ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Open Challenges in Deep Stereo: the Booster Dataset

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Ramirez, P.Z., Tosi, F., Poggi, M., Salti, S., Mattoccia, S., Di Stefano, L. (2022). Open Challenges in Deep Stereo: the Booster Dataset. IEEE [10.1109/CVPR52688.2022.02049].

Availability:
This version is available at: https://hdl.handle.net/11585/895291 since: 2022-10-03

Published:

DOI: http://doi.org/10.1109/CVPR52688.2022.02049

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

16 November 2026

This is the final peer-reviewed accepted manuscript of:

P. Z. Ramirez, F. Tosi, M. Poggi, S. Salti, S. Mattoccia and L. Di Stefano, "Open Challenges in Deep Stereo: the Booster Dataset," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 21136-21146

The final published version is available online at ***https://dx.doi.org/10.1109/CVPR52688.2022.02049***

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

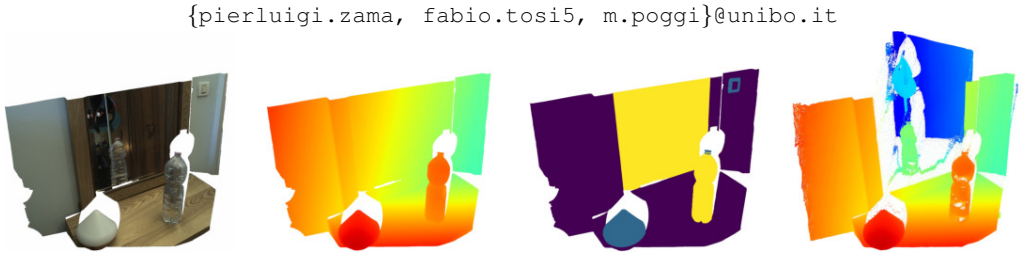
This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Open Challenges in Deep Stereo: the Booster Dataset

Pierluigi Zama Ramirez* Fabio Tosi* Matteo Poggi*
Samuele Salti Stefano Mattoccia Luigi Di Stefano
CVLAB, Department of Computer Science and Engineering (DISI)
University of Bologna, Italy
{pierluigi.zama, fabio.tosi5, m.poggi}@unibo.it



(a) (b) (c) (d)

Fig. 1. **A scene in the Booster dataset.** We collect images in a variety of indoor environments featuring challenging objects such as the mirror shown in (a). We provide dense ground-truth disparities (b) and segmentation masks that identify the most challenging materials (c). As state-of-the-art deep stereo networks [25] struggle on these scenes (d), our benchmark highlights the open challenges in deep stereo.

Abstract

We present a novel high-resolution and challenging stereo dataset framing indoor scenes annotated with dense and accurate ground-truth disparities. Peculiar to our dataset is the presence of several specular and transparent surfaces, i.e. the main causes of failures for state-of-the-art stereo networks. Our acquisition pipeline leverages a novel deep space-time stereo framework which allows for easy and accurate labeling with sub-pixel precision. We release a total of 419 samples collected in 64 different scenes and annotated with dense ground-truth disparities. Each sample include a high-resolution pair (12 Mpx) as well as an unbalanced pair (Left: 12 Mpx, Right: 1.1 Mpx). Additionally, we provide manually annotated material segmentation masks and 15K unlabeled samples. We evaluate state-of-the-art deep networks based on our dataset, highlighting their limitations in addressing the open challenges in stereo and drawing hints for future research.

1. Introduction

Depth estimation from images has long been deemed a favourable alternative compared to expensive and intrusive active sensors. Among several image-based approaches, stereo vision [32,36] is arguably the most popular and heavily researched technique. In the years, huge progresses have been made in this field, also thanks to the availability of challenging stereo benchmarks [15, 29, 34, 38] where the community competes for the higher ranks. Moreover, the abundance of stereo images paved the way for deep learning to succeed also in this field [22,28,54]. Indeed, by browsing the most popular benchmarks, one can notice how nowadays all the top-ranking proposals consist in end-to-end deep networks that can reach sub-pixel precision in most cases. Just to name a few, KITTI 2012 and 2015 [15, 29] or ETH3D [38] seem solved, with top entries achieving average error rates near to 1%. Should this evidence suggest that, thanks to deep learning, stereo vision is a solved problem? As shown in Fig. 1, we believe that this is definitely not the case and, rather, it is time for the community to focus on the **open-challenges** left unsolved in the field. In particular, we identify two of such challenges, namely i) non-Lambertian surfaces and ii) high-resolution images.

As for non-Lambertian reflectivity, a variety of materials and surfaces still represent a hard challenge to most computer vision methodologies and to deep stereo alike. Specifically, matching pixels dealing with transparent or specular surfaces is extremely difficult and may consist in an inherently ill-posed problem in many cases. Yet, we reckon that objects with such properties are almost absent or unlabeled in most stereo benchmarks, except for KITTI 2015, where cars have been replaced with CAD models providing supervision on some specular/transparent surfaces on cars. As reported in the KITTI 2015 online benchmark, deep learning has the potential to tackle this challenge as well, if properly

* Joint first authorship.

annotated samples are available.

Concerning the second challenge, when considering higher-resolution images, for instance in the Middlebury 2014 benchmark [34], we can notice in general higher errors. These are caused by the much larger image dimensions (and thus disparity range) and, consequently, by a larger number of occluded and untextured pixels in the images framed in this dataset. Besides, processing images at high resolution sets forth computational complexity issues, in particular when deploying deep networks. Indeed, most of the entries in the Middlebury benchmark can only process input images downsampled to half or quarter of the original 6 Mpx resolution. Moreover, an additional challenge emerges due the peculiar camera setup featured by modern smartphones, typically equipped with both a high resolution and a much lower resolution image sensors. In such a setting, one may wish to recover a high resolution depth map despite the different resolution of the input pair, i.e., solve an *unbalanced* stereo problem. However, such a research direction has been only barely explored so far [1, 26].

