

Real-Time Highly Accurate Dense Depth on a Power Budget using an FPGA-CPU Hybrid SoC

Oscar Rahnama

Tommaso Cavallari

Stuart Golodetz

Alessio Tonioni

Tom Joy

Luigi Di Stefano

Simon Walker

Philip H. S. Torr

Abstract—Obtaining highly accurate depth from stereo images in real time has many applications across computer vision and robotics, but in some contexts, upper bounds on power consumption constrain the feasible hardware to embedded platforms such as FPGAs. Whilst various stereo algorithms have been deployed on these platforms, usually cut down to better match the embedded architecture, certain key parts of the more advanced algorithms, e.g. those that rely on unpredictable access to memory or are highly iterative in nature, are difficult to deploy efficiently on FPGAs, and thus the depth quality that can be achieved is limited. In this paper, we leverage a FPGA-CPU chip to propose a novel, sophisticated, stereo approach that combines the best features of SGM and ELAS-based methods to compute highly accurate dense depth in real time. Our approach achieves an 8.7% error rate on the challenging KITTI 2015 dataset at over 50 FPS, with a power consumption of only 4.5W.

Index Terms—Heterogeneous, FPGA, real-time, stereo, depth

I. INTRODUCTION

Obtaining information about the 3D structure of a scene is important for many computer vision and robotics applications, e.g. 3D scene reconstruction [1]–[3], camera relocalisation [4]–[6], navigation and obstacle avoidance [7]. Often, this information will be obtained in the form of a depth image, and various options for acquiring such images exist. Passive approaches, which rely only on one or more image sensors, are popular due their low cost, low weight and size, lack of active/moving components, ability to work at longer ranges, deployability in a wider range of operating environments and lack of interference. Among them, binocular stereo relies on a pair of synchronised cameras to acquire the same scene from two different points of view. Given the two frames, a dense and reliable depth map can be computed by finding correspondences between the pixels in the two images [9]. State-of-the-art algorithms for this problem usually rely on costly global image optimisations or on massive convolutional neural networks that involve significant computational costs, making them hard to deploy on resource-limited systems such as embedded devices [10]. Two popular solutions offering a good trade-off between speed and accuracy are Semi-Global Matching (SGM) [11] and ELAS [8]. SGM computes initial matching hypotheses by comparing patches around pixels in

the left and right images, then approximates a costly image-wide smoothness constraint with the sum of several directional minimisations over the disparity range. By contrast, ELAS first identifies a set of sparse but reliable correspondences to provide a coarse approximation of the scene geometry, then uses them to define slanted plane priors that guide the final dense matching stage. We propose a novel stereo pipeline that efficiently combines the predictions of these two algorithms, achieving high accuracy and overcoming some of the limitations of each algorithm. First, we use multiple passes of a fast SGM variant [12], left-right consistency checking and decimation to obtain a sparse but reliable set of correspondences. Then, we use these as the support points for ELAS to obtain disparity priors from slanted planes. Finally, we incorporate these disparity priors into a final SGM-based optimisation (again based on [12]) to achieve dense predictions with high accuracy.

Our pipeline targets not only accuracy, but also speed, aiming for real-time execution (30 FPS) on an embedded platform. Recent works have deployed SGM successfully in real time both on multi-core CPUs [13] and GPUs [14], [15], but in real-world scenarios, power constraints often force us to rely on low-power devices like FPGAs. The development of reliable stereo pipelines for FPGAs is an active research field [10], [16]–[22], with recent works proposing FPGA-friendly variants of SGM [15], [23]–[27] or ELAS [28]. However, FPGA implementations of stereo algorithms usually perform some kind of approximation to deal with the limited resources available and to traverse the pixels in raster order.

We show how some of the intrinsic limitations of a pure FPGA-based implementation can be mitigated by appropriately leveraging a new-generation hybrid system on a chip (SoC), e.g. the Xilinx ZCU104, which combines both an ARM processor and an FPGA, with shared direct memory access, into a single chip. Recently, several works have explored the deployment of stereo methods on such platforms: both [26] and [19] use the CPU mainly for handling communication and controlling peripherals, while [28] actively leverages the CPU to execute iterative steps that would be infeasible on an FPGA (e.g. Delaunay triangulation). Similar to [28], we propose to actively use the elaboration capability of the built-in CPU to handle I/O and to execute part of the ELAS pipeline, while deploying all the other elaboration blocks on the FPGA. We show how our pipeline can outperform all previously published works by achieving an 8.7% error rate on the challenging KITTI 2015 dataset [29], [30], while still achieving real-time performance and low power consumption.

