScape* dataset [20]. We show d_y for corresponding points with respect to the x-coordinate (Fig. 1b, left) and color-coded with respect to the (x/y) position (Fig. 1b, right) in the left stereo image. The example shows a large systematic error with $d_y < 0$ in the bottom left image region and $d_y > 0$ in the top right part of the image. We can infer that this is caused by a suboptimal stereo camera calibration. This paper evaluates the stereo calibration accuracies of datasets. Secondly, we show that the proposed measure for stereo calibration accuracy can be used to optimize the calibration parameters. A simple but effective procedure for the optimization is followed. Based on the new calibration parameters the stereo images are subject to an additional rectification step. We demonstrate that the resulting images have significantly reduced stereo calibration error.

In the following sections, the proposed accuracy evaluation is done for current highly influential stereo vision datasets targeting automotive applications. We show that significant calibration errors occur in current state of the art datasets. In Sec. 2, the evaluation procedure is derived in detail. Results and comparisons are shown in Sec. 3. The optimization of camera parameters using the derived measure is shown and discussed in Sec. 4. In Sec. 5, conclusions are drawn.

2 Accuracy of Stereo Vision Datasets

In [9], the sensitivity of 3D reconstruction to erroneous camera calibration in derived. Erroneous camera parameters affect the quality of both the rectification and the 3D reconstruction. For given disparity error $\Delta\mathbf{d}$ in pixels (or normalized coordinates), the range uncertainty increases quadratically with distance. A consequence of vertical misalignment in the rectified stereo image is that corresponding pixels no longer have the same y-coordinate (same *scanline*) and most stereo matching algorithms deteriorate since the search space may not contain corresponding regions. Thus, in practical applications no correspondence might be found at all. Hence, the vertical misalignment should be small.

In all considered datasets (cf. Sec. 2.1) the stereo images are given in rectified stereo configuration, i.e., the extrinsic parameters of the cameras share the same rotation angles and image target plane. Their extrinsics only differ by a translation vector. It follows, that corresponding image points in left and right image have the same y-coordinate which eases the analysis significantly. Then, the depth z is calculated from the horizontal disparity d_x, the baseline b, and the focallength f as $z = f \cdot \frac{b}{d_x}$. Since real cameras are not installed accurately in rectified stereo configuration, camera parameters and images are transformed using *rectification* [24]. For the rectification, accurately calibrated cameras are required. Otherwise, a vertical offset d_y is encountered when comparing corresponding points. Inversely, this offset can be used to quantify the accuracy of the original calibration.

Table 1: Overview on current highly-influential stereo datasets. All datasets provide rectified images based on the respective calibration procedure. NI: *No Information*

Dataset	Resolution		Baseline	Calibration	publ.
Middelbury [26]	5.4 MP	2820×1920	14-40 cm	2D patterns + struct. light	2014
KITTI [16]	0.5 MP	1242× 375	54 cm	2D patterns [17]	2015
Cityscapes [8]	2.0 MP	2048×1024	22 cm	2D patterns [22]	2016
Driving Stereo [33]	0.4 MP	881× 400	54 cm	2D patterns + *MATLAB* toolbox	2019
ApolloScape [20]	2.9 MP	3130× 960	29.9 cm	NI	2019
Argoverse 1.1 [4]	5.0 MP	2464×2056	29.9 cm	NI	2019
Virtual KITTI 2 [3]	0.5 MP	1242× 375	54 cm	synthetic dataset	2020
DSEC [15]	1.6 MP	1440×1080	55 cm	2D patterns + *Kalibr* toolbox [14]	2021

In this section, we show an overview on the stereo vision datasets in Sec. 2.1. Then, the measure for the accuracy of the stereo calibration is derived in Sec. 2.2.

2.1 Stereo Dataset overview

Numerous datasets were published for the evaluation of *Stereo Vision* approaches. In Tab. 1, we show an overview on current high-impact datasets which focus on automated driving and, as a reference, *Middlebury 2014*. All considered datasets include rectified stereo images. *KITTI* [16], *Virtual KITTI 2* [3], *Cityscapes* [8], *DrivingStereo* [33], and DSEC [15] focus on the stereo perception task. In *KITTI*, *Driving Stereo*, and *DSEC*, ground truth depth information is derived using *SGM* [19] and a filtering approach based on LiDAR point clouds [27]. The *Cityscapes* dataset comes with depth information derived from the stereo camera system. In the *DSEC* dataset, dense depth maps are used for the evaluation of event cameras. *Argoverse 1.1* [4] and *ApolloScape* [20] target 360° view with multiple cameras. Additional stereo cameras face to the front and rectified stereo images are provided. Similar to the comprehensive KITTI stereo evaluation leaderboard, stereo competitions are provided.

Due to the possibility of mechanical variabilities, the camera extrinsic parameters may change slightly throughout the dataset acquisition. Thus, a recalibration is recommended to account the possibly changing camera orientation during the capturing process of the whole dataset. The *KITTI* and *Cityscapes* datasets provide new calibration parameters for each subset. The *DrivingStereo* dataset includes three acquisition periods from July to October in 2018 (2018-07, 2018-08, and 2018-10) with one separate calibration for each period. The data recording is done on 42 different days. DSEC includes five different calibration sets for 53 sequences. The *Argoverse 1.1* dataset includes one set of stereo camera parameters only. The follow-up work *Argoverse 2* [30] contains stereo cameras, but the provided images are unrectified. The *ApolloScape* dataset does not include the stereo calibration parameters, but rectified images.