To this aim, in this paper we present a novel high-resolution challenging stereo benchmark. Each image in our dataset, collected in indoor environments, features a set of objects and surfaces that are either specular or transparent, as well as very large untextured regions. To accurately annotate each collected sample, we implement a novel *deep* space-time stereo pipeline [10] which combines disparity estimates computed from multiple static images – up to 100 – acquired under a variety of texture patterns projected onto the scene from different directions and after having carefully painted all non-Lambertian surfaces. Peculiar to our pipeline is the use of a state-of-the-art, pretrained deep network [25] to compute the individual disparity maps accumulation through time within the space-time framework. Furthermore, a final careful manual cleaning is carried out to remove outliers/artefacts and ensure high-quality disparity labels. We point out that for some non-Lambertian surfaces it might be possible to provide multiple depth ground-truths: for instance, for transparent surfaces we might provide both depths for the surface itself and the objects seen through the surface. Yet, in our dataset we provide depth labels for the closest surfaces only, thereby enabling evaluation and training of stereo methods designed to return a single depth prediction per pixel. As such, our dataset mainly addresses scenarios dealing with autonomous driving, obstacle avoidance and robotic manipulation, while being less amenable to applications such as AR and novel view synthesis. The main contributions of our paper are:

• We propose a novel dataset consisting of both high-resolution as well as unbalanced stereo pairs featuring a large collection of labeled non-Lambertian objects. In particular, we have acquired a total of 64 scenes under different illuminations, yielding 419 *balanced* stereo pairs at 12 Mpx and 419 *unbalanced* pairs, each consisting in a 12 Mpx and 1.1 Mpx image. The latter setup provides the **first-ever** dataset for unbalanced stereo matching, as prior work is limited to simulation experiments [1, 26]. In both setups, samples are annotated with dense ground-truth disparities and grouped into 228 training images and 191 test images – for which ground-truth is withheld.

• Data annotation is performed in a semi-automatic manner based on a novel deep space-time stereo framework, which enables to deploy modern stereo networks [25] within the well-known space-time stereo framework [10].

• Alongside with ground-truth disparities, we provide manually annotated segmentation maps that identify and rank the hard-to-match materials based on specularity and transparency. This is conducive to focus on the open-challenges addressed in this paper when analyzing the behaviour of state-of-the-art networks. Moreover, we provide an additional set of 15K raw pairs, both in balanced and unbalanced settings, to encourage the development of weakly-supervised solutions to the open challenges in stereo.

• We evaluate the prominent state-of-the-art stereo networks [5, 9, 49, 55], as trained by their authors, on the test split of our dataset. The experimental findings highlight the open-challenges that need to be faced by the stereo community and provide hints on possible future research directions.

Our Benchmark on open-challenges in stereo (**Booster**) is available at https://cvlab-unibo.github.io/booster-web/.

2. Related Work

We briefly review the literature relevant to our work.

Traditional and Deep Stereo. For years, most algorithms have been developed following a common pipeline sketched in [36], starting with matching cost computation and successive optimization strategies. Among the vast literature on traditional algorithms [11, 12, 20, 52, 53], Semi-Global Matching (SGM) [18] is by far the most popular. With the advent of deep learning, the first research efforts focused on formulating the individual steps of the conventional pipeline [36] as learnable neural networks, *e.g.* matching cost computation [8, 27, 54], optimization [40, 41] and refinement [1, 3, 17]. Then, end-to-end deep stereo networks rapidly gained the main stage [22, 28, 30], thanks to the top-positions achieved on the KITTI 2012 [15] and 2015 [29] benchmarks.

This research direction produced a large variety of deep stereo architectures [5, 9, 13, 24, 43, 46, 49, 51, 55], as surveyed in [32], as well as investigations on self-supervised learning strategies [2, 23, 31, 44, 45, 48, 57], zero-shot generalization across datasets [1, 4, 56] and, more recently, unbalanced stereo setups [1, 26].

Stereo benchmarks. Among the factors behind the intensive research on stereo vision, the increasing availability

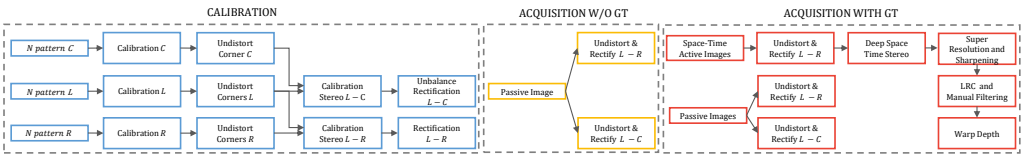


Figure 2. **Dataset acquisition overview.** Our dataset acquisition procedure can be divided into 3 main parts. Left (blue): initial calibration of our trinocular rig and the two stereo systems $L-C$ and $L-R$. Middle (yellow): image acquisition without ground-truth. Right (red): acquisition with ground-truth.

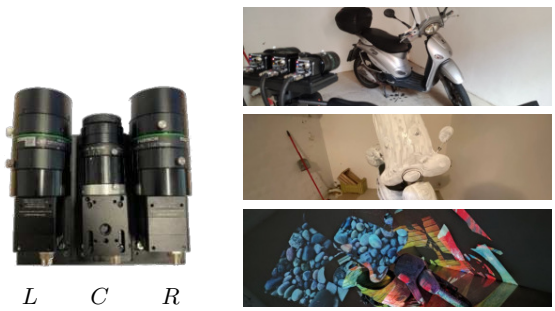


Figure 3. **Cameras setup and acquisition stages.** On left, we show our camera rig, in which L and R are two 12 Mpx cameras, and C is a wide-angle 2.3 Mpx camera. On right: i) acquisition of passive stereo pairs, ii) painting of reflective/transparent surfaces, iii) acquisition of textured stereo pairs.

of datasets and benchmarks plays a crucial role. For the first decades, the dataset were limited to few dozen samples, acquired in controlled environments and mostly made available by the Middlebury benchmark [19, 35–37]. In the '10s, more and more stereo datasets appeared, starting with KITTI 2012 [16] and 2015 [29], collected in driving environments and annotated by means of a Velodyne LiDAR sensor, then Middlebury 2014 [34], framing indoor environments at up to 6 Mpx and annotated through pattern projection, and ETH3D [39] which includes both indoor and outdoor scenes. Recently, other large-scale stereo benchmarks dealing with driving scenarios have been released, although not yet well-established as KITTI. Among them, we mention DrivingStereo [50], Argoverse [6], Apolloscape [21] and DSEC [14]. However, none of these more recent stereo datasets focus on the hardest open challenges for stereo matching as, instead, it is the case of our Booster dataset. Indeed, architectures ranking on top of KITTI perform remarkably well also on the above datasets. On the contrary, we show that state-of-the-art networks struggle on Booster.

