IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 5, 2021, accepted August 19, 2021, date of publication August 30, 2021, date of current version September 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3108626

Beyond the Baseline: 3D Reconstruction of Tiny Objects With Single Camera Stereo Robot

DANIELE DE GREGORIO[1], (Member, IEEE), MATTEO POGGI[2], (Member, IEEE), PIERLUIGI ZAMA RAMIREZ[2], (Member, IEEE), GIANLUCA PALLI[3], (Senior Member, IEEE), STEFANO MATTOCCIA[2], (Member, IEEE), AND LUIGI DI STEFANO[2], (Member, IEEE)

[1]EYECAN.ai S.r.l, 40127 Bologna, Italy
[2]DISI, Università degli Studi di Bologna, 40136 Bologna, Italy
[3]DEI, Università degli Studi di Bologna, 40136 Bologna, Italy

Corresponding author: Daniele De Gregorio (daniele.degregorio@eyecan.ai)

This work was supported in part by the European Commission‘s Horizon2020 Framework Programme with the Project REMODEL—Robotic Technologies for the Manipulation of Complex Deformable Linear Objects under Grant 870133.

ABSTRACT Self-aware robots rely on depth sensing to interact with the surrounding environment, e.g. to pursue object grasping. Yet, dealing with tiny items, often occurring in industrial robotics scenarios, may represent a challenge due to lack of sensors yielding sufficiently accurate depth measurements. Existing active sensors fail at measuring details of small objects (<1cm) because of limitations in the working range, e.g. usually beyond 50 cm away, while off-the-shelf stereo cameras are not suited to close-range acquisitions due to the need for extremely short baselines. Therefore, we propose a framework designed for accurate depth sensing and particularly amenable to reconstruction of miniature objects. By leveraging on a single camera mounted in eye-on-hand configuration and the high repeatability of a robot, we acquire multiple images and process them through a stereo algorithm revised to fully exploit multiple vantage points. Using a novel dataset addressing performance evaluation in industrial applications, our Single camera Stereo Robot (SiSteR) delivers high accuracy even when dealing with miniature objects. We will provide a public dataset and an open-source implementation of our proposal to foster further development in this field.

INDEX TERMS Intelligent robots, robot learning, robot vision systems.

I. INTRODUCTION

Is monocular vision enough for *self-aware* robots? The answer is no. Although there exist methods which attempt to estimate 3D information using only 2D images [1], they usually focus on the perception of a specific set of objects of interest. In the *random bin picking* task, for example, the robot should not only determine the location of the objects but also perceive the surrounding environment to interact safely with it (e.g. to avoid obstacles).

But well before the robots this need for advanced perception was obviously also felt in the animal kingdom. That is why, during the evolution, animals have acquired *stereopsis*, i.e. the capability to infer 3D information by triangulation between a pair of images sensed from two vantage points. Not only primates, as previously thought [2], but also all mammals, birds, amphibians, invertebrates and, according to a recent discovery, even insects [3], deploy multiple eyes, i.e. images from different vantage points, to perceive depth through triangulation. Beyond the stereopsis, which can be considered a form of *passive* ranging strategy, other animals, such as whales or bats, developed *active* methods, e.g. sonar, to actively reconstruct 3D shapes by emitting and receiving suitable signals.

Inspired by the animal kingdom, as often occurs, research on visual sensing has made progress by emulating both the above mentioned *passive* and *active* techniques. Sensors mimicking *stereopsis* by deploying two or multiple cameras are, in most cases, passive. Concerning active sensors, there exist different technologies. *Laser scanners* illuminate the scene through a laser beam swept by a rotating mirror and estimate depth by analyzing the back-scattering. *Time-of-flight* sensors measure distances by computing the return time of the signal emitted by a source. *Structured-light* sensors project a known pattern and estimate depth by observing the deformations induced by the 3D structure of the scene into an

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

image acquired by a 2D camera. Finally, sometimes, a stereo setup is enriched by a pattern projector, like in the Ensenso active stereo camera [4], in such a way as to deploy a hybrid approach.

The techniques mentioned above have been widely used to address many diverse tasks related to visual perception. However, in this paper, we wish to investigate an alternative and quite less popular depth sensing strategy inherited from the animal kingdom. Indeed, some insects (and others species), perform *side-to-side peering head movements* [5] in order to achieve a 3D perception of the environment. They move the head in a controlled manner to estimate the distance of objects through motion parallax by generating multiple viewpoints according to the movements of the eye. Although this strategy may sometimes turn out detrimental to animals causing exposure to predators, we feel comfortable in gathering inspiration by this behavior to tackle – safely – robotic scenarios.

Hence, the idea proposed in this work is to perceive depth using a single camera mounted on a robotic arm in an *eye-on-hand* configuration to gather multiple views in a precisely controlled manner. Taking advantage of the high repeatability of the robot, it is possible not only to emulate *peering head* movements, but also to go beyond by generating more articulated viewpoint patterns of the 2D camera (e.g. spiral patterns), where we can know the 3D poses with millimetric precision. Thus, the proposed solution may be seen as an evolution of the concept of *multi-baseline stereo*. Thanks to the dexterity and precision of a robot, we can generate seamlessly multiple vantage points to adapt the *baseline* between views according to the specific requirements of the task at hand, thereby obtaining high precision whatever the size of the sensed object in the scene is. Thanks to its peculiarities, our system, which we dubbed SiSteR (Single camera Stereo Robot), can effectively replace both the active sensors often deployed in robotics (e.g. Kinect or Asus Xtion) and stereo cameras whenever the scene is stationary. As it will be shown in our evaluation, we can reconstruct objects of varying sizes, even *tiny* components prohibitive for traditional sensors such as Kinect v2 as shown in Figure 1, and feed this information to robotic vision modules in applications like *pick&place*, *random bin picking* and *inspection*.

We have developed two variants of SiSteR: one based on a traditional stereo matching approach and another based on deep learning. The code will be released as an open-source ROS package[1] publicly available online.

II. METHOD AND MOTIVATIONS

This work aims at designing an effective robotic perception pipeline for 3D reconstruction. Despite availability of accurate active 3D sensors, limitations in the working range (usually, beyond 50 cm [6]) and encumbrance render these devices often inappropriate for the eye-on-hand configuration. In particular, when dealing with the reconstruction of tiny objects such as screws, washers or industrial components, extremely close acquisition and measurements are required to achieve high enough accuracy. Moreover, active technologies struggle in the presence of reflective or dark surfaces. Stereo cameras represent an attractive alternative, due to their potentially unconstrained working range. In particular, by knowing the focal length f and the distance between the two cameras (baseline b), depth is estimated the disparity d (the horizontal displacement of corresponding pixels between two images) through triangulation as $z = \frac{b \cdot f}{d}$. For this purpose, images are rectified according to camera parameters obtained through offline calibration. Given two imaging devices with focal length f, the depth error Δ_z depends on the baseline: the higher b, the lower Δ_z. However, according to the distance of the sensed objects from the stereo rig, the baseline affects also the common field-of-view between the two cameras. Therefore, the baseline should be small for close range measurements and large at more considerable distances, so to keep Δ_z comparable. Unfortunately, a standard stereo rig does not allow such behavior due to the fixed baseline. Figure 2 plots Δ_z as a function of the distance z from the camera for different baselines:

$$\Delta_z = \frac{z^2}{b \cdot f} \cdot \Delta_d \tag{1}$$

The plot shows how a disparity error Δ_d of 1 translates into a tiny or large gap between estimated and real depth according

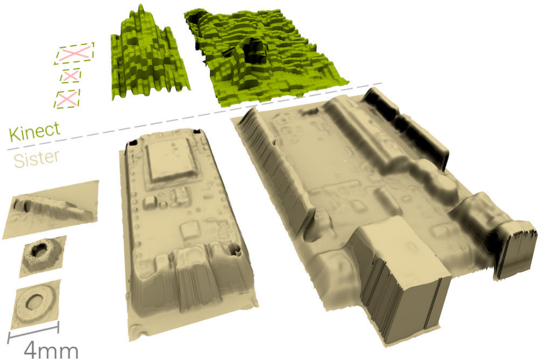


FIGURE 1. Qualitative reconstructions of *tiny* and *small* objects with SiSteR. In green, the inaccurate reconstructions by a Kinect v2, available for the latter components only (red crosses are in place of missing objects).