The *Virtual KITTI 2* dataset is a synthetically generated dataset. Thus, the calibration is correct by design. This dataset serves as a baseline for the proposed accuracy evaluation approach.

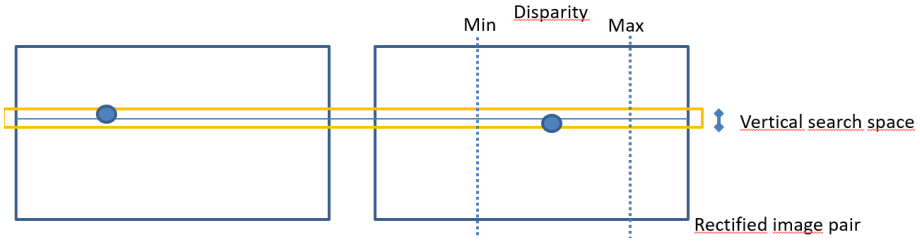


Fig. 2: Corresponding points (blue) for a rectified stereo image. The search space in the right image is determined in horizontal direction by the minimal/maximal disparity and in vertical direction by a small vertical boundary. Common approaches assume that corresponding points have the same vertical coordinate.

2.2 Vertical Disparity

For the stereo calibration accuracy measure, we assume rectified image pairs. In this configuration, the correspondence analysis can be reduced to a limited search space. We make use of a classical combination of scale invariant keypoint detection and descriptor computation and select A-KAZE keypoints and descriptors [1] since they are known for the highly accurate subpixel localization [7]. Recent methods, e.g., [10], using machine learning for the correspondence analysis have shown their strength in the keypoint matching, even for challenging scenarios, but not in localization accuracy.

For rectified stereo images, corresponding image points in the left and right image should be located in the same *scanline*, i.e., they have the same y-coordinate and their positions only differ by a horizontal disparity d_x. We assume a small, but non-zero vertical offset d_y as shown in Fig. 2. Since the search region is small, the probability of outliers, i.e., wrongly established correspondences is small. We will later see that there are a few outliers, but they are negligible for the evaluation.

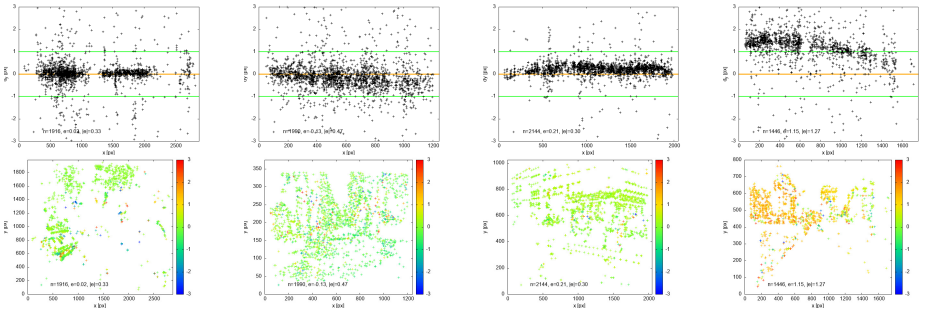
For each corresponding feature point pair $\mathbf{p}_l, \mathbf{p}_r$, the disparity $\Delta\mathbf{d} = (d_x, d_y)^t = \mathbf{p}_r - \mathbf{p}_l$ is determined. We evaluate the mean of vertical disparities for all corresponding n features points:

$$\epsilon = \frac{1}{n}\sum_{i=1}^{n} d_y^{(i)} \tag{1}$$

Additionally, the mean of the absolute vertical disparities for all corresponding n features points is considered:

$$|\epsilon| = \frac{1}{n}\sum_{i=1}^{n} |d_y^{(i)}| \tag{2}$$

While the ϵ in Eq. (1) indicates the systematic offset resulting from the calibration inaccuracies, $|\epsilon|$ in Eq. (2) provides the error magnitude. Clearly, inaccuracies in the feature point positions and mismatched feature points will add a bias which do not belong to a calibration issue. The ϵ in Eq. (1) should be rather unaffected from feature correspondence inaccuracies since we can assume that their errors mitigate. The $|\epsilon|$ in Eq. (2) sums up all unwanted error contributions and, therefore, is our proposed measure



(a) Middlebury 2014: Adirondack **(b)** *KITTI*: $i = 18$ from *0011* **(c)** *Cityscapes*: $i = 3$ from *Aachen* **(d)** *DrivingStereo 2018-10-11*, $i = 369$

Fig. 3: Exemplary visualizations of the vertical disparities d_y over x-coordinate (top row) and the color-coded d_y for each keypoint position (x/y) in the image (bottom row). The examples show small mean vertical disparities ϵ such as $\epsilon = 0.02$ for *Middlebury* (Fig. 3a), $\epsilon = 0.13$ for *KITTI* (Fig. 3b), and $\epsilon = 0.21$ for *Cityscapes* (Fig. 3c). For the *DrivingStereo* example image pair, a larger $\epsilon = 1.15$ is obtained. A large ϵ, indicates a systematic error induced by the suboptimal stereo calibration parameters (cf. Eq. (1)). This is visualized by the deviation from zero (top row) and by red / blue color (bottom row). See also the example in Fig. 1.