3. Processing pipeline

Camera setup and calibration. To collect our dataset, we have built a custom stereo rig made of 2 high resolution cameras featuring a Sony IMX253LQR-C 12.4 Mpx sensor and a lower resolution camera equipped with a Sony IMX174LQJ-C 2.3 Mpx sensor mounted between the former two, as shown in Fig. 3 (left picture). From left to right we denote as L, C, and R the three cameras, with L providing the reference image for both the balanced (L,R) and unbalanced (L,C) stereo pairs, and the baselines of these two setups being ~ 8 and 4 centimeters, respectively.

Before acquiring the dataset, we need to calibrate our rig, in particular the two stereo systems $L-C$ and $L-R$. Fig. 2 includes an overview of the calibration procedure, with a more detailed description provided in the **supplement**.

Image acquisition. Our trinocular rig has been embodied into a portable setup in order to acquire a variety of scenes across different environments. In addition, our setup includes six portable projectors used to enrich the scene with random textures during the acquisition of the stereo pairs endowed with ground-truth (red block of Fig. 2). For each ground-truth acquisition, before starting, we properly setup the stage in order to capture one or more objects/surfaces embodying some of the open-challenges peculiarly addressed by our dataset. Then, the image acquisition pipeline follows three main steps, visually resumed in Fig. 3 (right pictures): i) passive images acquisition – we collect a set of balanced and unbalances stereo images under different lightning conditions. ii) scene painting – we carefully cover any specular/transparent surface in the scene with paint, thus allowing to properly project texture over them. iii) textured images acquisition – we project random patterns from multiple directions and acquire a hundred images with varying textures. Differently from the white-black banded patterns often used for this purpose [10], we project color textures since we exploit state-of-the-art deep stereo networks to label the scene. We empirically observed that color patterns result more distinctive for a deep stereo network, that is used to process bright colors typical of the synthetic datasets [28] where they have been trained. The outcome of this procedure consists of a set of passive stereo pairs – both unbalanced and at high-resolution – with different illumination conditions, representing the actual images that will be released with the dataset, and a larger set of textured images – these latter used to produce ground-truth disparities only, as detailed in the next paragraph, and thus acquired at high-resolution only in order to produce the finest annotations.

Deep space-time stereo processing. Once a set of multiple high-resolution stereo pairs – augmented with distinctive colorful textures according to the aforementioned strategy – has been acquired for a scene, we deploy our deep space-time stereo pipeline to infer a dense and accurate disparity map for the passive pair. Purposely, we leverage a pre-trained deep stereo network achieving high zero-shot generalization accuracy. We expect that, in the presence of the distinctive colorful texture we project in the scene as described before, the deep network can correctly infer a reliable disparity map. Moreover, we exploit the availability of multiple stereo pairs to further improve the outcome.

Driven by the observation that most stereo networks process a *cost-volume*, we accumulate all the cost volumes computed from each textured single stereo pair into an aggregated one. The resulting volume will reduce the effect of noise due to portions of the scene that may turn out not properly textured in a single acquisition. Purposely, we select RAFT-Stereo [25], as the top-1 method on the Middlebury 2014 stereo benchmark at the time of writing. Specifically, it uses the dot product as a measure of visual similarity between features $\mathbf{f}$ and $\mathbf{g}$ extracted respectively from the reference and target images. Thus, RAFT-Stereo computes a correlation volume storing the inner product between any pixel features in the reference image and all those at the same y-coordinate on the target image:

$$\mathbf{C}_{ijk} = \sum_{h} \mathbf{f}_{ijk} \cdot \mathbf{g}_{ikh}, \qquad \mathcal{C} \in \mathbb{R}^{H \times W \times W} \tag{1}$$

Then, the network recursively estimate a disparity map $\mathbf{d}_i$ by means of a correlation look-up mechanism, implemented as a recurrent neural network Θ processing reference image features $\mathbf{f}$, some additional context features $\mathbf{c}$, the correlation volume $\mathbf{C}$ and the disparity $\mathbf{d}_{i-1}$ estimated at the previous iteration

$$\mathbf{d}_i = \Theta(\mathbf{f}, \mathbf{c}, \mathbf{d}_{i-1}, \mathbf{C}) \tag{2}$$

until a final disparity map $\mathbf{d}$ is estimated after a fixed number of iterations. We exploit the availability of T stereo pairs and build an accumulated correlation volume $\mathbf{C}^*$ by averaging the correlation volumes computed from $\mathbf{f}^t$ and $\mathbf{g}^t$ extracted from a single stereo pair t

$$\mathbf{C}^*_{ijk} = \frac{1}{T} \sum_{t} \sum_{h} \mathbf{f}^t_{ijk} \cdot \mathbf{g}^t_{ikh}, \qquad \mathcal{C} \in \mathbb{R}^{H \times W \times W} \tag{3}$$

Then, we exploit this enriched volume to estimate a set of disparity maps from any given stereo pair

$$\mathbf{d}^t_i = \Theta(\mathbf{f}^t, \mathbf{c}^t, \mathbf{d}^t_{i-1}, \mathbf{C}^*) \tag{4}$$

Once the disparity maps $\mathbf{d}^t$ have been estimated, we finally compute their average to obtain an initial, ground-truth disparity map $\mathbf{d}^*$ as well as an uncertainty guess $\mathbf{u}^*$ through their variance. The pipeline sketched so far is effective at estimating accurate ground-truths up to half the resolution of our textured images, *i.e.* about 6 Mpx, since RAFT-Stereo has never observed samples at such higher resolution and with such higher disparity range. Thus, the outcome of our deep space-time stereo pipeline is a set of accurate disparity maps which yet require additional processing.

