The final version is available at http://dx.doi.org/10.1109/ICIP.2018.8451500

# A NOVEL CONFIDENCE MEASURE FOR DISPARITY MAPS BY PIXEL-WISE COST FUNCTION ANALYSIS

*Ron Op het Veld*[⋆]*, Tobias Jaschke*[⋆]*, Michel Bätz*[⋆]*, Luca Palmieri*[†] *and Joachim Keinert*[⋆]

[⋆] Department Moving Picture Technologies, Fraunhofer IIS, Erlangen, Germany
[†] Department of Computer Science, Kiel University, Germany

## ABSTRACT

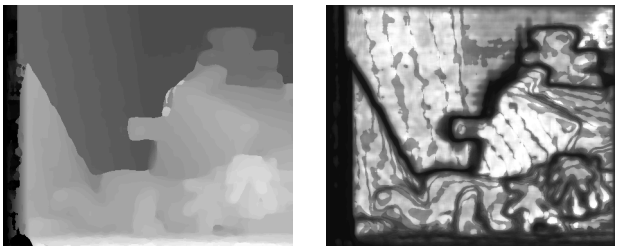
Disparity estimation algorithms mostly lack information about the reliability of the disparities. Therefore, errors in initial disparity maps are propagated in consecutive processing steps. This is in particularly problematic for difficult scene elements, e.g., periodic structures. Consequently, we introduce a simple, yet novel confidence measure that filters out wrongly computed disparities, resulting in improved final disparity maps. To demonstrate the benefit of this approach, we compare our method with existing state-of-the-art confidence measures and show that we improve the ability to detect false disparities by $54.2\%$.

***Index Terms*—** Stereo Vision, 3D reconstruction, disparity estimation, confidence measure, CNN.

## 1. INTRODUCTION

Stereo disparity estimation is one of the most researched and active fields within computer vision. This is mainly because estimated disparities using existing algorithms are not accurate enough and the computational costs are often too high [1, 2, 3]. In recent years, deep-learning methods increased the accuracy of such algorithms [4, 5]. Overall, more accurate disparity maps can improve the results for depth-image-based-rendering methods. Initial disparity maps are mostly computed from two stereo images, and later combined with disparity maps computed from other camera pairs. Fusion of multiple disparity seems to be straightforward, however, due to false disparities, it is not. False disparities are being propagated and thus result in unreliable disparity maps. We propose a new confidence measure to filter out these initially false disparities.

In this paper, we introduce a novel confidence measure based on conventional approaches [6, 7, 8, 9, 10, 11, 12], in which confidences are assigned by examining the cost curves. When we assume that truly corresponding pixels have the minimal matching cost, the ideal cost curve as a function of disparity for a pixel, has a single, distinct minimum. However, most cost curves are ambiguous because they have multiple minima or multiple adjacent disparities with similar costs, making localization of the exact disparity hard. The shape of the cost curve heavily depends on the stereo algorithm used. Some algorithms tend to be more sensitive to noise and produce cost curves without an easy to distinguish global minimum. We examine our proposed method and compare it to two state-of-the-art methods according to their ability to rank potential matches. As Spyropoulos and Mordohai already envisioned in their paper [13], we use the stereo method developed by Žbontar and Le Cun [4] to compute the matching costs. They trained a convolutional neural network (CNN) to predict whether two image patches match or not.



(a) Estimated disparities (b) Our confidences

**Fig. 1**. (a) Estimated disparities MC-CNN + box-filter (brighter is closer) for Teddy (Fig. 2 (a)) image from MB03 and (b) the confidences from our proposed confidence measure (brighter is higher, scaled for better visualization).

The remainder of this paper is structured as follows. In Section 2, we list related work, while in Section 3, our proposed method is introduced. An in-depth discussion about the experimental results is provided in Section 4, followed by the conclusion and future work in Section 5.