Correspondence: {oscar@robots.ox.ac.uk}

O. Rahnama is with the University of Oxford and FiveAI Ltd.

T. Joy and P. Torr are with the University of Oxford.

A. Tonioni and L. Di Stefano are with the University of Bologna.

T. Cavallari, S. Golodetz and S. Walker are with FiveAI Ltd.

Work done whilst A. Tonioni was visiting the University of Oxford.

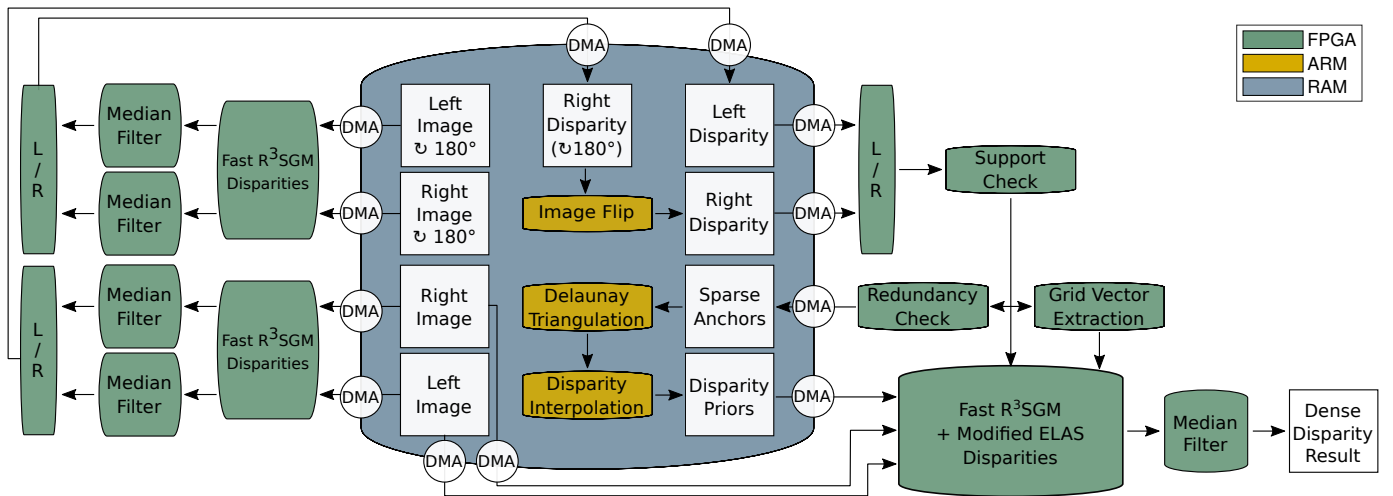


Fig. 1: **Overview of our approach.** First, we use Fast R³SGM (see §II-A1) to compute disparity images for the input stereo pair (in raster and reverse-raster order). We then flip the right result and perform a left-right consistency check to obtain an accurate but sparse disparity map for the left input image (see §II-A2). Next, as ELAS [8] does, we perform *support checking* (see §II-B1) to remove points whose disparities appear abnormal relative to neighbouring pixels: this yields a sparser support point image that contains only points with confident disparities. This support point image is subsequently used in multiple ways: (i) it is further sparsified via a *redundancy check*, producing sparse anchors that are then used to generate plane disparity priors through a *triangulation* and *interpolation* process (see §II-B2); (ii) it is split into a grid where, for each grid cell, a binary vector representing the set of viable disparities is computed (see §II-B3). Finally, the support point image is combined with the outputs of (i) and (ii) in a *disparity optimisation* that combines R³SGM and ELAS to produce a dense disparity image (see §II-C). We then median filter this image for robustness to produce the final result.