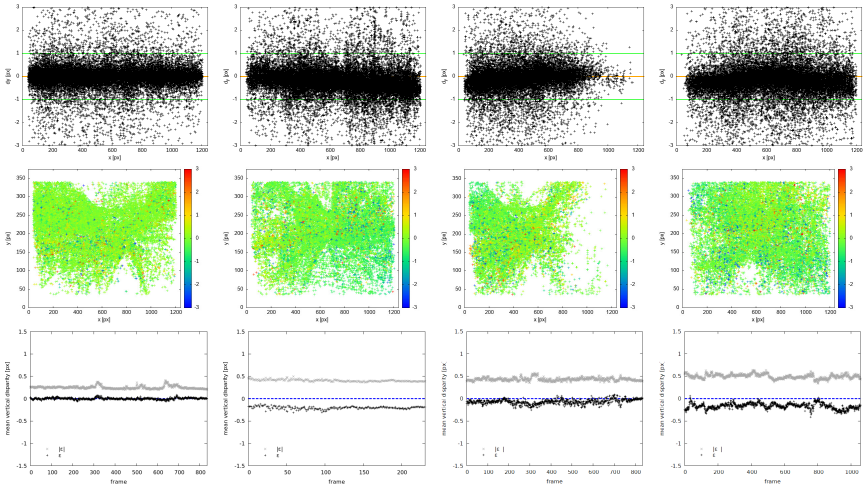
for the stereo calibration accuracy. For simplicity, we measure d_y in pixels and keep in mind that the image resolution has an influence on the resulting error in the respective application.

3 Accuracy Evaluation of Stereo Vision Datasets

To quantify the stereo calibration accuracy, the mean of the vertical disparities ϵ (Eq. (1)) and the mean of the absolute vertical disparities $|\epsilon|$ (Eq. (2)) as derived in Sec. 2.2 are used. For the correspondence analysis, A-KAZE keypoints and descriptors [1] are used due to the superior subpixel accuracy of detected keypoints. Visualizations of the vertical disparities for a single stereo image pair are shown in Fig. 1b and in Fig. 3. The examples in Fig. 3 show optimal calibration (Fig. 3a), small calibration errors (Figs. 3b and 3c), and slightly increased errors (Fig. 3d). In Fig. 1b, large errors occur.

It is known that the spatial distribution of feature points of a detector is dependent on the image content [12]. To account for different local distributions of the feature point positions, several images of a dataset sequence are considered, and the results are accumulated. The aim is to provide a dense spatial feature distribution for visualizations as shown in Fig. 3. As a compromise, we visualize 25 images of a sequence sampled equidistantly over time for Figs. 4 to 7, top and center row.

The following *Stereo Vision* datasets are considered: *Virtual KITTI 2* (serves as a baseline), *KITTI*, *Cityscapes*, *DSEC*, *DrivingStereo*, *ApolloScape*, and *Argoverse 1.1*. From each dataset, representative sequences, i.e., with varying recording dates, are selected for the evaluation. In Secs. 3.1 and 3.2, the results are visualized (Figs. 4 to 7) and discussed. Tab. 2 summarizes the calibration errors and provides the comparison between the selected stereo datasets.



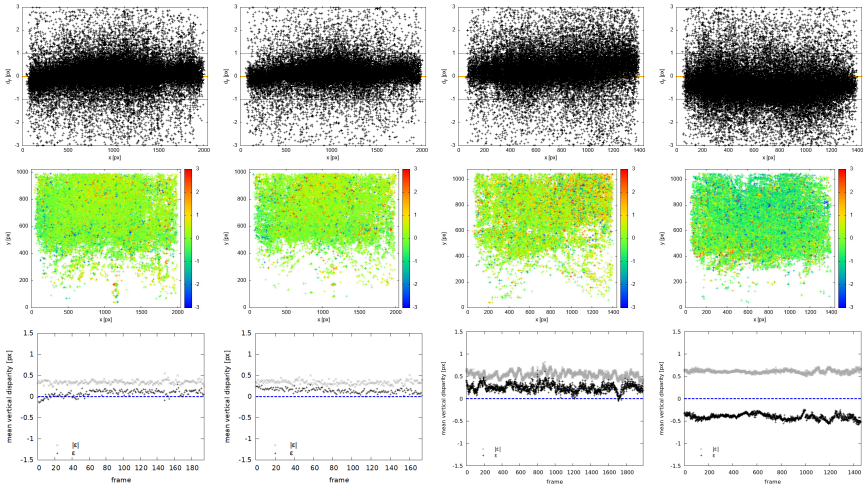
(a) VKITTI - Scene 0020 **(b)** KITTI - Scene 0011 **(c)** KITTI - Scene 0047 **(d)** KITTI - Scene 0071

Fig. 4: Visualization of the vertical disparities d_y for training sequences from *Virtual KITTI 2* (Fig. 4a) and *KITTI* (Figs. 4b to 4d). For Fig. 4a, there is no calibration error and d_y should be near zero for all positions (top) and have green color (center). The mean values ϵ for the full sequence are near zero (bottom). The *KITTI* examples show only small deviations and a slight imbalance (top row). Consequently, nearly all points in the diagrams in the center row are green. The mean values (bottom row) show vertical disparities below 0.25 px.

In Sec. 4, a method for the optimization of the stereo calibration parameters is proposed and demonstrated. The optimized parameters are used to generate new rectified images and lead to significantly increased stereo calibration accuracy validated with ϵ and $|\epsilon|$. The proposed optimization procedure can be used for *Online Calibration*. The idea is to modify the calibration if camera parameters change due to small changes in their relative orientation during the capturing process. The optimization procedure adapts the current camera parameters, and the time-consuming recalibration of the camera system with calibration patterns can be avoided.