## 2. RELATED WORK

In recent years, many confidence measures aiming at detecting unreliable disparity assignments, proved to be very effective cues when combined with state-of-the-art stereo algorithms [5, 14, 15, 16]. The ability to reliably detect failures of a stereo algorithm by means of a confidence measure is

⋆ron.ophetveld@iis.fraunhofer.de
The work in this paper was funded from the European Unions Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 676401, European Training Network on Full Parallax Imaging.

fundamental and many approaches have been proposed for this purpose. Hu and Mordohai [17] were the first ones to exhaustively review and compare confidence measures available at that time and defined an effective metric to evaluate the performance of the different measures. New confidence measures have been introduced and evaluated, most measures are based on deep learning [5, 18] and other machine learning methods [13, 14, 15, 16, 19]. The latest thorough evaluation of 76 state-of-the-art confidence measures has been performed by Poggi *et al.* [20] in 2017. In this, a similar trend as in the evaluation of stereo algorithms can be seen, machine learning based approaches clearly outperform conventional approaches.

Based on previous evaluations, we selected the state-of-the-art confidence measures Left-Right Difference (**LRD**) [17] and Confidence CNN (**CCNN**) [18], and compare their performance to the performance of our proposed method. Following [17], to better clarify which cues are processed by each single measure we use the following notation.

Given a stereo pair of rectified left ($L$) and right ($R$) images, we compute the cost volumes $c_{\{\mathrm{R,L}\}}(x, y, d)$ that contain cost values for each possible match within the defined disparity range from a pixel in the left image $I_L(x_L, y)$ to a pixel in the right image $I_R(x_R, y)$, and vice versa. Disparity is defined conventionally as $d = x_L - x_R$. The minimum and maximum disparity values, $d_{\text{min}}$ and $d_{\text{max}}$, are provided by the dataset (Section 4.1). Thus, the cost curve of a pixel is the set of cost values for all allowable disparities for the pixel. $c_{\{\mathrm{R,L}\}1}(x, y)$ and $c_{\{\mathrm{R,L}\}2}(x, y)$ indicate the minimum and second minimum values of the cost curve, respectively, $c_{\{\mathrm{R,L}\}2}(x, y)$ does not have to be a local minimum. The disparity value $d(c_{\{\mathrm{R,L}\}1}(x, y))$ is denoted by $d_{\{\mathrm{R,L}\}1}(x, y)$.

We will now describe the two state-of-the-art methods in more details. The **LRD** confidence measure $C_{\mathrm{LRD}}(x_L, y)$ favors a large margin between the two smallest minima of the cost for pixel $(x_L, y)$ in the left image. Also consistency of the minimum costs between the left-to-right and right-to-left disparity maps:

$$C_{\mathrm{LRD}}(x_{\mathrm{L}}, y) = \frac{c_{\mathrm{L2}}(x_L, y) - c_{\mathrm{L1}}(x_L, y)}{|c_{\mathrm{L1}}(x_L, y) - c_{\mathrm{R1}}(x_{\mathrm{R}}, y)| + \epsilon}, \qquad (1)$$

with $x_{\mathrm{R}} = x_{\mathrm{L}} - d(x_{\mathrm{L}}, y)$ and $\epsilon$ a very small value to avoid zero-devision.

The intuition is that truly corresponding pixels should result in similar cost values and thus a small denominator. This formulation provides safeguards against two failure modes. If the margin $c_{\mathrm{L2}}(x_L, y) - c_{\mathrm{L1}}(x_L, y)$ is large, but the pixel has been mismatched the denominator will be large. If the margin is small, the match is likely to be ambiguous. In this case, a small denominator indicates that a correspondence between two similar pixels has been established. According to [17], LRD is one of the best overall confidence measures for stereo inputs.

As a second confidence measure, we use **CCNN**. In this approach, confidence prediction is regressed by a CNN without extracting any cue from the stereo input images. The deep network, trained on patches, learns from scratch a confidence measure by processing only the left disparity map, normalized with respect to the maximum disparity, to values between zero and one. For the evaluation we used the source code provided by the authors (using 8 bit confidence maps). This confidence measure has been identified by Poggi *et al.* [20] as the best performing one. However, training of such a neural network is an additional issue.