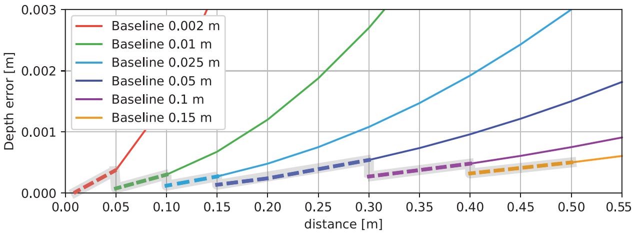


FIGURE 2. Given a stereo rig of fixed f, by varying b the depth error Δ_z (y-axis) changes according to the distance z in the scene (x-axis). A dashed line highlights the initial portion of each curve that linearizes the error by modulating b at will, a feature peculiarly achievable by SiSteR.

[1] https://cvlab-unibo.github.io/sister/

to sensed distance. However, despite ideally unconstrained to the working range, the extent of the overlapping area sensed by the two cameras poses an upper bound to the baseline length depending on the working distance. This fact limits the adoption of off-the-shelf sensors, e.g. Duo MLX,[2] when dealing with the accurate reconstruction of a variety of medium-sized or tiny objects, typically enabling precise reconstruction of the former, but being not able to sense the latter adequately.

For the reasons outlined, our SiSteR framework provides a versatile solution to infer accurate depth reconstruction from robotic arms by employing a single camera mounted as the robot end-effector. Moreover, it also tackles a severe limitation of off-the-shelf 3D sensors when dealing with the accurate 3D reconstruction of tiny objects. Specifically, our proposal has the following advantages:

- A single camera is much more compact than a stereo rig or an active 3D sensor, thus better suited to the *eye-on-hand* configuration.
- The precise movement of the robotic arm allows emulating the acquisition from a stereo rig, thus ensuring epipolar geometry without offline calibration between viewpoints.
- The baseline can be optimally adapted at acquisition time so as to i) enable accurate reconstruction of both tiny and larger objects according to the piecewise linear dashed curve plotted in Figure 2 and ii) maximize the overlapping area between multiple viewpoints.
- The proposed approach allows for addressing some well-known limitations of binocular stereo, e.g. occlusions, by acquiring images from multiple viewpoints. We achieve this by deploying a novel algorithm, referred to as SiSteR Semi-Global Matching.

III. RELATED WORK

In subsection III-A and subsection III-B, we review popular solutions for 3D perception in robotics and stereo vision, since both topics are relevant to our work.

A. 3D PERCEPTION IN ROBOTICS

Nowadays, 3D perception is essential in several robotic applications. However, a fundamental distinction has to be drawn between traditional machine vision approaches and the latest robotic vision paradigms leveraging on an active synergy between the robot and the vision sensor. Among the most compelling examples in this field is the *eye-on-hand* configuration, with the optical sensor positioned on a robotic arm as its end-effector, which enables to go beyond all the limitations of a fixed point-of-view on the scene (e.g. active occlusions avoidance and/or controllable interest region distance). Although there exist several solutions addressing robotic applications [7] with a stationary 3D sensor (*eye-to-hand*), we will mainly deal with cameras mounted directly on a robot because, when feasible, it is a much more flexible setup. In [8] and [9] a stereo camera is directly mounted on the agent to tackle the problem of occlusions in unstructured environments for assistive robotics applications. More recently, with the introduction of consumer 3D active sensors, this field has rapidly evolved further and one of the most popular sensors, probably due to its small size, is the *RealSense*,[3] widely used, e.g,, in the Amazon Picking Challenge for 3D reconstruction and object detection [10]–[12]. Equally compact is the *PrimeSense*[4] (with its branded variant *Asus Xtion*), also widely used for the above mentioned challenge [13] and other applications dealing with robotic manipulation alike [14]–[16]. Another – more expensive and sophisticated – widely deployed sensor is the *Ensenso 3D Camera* [4] thanks to its reduced encumbrance for detection in narrow spaces [17].

One of the most accurate sensors, recently adopted in robotics, is the *Kinect-v2*[5] (successor of *Kinect-v1*). However, being not small in size and quite heavy, it usually does not fit on the final links of the robot, such as the *hand*, but on less dexterous portions of it, as the *head*, with reduced degrees of freedom [18], or – even – on the back of quadruped robots [19]. With even greater accuracy, arm-mounted *Linear Laser Scanners* can also be used in such a way as to integrate linear measurements – during arm movement – to obtain very accurate depth reconstructions [20].

Unfortunately, all previous methods have severe limitations. First of all, most of them have a minimum operating distance (i.e. from 20 to 50 cm), thus preventing a sufficiently accurate perception of small objects. Regarding high-precision sensors, such as laser scanners, their weight and footprint hinders deployment on small and lightweight collaborative robots. Conversely, with SiSteR we propose an easy-to-integrate and flexible solution enabling to accurately reconstruct even objects a few millimeters in size.

B. STEREO VISION

Binocular Stereo: Scharstein and Szeliski [21] classified stereo algorithms into two main broad categories, namely *local* and *global* approaches, according to the different steps carried out: i) cost computation, ii) cost aggregation, iii) disparity optimization/computation and iv) disparity refinement. While local algorithms are typically fast, they are at the same time ineffective in the presence of low-texture regions. On the other hand, global algorithms typically perform better at the cost of higher complexity and runtime. More often than not, Hirschmuller's SGM [22] yields the preferred trade-off between speed and accuracy and thus deployed in most practical applications. Because of its popularity, several works were aimed at improving its accuracy by acting on different steps of the pipeline [23]–[25], so as to address some well-known weaknesses such as streaking artifacts [24] by leveraging confidence measures [26].

[2] https://duo3d.com/product/duo-minilx-lv1

[3] https://en.wikipedia.org/wiki/Intel_RealSense

[4] https://en.wikipedia.org/wiki/PrimeSense

[5] https://it.wikipedia.org/wiki/Microsoft_Kinect

The advent of deep learning in computer vision hit stereo matching as well. Zbontar and LeCun [27] were the first to propose learning a matching function by a Convolutional Neural Network (CNN). This strategy, requiring a reasonable amount of training samples (i.e. few hundreds of images [28]), allows for a more robust similarity estimation between pixels. Moreover, it also allows for improving well-established pipelines such as SGM when deployed in place of traditional matching functions such as absolute difference (AD) or census [29]. Currently, end-to-end paradigms represent the most effective solution to tackle stereo matching [30]–[36]. However, to attain accurate results, a large amount of training data, i.e. thousand of stereo pairs, is required, such amount of samples usually obtained through image synthesis [30]. Furthermore, hundreds of real samples [28] are still necessary to tackle the synthetic-to-real domain shift, which, indeed, does limit the practical deployment of end-to-end stereo networks quite significantly.