3.1 Experimental Results

The analysis results are visualized in Figs. 4 to 7. The diagrams on top show the vertical disparities d_y over x-coordinate (left image) for corresponding keypoints. The center rows show the color-coded d_y for each position (x/y) in the respective image. These two visualizations employ accumulated results of 25 equidistantly sampled stereo image pairs of a sequence. The accumulated visualizations shown that the vertical disparity results are independent of image content. Some outliers are visible as points with a large color distance to their neighbours (center row). The number of outliers is small, and they are neglected for the analysis. The bottom row of Figs. 4 to 7 shows the error measures ϵ and $|\epsilon|$ for all images of a dataset sequence. In Tab. 2, mean and variance



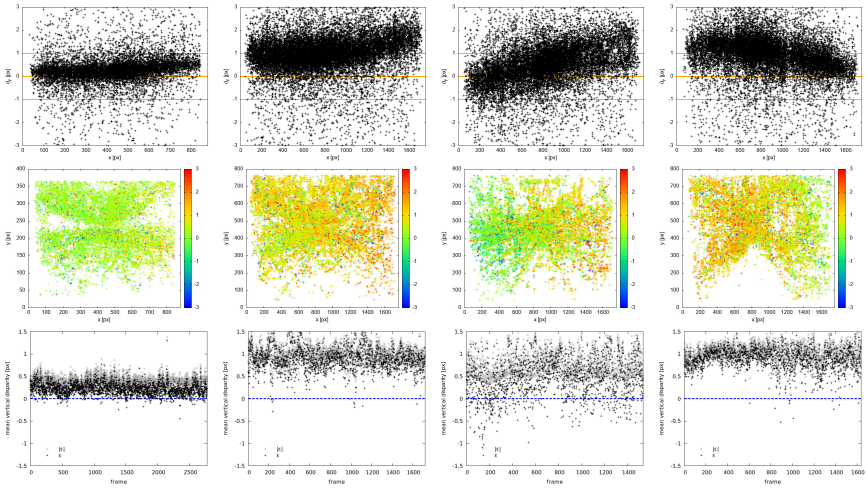
(a) Cityscapes - Stuttgart **(b)** Cityscapes - Aachen **(c)** DSEC - Interlaken 00e **(d)** DSEC - ZurichCity 00b

Fig. 5: Visualization of the vertical disparities d_y and their mean values for training sequences from *Cityscapes* (Figs. 5a and 5b) and *DSEC* (Figs. 5c and 5d). For *Cityscapes*, small systematic errors occur as small deviations for ϵ from $y = 0$. For the two *DSEC* sequences, larger errors occur with different shape and $\epsilon < 0$ for Fig. 5c versus $\epsilon > 0$ for Fig. 5d. Compared *KITTI* in Fig. 4, *Cityscapes* shows similar error magnitude while the error of *DSEC* is larger.

of ϵ for the selected sequences are listed. The results are discussed separately for each dataset in the following paragraphs.

Virtual KITTI 2: no systematic error This sequence serves as a baseline. The results for the *Virtual KITTI 2 - scene 20* are shown in Fig. 4a. Since this is a synthetic dataset, the calibration is correct. Thus, errors are caused by inaccurate keypoint localization and mismatched keypoints. The mean vertical disparity ϵ is near zero for all 836 image pairs of the sequence. The mean over all images is small (0.083 px), cf. Tab. 2. Thus, there is no systematic error in the stereo calibration. The mean for the absolute vertical disparity $|\epsilon|$ is 0.391. This is surprisingly high and shows that the localization accuracy of the features detector is not optimal for synthetic image content with a lack of distinctive texture details. Nevertheless, the results demonstrate a diminishing systematic error. The proposed metric provides reasonable output. The result of the sequence *scene 20* is representative for the whole dataset *Virtual KITTI 2*.

KITTI: low error, mean vertical disparity below 0.25 px The KITTI dataset includes new calibration parameters for each recording. Our results in Figs. 4b to 4d show small mean vertical disparities with a similar magnitude throughout the sequences. For sequence 0011 (Fig. 4b), the mean of ϵ for the sequence of 232 images is 0.203 px while the mean of $|\epsilon|$ for this sequence is 0.455 px (cf. Tab. 2). Both examples 0047 (Fig. 4c) and 0071 (Fig. 4d) show similar results with low error. The mean vertical disparity ϵ



(a) DrivingStereo - 2018-07-09 **(b)** DrivingStereo - 2018-07-11 **(c)** DrivingStereo - 2018-08-01 **(d)** DrivingStereo - 2018-10-11