Super-resolution and sharpening. The quality of the disparity labels produced so far is dampened by two main causes, i) the resolution, being half of the real image resolution and ii) the presence of over-smoothed depth discontinuities, a common concern in disparity maps predicted by deep networks [7, 47]. To address both at once, we deploy the neural disparity refinement architecture proposed in [1]. However, being our images at a much higher-resolution compared to existing datasets, we pretrain the refinement network following [1], then we overfit a single instance of it on each scene, assuming the disparity map as both input and ground-truth. This strategy allows us to preserve accurate disparity values at high-resolution while sharpening depth boundaries thanks to the network output formulation.

Besides, we replaced the sub-pixel prediction mechanism described in [1] with the SMD head proposed by Tosi *et al.* [47], since we empirically observed that the former introduces undesired artefacts in our setting. Thus, each neural disparity refinement network is optimized to infer a bimodal Laplacian distribution

$$\mathbf{p}(d) = \frac{\pi}{2\mathbf{b}_1} e^{-\frac{\mathbf{d}^* - \mu_1}{\mathbf{b}_1}} + \frac{1-\pi}{2\mathbf{b}_2} e^{-\frac{\mathbf{d}^* - \mu_2}{\mathbf{b}_2}} \tag{5}$$

Once the network is trained, a sharpened disparity map $\mathbf{d}^*$ is obtained at full resolution by exploiting the continuous representation enabled by the refinement network, selecting the mode with highest density value. Concerning the implementation, a shared refinement network is pre-trained on SceneFlow following the guidelines in [1]. Then, a single instance is overfitted on each scene for about 300 steps before inferring the refined disparity map.

Manual cleaning and filtering. Once a full-resolution disparity map has been obtained, we manually clean it from any remaining artefact. To this aim, we project it into a 3D point cloud to better visualize structural errors in the geometry of the scene. We use the variance map $\mathbf{u}^*$ as a guidance during this operation, allowing to easily detect most of the artifacts. Points removed from the point cloud are then filtered out from the disparity map as well. Finally, we apply a 35×35 bilateral filter – with $\sigma_{\text{color}} = 5$ and $\sigma_{\text{dist}} = 50$ – to smooth objects surfaces and obtain the final map $\mathbf{d}^*$.

Fig. 4 illustrates the pipeline described so far, showing the increasing quality of 3D reconstruction yielded by our

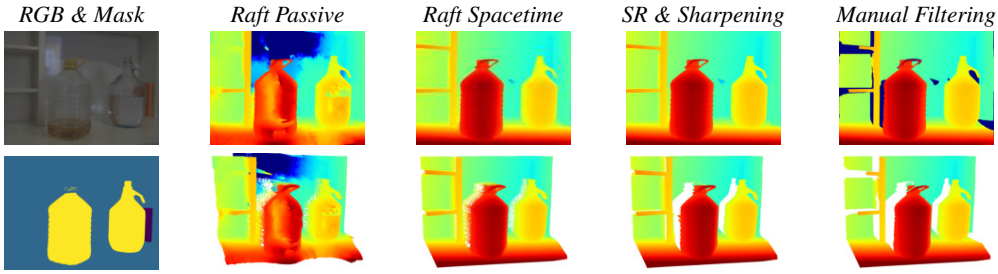


Figure 4. **Data annotation pipeline.** From left to right: reference image (top) and material segmentation mask (bottom), disparity maps (top) and point clouds (bottom) obtained by RAFT-Stereo on the passive pairs, by our deep space-time stereo algorithm, by the super resolution & sharpening procedure, and after manual cleaning.

annotations after each step.

Accuracy assessment. We follow the strategy used by Scharstein *et al*. [34] in the Middlebury 2014 dataset to measure the accuracy of our ground-truth annotations. Accordingly, we manually select planar regions from the images and fit a plane to the recovered disparities over each of them, then we measure the residuals between the fitted plane equation and the actual disparities. We perform this evaluation over 153 planar regions achieving an average residual error of 0.053, which turns out comparable to that reported for the Middlebury 2014 dataset (0.032), yet without applying an explicit sub-pixel refinement based on plane fitting.

Left-right consistency (balanced setup). We also filter out occluded pixels by performing a left-right consistency check. Purposely, the processing pipeline described so far is performed twice for each scene, producing two disparity maps, $\mathbf{d}_L^*$ and $\mathbf{d}_R^*$, for the left and right images, respectively. Then, any pixel at coordinates (x, y) in $\mathbf{d}_L^*$ is filtered out in case the absolute difference with its match $x - \mathbf{d}_L(x, y), y$ in $\mathbf{d}_R^*$ is larger than a threshold, set to 2 pixels in our case

$$|\mathbf{d}_L(x, y) - \mathbf{d}_R(x - \mathbf{d}_L(x, y), y)| > 2 \qquad (6)$$

The same procedure is performed on top of $\mathbf{d}_R^*$, removing any pixel at coordinate (x, y) after comparison with pixel $(x + \mathbf{d}_R^*(x, y), y)$ on the left disparity map.

The output of our overall annotation pipeline consists of three high-resolution ground-truth disparity maps per scene: two for the left and right images of the balanced setup, one for the unbalanced setup.

Segmentation masks. Finally, we manually label images to annotate challenging surfaces, *i.e*. transparent or specular, with segmentation masks. We cluster object surfaces into 4 classes (from 0 to 3) with increasing level of transparency and/or specularity, with class 0 identifying very opaque materials (*e.g*., a wood table) and class 3 those highly transparent/specular (*e.g*., window glasses/mirrors). An example of segmentation mask is shown in Fig. 4.

Warping (unbalanced setup). The ground-truths obtained so far are aligned with images of $L - R$. However, we want ground-truths also for the unbalanced $L - C$ stereo system. Being the rectification transformation an homography (i.e., only a change of intrinsic parameters and a rotation), we can easily perform a backward warping of the ground-truths of the left images of $L - R$ to align them to the left images of $L - C$. When warping disparity maps, we take into account the rotation of the camera reference frame and the different baselines of the two stereo systems before performing the warping. Additional details about the warping procedure can be found in the **supplement**.