II. METHOD

Our overall pipeline is shown in Figure 1. It consists of several different components which we describe in the subsections that follow. The system leverages both parts of the FPGA-CPU hybrid SoC to achieve optimal results. Tasks that are very data intense, but which access that data in a predictable manner, are run on dedicated FPGA accelerators to benefit from their parallel processing capability. In addition, they can take advantage of the FPGA accelerators' internal ability to pipeline data so that multiple inputs are processed together in staggered fashion. Tasks that are very dynamic and unpredictable, which often involve many unforeseen or random accesses to external memory, are run on the CPU since they benefit both from the significantly faster clock frequency of the CPU as well as its ability to access memory in constant time (CPU memory accesses can be sped up via appropriate use of the cache). To minimise the amount of FPGA resources used by our method, as well as allow the deployment of the design on a real platform, we reuse some accelerators whilst buffering intermediate results in RAM. We will detail which blocks are reused in our final design in the rest of this section.

A. Sparse Disparity Computation

1) *Fast R³SGM*: Initially, we use a modified version of R³SGM [12] (a memory-efficient adaptation of classic SGM [11] to FPGAs), which we call Fast R³SGM, to compute disparity images for input stereo pairs. We process each input pair twice: once in raster order, and once in reverse raster order, yielding two disparity images overall. The advantage of this is that the raster and reverse-raster passes of R³SGM will

base the disparity for each pixel on the disparities of pixels in different regions of influence: this means that we can later check for consistency between the two, improving the accuracy of our results.

The original version of R³SGM [12] aggregated contributions to the disparity of each pixel along four different scanlines: three above the pixel, and one to the left. However, as mentioned in [12], using the left scanline severely limits the overall throughput of the system (one disparity value is output every three clock cycles) due to a blocking dependency between immediately successive pixels. To avoid this, we modify the approach to use only the scanlines above the pixel, allowing us to output one disparity per clock cycle. The mild loss in accuracy this causes is more than compensated for by the improvements yielded by the rest of our pipeline.

In our implementation, we deploy a single instance of the Fast R³SGM block, together with the associated median filtering and L/R consistency checking blocks. We first feed the blocks with the raster-order stereo pair, then with the reverse-raster-order pair, storing the disparities resulting from each pair back into RAM between the computations.

2) *Consolidating Consistency Checking*: Each pass of Fast R³SGM outputs a disparity map that has been checked for consistency using the first input as the reference image [12]. The raster pass outputs a disparity map for the left input image; the reverse-raster pass outputs one for the (reversed) right input image. Due to the streaming nature of the disparity computation, however, the results suffer from a raster or reverse-raster scan bias, i.e. the disparity value of any given pixel is encouraged to be similar to those computed before it. To reconcile the inconsistencies between these two disparity

maps, we perform a further left-right consistency check, which yields an accurate but sparse disparity map for the left input image as its result (see Figure 1). The memory access pattern of such a process is problematic, however, as the first pixels in the left disparity map need to be checked against the last pixels in the right disparity map. To overcome this problem, we first reverse the latter image on the CPU (since this is an inherently sequential process, it benefits from the higher clock rate provided by the ARM core), then perform a standard left-right consistency check (on the programmable logic).

B. Generation of Priors

Using the sparse disparity map output by the consolidating consistency check, we adapt the ELAS method described in [8] to generate priors that can be fed into a combined disparity optimisation process (see §II-C) to produce a more accurate and dense final result. The prior generation process begins by taking the disparity map produced by §II-A2 as input and producing a support point image (see §II-B1) containing sparse but confident disparities. The support points are then fed to two more blocks before being used by the final disparity optimisation process: (i) a redundancy checking and disparity prior generation block, which first computes a sparse anchor points image and then triangulates such anchors to generate disparity priors for all pixels in the image (see §II-B2); and (ii) a grid vector extraction block that divides the support points image into a grid and then determines the set of possible disparities for each cell (see §II-B3).

1) *Support Checking*: To produce the support point image, we filter the sparse disparity map to remove any pixels whose disparity is not sufficiently supported by the pixels in their immediate neighbourhood (in practice, a square window centred on the pixel). For a pixel to be considered “supported”, there must exist, in its neighbourhood, another predefined number of pixels that have very similar disparity values (e.g. at least 10 pixels within a 5×5 window that differ by less than 5 from that of the center pixel). The disparities of all other pixels are marked as invalid. The resulting support point image will evidently be sparser than the original disparity map, since we have kept only those pixels about whose disparities we can be reasonably confident.