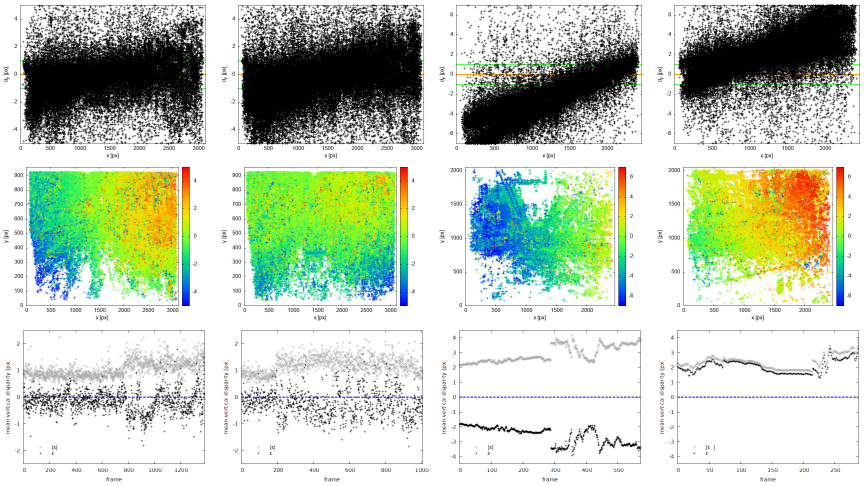
Fig. 6: Visualization of the vertical disparities d_y and their mean values for training sequences from *DrivingStereo*. The diagrams show different error shapes. Compared to *DSEC* (Figs. 5c and 5d), *DrivingStereo* shows similar error magnitude.

is below 0.25 px for all images. The error is independent of the image region. The error magnitude is representative for the whole dataset. Thus, the KITTI dataset provides high quality relative stereo calibrations.

Cityscapes: low error, mean vertical disparity below 0.25 px The *Cityscapes* dataset includes new calibration parameters for each recording. We show results for two sequences in Figs. 5a and 5b. Both provide small errors with a magnitude similar to the KITTI example (mean vertical disparities below 0.25 px, no dependency on image region). We verified the result for many sequences of the set. The *Cityscapes* dataset provides high quality relative stereo calibrations.

DSEC: varying error up to 1.5 px The *DSEC* dataset targets the validation of stereo event cameras. It includes 53 sequences in three different areas of Switzerland. Five different calibration sets are provided. In Figs. 5c and 5d, example sequences from two different sets are visualized. While the magnitude of the measured vertical disparities are similar, the structure is different. For *Interlaken*, we have vertical disparities $d_y < 0$ for all image positions while obtaining $d_y > 0$ for all image positions in *Zurich City*. The error magnitude is up to 1.5 px which is mediocre accuracy for the datasets considered in this paper. For demonstration, we optimize one set of camera parameters (*DSEC - ZurichCity 00b*) and show the results in Sec. 4.

DrivingStereo: varying error up to 1.5 px The *DrivingStereo* dataset was recorded on 42 different days in 3 recording periods {2018-07, 2018-08, 2018-10}. Three dif-



(a) ApolloScape - Trainset 01 **(b)** ApolloScape - Testset **(c)** Argoverse 1.1 - Trainset, parts 06, 07 **(d)** Argoverse 1.1 - Trainset, parts 10, 11

Fig. 7: Visualization of the vertical disparities d_y and their mean values for sequences from *ApolloScape* and *Argoverse*. Large vertical disparities in positive (red) and in negative direction (blue) are clearly visible. Compared to the previous visualizations (Figs. 4 to 6), *ApolloScape* and *Argoverse* show larger errors. The Argoverse examples show different error shape with $\epsilon < 0$ for Fig. 7c versus $\epsilon > 0$ for Fig. 7d.

ferent calibrations are provided for the sequences of the dataset. In Fig. 6, we show the evaluation results for sequences from different periods of *DrivingStereo*. The evaluation for the vertical error shows differing results for these sequences. Sequence 2018-07-09 is the first sequence of one of the three recording periods and shows mean disparity values of < 0.6 px, cf. Fig. 6a. Two days later (Fig. 6b), a significant increase of the vertical disparities occurs (up to 1.5 px). Likewise, the other two examples (Figs. 6c and 6d) show mean vertical disparities of up to 1.5 px. We can infer that a more frequent calibration would have been advantageous for the stereo calibration accuracy of the dataset. Additionally, the mean values have a large variation throughout a sequence, verified by large standard deviation compared to *KITTI* and *Cityscapes* as shown in Tab. 2. The cameras have a jitter in their relative orientation during the drive. As documented in [33], the stereo system consists of two single cameras, one is mounted on the top center, and the other mounted on the top right. The setup is similar to the *KITTI* setup but there, the jitter does not occur.

ApolloScape: varying, large errors up to 5px The *ApolloScape* dataset targets 360° view with six video cameras. The two front facing cameras build the stereo perception. Our analysis of the provided rectified stereo images are visualized in Figs. 7a and 7b. Both examples show similar magnitude and structure of the resulting vertical disparities with a magnitude of up to 5px (top row). The largest errors are visible in the bottom third ($d_y < 0$) and in the top right corner ($d_y > 0$) of the image, cf. center row. The

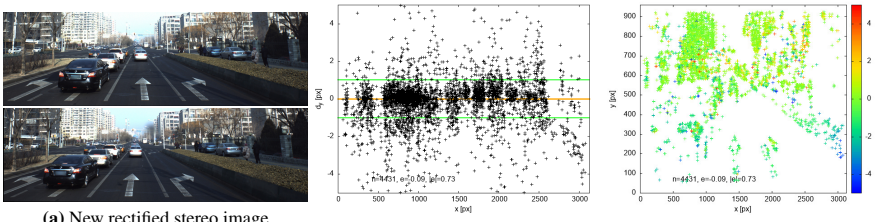
Table 2: Mean and variance of ϵ and $|\epsilon|$ for selected sequences. While the mean results of ϵ for the *Cityscapes* sequences are similar and near zero, very different values are obtained for other sequences, in particular for the Argoverse examples. This indicates a change in the camera configuration which is not captured by the provided calibration. The deviation from zero shows the amount of the calibration error in pixels. A small standard deviation shows the stability of the proposed method for calibration error measurement throughout the respective stereo image sequence. The *Virtual KITTI 2* results demonstrate the baseline for the measure.