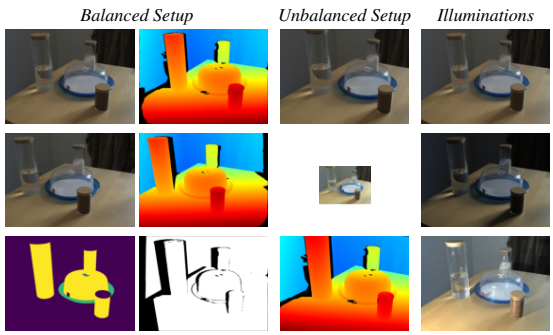


Figure 5. **A scene from the Booster testing split.** First two columns: data made available in the balanced setup (12 Mpx stereo pair, material segmentation mask, left and right disparity maps and left-right consistency mask). Third column: data dealing with the unbalanced setup (12 Mpx - 1.1 Mpx image pair, high-res disparity map associated with the 12 Mpx image). Last columns: additional 12 Mpx images acquired under different illuminations.

4. The Booster Dataset

Composition. To build up the dataset, we set the stage in 64 different indoor scenes. Then, we collected a variety of passive stereo images under different illumination conditions, leading to a total of 419 stereo samples for which we obtain dense annotations through the pipeline detailed in Sec. 3. We split the 64 scenes into 38 and 26 for training and testing purposes, respectively. As a result, Booster counts

228 training images and 191 testing images. In defining the split we aimed at having diversity of environments between the training and test scenes as well as at achieving a balanced distribution of challenging objects and materials (*e.g.*, both splits contain a scene framing a mirror). Two main benchmarks are defined in Booster: the **Balanced** benchmark, including 419 stereo pairs at 12 Mpx, and the **Unbalanced** one, featuring equally many 12 Mpx - 1.1 Mpx pairs. The latter represents the first-ever real dataset for unbalanced stereo matching, a task studied so far only by simulating the unbalanced setup by resizing one of the two images of a balanced pair same-resolution stereo images [1, 26]. More details regarding dataset images are reported in the **supplement**. Fig. 5 concerns a sample from the testing split and shows the data made available for any acquired scene.

Unlabeled samples. To encourage research on weakly-supervised approaches, *i.e.* not requiring ground-truth labels at training time, we release 15K additional samples collected –in both balanced and unbalanced settings – in a variety of indoor and outdoor environments.

Evaluation metrics. To assess the accuracy of stereo algorithms and networks, we adopt a set of metrics inspired by Middlebury 2014 [34]. Specifically, we compute the amount of pixels having error larger than a threshold τ (bad-τ). As initially our ground-truth maps are inferred at half the input resolution, we assume 2 pixels as the lowest threshold. Then, given the much higher resolution of our images, we compute error rates up to bad-8. We also measure Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). All metrics are computed on any valid pixel (*All*), or in alternative, on pixels belonging to the material class i (*Class* i) to evaluate the impact of non-Lambertian objects. In the case of the balanced setup, we also evaluate on not-occluded pixels identified by the left-right check of our annotation pipeline (*e.g.*, the bottom image in the third column of Fig. 5).

5. Experiments

5.1. Balanced Stereo Benchmark

We start by considering the Balanced split of Booster and perform a set of different experiments.

Off-the-shelf deep networks. We run a set of off-the-shelf, state-of-the-art deep stereo networks on the test set of Booster in order to assess their accuracy. We select networks with freely available implementations and pre-trained weights providing good performance on the Middlebury 2014 dataset, *i.e.* the most challenging among existing benchmarks. This constraint limits our selection to HSM-Net [49] LEAStereo [9], CFNet [42], RAFT-Stereo [25] and Neural Disparity Refinement [1]. We also evaluate, as references, the popular Semi-Global Matching algorithm (SGM) [18] and the pivotal MC-CNN network [54] in its fast variant because of memory constraints.

Tab. 1 collects the outcome of this evaluation. In the top portion of the table, we compare the predicted disparity maps with full-resolution ground-truths, on *All* (left) and *Cons* (right) pixels, the latter being the pixels of the left image that turn out consistent upon performing the left-right check and, as such, are considered as not-occluded. Each method processes input images either at the original resolution (F) or scaled to half (H) or quarter (Q) resolution. Deep networks inferences are performed on a single 3090 RTX GPU. We can notice how most methods can run only at Q resolution, mainly because of memory constraints. Consequently, their output is upsampled with nearest-neighbor interpolation in order to perform the comparison with the full-resolution ground-truth maps, with disparities scaled by the upsampling factor itself. We can notice how all methods struggle at achieving good results at such high resolution, with RAFT-Stereo achieving the best results – not surprisingly, perhaps, given its top-rank on Middlebury. Error metrics computed on *All* and *Cons* pixels yield similar scores, proving that occlusions do not represent the main difficulties in our benchmark. In the bottom portion of Tab. 1, predicted disparities are compared with ground-truth disparity maps downsampled to a quarter (Q) of the original resolution. Although the error metrics are much lower in general, we point out how they are still very far from those observed on existing benchmarks [15, 29, 34, 38], confirming that resolution is certainly a challenge in our benchmark, yet not the only one – due to the large presence of transparent and specular surfaces framed during acquisition.

Evaluation on challenging regions. We dig deeper into the unique features of Booster by evaluating the accuracy of the predicted disparities in regions of increasing level of difficulty, as defined by means of the material segmentation masks. Purposely, we select the top-performing network from the previous evaluation, *i.e.* RAFT-Stereo, and evaluate it on subsets of pixels defined by our manually annotated masks. Tab. 2 collects the outcome of this evaluation, together with results on all valid pixels as a reference. Starting from the least challenging category, we observe much lower error scores – in particular, by evaluating on quarter resolution ground-truths (bottom), we achieve results comparable with those of existing benchmarks [34]. By gradually increasing the degree of difficulty of the considered pixels, we witness a large increase of the errors. This confirms both our claims on the open-challenges in deep stereo as well as the significance of our segmentation masks.