Leveraging on more than two viewpoints has the potential to overcome the limits of a binocular setup. For instance, [37], [38] deployed a triangular rig to deal with occlusions, while [39]–[41] used multiple horizontally aligned cameras in order to combine the strengths of short and wide baselines. A virtual trinocular setup was recently proposed in [42] to improve self-supervised monocular depth estimation. Although not very popular, there exist also some off-the-shelf multi-baseline stereo systems, such as the Bumblebee XB3.

In contrast, our proposal relies on a single camera onboard of a robotic arm which is precisely moved to emulate a multi-baseline setup similar to [43], with the difference that we use not only millimeter baselines (in order to perform reconstruction of very tiny objects), but also vertical movements in addition to horizontal ones (to minimize occlusions sources). The very high precision enabled by the robot allows to acquire aligned frames and to run stereo correspondence algorithms, as traditionally performed on rectified images, by choosing the baseline according to the particular task.

IV. SiSteR SEMI-GLOBAL MATCHING

By leveraging on the setup mentioned above to perform multiple single-camera acquisitions, in analogy to traditional binocular stereo, we look for corresponding pixels between the different views to estimate disparity and triangulate depth. Our algorithm extends the SGM pipeline [22], the preferred choice for most binocular stereo frameworks, so as to process an unconstrained number of images. In particular, we will refer to a set of images $\mathcal{I}$ acquired by our framework and to the reference image as the origin O of our scene, on which we will compute the final depth map. Our pipeline consists of three main steps: i) Binocular Matching Cost Computation, ii) Multi-Frame Cost Fusion, iii) Semi-Global Optimization.

A. BINOCULAR MATCHING COST COMPUTATION

At the very beginning of any binocular stereo pipeline, raw matching scores $C^{L,R}(p,d)$ are computed between each pixel p in the left image L and possible candidates $p-d$, $d \in [0, D_{\max}]$ in the right image R. A popular choice consists in using the census transform to assign a binary string to each pixel p computed on a patch $P(p)$, where each bit encodes the relationship between the intensities of p and the remaining N pixels in $P(p)$. Then, matching scores are obtained using the Hamming distance on census transformed images $\mathcal{L}$ and $\mathcal{R}$.

$$C^{L,R}_{\text{census}}(p,d) = \sum_{i=0}^{N} \mathcal{L}(p)[i] \neq \mathcal{R}(p-d)[i] \tag{2}$$

Often, a 9 × 7 patch is chosen to compute the census transform, resulting in 63bit strings.

A recent alternative deals with training a CNN to estimate similarity scores between image patches [27], which provides a much more reliable matching function. Being $s(< P^L(p), P^R(p-d)\ >)$ the output of a CNN trained for this purpose, matching scores are obtained as

$$C^{L,R}_{\text{CNN}}(p,d) = -s(< P^L(p), P^R(p-d) >) \tag{3}$$

By using a fully convolutional network [27] and deploying a high-end GPU (e.g. Titan X), the runtime can be kept equivalent to traditional matching functions.

According to either of two outlined methods, in the very first phase of our pipeline, we compute binocular matching costs $C^{O,T}(p,d)$ between the origin frame O and each additional image $T \in \mathcal{I}$ acquired by our setup.

B. MULTI-FRAME COST FUSION

Traditionally, binocular stereo suffers in the presence of occlusions occurring between the left and the right views. In particular, assuming the former as the reference image, both the left border and the left-most part behind depth discontinuities have no correspondences in the other frame, making the matching assignment for such pixels ill-posed. To overcome this issue, we leverage the setup mentioned above to acquire a set $\mathcal{I}$ made out of pairs of images that are complementary to the origin O of the scene. By keeping a constant baseline between O and each image of a complementary pair, matching pixels for a given p in the O image will be found at the same disparity on both views. This strategy can be extended to all the images in our set $\mathcal{I}$ if such radius (i.e. the baseline) is kept constant, thereby allowing for seamless integration of the matching scores between O and any image in $\mathcal{I}$ on the same cost volume representation centered in O. Thanks to the accurate movement of the robot, the search domain for corresponding points will be 1D as for conventional binocular rectified images. Only the search direction will change according to the relative position between O and the acquired image. For instance, Figure 3 shows a set of three images, made of an origin frame O in between two other images T^x and T^y on its left and right, respectively. Binary costs C^{O,T^x} and C^{O,T^y} (respectively in red and blue) are obtained by searching for the horizontal displacement on the right and left of T^x and T^y, respectively. The combination of such costs into the green

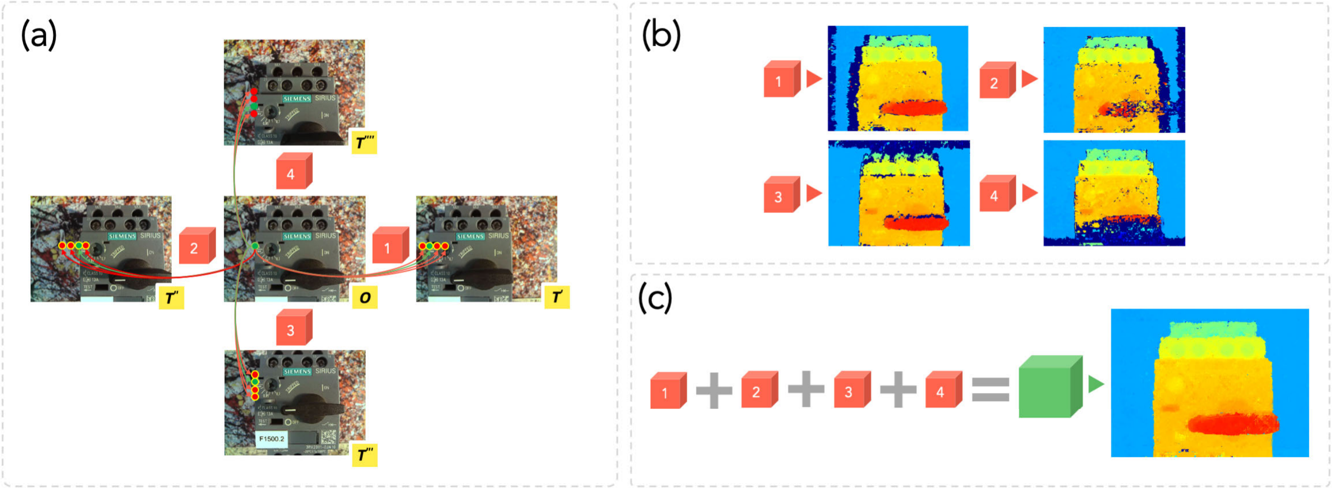


FIGURE 3. SiSteR in action. We acquire a set of images (a), running binocular stereo between origin image *O* and one of the four targets produces inaccurate reconstructions (b), but SiSteR effectively combines all the views to get a high accuracy (c).