2) *Redundancy Checking and Disparity Prior Generation*: To produce the anchor image, we further sparsify the support point image produced in §II-B1 by processing it in raster order and invalidating any pixel whose disparity has already been seen within a window behind and above the pixel. Unlike [28], which for each pixel (x, y) used a window of

$$\{(x, y - \delta_y) : 0 < \delta_y \leq 2K\} \cup \{(x - \delta_x, y) : 0 < \delta_x \leq 2K\},$$

where K was set to 5, which only encompassed points in the same row or same column as the pixel being processed, here we use a larger window of

$$\begin{aligned} & \{(x + \delta_x, y + \delta_y) : -K \leq \delta_x \leq K, -2K \leq \delta_y < 0\} \\ \cup & \{(x - \delta_x, y) : 0 < \delta_x \leq K\}. \end{aligned}$$

This has the effect of creating a sparser anchor image than that used in [28], significantly speeding up the subsequent Delaunay triangulation process. Whilst this inevitably reduces

the accuracy of the depth priors, we found that the final disparities produced by the combined optimisation (see §II-C) are in practice only loosely affected by the depth prior quality; as a result, and since the Delaunay triangulation process is a key bottleneck [28], it makes sense to generate slightly less accurate depth priors faster, rather than spending the extra time to compute more accurate depth priors that make little difference to the final result.

Finally, to produce the disparity priors, we first move the anchor points image back to RAM, then perform a Delaunay triangulation of those points, and finally compute the disparity of each non-anchor point located within one of the Delaunay triangles by interpolating the disparities of the triangle's vertices. The entirety of this process is performed by the CPU, since the triangulation and interpolation algorithms are inherently non-sequential in their memory access patterns, and can benefit from both the availability of memory caches and the higher speed of the ARM core.

3) *Grid Vector Extraction*: The final input to the combined disparity optimisation we describe in §II-C is a set of binary *grid vectors* used to determine which disparities are suitable for each part of the image. To produce such vectors, we first divide the support point image into a regular grid (with cells of size 50×50 in our implementation). Then, for each cell, we find the valid disparity values within it, and store both those and their neighbouring disparities (± 1) into a binary grid vector for that cell. See [28] for more details.

C. Combined Disparity Optimisation

Finally, we perform a combined disparity optimisation that takes into account not only the original pair of input images, but also the plane priors, grid vectors and support points. Essentially, we perform Fast R³SGM, as in §II-A1 (once again reusing the corresponding FPGA block), but first modifying the cost vectors of the pixels to take the various different priors we have available into account.

The disparities of the support points are fixed throughout and not recomputed. Every cost vector element for a support point (each of which corresponds to a specific disparity) is set to a large arbitrary value, except for the element corresponding to the disparity of the support point, which is set to zero instead. Through the Fast R³SGM smoothing process, pixels near the support point will then naturally be encouraged to adopt disparities similar to that of the support point itself, with the influence of this effect attenuating with distance. To take the disparity prior for each pixel into account, we decrease those elements of its cost vector that correspond to disparities close to the prior (more specifically, we superimpose a negative Gaussian over the cost vector, centred on the prior, and decrease the relevant elements within a certain radius accordingly). To make use of the grid vectors, we set all elements of the cost vectors for the pixels within each grid cell that do not appear in the grid vector for that cell to an arbitrarily large value, thus strongly encouraging them not to be selected. As with the effects of the support points, these cost vector modifications are similarly propagated by the Fast R³SGM smoothing process.