Fine-tuning by the Booster training data. Finally, we fine-tune RAFT-Stereo on the Booster training set to show that the availability of annotated scenes can be effective in improving the result in presence of the open challenges addressed in this paper. We run 100 epochs on batches of two 884×456 crops, extracted from images randomly resized to

	Model	Input Res.	All pixels bad-2 (%)	bad-4 (%)	bad-6 (%)	bad-8 (%)	MAE (px.)	RMSE (px.)	Cons pixels bad-2 (%)	bad-4 (%)	bad-6 (%)	bad-8 (%)	MAE (px.)	RMSE (px.)
Full res.	SGM [18]	Q	80.35	66.89	58.09	52.21	57.01	119.21	78.40	63.70	54.13	47.79	41.28	91.86
	MC-CNN [54]	Q	88.09	66.30	47.77	40.53	31.23	62.98	87.64	64.20	44.24	36.70	27.56	57.34
	LEAStereo [9]	Q	70.86	55.41	47.56	42.25	27.61	51.72	69.15	53.17	45.42	40.24	26.36	49.52
	CFNet [42]	Q	61.34	48.33	42.22	38.34	27.60	51.62	59.13	46.02	40.08	36.36	25.72	48.55
	HSMNet [49]	Q	66.95	48.05	37.46	31.14	20.97	42.72	65.23	45.86	35.36	29.31	20.93	42.42
	RAFT-Stereo [25]	Q	**40.27**	**27.54**	**22.83**	**20.13**	**17.08**	**36.30**	**38.65**	**26.49**	**22.25**	**19.84**	**17.13**	**35.76**
	SGM [18]	H	76.61	64.72	58.34	54.37	71.68	133.35	74.18	61.17	54.25	49.99	55.25	106.55
	HSMNet [49]	H	53.75	36.47	28.71	24.50	19.17	42.00	51.25	34.06	26.78	23.01	18.92	41.28
	SGM+Neural Ref. [1]	H	78.54	63.20	53.77	46.87	31.82	67.02	78.35	60.59	49.59	42.50	30.92	68.37
	RAFT-Stereo [25]	H	46.31	35.49	30.98	28.15	23.95	49.94	44.02	33.59	29.49	26.95	23.25	48.11
	HSMNet [49]	F	50.85	36.53	30.77	27.56	30.82	68.97	48.11	33.88	28.50	25.61	30.02	66.79

	Model	Input Res.	All pixels bad-2 (%)	bad-4 (%)	bad-6 (%)	bad-8 (%)	MAE (px.)	RMSE (px.)	Cons pixels bad-2 (%)	bad-4 (%)	bad-6 (%)	bad-8 (%)	MAE (px.)	RMSE (px.)
Quarter res.	SGM [18]	Q	52.76	39.43	33.11	29.26	14.64	30.68	48.42	34.18	27.58	23.63	10.75	24.05
	MC-CNN [54]	Q	40.33	30.36	25.64	22.25	7.82	15.85	36.50	26.50	21.84	18.79	6.90	14.43
	LEAStereo [9]	Q	42.21	30.23	24.37	20.43	6.89	12.92	40.19	28.68	23.21	19.50	6.58	12.36
	CFNet [42]	Q	38.31	29.53	24.70	21.34	6.89	12.89	36.32	27.85	23.24	20.05	6.42	12.11
	HSMNet [49]	Q	31.11	20.25	15.92	13.23	5.24	10.67	29.25	19.47	15.70	13.23	5.22	10.59
	RAFT-Stereo [25]	Q	**20.13**	**15.13**	**12.85**	**11.05**	**4.27**	**9.05**	**19.82**	**15.19**	**12.98**	**11.17**	**4.28**	**8.91**

Table 1. **Results on the Booster Balanced testing split.** We run off-the-shelves stereo networks, using weights provided by their authors. We evaluate on full resolution ground-truth maps, or by downsampling them to quarter resolution. Best scores in **bold**.

	Category	All pixels bad-2 (%)	bad-4 (%)	bad-6 (%)	bad-8 (%)	MAE (px.)	RMSE (px.)
Full res.	All	40.27	27.54	22.83	20.13	17.08	36.30
	Class 0	32.81	16.67	11.11	7.92	3.72	9.38
	Class 1	42.95	27.47	21.60	18.21	10.20	19.96
	Class 2	73.59	60.69	51.03	44.51	36.67	47.44
	Class 3	81.54	71.93	65.22	59.62	47.73	59.38

	Category	All pixels bad-2 (%)	bad-4 (%)	bad-6 (%)	bad-8 (%)	MAE (px.)	RMSE (px.)
Quarter res.	All	20.13	15.13	12.85	11.05	4.27	9.05
	Class 0	7.97	3.72	2.33	1.82	0.93	2.28
	Class 1	18.22	11.22	7.84	6.68	2.55	4.97
	Class 2	44.47	32.14	27.92	25.43	9.17	11.88
	Class 3	59.65	43.99	34.97	28.51	11.92	14.82

Table 2. **Results on the Booster Balanced testing split – material segmentation.** We run RAFT-Stereo [25], using weights made available by their authors and process quarter resolution images. We evaluate on full resolution ground-truth maps, or by downsampling them to quarter resolution.

	Category	All pixels bad-2 (%)	bad-4 (%)	bad-6 (%)	bad-8 (%)	MAE (px.)	RMSE (px.)
Full res.	All	38.68	23.33	17.66	14.55	7.56	17.39
	Class 0	37.50	20.47	13.75	10.40	4.43	10.07
	Class 1	42.48	23.35	16.15	12.22	5.24	12.05
	Class 2	61.84	42.37	33.23	27.37	13.08	18.08
	Class 3	65.59	48.74	39.19	32.93	14.91	21.75

	Category	All pixels bad-2 (%)	bad-4 (%)	bad-6 (%)	bad-8 (%)	MAE (px.)	RMSE (px.)
Quarter res.	All	14.46	9.47	7.32	5.76	1.87	4.23
	Class 0	10.29	4.61	2.76	2.00	1.08	2.33
	Class 1	12.09	6.35	4.62	3.58	1.28	2.82
	Class 2	27.22	16.83	13.06	10.65	3.25	4.42
	Class 3	32.91	21.08	15.36	10.46	3.70	5.32

Table 3. **Results on the Booster Balanced testing split after fine tuning on the training split – material segmentation.** We run RAFT-Stereo, **fine-tuned** on the Booster training split, processing quarter resolution images. We evaluate on full resolution ground-truth maps, or by downsampling them to quarter resolution.

half or quarter of the original resolution, using the optimization procedure from [25] and initial learning rate set to 1e-5.