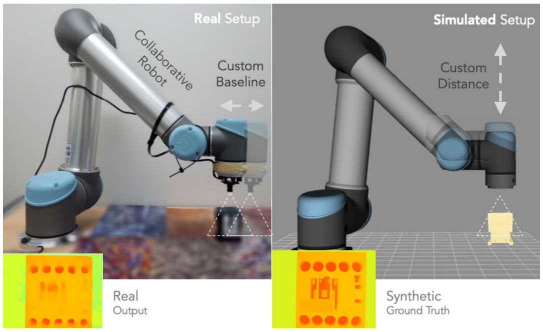


FIGURE 4. On the left, the actual setup used for our experiments. On the right, the simulated counterpart. Synthetic views of real objects are generated in the simulated environment based on their 3D CAD model and used as ground-truth in the experiments.

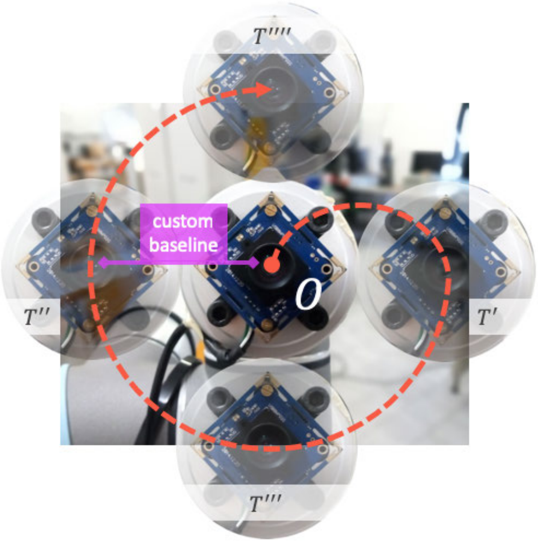


FIGURE 5. Front-facing image of the Camera Module mounted on the robot's flange. By a semi-transparent overlay we also show the four additional vantage points which contribute to gather the images set *I* in Equation 4. The dashed line shows a suitable trajectory to obtain the 5 images.

volume, as outlined below, yields a more accurate disparity map while enabling to handle occluded areas. Consequently, $C^{O,\mathcal{I}}$ is obtained by summing all the pairwise matching costs between O and any other views in $\mathcal{I}$ as follows

$$C^{O,\mathcal{I}}(p,d) = \sum_{T \in \mathcal{I}} \omega^{O,T}(p) \cdot C^{O,T}(p,d) \tag{4}$$

where $\omega^{O,T}(p)$ is a binary confidence score assigned to the matching curve of each pixel p computed between (O, T). This latter term is crucial to neglect the contribution of unreliable pixels that would introduce noise in the matching cost volume. In particular, being occlusions significant sources of mismatches, we obtain $\omega^{O,T}$ by enforcing an Origin-to-Target consistency check (OTC) between disparity computed with respect to both O and T. This is traditionally known as *left-right consistency check* (LRC) in binocular stereo and it is generalized to our setup as follows

$$\omega^{(O,T)}(p) = \begin{cases} 1 & \text{if } |D^O(p) - D^T(p - D^O(p))| \le \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

We will show in the experimental results how leveraging multiple images acquired from the same, moving camera provides higher accuracy compared to a traditional binocular stereo setup.

C. SEMI-GLOBAL OPTIMIZATION

The multi-frame sourced $C^{O,\mathcal{I}}$ term is then further optimized by the subsequent steps of the pipeline to obtain the final disparity map. The SGM framework [22] aims at regularizing a cost volume by means of minimization of an energy function $E(p, d)$, sum of data and smoothness terms $E_{\text{data}}(p, d)$ and $E_{\text{smooth}}(p, d)$

$$E(p,d) = E_{\text{data}}(p,d) + E_{\text{smooth}}(p,d) \tag{6}$$

The data term E_{data} consists of a matching cost $C^{L,R}(p, d)$ in traditional binocular stereo, replaced by $C^{O,\mathcal{I}}(p, d)$ in our

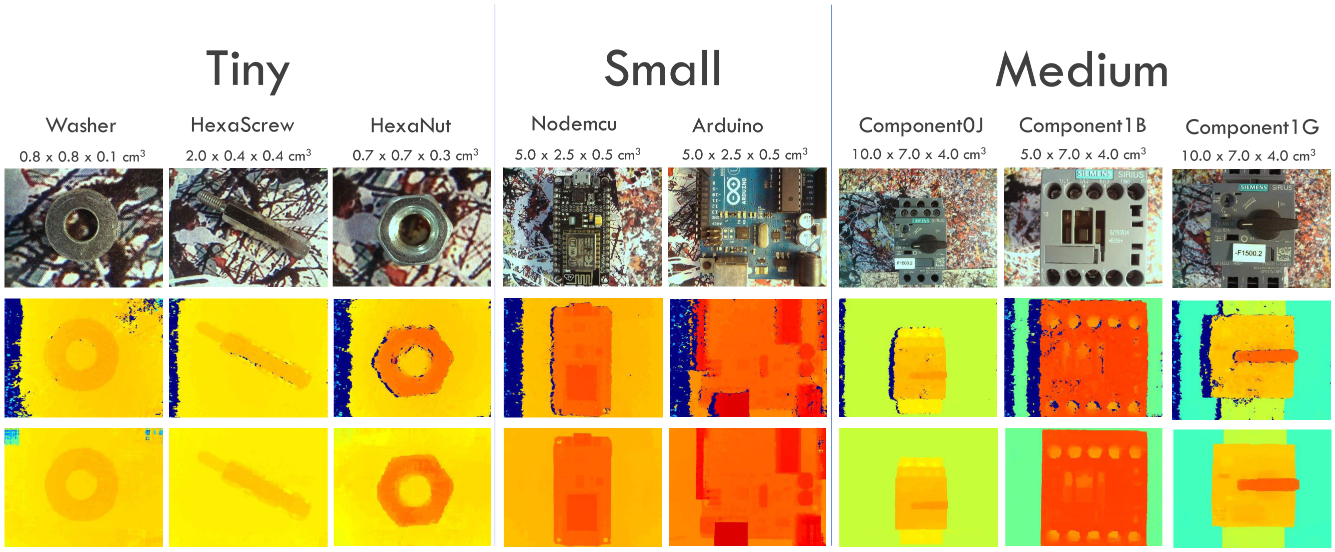


FIGURE 6. Proposed dataset and examples of estimated disparity maps (colormap jet). From left to right, three main categories of objects according to size: Tiny (< 1 cm^3), Small (6.5 cm^3) and Medium (up to 280 cm^3). From top to bottom, reference images, reconstructions by means of binocular SGM (choosing the baseline according to Figure 2) and SiSteR.

setup. The smoothness term $E_{\text{smooth}}(p, d)$ enforces spatial continuity in the disparity domain as follows

$$E_{\text{smooth}}(p, d) = \min_{q>1}[C^{O,\mathcal{I}}(p', d), C^{O,\mathcal{I}}(p', d \pm 1) + P1, C^{O,\mathcal{I}}(p', d \pm q) + P2] - \min_{k<D_{\max}} (C^{O,\mathcal{I}}(p', k)) \tag{7}$$

being P_1 and P_2 two smoothness penalties, discouraging significant disparity gaps between p and previous pixel p' along the scanline path. Finally, a Winner Takes All (WTA) strategy is applied to select for each pixel p the disparity hypothesis d with the minimum cost.