Method	Background	Foreground	All	Density	Runtime (s)	Environment	Power Used (W)
Ours	7.2%	17.3%	8.7%	99.7%	0.019	FPGA (Xilinx ZCU104)	4.5
R ³ SGM [12]	–	–	9.9%	85.0%	0.014	FPGA (Xilinx ZC706)	≈ 4
ELAS-FPGA [28]	–	–	13.6%	–	0.095	FPGA (Xilinx ZC706)	≈ 3
DeepCostAggr [31]	5.3%	11.4%	6.3%	99.98%	0.03	Nvidia GTX Titan X	≈ 250
CSCT+SGM+MF [15]	6.9%	14.9%	8.2%	100%	0.006	Nvidia GTX Titan X	≈ 250

TABLE I: The quantitative results of our approach, in comparison to state-of-the-art GPU-based real-time methods, on the Stereo 2015 subset of the KITTI benchmark [29], [30]. As in the official evaluation protocol, we report the percentage of accurate disparities (using a threshold of < 3 disparity values or 5%, whichever is greater) after an interpolation step (meant to assign a disparity value to all pixels in the image), on respectively the subsets of background, foreground and all pixels. We additionally report the density of valid disparity values. As can be seen, with the exception of R³SGM [12], all methods provide almost dense disparity images, therefore the extra interpolation step mandated by the benchmark is not strictly required to obtain usable disparity images. Finally, for each method, we report the typical time required to process a stereo pair, as well as the approximate power consumption of the platform used. Whilst all approaches can process images in real-time, only the FPGA-based methods (ours and [12]) can do so in a power-efficient manner, with ours providing $\approx 12\%$ additional accuracy and much higher density w.r.t. [12], at the expense of slightly higher power usage and processing time.

	Ours	[12]	[28]
Platform	ZCU104	ZC706	ZC706
LUT Utilisation (%)	87.5	75.7	37.3
FF Utilisation (%)	24.1	40.5	22.6
BRAM Utilisation (%)	70.7	30.4	11.9
FPGA Power Consumption (W)	1.72	3.94	1.21
CPU Power Consumption (W)	2.78	–	1.7
Total Power Consumption (W)	4.5	–	2.91

TABLE II: Resources (programmable logic units) and power (as estimated by the Xilinx Vivado tool) used by the proposed approach, when deployed on a Xilinx ZCU104, in comparison to the FPGA-based methods from which we draw inspiration.

At the end of this process, we perform a final median filter on the Fast R³SGM result to further mitigate impulsive noise, ultimately yielding a dense, accurate disparity map.

III. RESULTS

We developed the FPGA accelerators using the Vivado High-Level Synthesis (HLS) tool, as this approach was quicker, and allowed for greater flexibility and reusability of the components. We deployed the system on a Xilinx ZCU104 board, and all of the power consumption results that we present for our method were estimated by the Xilinx Vivado tool.

We quantitatively evaluate the disparities produced by our approach on the standard KITTI 2015 stereo benchmark [29], [30]. In Table I, we report the average percentages of pixel disparities estimated correctly for background, foreground and all pixels, respectively. We also report average runtimes and power consumptions for both our and alternative methods that achieve real-time processing speeds on the images used in the benchmark (which have a resolution of 1242×375). Whilst the proposed method results in slightly less accurate disparities than the DeepCostAggr [31] and CSCT-SGM-MF [15] methods, it is worth pointing out that both [15], [31] rely on powerful GPUs to achieve real-time processing speed, whereas our approach does so in a much more power-efficient manner, relying only on a hybrid FPGA-CPU board. We also compare favourably to R³SGM [12], the underlying method on which we base our approach for the estimation of the initial disparities (see §II-A1), providing more accurate

and denser results at a similar speed and with similar power consumption. We similarly outperform the FPGA variant of ELAS [28], achieving a lower error rate at a much higher speed, and with similarly low power consumption.

In Table II, we detail the hardware resources used by our approach when deployed on our Xilinx ZCU104 board. We break down the amount of logic resources used in the FPGA chip, as well as the power consumption of both the programmable logic and the ARM core. We also report the amount of resource and power used by the methods from which we draw inspiration [12], [28]. Notably, despite making full use of many of the logic resources available on the FPGA, our power consumption remains very low. More specifically, breaking down the resource utilization of the programmable logic amongst the different accelerators, the largest share is taken by the Fast R³SGM block which, alone, consumes about 65% of the FPGA power. The next most resource heavy blocks are the ones which perform the median filtering of the disparities, which require approximately 30% of the power. The remaining blocks have much smaller resource requirements, which altogether account for the remaining 5% of the power.