Dataset	Sequence	ϵ		$\lvert\epsilon\rvert$		Visualization
		mean	std.dev	mean	std.dev	
Virtual KITTI 2	Scene 20	0.083	0.077	0.391	0.056	Fig. 4a
KITTI	Drive 0011	−0.203	0.354	0.455	0.027	Fig. 4b
KITTI	Drive 0047	−0.069	0.054	0.432	0.036	Fig. 4c
KITTI	Drive 0071	−0.167	0.063	0.498	0.048	Fig. 4d
Cityscapes	Stuttgart	0.083	0.077	0.391	0.056	Fig. 5a
Cityscapes	Aachen	0.146	0.056	0.381	0.061	Fig. 5b
DSEC	Interlaken 00e	0.279	0.095	0.642	0.097	Fig. 5c
DSEC	Zurich City 00b	−0.411	0.071	0.676	0.044	Fig. 5d
DrivingStereo	2018-07-09	0.229	0.136	0.467	0.104	Fig. 6a
DrivingStereo	2018-07-11	0.944	0.288	1.100	0.242	Fig. 6b
DrivingStereo	2018-08-01	0.494	0.494	0.840	0.299	Fig. 6c
DrivingStereo	2018-10-11	0.965	0.272	1.155	0.207	Fig. 6d
ApolloScape	TrainSet 01	−0.181	0.412	1.056	0.327	Fig. 7a
ApolloScape	TestSet	−0.268	0.490	1.202	0.315	Fig. 7b
ArgoVerse 1.1	TrainSet pt. 06,07	−2.583	0.566	2.904	0.553	Fig. 7c
ArgoVerse 1.1	TrainSet pt. 10,11	2.073	0.44	2.366	0.451	Fig. 7d

teaser figure Fig. 1 depicts the same tendency. Like in the results of *DrivingStereo*, jitter occurs in the ApolloScape stereo data resulting in a standard deviation of larger that 0.4 (Tab. 2).

For demonstration, we optimize one set of camera parameters and show the results in Sec. 4. Our analysis and the rectified images in Fig. 1 are used to estimate optimized camera parameters.

ArgoVerse 1.1: varying, large errors up to 5px The *ArgoVerse 1.1* dataset provides surround view video with seven ring cameras. Additional stereo camera systems are provided. Although the focus is on surround view, there are challenges and leaderboards using the rectified stereo images. Our results for the vertical disparities of these images are visualized in Figs. 7c and 7d. The top row depicts a slanted distribution. The vertical disparity clearly increases with the x-coordinate. For Fig. 7c, most d_y lie below the x-axis, for Fig. 7d, most d_y are above the x-axis. Thus, large regions in the visualizations in the second row are colored red and blue. The sequence evaluations (bottom row) show the different shape with mean values $\epsilon > 0$ and $\epsilon < 0$ for different sections of the train set. The resulting disparities have the largest standard deviation (cf. Tab. 2) in the test field. We infer that the relative camera orientation changed after the initial calibration and that the change has not been covered by an adapted stereo



(a) New rectified stereo image

(b) Vertical disparity error for corresponding image points.

Fig. 8: Error analysis of the optimized parameters. Here, the rectified images (Fig. 8a) are computed based on the optimized camera parameters from our optimization procedure (Sec. 4). The vertical disparity errors d_y decreased significantly (Fig. 8b). The remaining vertical disparities are distributed around zero independent of the image coordinate (Fig. 8b, left). Thus, nearly all data points have green color (Fig. 8b, right).

calibration. We optimize one set of camera parameters (*TrainSet pt.06*) and show the results in Sec. 4.

3.2 Summary and Discussion

The proposed measure for vertical disparity reveals the misalignment of rectified images of a stereo camera system. Since all considered datasets provide rectified images, the causes for the errors are erroneous extrinsic camera parameters. The considered datasets show varying error magnitude and shape (cf. Figs. 4 to 7 and Tab. 2). In some examples, the result varies largely within a dataset, e.g., Fig. 6a versus Fig. 6b and Fig. 7c versus Fig. 7d.

The error magnitude is small for *KITTI* and *Cityscapes*, medium for *DrivingStereo* and *DSEC*, and rather large for *ApolloScape* and *Argoverse 1.1*. In the latter cases, classical stereo algorithms suffer from the biased rectified images. For *Machine Learning* based approaches, an erroneous stereo configuration is learned limiting the transferability of the resulting model. These errors are larger than expected, even when considering the larger image resolutions of the newer datasets.

Using the measure for stereo calibration accuracy, the calibration parameters can be optimized using the error as cost function. For the optimization, camera parameters are treated as variables and the mean vertical disparity in Eq. (2) is minimized. In case of convergence, the parameters are considered as improved and new rectified images are generated. We follow this idea in Sec. 4.

4 Optimization of the Stereo Calibration

After demonstrating that the proposed methodology yields a valid measure for stereo calibration accuracy, we now use it for the optimization of the calibration parameters. A stereo calibration with minimal vertical disparity as expressed in Eq. (2) shall be favoured.