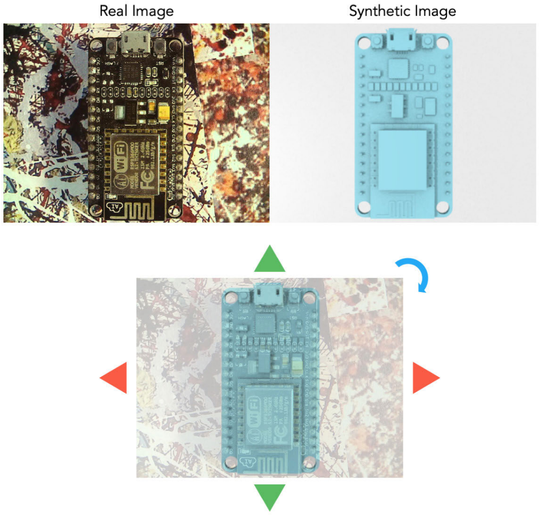


FIGURE 7. Manual alignment (below) between a real image (top left) of an object belonging to our dataset (in this case the *NodeMcu*) and its synthetic model (top right).

V. EXPERIMENTS

A. DATASET ACQUISITION

The setup used in our experiments consists of a *UR5 Universal Robot* with a flange-mounted *VGA Camera Module*, as depicted in Figure 4. Although the camera has been mounted with CAD-aided precision on the flange axis (Figure 5), the extrinsic parameters are further refined using the approach described in [44]. In practice, using a chessboard calibration pattern, stationary in the world, and framing it from multiple viewpoints with the *eye-on-hand* camera, one can estimate the pose of the camera with respect to the end-effector by solving a bundle adjustment problem [45].

Thanks to the high precision and accuracy provided by the robot, it is feasible to control the camera 6-DoF pose arbitrarily, i.e. as if it were the end-effector of the manipulator. With these operational conditions, it was possible to collect a dataset (shown in Figure 6), by varying the object-camera distance and producing several images sets $\mathcal{I}$, as described in section IV. At the same time, due to availability of 3D CAD models for all objects, it was possible to reproduce a simulated counterpart (often referred to as *digital twin*) of the real scenes, as illustrated in Figure 4. This simulation made it possible to move a virtual camera according to the robot feedback and then render the Z-Buffer of the corresponding viewpoint so as to obtain a synthetic *Depth ground-truth*, as long as the 3D pose of each object, in the robot's reference frame, is computed. Although calculating this 3D pose may seem like a hard-working approach, it is straightforward if the extrinsic parameters of the camera and the precise elevation of the worktable are known. The latter can be easily inferred upon installation of the robot. In our settings, we have chosen to place all the objects in an upright position. Thus, by knowing the exact elevation of the plane on which they lay, estimating their 3D pose boils down to finding the

TABLE 1. Experiments on the proposed dataset. We collected eight objects, grouped into Medium (a), Small (b) and Tiny (c) according to their size. Each object was acquired from two different distances.

	Component0J		Component1B		Component1G	
	5 cm	10 cm	5 cm	10 cm	5 cm	10 cm
SGM - COTS	1.6499	0.4395	1.8028	1.1047	1.3098	0.2779
SGM - BTB	0.1870	0.4395	0.1158	1.1047	0.1810	0.2779
SiSteR-SGM w/o ω	0.1297	0.2037	0.1334	0.4810	0.0556	0.2138
SiSteR-SGM	0.0619	0.0496	0.0324	**0.3261**	0.0216	0.0230
SiSteR-MC-CNN-fst	**0.0193**	**0.0273**	**0.0091**	0.3732	**0.0128**	**0.0188**

(a) Medium split: *Component0J*, *Component1B* and *Component1G*, acquired at 5cm and 10cm.

	Arduino		Nodemcu	
	1 cm	5 cm	1 cm	5 cm
SGM - COTS	2.4032	1.4700	2.1424	1.5253
SGM - BTB	0.1059	0.3213	0.0864	0.3326
SiSteR-SGM w/o ω	0.0548	0.2142	0.0940	0.0797
SiSteR-SGM	**0.0310**	0.0748	**0.0371**	0.0068
SiSteR-MC-CNN-fst	0.0759	**0.0549**	0.0512	**0.0050**

(b) Small split: *Arduino* and *Nodemcu*, acquired at 1cm and 5cm.

	Hexa Nut		Hexa Screw		Washer	
	1 cm	5 cm	1 cm	5 cm	1 cm	5 cm
SGM - COTS	2.1076	1.9321	1.9135	1.3534	2.0542	1.8016
SGM - BTB	0.0539	0.5522	0.0419	0.4223	0.0908	0.3864
SiSteR-SGM w/o ω	0.0534	0.1499	0.0160	0.1294	0.0532	0.1429
SiSteR-SGM	0.0090	0.0418	0.0057	0.0286	**0.0126**	0.0174
SiSteR-MC-CNN-fst	**0.0013**	**0.0009**	**0.0027**	**0.0080**	0.0258	**0.0070**

(c) Tiny split: *Hexa Nut*, *Hexa Screw* and *Washer*, acquired at 1cm and 5cm.

3 degrees of freedom (*x*-*y* plane coordinates and *yaw* angle) with which they can be placed on the working table. Finding these 3 degrees of freedom manually by comparing the two RGB images, i.e. the real and the synthetic ones, turns out a trivial problem, as illustrated in Figure 7.

Hence, by determining the exact 3D pose of our objects in the real and virtual world, we can move the real and virtual cameras simultaneously to collect any viewpoint of our targets with a depth ground truth perfectly aligned to the real RGB image. We collected images according to a set of baselines (listed in Figure 2) and from different distances to the target object (1 cm, 5 cm, 10 cm). This setup will allow for highlighting the limitations of standard off-the-shelf stereo cameras, ineffective for very close acquisitions whereas our framework succeeds. The dataset is made of eight objects, depicted in Figure 6, organized into three main splits: *Tiny*, *Small* and *Medium* according their size. In the figure, we report name and dimensions on top of each component.

B. 3D RECONSTRUCTION ACCURACY

We evaluate the effectiveness of SiSteR by reconstructing the objects included in our dataset. Purposely, we measure the difference between the estimated depth maps and the ground-truth, obtained through the alignment procedure described in subsection V-A, by computing the Root Mean Square Error (RMSE).

We report the outcome of these experiments in Table 1a, Table 1b and Table 1c for Medium, Small and Tiny objects, respectively. As reported in the tables, for each object we acquired images from two distances, with the best result for each distance highlighted in bold and the best for each object in red. Besides, in each table, we report results obtained by the following approaches and configurations:

- **SGM - COTS**, by processing image pairs acquired with a fixed stereo baseline (25mm) and the SGM algorithm [22], with the purpose of emulating a *commercial off-the-shelf* (COTS) device, e.g. the Duo MLX camera.
- **SGM - BTB**, by processing image pairs acquired *beyond the baseline* (BTB), traditionally fixed in stereo cameras. The baseline is set in order to minimize the error with respect to the distance according to Equation 1, i.e. $2mm$, $10mm$ and $25mm$ at distances $1cm$, $5cm$ and $10cm$, respectively.
- **SiSteR-SGM w/o** ω, our baseline implementation of SiSteR leveraging four target views, as in Figure 3, without occlusion handling during cost aggregation.
- **SiSteR-SGM** in its full configuration, i.e. with occlusion handling.
- **SiSteR-MC-CNN-fst**, replacing in the previous configuration the conventional matching cost computation by a Convolutional Neural Network [27] in charge of estimating initial matching costs between pixels across the five images.