IV. CONCLUSIONS

In this paper, we have presented a novel approach to computing depth from stereo images on a hybrid FPGA-CPU chip. Our approach uses an adapted version of ELAS [8] to refine the initial sparse disparity map produced by a fast variant of R³SGM [12], and achieves an impressive 8.7% error rate on the challenging KITTI 2015 dataset [29], [30]. By fully leveraging the capabilities of our hybrid board, we are able to produce highly accurate dense depth at over 50 FPS, with a power consumption of only 4.5W, making our approach attractive for applications in mobile, real-time computing.

ACKNOWLEDGEMENTS

This work was supported by Innovate UK/CCAV project 103700 (StreetWise), EPSRC grant Seebibyte EP/M013774/1 and EPSRC/MURI grant EP/N019474/1. We would also like to acknowledge the Royal Academy of Engineering and FiveAI.

REFERENCES

- [1] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger, "ElasticFusion: Real-Time Dense SLAM and Light Source Estimation," *IJRR*, vol. 35, no. 14, pp. 1697–1716, 2016. 1
- [2] V. A. Prisacariu, O. Kähler, S. Golodetz, M. Sapienza, T. Cavallari, P. H. S. Torr, and D. W. Murray, "InfiniTAM v3: A Framework for Large-Scale 3D Reconstruction with Loop Closure," *arXiv preprint arXiv:1708.00783v1*, 2017. 1
- [3] S. Golodetz*, T. Cavallari*, N. A. Lord*, V. A. Prisacariu, D. W. Murray, and P. H. S. Torr, "Collaborative Large-Scale Dense 3D Reconstruction with Online Inter-Agent Pose Optimisation," *TVCG*, vol. 24, no. 11, 2018. 1
- [4] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images," in *CVPR*, 2013, pp. 2930–2937. 1
- [5] T. Cavallari, S. Golodetz*, N. A. Lord*, J. Valentin, L. D. Stefano, and P. H. S. Torr, "On-the-Fly Adaptation of Regression Forests for Online Camera Relocalisation," in *CVPR*, 2017, pp. 4457–4466. 1
- [6] T. Cavallari*, S. Golodetz*, N. A. Lord*, J. Valentin*, V. A. Prisacariu, L. D. Stefano, and P. H. S. Torr, "Real-Time RGB-D Camera Pose Estimation in Novel Scenes using a Relocalisation Cascade," *arXiv preprint arXiv:1810.12163*, 2018. 1
- [7] S. L. Hicks, I. Wilson, L. Muhammed, J. Worsfold, S. M. Downes, and C. Kennard, "A Depth-Based Head-Mounted Visual Display to Aid Navigation in Partially Sighted Individuals," *PLOS ONE*, 2013. 1
- [8] A. Geiger, M. Roser, and R. Urtasun, "Efficient Large-Scale Stereo Matching," in *Computer Vision—ACCV 2010*. Springer, 2010, pp. 25–38. 1, 2, 3, 4
- [9] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *International Journal of Computer Vision*, vol. 47, no. 1–3, pp. 7–42, 2002. 1
- [10] B. Tippetts, D. J. Lee, K. Lillywhite, and J. Archibald, "Review of Stereo Vision Algorithms and their Suitability for Resource-Limited Systems," *Journal of Real-Time Image Processing*, vol. 11, no. 1, pp. 5–25, 2016. 1
- [11] H. Hirschmüller, "Stereo Processing by Semiglobal Matching and Mutual Information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008. 1, 2
- [12] O. Rahnama, T. Cavallari*, S. Golodetz*, S. Walker, and P. H. S. Torr, "R³SGM: Real-time Raster-Respecting Semi-Global Matching for Power-Constrained Systems," in *FPT*, 2018. 1, 2, 4
- [13] S. K. Gehrig and C. Rabe, "Real-Time Semi-Global Matching on the CPU," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010 IEEE Computer Society Conference on. IEEE, 2010, pp. 85–92. 1
- [14] C. Banz, H. Blume, and P. Pirsch, "Real-Time Semi-Global Matching Disparity Estimation on the GPU," in *Computer Vision Workshops (ICCV Workshops)*, 2011 IEEE International Conference on. IEEE, 2011, pp. 514–521. 1