Tab. 3 collects the results on *All* pixels, as well as on each segmentation class. Compared to the results in Tab. 2, all error metrics tend to improve. More specifically, we can notice how the metrics do improve significantly for the most challenging materials, at the cost of a minimal decrease in accuracy within the simpler regions (Class 0). Overall we reckon that, although our experiments show that availability of annotated data can help to better handle specular/transparent objects by deep stereo networks, the accuracy level turns out still much worse compared to opaque surfaces. Hence, we observe that these kinds of materials set forth really hard open challenges in stereo which, hopefully, may be addressed in future research thanks also to the availability of the annotated data provided by Booster.

In Fig. 6 we provide some qualitative results dealing with the predictions obtained by the networks evaluated in Tab. 1 as well as, in the rightmost column, by RAFT-Stereo after fine-tuning on the Booster training set (Tab. 3). After fine-tuning on the Booster training split, RAFT-Stereo has learned to handle transparent objects much better.

5.2. Unbalanced Stereo Benchmark

Here, we evaluate the considered stereo methods on the Booster Unbalanced testing split. Tab. 4 collects the outcome of this experiment. For most methods, we follow the baseline approach defined in [1] and downsample the reference high-resolution image to the same resolution as the second image. Yet, as HSMNet is designed to handle high-resolution stereo pairs, for this network we upsample the target up to the reference image size. We point out that these results are not directly comparable to those in Tab. 1, since the baseline length (and thus disparity values) in this setup is halved, thus making the matching problem easier (i.e., smaller research range). Therefore, being errors larger than those of the Balanced split, it is evident the major difficulty of this scenario. Moreover, we highlight that, sim-

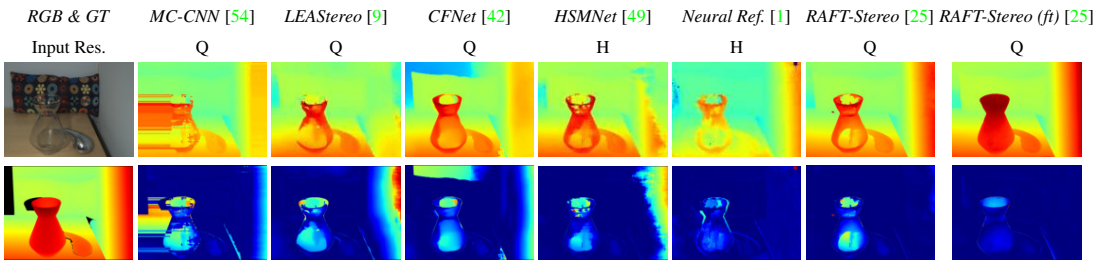


Figure 6. **Qualitative results on Booster Balanced testing split.** We show the reference image (top) and the ground-truth map (bottom) on leftmost column, followed by disparity (top) and error maps (bottom) for the deep models evaluated in our benchmark.

	All pixels					
Model	bad-2 (%)	bad-4 (%)	bad-6 (%)	bad-8 (%)	MAE (px.)	RMSE (px.)
SGM [18]	78.47	62.74	52.62	45.97	42.63	97.62
MC-CNN [54]	86.30	68.67	54.20	44.78	23.64	45.46
LEAStereo [9]	74.31	57.70	47.11	39.88	17.68	31.29
CFNet [42]	70.22	53.20	43.61	37.10	16.19	28.78
HSMNet [49] †	63.20	43.22	32.87	26.55	11.96	22.82
SGM+Neural Ref. [1] †	70.90	52.15	41.71	35.40	24.27	52.52
RAFT-Stereo [25]	**55.96**	**36.81**	**27.87**	**22.33**	**9.86**	**19.36**
RAFT-Stereo [25] (ft)	58.67	32.83	22.96	17.65	6.31	11.11

Table 4. **Results on Booster Unbalanced testing split.** We run stereo networks, using weights made available by their authors. We evaluate on full resolution ground-truth maps. † denotes images being resized to half the reference resolution (about 6 Mpx). (ft) denotes fine-tuned on Booster Unbalanced training split.

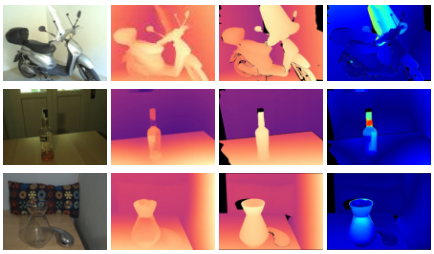


Figure 7. **Qualitative results for monocular depth estimation.** From left to right: reference images, depth maps predictions by DPT [33], ground-truth depth maps, error maps.

ilarly to the Balanced setup, by fine-tuning RAFT-Stereo on the Unbalanced training split, we can improve its performance on nearly all metrics. Thus, future research on stereo may leverage the finding that state-of-the-art deep models hold the potential to better learn to match specular/transparent surfaces even in unbalanced setting when properly fine-tuned with carefully annotated data.

5.3. Challenges in Monocular Depth Estimation

We argue that most of the difficult surfaces featured by Booster set forth open-challenges to be addressed in future research also for other image-based depth estimation approaches, such as, in particular, monocular depth estimation. As a side experiment, thus, we run DPT [33] – a transformer for single image depth estimation – on the images present in our dataset. Fig. 7 shows some qualitative examples, highlighting how scale-aligned DPT predictions are very inaccurate on transparent surfaces.

6. Conclusion, Limitations and Future work

In this paper, we have presented the Benchmark on open-challenges in stereo (**Booster**), a novel stereo dataset collecting 419 images – acquired both in *balanced* and *unbalanced* setups – featuring extremely challenging environments and kinds of objects. It comes with dense and accurate ground-truth disparities, obtained through a novel deep space-time stereo pipeline, as well as with manually annotated material segmentation masks. Compared to recent stereo datasets targeting autonomous/assisted driving, such as DrivingStereo [50], Booster includes a much smaller number of annotated images and, hence, cannot be considered a *large-scale* dataset. Moreover, the deep space-time pipeline and the small baseline used for annotations constraints the collected scenes to frame indoor environments.