Except for the latter, tuned as in [27] and deployed without additional optimizations, such as CBCA and post-processing to avoid over-smoothing, all other configurations rely on a 7×9 census transform to compute the initial matching costs,

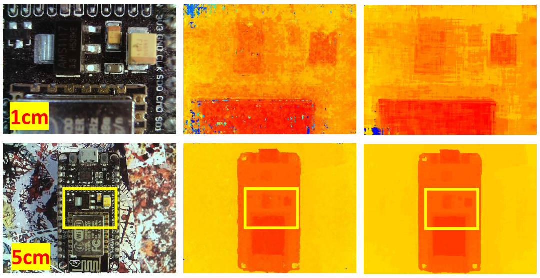


FIGURE 8. Nodemcu acquisitions at 1cm (top) and 5cm (bottom). From left to right, origin image *O*, SiSteR-SGM and SiSteR-MC-CNN-fst depth estimations. The low textured appearance from a very close viewpoint leads to noisy estimates while acquiring the object from a higher distance and with an appropriate baseline leads to more accurate reconstructions.

P1 and P2 are set according to [46], and ε is set to 3. For all methods, pixels without valid disparities (e.g. discarded by LRC in the binocular case) are interpolated as in [28]. In the reminder, we will analyze in detail how each of the configurations listed above does behave across the three splits of the dataset.

1) BEYOND THE BASELINE

By looking at the results achieved by SGM-COTS across the three splits, we can perceive how an off-the-shelf stereo camera is not suited to close range accurate depth perception. In particular, although it is quite reliable at 10 cm when dealing with Medium components, as shown in Table 1a, the 25 mm baseline chosen to emulate such setup fails at estimating reliable depth at 1cm and 5cm distance from the sensors, these distances being required to achieve enough resolution in the case of Small and Tiny objects, as shown in Table 1b and Table 1c. This behaviour is mainly due to the chosen baseline being too wide for these very close acquisitions, dramatically reducing the overlapping regions in the two images and thus preventing many correspondences between pixels to be established. This results in noisy disparity maps and much higher RMSE, as reported in row 1 in the tables.

In contrast, removing the fixed baseline constraint (BTB) leads to much more accurate results on the very close acquisitions at 1 cm and 5 cm. Although the accuracy achieved at 10 cm by a COTS solution is not improved, confirming that the 25 mm baseline is effective in this latter case, the capability of going beyond the fixed baseline enabled by our single camera configuration allows for much more precise reconstructions at closer distances, crucial to accurately reconstruct details in small and tiny objects.

2) SiSteR AND OCCLUSIONS HANDLING

Considering the complete acquisition setup enabled by SiSteR, we can further improve the quality of reconstructed objects. At first, we focus on Medium objects and acquisitions at 10cm distance, i.e. the case for which the considered COTS solution is effective. Although the baseline distance is not an issue under this hypothesis, occlusions still limit the effectiveness of a binocular stereo algorithm. Running SiSteR-SGM even without explicit occlusion handling (w/o ω) already soften the errors in these regions (rows 1 vs 3 in the table), further improved when explicit handling is enabled (row 4). Thus, improvements achieved by SiSteR setup are not limited by the possibility of choice of the baseline distance.

Moving to close acquisition distances, i.e. 5 and 1 cm, COTS baseline is no longer effective and thus we compare the SiSteR framework with SGM-BTB, observing moderate improvements introduced without explicit occlusion handling (w/o ω), with just few exceptions as reported by rows 2 and 3. In particular, this occurs in the presence of large occlusions in Medium and Small objects, i.e. Component1B-5 cm and Nodemcu-1 cm, where the cost fusion does not filter out noisy matching scores near boundaries and a proper strategy is necessary to overcome this issue. Indeed, the accuracy consistently improves on all configurations when enabling the OTC check, as it can be perceived by comparing rows 3 and 4 of each table. In particular, all SiSteR-SGM setups outperform both binocular strategies when handling occlusions by the OTC check.

3) SiSteR AND DEEP LEARNING

Finally, we embed a learning-based matching function in SiSteR when running the binocular matching cost computation step, as this this strategy is known to be more accurate than traditional matching measures like the census transform. We choose the MC-CNN-fst network by Zbontar and LeCun [27] because of its fast processing time on a high-end GPU hardware. Although such setup is currently not common in most practical robotic applications, this experiment proves that SiSteR is amenable to the more recent trends in computer vision. However, obtaining enough samples to train the network from scratch may require a substantial overhead. Thus, in our experiments, we use the weights made available by the authors [27] and trained on the KITTI dataset [28], as they better generalize to our environment compared to the Middlebury v3 weights. The last row in each table report the accuracy achieved with this configuration, namely *SiSteR-MC-CNN-fst*. We can see how it produces the most accurate results on any object in our dataset in most cases, with very few exceptions such as for Arduino-1 cm on which SiSteR-SGM results much more effective. We ascribe this to the challenging, reflective surface shown in Figure 9. In fact, despite the excellent quality of the overall reconstruction, the magnitude of errors onto the reflective surfaces is much higher deploying the learning-based matching cost MC-CNN-fst. Moreover, sometimes *SiSteR-MC-CNN-fst* performs worse when dealing with close-range acquisitions, where it is outperformed by *SiSteR-SGM*, like in the case of Nodemcu-1 cm and Washer-1 cm. This is caused by the very different image content observed at such close distance with respect to what observed at training time.

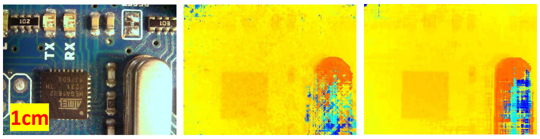


FIGURE 9. Arduino acquisition at 1cm. From left to right, origin image *O*, SiSteR-SGM and SiSteR-MC-CNN-fst depth estimates.

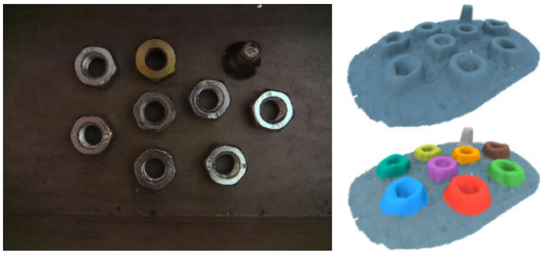


FIGURE 10. Reconstruction (top right) of a *Bunch of Nuts* (left image). The bottom-right picture shows the 3D instance segmentation obtained by a Plane Segmentation Algorithm [47].

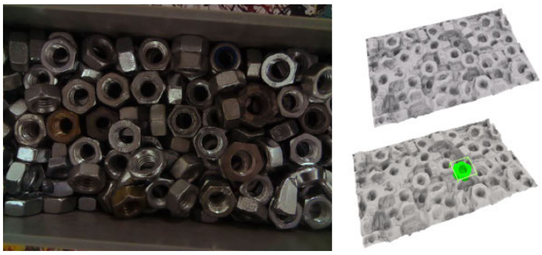


FIGURE 11. Reconstruction (top right) of a *Bin of nuts* (left image). The bottom-right picture shows detection of an item by a 3D Object Detection pipeline [48].

4) RUNTIME ANALYSIS

Compared to a classical binocular stereo algorithm, the execution time required to run SiSteR is bound to the number of cost volumes computed. For instance, to run SiSteR-MC-CNN-fst on the whole set of targer views requires to compute four cost volumes, thus requiring 3.2 secs instead of 0.8 on a Titan X GPU [27].