In three experiments, we derive improved parameters based on the rectified image pairs for

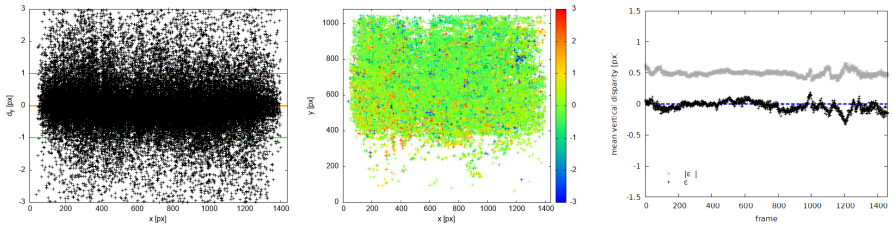


Fig. 9: Visualizations of the optimized parameters for the *DSEC - ZurichCity 00b* sequence (cf. Fig. 5d). The systematic error decreased significantly.

(a) the *ApolloScape* example image pair used in Fig. 1,
(b) the *DSEC* ZurichCity 00b sequence example in Fig. 5d, and
(c) the *Argoverse 1.1 Trainset part 06* example in Fig. 7c.

For the optimization of camera parameters, we aim at minimizing the mean vertical disparity in Eq. (2). Therefore, the *Differential Evolution* (DE) algorithm [25] is used. It is known as a simple and efficient global optimization method for continuous problem spaces. Here, we only optimize 3 parameters (angles of the right camera) within reasonable boundaries. We assume that the stereo camera baseline is correct, i.e., the relative position between left and right camera is unchanged. Thus, we have an easy optimization problem which quickly convergences to a solution with the desired camera angles. The resulting relative angles for the stereo cameras in the three experiments are as follows:

(a) $(pan, tilt, roll) = (0.00832°, 0.00999°, 0.00053°)$
(b) $(pan, tilt, roll) = (0.16591°, -0.00286°, 0.12576°)$
(c) $(pan, tilt, roll) = (-0.00027°, 0.02195°, 0.16709°)$

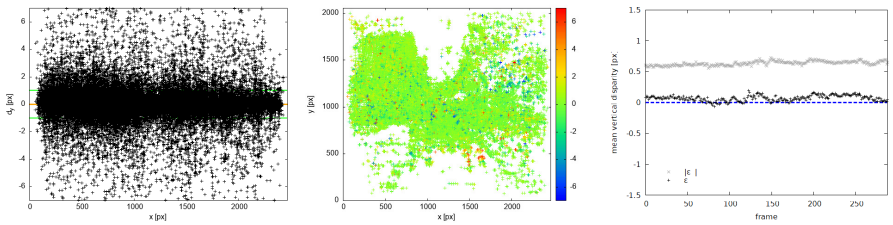


Fig. 10: Visualizations of the optimized parameters for the *Argoverse 1.1 - Trainset, part 06* (cf. Fig. 7c, left part). The systematic error decreased significantly.

With the derived camera parameters, new rectified images are generated. For (a), the new rectified images are shown in Fig. 8a. The difference to Fig. 1a is tiny, but its impact is huge. For validation, the error measures for these images are shown in Fig. 8b for experiment (a), in Fig. 9 for experiment (b), and in Fig. 10 for experiment (c).

For experiment (b) and (c), the full sequence is evaluated with 1462 and 289 images, respectively. Compared to the original versions (Fig. 1, Fig. 5d, Fig. 7c, left part), a significant decrease of the vertical disparity error is visible. These experiments show the practicability of the proposed measure for the relative recalibration of a stereo camera. This procedure can be used to adjust the extrinsic parameters during acquisition (*Online-calibration*) since no calibration patterns are needed.

5 Conclusions

The calibration accuracy of high-impact *Stereo Vision* datasets is evaluated using the proposed measure for misalignment of the rectified images. Therefore, a keypoints correspondence analysis with high localization accuracy is employed. From the keypoints, vertical disparities are computed. For rectified images computed with accurate calibration, the vertical disparity is zero. For several *Stereo Vision* datasets, we obtain systematic errors.

We evaluate the stereo calibration accuracy of datasets targeting automated driving: *KITTI*, *Cityscapes*, *DSEC*, *DrivingStereo*, *ApolloScape*, and *Argoverse 1.1*. The data from *Middlebury 2014* and *Virtual KITTI 2* serve as baselines. The comparison shows varying accuracies and error shapes. While *KITTI* and *Cityscapes* provide reasonable accuracies, *ApolloScape* and *Argoverse 1.1* show surprisingly large errors.

We infer that small changes in the camera orientation are likely to occur during data acquisition, especially when there are large temporal distances between recordings. These situations are found quite often in the considered datasets. As a conclusion, we recommend frequent on-site recalibration when benchmark data is generated. The proposed accuracy measure provides the possibility for error control and indicates the need for a recalibration. Furthermore, an optimization scheme is proposed which improves the stereo camera calibration. Rectified images computed with the new calibration parameters do not show a systematic error. The proposed methodology enables *Online Calibration* since calibration patterns are not needed.

More visualizations of the results are shown in our demonstration video: https://youtu.be/QDXiGDdth1o.