- [15] D. Hernandez-Juarez, A. Chacón, A. Espinosa, D. Vázquez, J. C. Moure, and A. M. López, "Embedded real-time stereo estimation via Semi-Global Matching on the GPU," *Procedia Computer Science*, vol. 80, 2016. 1, 4
- [16] S. Perri, F. Frustaci, F. Spagnolo, and P. Corsonello, "Design of Real-Time FPGA-based Embedded System for Stereo Vision," in *Circuits and Systems (ISCAS)*, 2018 IEEE International Symposium on. IEEE, 2018, pp. 1–5. 1
- [17] O. Rahnama, A. Makarov, and P. Torr, "Real-time depth processing for embedded platforms," in *Real-Time Image and Video Processing 2017*, vol. 10223. International Society for Optics and Photonics, 2017, p. 102230N. 1
- [18] L. Zhang, K. Zhang, T. S. Chang, G. Lafruit, G. K. Kuzmanov, and D. Verkest, "Real-time high-definition stereo matching on FPGA," in *Proceedings of the 19th ACM/SIGDA International Symposium on Field Programmable Gate Arrays*. ACM, 2011, pp. 55–64. 1
- [19] S. Perri, F. Frustaci, F. Spagnolo, and P. Corsonello, "Stereo vision architecture for heterogeneous systems-on-chip," *Journal of Real-Time Image Processing*, pp. 1–23, 2018. 1
- [20] C. Ttofis and T. Theodoridis, "High-quality real-time hardware stereo matching based on guided image filtering," in *Proceedings of the Conference on Design, Automation & Test in Europe*. European Design and Automation Association, 2014, p. 356. 1
- [21] M. Dehnavi and M. Eshghi, "FPGA based real-time on-road stereo vision system," *Journal of Systems Architecture*, vol. 81, pp. 32–43, 2017. 1
- [22] D. Zha, X. Jin, and T. Xiang, "A real-time global stereo-matching on FPGA," *Microprocessors and Microsystems*, vol. 47, pp. 419–428, 2016. 1
- [23] S. K. Gehrig, F. Eberli, and T. Meyer, "A Real-Time Low-Power Stereo Vision Engine Using Semi-Global Matching," in *International Conference on Computer Vision Systems*. Springer, 2009, pp. 134–143. 1
- [24] C. Banz, S. Hesselbarth, H. Flatt, H. Blume, and P. Pirsch, "Real-Time Stereo Vision System using Semi-Global Matching Disparity Estimation: Architecture and FPGA-Implementation," in *Embedded Computer Systems (SAMOS)*, 2010 International Conference on. IEEE, 2010, pp. 93–101. 1
- [25] S. Mattoccia and M. Poggi, "A passive RGBD sensor for accurate and real-time depth sensing self-contained into an FPGA," in *Proceedings of the 9th International Conference on Distributed Smart Cameras*. ACM, 2015, pp. 146–151. 1
- [26] D. Honegger, H. Oleynikova, and M. Pollefeys, "Real-time and Low Latency Embedded Computer Vision Hardware Based on a Combination of FPGA and Mobile CPU," in *Intelligent Robots and Systems (IROS 2014)*, 2014 IEEE/RSJ International Conference on. IEEE, 2014, pp. 4930–4935. 1
- [27] W. Wang, J. Yan, N. Xu, Y. Wang, and F.-H. Hsu, "Real-Time High-Quality Stereo Vision System in FPGA," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 10, pp. 1696–1708, 2015. 1
- [28] O. Rahnama, D. Frost, O. Miksik, and P. H. Torr, "Real-Time Dense Stereo Matching With ELAS on FPGA-Accelerated Embedded Devices," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2008–2015, 2018. 1, 3, 4
- [29] M. Menze, C. Heipke, and A. Geiger, "Joint 3D Estimation of Vehicles and Scene Flow," in *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015. 1, 4
- [30] —, "Object Scene Flow," *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)*, 2018. 1, 4
- [31] A. Kuzmin, D. Mikushin, and V. Lempitsky, "End-to-end Learning of Cost-Volume Aggregation for Real-time Dense Stereo," in *MLSP*, 2017. 4