Our experiments show that Booster unveils some of the most intriguing challenges in deep stereo and provides hints on promising research directions. In particular, follow-up work fostered by Booster may be devoted to i) investigating on the ability of deep models suitably fine-tuned on Booster to generalize to outdoor setting featuring similar difficult surfaces and materials, ii) devising a pipeline, e.g, leveraging on Lidar sensors, to collect annotated data with transparent/specular surfaces also in outdoor setting, iii) build large scale synthetic datasets specifically addressing the open-challenges highlighted by Booster to enable more effective pre-training and iv) building and scanning scenes through successive depth layers, to gather multiple depths at transparent/reflective objects which can be useful for applications such as augmented reality.

Therefore, we are lead to believe that Booster holds the potential to *boost* future research in deep stereo.

Acknowledgements. We gratefully acknowledge the funding support of Huawei Technologies Oy (Finland).

References

[1] Filippo Aleotti, Fabio Tosi, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Luigi Di Stefano, and Stefano Mattoccia. Neural disparity refinement for arbitrary resolution stereo. In *International Conference on 3D Vision*, 2021. 3DV. 2, 4, 6, 7, 8

[2] Filippo Aleotti, Fabio Tosi, Li Zhang, Matteo Poggi, and Stefano Mattoccia. Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation. In *16th European Conference on Computer Vision (ECCV)*. Springer, 2020. 2

[3] Konstantinos Batsos and Philipos Mordohai. Recresnet: A recurrent residual cnn architecture for disparity map enhancement. In *In International Conference on 3D Vision (3DV)*, 2018. 2

[4] Changjiang Cai, Matteo Poggi, Stefano Mattoccia, and Philippos Mordohai. Matching-space stereo networks for cross-domain generalization. In *2020 International Conference on 3D Vision (3DV)*, pages 364–373, 2020. 2

[5] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5418, 2018. 2

[6] Ming-Fang Chang, John W Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[7] Chuangrong Chen, Xiaozhi Chen, and Hui Cheng. On the over-smoothing problem of cnn based disparity estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8997–9005, 2019. 4

[8] Zhuoyuan Chen, Xun Sun, Liang Wang, Yinan Yu, and Chang Huang. A deep visual correspondence embedding model for stereo matching costs. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2

[9] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 6, 7, 8

[10] James Davis, Ravi Ramamoorthi, and Szymon Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–359. IEEE, 2003. 2, 3

[11] Leonardo De-Maeztu, Stefano Mattoccia, Arantxa Villanueva, and Rafael Cabeza. Linear stereo matching. In *2011 International Conference on Computer Vision*, pages 1708–1715. IEEE, 2011. 2

[12] Luigi Di Stefano, Massimiliano Marchionni, and Stefano Mattoccia. A fast area-based stereo matching algorithm. *Image and vision computing*, 22(12):983–1005, 2004. 2

[13] Pier Luigi Dovesi, Matteo Poggi, Lorenzo Andraghetti, Miquel Martí, Hedvig Kjellström, Alessandro Pieropan, and Stefano Mattoccia. Real-time semantic stereo matching. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10780–10787. IEEE, 2020. 2

[14] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 2021. 3

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 2, 6

[16] Andreas Geiger, Martin Roser, and Raquel Urtasun. Efficient large-scale stereo matching. In *Asian conference on computer vision*, pages 25–38. Springer, 2010. 3

[17] Spyros Gidaris and Nikos Komodakis. Detect, replace, refine: Deep structured prediction for pixel wise labeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5248–5257, 2017. 2

[18] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007. 2, 6, 7, 8

[19] Heiko Hirschmuller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 3

[20] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):504–511, 2012. 2

[21] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2702–2719, 2019. 3

[22] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2

[23] Hsueh-Ying Lai, Yi-Hsuan Tsai, and Wei-Chen Chiu. Bridging stereo matching and optical flow via spatiotemporal correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[24] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[25] Lahav Lipson, Zachary Teed, and Jia Deng. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. *arXiv preprint arXiv:2109.07547*, 2021. 1, 2, 4, 6, 7, 8

[26] Yicun Liu, Jimmy Ren, Jiawei Zhang, Jianbo Liu, and Mude Lin. Visually imbalanced stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2029–2038, 2020. 2, 6

[27] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition, pages 5695–5703, 2016. 2

[28] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2, 3

[29] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 3, 6

[30] Jiahao Pang, Wenxiu Sun, Jimmy SJ. Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[31] Matteo Poggi, Alessio Tonioni, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Continual adaptation for deep stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 2

[32] Matteo Poggi, Fabio Tosi, Konstantinos Batsos, Philippos Mordohai, and Stefano Mattoccia. On the synergies between machine learning and binocular stereo for depth estimation from images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2

[33] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, October 2021. 8

[34] 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*, pages 31–42. Springer, 2014. 1, 2, 3, 5, 6

[35] Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 3

[36] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002. 1, 2, 3

[37] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003. 3

[38] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 6

[39] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3260–3269. IEEE, 2017. 3

[40] Akihito Seki and Marc Pollefeys. Patch based confidence prediction for dense disparity map. In *BMVC*, pages 23.1–23.13, 2016. 2

[41] Akihito Seki and Marc Pollefeys. SGM-Nets: Semi-global matching with neural networks. In *CVPR*, pages 231–240, 2017. 2

[42] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2021. 6, 7, 8

[43] Xiao Song, Xu Zhao, Hanwen Hu, and Liangji Fang. Edgestereo: A context integrated residual pyramid network for stereo matching. In *ACCV*, 2018. 2

[44] Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Unsupervised adaptation for deep stereo. In *The IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct 2017. 2

[45] Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Unsupervised domain adaptation for depth prediction from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[46] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019. 2

[47] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[48] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8071–8081, 2019. 2

[49] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2019. 2, 6, 7, 8

[50] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2019. 3, 8

[51] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *ECCV*, pages 636–651, 2018. 2

[52] Qingxiong Yang. A non-local cost aggregation method for stereo matching. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1402–1409. IEEE, 2012. 2

[53] Kuk-Jin Yoon and In So Kweon. Adaptive support-weight approach for correspondence search. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):650–656, 2006. 2

[54] Jure Zbontar, Yann LeCun, et al. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(1):2287–2318, 2016. 1, 2, 6, 7, 8

[55] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. GA-Net: Guided aggregation net for end-to-end stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[56] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *Europe Conference on Computer Vision (ECCV)*, 2020. 2

[57] Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In *The IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017. 2