We evaluate these confidence measures using the stereo method Matching Cost Convolutional Neural Network (**MC-CNN**) developed by Žbontar and Le Cun [4]. An eight-layer network is trained on pairs of patches to compute a measure of similarity between them. These outputs represent matching scores for every possible disparity of each pixel. The scores are adaptively aggregated [21] and optimized using semi-global matching (SGM) to obtain the highly ranked results on the KITTI benchmark [22]. Žbontar and Le Cun proposed an accurate architecture and a faster/simplified one, skipping cross-based aggregation. The latter showed a remarkable speed-up with respect to the accurate CNN architecture (0.8 sec vs 67 sec) with an increase of the error rate smaller than $1\%$ on both KITTI datasets. We compute our cost volumes using the code provided by the authors, using their fast architecture. We use the network that is pre-trained on the KITTI 2012 dataset [22], which is different from our test set, to avoid a biased evaluation.

## 3. PROPOSED METHOD

Our proposed method computes a confidence solely based on the cost curve for each pixel in the disparity map. The confidence value for each pixel indicates how likely the assigned disparity is correct. Our confidence measure is defined as

$$C(x, y) = \frac{1}{\sum_{d=d_{\text{min}}}^{d_{\text{max}}} \frac{\max(\min(\Delta d(x,y,d)-1, \frac{d_{\text{max}}-d_{\text{min}}}{3}), 0)^2}{\max(\Delta c(x,y,d) - \frac{c_{\text{mean}}(x,y)}{3}, 1)}}, \qquad (2)$$

with $\Delta d(x, y, d) = |d - d_1(x, y)|$ and $\Delta c(x, y, d) = c(x, y, d) - c_1(x, y)$.

Multiple local minima, corresponding to multiple small values of $\Delta c(x, y, d)$, in the cost curve indicate uncertainty about the disparity value of the pixel, therefore, the confidence should be low. Empirical tests indicated the importance of the distance between multiple minima. A higher distance, i.e. larger $\Delta d(x, y, d)$, indicates a higher uncertainty for the disparity value. We subtract 1 from $\Delta d(x, y, d)$, to not penalize two minima next to each other, as this is most likely a quantization error and will be fixed in post-processing steps. To avoid negative penalties, the maximum with 0 is taken. If there are multiple local minima more than 1 pixel apart, the confidence decreases. The decrease in confidence is clipped at $\frac{d_{\text{max}}-d_{\text{min}}}{3}$. A large margin between the global minimum and all other costs is favored, as then it's most likely to be the cor-

rect disparity. This margin is empirically defined as $\frac{c_{\text{mean}}(x,y)}{3}$, where $c_{\text{mean}}(x, y)$ is the average of the costs within the defined disparity range. Costs within this margin will have a negative influence on the confidence value, whereas costs outside this margin, even when belonging to a local minimum, do not have any influence. To incorporate the influences of all costs, we sum over the complete disparity range $d_{\text{min}}$ to $d_{\text{max}}$.

## 4. EVALUATION AND RESULTS

We evaluate the two state-of-the-art methods **LRD** and **CCNN** and compare the performance to our proposed method, using **MC-CNN** as basis. We maintain the same evaluation procedure as first described in [17]. For our evaluation, we use the following dataset.

### 4.1. Dataset

**Table 1**. Details of Middlebury datasets used.

| | MB03Q | MB05T | MB06T | MB14Q |
|---|---|---|---|---|
| # pairs | 2 | 6 | 21 | 15 |
| Resolution | Quarter | Third | Third | Quarter |
| $d_{\text{max}}$ | 59px | 80px | 80px | As provided |

For an easier comparison with previous evaluations, we use a combination of available Middlebury datasets for our experiments. This extended Middlebury stereo dataset consists of the two stereo pairs from the 2003 dataset [23] (MB03Q), six stereo pairs from the 2005 dataset [24, 25] (MB05T, the remaining three do not have ground-truth disparity maps available), all 21 image pairs from the 2006 dataset [24, 25] (MB06T), and all image pairs from the 2014 training dataset [3] (MB14Q), leading to a total of 44 stereo pairs. The images were captured indoors in a lab environment and depict objects with varying complexity. For each dataset, we evaluate on the smallest spatial resolution available and use maximum disparities as provided (see Table 1 for details). The minimum disparity is always set to 0 pixels. As per the datasets specifications, the values of the calculated disparities are considered correct if the difference to the ground-truth is within 1 pixel. We always evaluated the algorithms using the confidences and disparity maps of left images.