C. QUALITATIVE RESULTS AND APPLICATIONS

Finally, we report some qualitative results obtained by SiSteR, showing some examples of potential applications enabled by our novel 3D reconstruction approach. Figure 10 depicts the outcome of a 3D segmentation algorithm applied to a bunch of nuts acquired by our robot. We can see how running a Plane Segmentation Algorithm [47] allows us to easily detect all the single nuts laying on the plane. Figure 11 illustrates the results obtained by means of a 3D object detection pipeline [48] run on a large set of nuts. We can perceive again how the fine details recovered by the SiSteR reconstruction allow for effective detection of a single instance of multiple objects in the scene.

D. SUPPLEMENTARY MATERIAL

In the supplementary material, we show a *proof-of-concept* robotic manipulation task, of a *tiny* object, in which the 3D reconstruction of the environment is entirely carried out by Sister. Figure 12 frames the experimental setup on which we deploy SiSteR to fully reconstruct the scene with multiple levels of detail, reducing the baseline according to the size of the target objects, in order to obtain a point cloud (b) or mesh (c) representation. We refer the reader to the supplementary video material for a full demonstration.

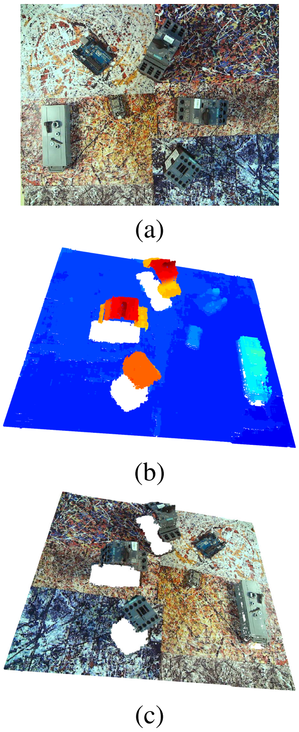


FIGURE 12. Example of experimental setup. (a) Top view RGB image, (b) reconstructed point cloud, (c) RGB mesh.

VI. CONCLUDING REMARKS

We have proposed a simple, yet effective, approach to 3D reconstruction based on robotic vision which is particularly amenable to small objects. Indeed, the proposed SiSteR approach enables 3D perception of the environment by leveraging a *camera* mounted in *eye-on-hand* configuration on a *robotic manipulator*. Moreover, our OpenSource implementation, available either as a ROS service [49] or a standalone C++ library, makes it straightforward to generate a Colored Point Cloud associated to the central vantage point of the cross-shaped – variable baseline – acquisition setup sketched in Figure 5.

We have assessed the performance of our proposal on a specific viewpoints configuration (the *cross-shape*) on a novel dataset with ground-truth aimed at reconstruction of Medium, Small and Tiny objects. Our experiments vouch for the superior quality of the 3D reconstructions enabled by SiSteR, thanks to i) its capability of adjusting the baseline to keep the error linear with respect to the distance from the acquired objects and maximize the overlap between images, and ii) the joint deployment of multiple viewpoints in order to handle occlusions.

As future work, we plan to carry out a more in-depth investigation on other possible viewpoint configurations as well as on the deployment of the minimum baseline allowed by the current set-up (2mm). As a further evolution of SiSteR, we will also consider the end-to-end deep learning approaches that proved to be very effective in standard binocular stereo.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

[1] D. Drover, M. V. Rohith, C.-H. Chen, A. Agrawal, A. Tyagi, and C. P. Huynh, "Can 3D pose be learned from 2D projections alone?" in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2018, pp. 78–94.

[2] V. Nityananda and J. C. A. Read, "Stereopsis in animals: Evolution, function and mechanisms," *J. Exp. Biol.*, vol. 220, no. 14, pp. 2502–2512, Jul. 2017.

[3] V. Nityananda, G. Tarawneh, S. Henriksen, D. Umeton, A. Simmons, and J. C. Read, "A novel form of stereo vision in the praying mantis," *Current Biol.*, vol. 28, no. 4, pp. 588–593, 2018.

[4] E. GmbH. *Stereo 3D Cameras for Industrial Applications*. [Online]. Available: https://www.ensenso.com/

[5] M. A. Lewis and M. E. Nelson, "Look before you leap: Peering behavior for depth perception," in *From animals to animats*, vol. 5, R. Pfeifer, B. Blumberg, J. A. Meyer, and S. W. Wilson, Eds. Citeseer, 1998, pp. 98–103.

[6] A. M. Pinto, P. Costa, A. P. Moreira, L. F. Rocha, G. Veiga, and E. Moreira, "Evaluation of depth sensors for robotic applications," in *Proc. IEEE Int. Conf. Auto. Robot Syst. Competitions*, Apr. 2015, pp. 139–143.

[7] A. Zeng, S. Song, K. T. Yu, E. Donlon, F. R. Hogan, M. Bauza, and D. Ma, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1–8.

[8] C. Dune, C. Leroux, and E. Marchand, "Intuitive human interaction with an arm robot for severely handicapped people–a one click approach," in *Proc. IEEE 10th Int. Conf. Rehabil. Robot.*, Jun. 2007, pp. 582–589.

[9] D.-J. Kim, R. Lovelett, and A. Behal, "Eye-in-hand stereo visual servoing of an assistive robot arm in unstructured environments," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 2326–2331.

[10] A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6D pose estimation in the Amazon picking challenge," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 1383–1386.

[11] M. Schwarz, A. Milan, C. Lenz, A. Munoz, A. S. Periyasamy, M. Schreiber, S. Schüller, and S. Behnke, "NimbRo picking: Versatile part handling for warehouse automation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 3032–3039.

[12] H. Zhang, P. Long, D. Zhou, Z. Qian, Z. Wang, W. Wan, D. Manocha, C. Park, T. Hu, C. Cao, Y. Chen, M. Chow, and J. Pan, "DoraPicker: An autonomous picking system for general objects," in *Proc. IEEE Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2016, pp. 721–726.

[13] C. Eppner, S. Höfer, R. Jonschkowski, R. M. Martin, A. Sieverling, V. Wall, and O. Brock, "Lessons from the Amazon picking challenge: Four aspects of building robotic systems," in *Robotics: Science and Systems*. 2016. [Online]. Available: http://www.roboticsproceedings.org/

[14] J. Stria, D. Prusa, V. Hlavac, L. Wagner, V. Petrik, P. Krsek, and V. Smutny, "Garment perception and its folding using a dual-arm robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2014, pp. 61–67.

[15] C. Matuszek, B. Mayton, R. Aimi, M. P. Deisenroth, L. Bo, R. Chu, M. Kung, L. LeGrand, J. R. Smith, and D. Fox, "Gambit: An autonomous chess-playing robotic system," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 4291–4297.

[16] D. De Gregorio, F. Tombari, and L. Di Stefano, "RobotFusion: Grasping with a robotic manipulator via multi-view reconstruction," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 634–647.

[17] C. Hernandez, M. Bharatheesha, W. Ko, H. Gaiser, J. Tan, K. van Deurzen, M. de Vries, B. Van Mil, J. van Egmond, R. Burger, M. Morariu, J. Ju, X. Gerrmann, R. Ensing, J. Van Frankenhuyzen, and M. Wisse, "Team delft's robot Winner of the Amazon picking challenge 2016," in *Robot World Cup*. Berlin, Germany: Springer, 2016, pp. 613–624.