References

1. Alcantarilla, P.F., Solutions, T.: Fast explicit diffusion for accelerated features in nonlinear scale spaces. IEEE Trans. Patt. Anal. Mach. Intell **34**(7), 1281–1298 (2011) 5, 6
2. Bansal, M., Jain, A., Camus, T., Das, A.: Towards a practical stereo vision sensor. In: Conference on Computer Vision and Pattern Recognition (CVPR) -Workshops. pp. 63–63. IEEE (2005) 2
3. Cabon, Y., Murray, N., Humenberger, M.: Virtual KITTI 2. arXiv preprint arXiv:2001.10773 (2020) 4
4. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., et al.: Argoverse: 3d tracking and forecasting with rich maps. In: IEEE/CVF Conference on computer vision and pattern recognition (CVPR). pp. 8748–8757 (2019) 1, 2, 4

5. Chen, Z., Long, W., Yao, H., Zhang, Y., Wang, B., Qin, Y., Wu, J.: Mocha-stereo: Motif channel attention network for stereo matching. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27768–27777 (2024) 2
6. Cheng, X., Zhong, Y., Harandi, M., Dai, Y., Chang, X., Li, H., Drummond, T., Ge, Z.: Hierarchical neural architecture search for deep stereo matching. Advances in Neural Information Processing Systems **33**, 22158–22169 (2020) 2
7. Cordes, K., Grundmann, L., Ostermann, J.: Feature evaluation with high-resolution images. In: Azzopardi, G., Petkov, N. (eds.) Computer Analysis of Images and Patterns. pp. 374–386. Springer International Publishing (2015) 5
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2, 4
9. Dang, T., Hoffmann, C., Stiller, C.: Continuous stereo self-calibration by camera parameter tracking. IEEE Transactions on image processing **18**(7), 1536–1550 (2009) 2, 3
10. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: IEEE conference on computer vision and pattern recognition (CVPR) workshops. pp. 224–236 (2018) 5
11. Diaz-Ruiz, C.A., Xia, Y., You, Y., Nino, J., Chen, J., Monica, J., Chen, X., Luo, K., Wang, Y., Emond, M., et al.: Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21383–21392 (2022) 2
12. Dickscheid, T., Schindler, F., Förstner, W.: Coding images with local features. International journal of computer vision **94**, 154–174 (2011) 6
13. Fan, R., Wang, L., Bocus, M.J., Pitas, I.: Computer stereo vision for autonomous driving. arXiv preprint arXiv:2012.03194 (2020) 1
14. Furgale, P., Rehder, J., Siegwart, R.: Unified temporal and spatial calibration for multi-sensor systems. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1280–1286. IEEE (2013) 2, 4
15. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. IEEE Robotics and Automation Letters **6**(3), 4947–4954 (2021) 2, 4
16. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3354–3361. IEEE (2012) 2, 4
17. Geiger, A., Moosmann, F., Car, O., Schuster, B.: A toolbox for automatic calibration of range and camera sensors using a single shot. In: International Conference on Robotics and Automation (ICRA) (2012) 1, 4
18. Guo, X., Lu, J., Zhang, C., Wang, Y., Duan, Y., Yang, T., Zhu, Z., Chen, L.: Openstereo: A comprehensive benchmark for stereo matching and strong baseline. arXiv preprint arXiv:2312.00343 (2023) 2
19. Hirschmüller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In: Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2, pp. 807–814. IEEE (2005) 2, 4
20. Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., Yang, R.: The apolloscape open dataset for autonomous driving and its application. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2702–2719 (2019) 2, 3, 4
21. Kolmogorov, V., Zabih, R.: Computing visual correspondence with occlusions using graph cuts. In: International Conference on Computer Vision (ICCV). vol. 2, pp. 508–515 (2001) 2
22. Kruger, L.E., Wohler, C., Wurz-Wessel, A., Stein, F.: In-factory calibration of multiocular camera systems. In: Optical Metrology in Production Engineering. vol. 5457, pp. 126–137. SPIE (2004) 2, 4

23. Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: International Conference on 3D Vision (3DV). pp. 218–227. IEEE (2021) 2
24. Loop, C., Zhang, Z.: Computing rectifying homographies for stereo vision. In: Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. 125–131. IEEE (1999) 3
25. Price, K.V., Storn, R., Lampinen, J.A.: Differential Evolution - A Practical Approach to Global Optimization. Natural Computing Series, Springer (2005) 13
26. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: German Conference on Pattern Recognition (GCPR). pp. 31–42. Springer (2014) 1, 2, 4
27. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant CNNs. In: international conference on 3D Vision (3DV). pp. 11–20. IEEE (2017) 4
28. Wang, X., Xu, G., Jia, H., Yang, X.: Selective-stereo: Adaptive frequency information selection for stereo matching. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19701–19710 (2024) 2
29. Weinzaepfel, P., Lucas, T., Leroy, V., Cabon, Y., Arora, V., Brégier, R., Csurka, G., Antsfeld, L., Chidlovskii, B., Revaud, J.: Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In: IEEE/CVF International Conference on Computer Vision (CVPR). pp. 17969–17980 (2023) 2
30. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., et al.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. arXiv preprint arXiv:2301.00493 (2023) 1, 4
31. Xu, G., Wang, X., Ding, X., Yang, X.: Iterative geometry encoding volume for stereo matching. arXiv preprint arXiv:2303.06615 (2023) 2
32. Xu, P., Xiang, Z., Qiao, C., Fu, J., Pu, T.: Adaptive multi-modal cross-entropy loss for stereo matching. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5135–5144 (2024) 2
33. Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., Zhou, B.: Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 899–908 (2019) 2, 4, 10
34. Zbontar, J., LeCun, Y., et al.: Stereo matching by training a convolutional neural network to compare image patches. J. Mach. Learn. Res. **17**(1), 2287–2318 (2016) 1
35. Zeng, J., Yao, C., Wu, Y., Jia, Y.: Temporally consistent stereo matching. arXiv preprint arXiv:2407.11950 (2024) 2