### 4.2. Evaluation

The ability to distinguish correct disparity assignments from wrong ones is the most desirable property of a confidence measure. To quantitatively evaluate this, the accuracy of disparity assignments based on confidences is evaluated using curves of error rate as a function of disparity map density (see Fig. 2 (d)), based on Gong and Yang [26]. The error rate is defined as the percentage of wrong pixels with respect to the density $p$. All disparities are sorted in decreasing order of confidence and disparity maps of increasing density are produced by selecting disparities according to rank. This measures the capability of removing errors from a disparity map according to the confidence values. The area under the curve (AUC) quantifies the capability of the confidence measure to effectively distinguish good matches from wrong ones. Better confidence measures result in lower AUC values.

Given a disparity map, a subset $\mathcal{P}$ of pixels is extracted in order of decreasing confidence (e.g., $5\%$ of the total pixels) and the error rate of this subset is computed as the percentage of pixels, with respect to the density $p$, with an absolute distance from ground-truth values (including occluded pixels) higher than a threshold. Then, the subset is increased by extracting more pixels (e.g., an additional $5\%$) and the error rate is computed, until all the pixels in the image are considered. When confidences have identical values, all disparities with equal confidences are included into the subsample. This increases the density, therefore the x-axis in Fig. 2 (d) is labeled with minimum density.

The theoretically optimal AUC can be achieved by selecting all correct disparities before starting to fill the quasi-dense disparity maps with the remaining wrong ones and is defined as in [17]:

$$A_{\text{opt}} = \int_{1-\epsilon}^{1} \frac{p-(1-\epsilon)}{p} dp = \epsilon + (1-\epsilon) ln(1-\epsilon), \quad (3)$$

where $p$ is the density and $\epsilon$ is the disparity error rate at full density as introduced in [17]. Following this protocol, we evaluate the three confidence measures on the extended Middlebury dataset, using the stereo algorithm MC-CNN as input. This method adopts a winner takes all (WTA) strategy and infers costs using a local method, comparing image patches using a convolutional neural network. We used the fast architecture network, trained by the authors Žbontar and Le Cun on the KITTI 2012 dataset. We also adopt our own post-processing method, consisting of a $9 \times 9$ box-filter (Eq. 4) operating on the cost volume, which improves the results even further.

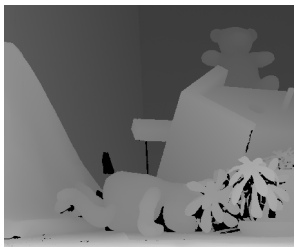
$$c'(x, y, d) = \sum_{i=-4}^{4} \sum_{j=-4}^{4} c(x+i, y+j, d) \quad (4)$$

### 4.3. Results

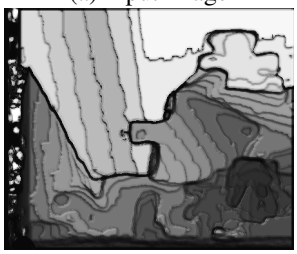
In Fig. 2, (a) one of the input images (Teddy, MB03Q), with (b) ground-truth disparities, and (c) confidences from CCNN are shown. Estimated disparities and the confidences computed by our proposed method can be found in Fig. 1. In Fig. 2 (d), the disparity density ($p$) vs the error rate for the Teddy image pair from the Middlebury 2003 dataset is shown. By combining these results for all image pairs into one graph, we end up with Fig. 3. For each stereo pair in the extended Middlebury dataset, the obtained AUC is depicted. The lower the value, the better the confidence measure. All results are sorted by AUC values with respect to our proposed method.