[18] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.*, vol. 34, nos. 4–5, pp. 705–724, 2015.

[19] P. Fankhauser, M. Bloesch, D. Rodriguez, R. Kaestner, M. Hutter, and R. Siegwart, "Kinect V2 for mobile robot navigation: Evaluation and modeling," in *Proc. Int. Conf. Adv. Robot. (ICAR)*, Jul. 2015, pp. 388–394.

[20] S. Sharifzadeh, I. Biro, N. Lohse, and P. Kinnell, "Abnormality detection strategies for surface inspection using robot mounted laser scanners," *Mechatronics*, vol. 51, pp. 59–74, May 2018.

[21] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, Apr. 2002.

[22] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 807–814.

[23] M.-G. Park and K.-J. Yoon, "Leveraging stereo matching with learning-based confidence measures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 101–109.

[24] M. Poggi and S. Mattoccia, "Learning a general-purpose confidence measure based on $O(1)$ features and a smarter aggregation strategy for semi global matching," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 509–518.

[25] A. Seki and M. Pollefeys, "Patch based confidence prediction for dense disparity map," in *Proc. Brit. Mach. Vis. Conf.*, 2016, p. 4.

[26] M. Poggi, F. Tosi, and S. Mattoccia, "Quantitative evaluation of confidence measures in a machine learning world," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5228–5237.

[27] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, nos. 1–32, p. 2, 2016.

[28] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[29] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. 3rd Eur. Conf. Comput. Vis.*, vol. 2. Secaucus, NJ, USA: Springer-Verlag, 1994, pp. 151–158.

[30] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4040–4048.

[31] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang, "Learning for disparity estimation through feature constancy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2811–2820.

[32] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "GA-Net: Guided aggregation net for end-to-end stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 185–194.

[33] M. Poggi, D. Pallotti, F. Tosi, and S. Mattoccia, "Guided stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 979–988.

[34] A. Tonioni, F. Tosi, M. Poggi, S. Mattoccia, and L. D. Stefano, "Real-time self-adaptive deep stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 195–204.

[35] M. Poggi, A. Tonioni, F. Tosi, S. Mattoccia, and L. Di Stefano, "Continual adaptation for deep stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 28, 2021, doi: 10.1109/TPAMI.2021.3075815.

[36] M. Poggi, F. Tosi, K. Batsos, P. Mordohai, and S. Mattoccia, "On the synergies between machine learning and binocular stereo for depth estimation from images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 5, 2021, doi: 10.1109/TPAMI.2021.3070917.

[37] M. Ito and A. Ishii, "Three-view stereo analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 4, pp. 524–532, Jul. 1986.

[38] N. Ayache and F. Lustman, "Trinocular stereo vision for robotics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 1, pp. 73–85, Jan. 1991.

[39] M. Okutomi and T. Kanade, "A multiple-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 4, pp. 353–363, Apr. 1993.

[40] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys, "Variable baseline/resolution stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[41] D. Honegger, T. Sattler, and M. Pollefeys, "Embedded real-time multi-baseline stereo," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 5245–5250.

[42] M. Poggi, F. Tosi, and S. Mattoccia, "Learning monocular depth estimation with unsupervised trinocular assumptions," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 324–333.

[43] H. Farid, S. W. Lee, and R. Bajcsy, "View selection strategies for multi-view, wide-baseline stereo," Tech. Rep., 1994.

[44] A. Tabb and K. M. A. Yousef, "Solving the robot-world hand-eye (S) calibration problem with iterative methods," *Mach. Vis. Appl.*, vol. 28, nos. 5–6, pp. 569–590, 2017.

[45] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—A modern synthesis," in *Proc. Int. Workshop Vis. Algorithms*. Berlin, Germany: Springer, 1999, pp. 298–372.

[46] R. Spangenberg, T. Langner, S. Adfeldt, and R. Rojas, "Large scale semi-global matching on the CPU," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2014, pp. 195–201.

[47] R. B. Rusu, "Semantic 3D object maps for everyday manipulation in human living environments," *KI-Künstliche Intelligenz*, vol. 24, no. 4, pp. 345–348, 2010.

[48] A. Aldoma, F. Tombari, L. Di Stefano, and M. Vincze, "A global hypotheses verification method for 3D object recognition," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 511–524.

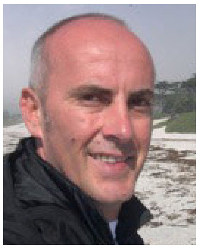
[49] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "ROS: An open-source robot operating system," in *Proc. ICRA Workshop Open Source Softw.*, 2009, vol. 3, no. 3. Kobe, Japan, p. 5.

DANIELE DE GREGORIO (Member, IEEE) received the B.Sc. and M.Sc. degrees from the University of L'Aquila, Italy, in 2008 and 2012, respectively, and the Ph.D. degree in software engineering from the University of Bologna, Italy, in 2018. He was a Postdoctoral Researcher with the University of Bologna in the field of Robotic Vision. He has been a Software Consultant outside of the University for over ten years. He is currently the Co-Founder and the CEO of EYECAN.ai S.r.l. with the University of Bologna spin-off company dealing with deep learning and robotics in the industrial field. He is the author of more than 15 publications and two patents.

MATTEO POGGI (Member, IEEE) received the master's and Ph.D. degrees in computer science and engineering from Alma Mater Studiorum, University of Bologna, in 2014 and 2018, respectively. He is currently a Postdoctoral Researcher with the Department of Computer Science and Engineering, University of Bologna. His research interest includes depth estimation from images. In this field, he published about 50 papers.

PIERLUIGI ZAMA RAMIREZ (Member, IEEE) received the master's and Ph.D. degrees in computer science and engineering from Alma Mater Studiorum, University of Bologna, in 2017 and 2021, respectively. He is currently a Postdoctoral Researcher at the University of Bologna. His research interests include deep learning and computer vision.

GIANLUCA PALLI (Senior Member, IEEE) received the Laurea and Ph.D. degrees in automation engineering from the University of Bologna, Bologna, Italy, in 2003 and 2007, respectively. He is currently an Associate Professor with the University of Bologna. He is the author or a co-author of over 80 scientific papers presented at conferences or published in journals. His research interests include the design and control of robotic hands, the modeling and control of robots with variable stiffness joints, the design of compliant structures and actuation systems for robotics applications, and the development of real-time systems for automatic control applications.



STEFANO MATTOCCIA (Member, IEEE) received the Ph.D. degree in computer science engineering from the University of Bologna, in 2002. He is currently an Associate Professor with the Department of Computer Science and Engineering, University of Bologna. His research interests include computer vision, depth perception, embedded vision, and deep learning. In these fields, he has authored about 100 scientific publications/patents.

LUIGI DI STEFANO (Member, IEEE) received the Ph.D. degree in electronic engineering and computer science from the University of Bologna, in 1994. He is currently a Full Professor with the Department of Computer Science and Engineering, University of Bologna, where he founded and leads the Computer Vision Laboratory (CVLab). He is the author of more than 150 papers and several patents. His research interests include image processing, computer vision, and machine/deep learning. He has been a Scientific Consultant for major companies in the fields of computer vision and machine learning. He is a member of the IEEE Computer Society and the IAPR-IC.