Observing these figures, we can see that our proposed method clearly outperforms LRD and CCNN for all image pairs in our dataset. Our proposed method improves $54.2\%$ on

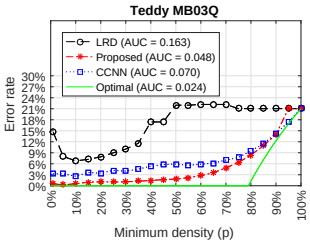
(a) Input image



(b) GT disparity



(c) CCNN confidences




(d) MC-CNN with box-filter

**Fig. 2**. (a) Teddy image from MB03, (b) GT disparities (brighter is closer) and (c) the confidences from CCNN (brighter is higher, scaled for better visualization). In (d), error rates for Teddy at different densities are shown, with curves for LRD, proposed, CCNN and optimal. Besides from the theoretically optimal curve, the proposed method has the lowest error rate for almost every density.

the CCNN measure, indicating that a non-learning based approach can outperform a machine learning-based one. However, existing machine learning-based confidence measures could benefit from including our confidence measure as an additional feature.

For completeness, we also integrated our confidence measure into the ADCensus [27] stereo algorithm. The cost function is a combination of Sum of Absolute Difference (SAD) and Census. Evaluating on the extended Middlebury dataset, we obtained the average AUC values as shown in Table 2. Our proposed confidence measure obtains similar results to the CCNN confidence measure. We believe we cannot outperform the state-of-the-art using this stereo algorithm as input, due to the noise present in the cost-curve.

To give some additional insight, we also measured the execution time, see Table 2. The LRD and proposed algorithms are not optimized for speed, both are implemented in Matlab. CCNN is a GPU implementation. As expected, LRD is the fastest, as it does not evaluate all cost values and utilizes some of the builtin Matlab optimizations for finding minima. CCNN is the slowest, as it cannot be integrated into the disparity estimation and an image has to be copied to the GPU memory.

**Table 2**. Average AUC values and execution time evaluating different confidence measures on the extended Middlebury dataset, using ADCensus to compute the cost function, compared to using the MC-CNN with box-filter as input.

| | LRD | CCNN | Proposed | Optimal |
|---|---|---|---|---|
| MC-CNN $AUC_{mean}$ | 0.188 | 0.168 | **0.077** | 0.039 |
| ADCensus $AUC_{mean}$ | 0.266 | **0.223** | 0.224 | 0.090 |
| Avg. execution time | **1.865s** | 28.373s | 5.386s | - |

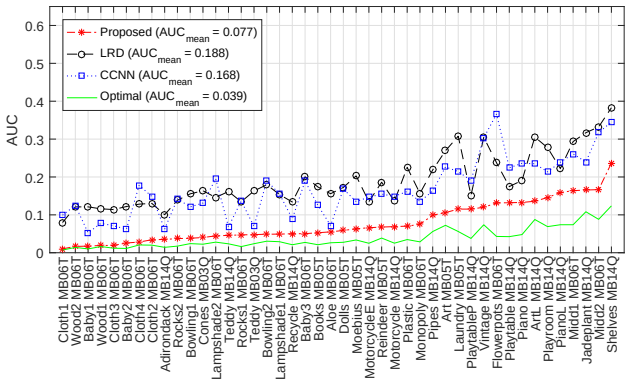



**Fig. 3**. AUC values for the three confidence measures using MC-CNN as input, evaluated on the extended Middlebury dataset. Lower values are better. Sorted by AUC with respect to proposed method. Our proposed method outperforms the state-of-the-art methods on all stereo pairs.

## 5. CONCLUSION AND FUTURE WORK

In this paper we proposed a novel confidence measure, reviewed and evaluated two state-of-the-art confidence measures and compared them to our proposed method. Our evaluation, using the MC-CNN stereo algorithm and the extended challenging Middlebury dataset, clearly highlights that our proposed method outperforms the currently best performing confidence measure CCNN by $54.2\%$. Our confidence computation does not need any machine learning and can be applied directly to most stereo algorithms (provided a cost volume is available). The execution time is of the same order of magnitude as LRD and several times smaller than for CCNN, while performance is better. This evaluation shows that learning-based methods can be outperformed by conventional approaches and that our proposed method would be an useful addition to machine learning-based confidence measures.

Future work includes the integration of the proposed confidence measure into different applications, e.g., disparity post-processing algorithms [13], multi-view-stereo, and data fusion. The improvement of initial disparity maps, could lead to improved depth-image-based-rendering results.

## 6. REFERENCES


[1] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[2] Moritz Menze and Andreas Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE*

*Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3061–3070.

[3] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German Conference on Pattern Recognition*. Springer, 2014, pp. 31–42.

[4] Jure Žbontar and Yann Le Cun, "Computing the stereo matching cost with a convolutional neural network," 2015, vol. 07-12-June, pp. 1592–1599.

[5] Akihito Seki and Marc Pollefeys, "Patch based confidence prediction for dense disparity map," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016, p. 23.

[6] Philippos Mordohai, "The self-aware matching measure for stereo," in *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, 2009, pp. 1841–1848.

[7] Larry Matthies, "Stereo vision for planetary rovers: Stochastic modeling to near real-time implementation," *International Journal of Computer Vision*, vol. 8, no. 1, pp. 71–91, 1992.

[8] Daniel Scharstein and Richard Szeliski, "Stereo matching with nonlinear diffusion," *International journal of computer vision*, vol. 28, no. 2, pp. 155–174, 1998.

[9] Zhengyou Zhang and Ying Shan, "A progressive scheme for stereo matching," in *European Workshop on 3D Structure from Multiple Images of Large-Scale Environments*. Springer, 2000, pp. 68–85.

[10] Heiko Hirschmüller, Peter R Innocent, and Jon Garibaldi, "Real-time correlation-based stereo vision with reduced border errors," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 229–246, 2002.

[11] Paul Merrell, Amir Akbarzadeh, Liang Wang, Philippos Mordohai, Jan-Michael Frahm, Ruigang Yang, David Nistér, and Marc Pollefeys, "Real-time visibility-based fusion of depth maps," in *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.

[12] Kuk-Jin Yoon and In So Kweon, "Distinctive similarity measure for stereo matching under point ambiguity," *Computer Vision and Image Understanding*, vol. 112, no. 2, pp. 173–183, 2008.

[13] Aristotle Spyropoulos and Philippos Mordohai, "Correctness Prediction, Accuracy Improvement and Generalization of Stereo Matching Using Supervised Learning," *International Journal of Computer Vision*, vol. 118, no. 3, pp. 300–318, 2016.

[14] Aristotle Spyropoulos, Nikos Komodakis, and Philippos Mordohai, "Learning to detect ground control points for improving the accuracy of stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1621–1628.

[15] Min-Gyu Park and Kuk-Jin Yoon, "Leveraging stereo matching with learning-based confidence measures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 101–109.

[16] Matteo Poggi and Stefano Mattoccia, "Learning a general-purpose confidence measure based on o (1) features and a smarter aggregation strategy for semi global matching," in *Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 509–518.

[17] Xiaoyan Hu and Philippos Mordohai, "A quantitative evaluation of confidence measures for stereo vision," 2012, vol. 34, pp. 2121–2133.

[18] Matteo Poggi and Stefano Mattoccia, "Learning from scratch a confidence measure," 2016, number Cv, pp. 46.1–46.13.

[19] Ralf Haeusler, Rahul Nair, and Daniel Kondermann, "Ensemble learning for confidence measures in stereo vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 305–312.

[20] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia, "Quantitative Evaluation of Confidence Measures in a Machine Learning World," 2017, vol. 2012, pp. 5238–5247.

[21] Ke Zhang, Jiangbo Lu, and Gauthier Lafruit, "Cross-based local stereo matching using orthogonal integral images," 2009, vol. 19, pp. 1073–1079, IEEE.

[22] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[23] Daniel Scharstein and Richard Szeliski, "High-accuracy stereo depth maps using structured light," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003, vol. 1.

[24] Daniel Scharstein and Chris Pal, "Learning conditional random fields for stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[25] Heiko Hirschmuller and Daniel Scharstein, "Evaluation of cost functions for stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[26] Minglun Gong and Yee-Hong Yang, "Fast unambiguous stereo matching using reliability-based dynamic programming," 2005, vol. 27, pp. 998–1003.

[27] Xing Mei, Xun Sun, Mingcai Zhou, Shaohui Jiao, Haitao Wang, and Xiaopeng Zhang, "On building an accurate stereo matching system on graphics hardware," 2011, pp. 467–474.