Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

RGB-Multispectral Matching: Dataset, Learning Methodology, Evaluation

06 May 2024

RGB-Multispectral Matching: Dataset, Learning Methodology, Evaluation

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

(a) RGB (3222×1605×3)   (b) MS (510×254×10)   (c) Ground-truth (3222×1605)   (d) Prediction (3222×1605)
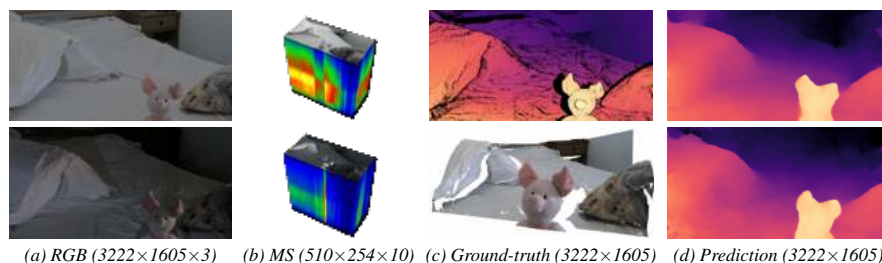
Figure 1. (a) RGB images and (b) low-resolution multi-spectral images acquired under varying lighting conditions. Visualization as a cube with on top the first band and on sides two stretched spectral slices of the bottom row and far-right column after the application of a color-map. (c) ground-truth disparity aligned with the RGB image and its point-cloud visualization while (d) prediction by our architecture.

Abstract

We address the problem of registering synchronized color (RGB) and multi-spectral (MS) images featuring very different resolution by solving stereo matching correspondences. Purposely, we introduce a novel RGB-MS dataset framing 13 different scenes in indoor environments and providing a total of 34 image pairs annotated with semi-dense, high-resolution ground-truth labels in the form of disparity maps. To tackle the task, we propose a deep learning architecture trained in a self-supervised manner by exploiting a further RGB camera, required only during training data acquisition. In this setup, we can conveniently learn cross-modal matching in the absence of ground-truth labels by distilling knowledge from an easier RGB-RGB matching task based on a collection of about 11K unlabeled image triplets. Experiments show that the proposed pipeline sets a good performance bar (1.16 pixels average registration error) for future research on this novel, challenging task.

1. Introduction

Traditional RGB sensors acquire images with three channels, approximately mimicking color perception in trichromatic mammals like humans, where three kinds of cone cells are sensitive to three different ranges of wavelengths in the visible spectrum. This is usually achieved by using bandpass optical filters corresponding to such colors, arranged in a Bayer pattern [6]. Multi-spectral (MS) imaging devices generalize such image acquisition mechanics and enable acquisition of images with a larger number of channels, i.e. ten or more, usually corresponding to narrower wavelength ranges. MS sensors may also be sensitive to wavelengths outside the visible spectrum, e.g.in the infra-red or ultra-violet bands. By extending the range of wavelengths as well as the granularity of their quantization into image channels, MS devices enable extraction of additional information about the sensed scene that human eyes fail to capture, which in turns forms the basis of peculiar applications. For instance, MS imaging devices are used to perform analysis of artworks [15], remote sensing for agriculture [81] and land use [79], target tracking [21], pedestrian detection [31], counterfeit detection, e.g. of banknotes [3], diagnostic medicine and skin inspection [39], food inspection [54] and contamination detection [20]. In spite of such broad range of applications, deployment of MS sensors is usually limited to industrial settings or expensive equipment like satellites. However, the ability to run several of these applications on mobile phones and other consumer devices directly operated by end users could open up interesting scenarios, like, e.g., enabling early diagnosis of skin diseases [35], democratizing food quality control [4], mak-

_____

* Joint first authorship.

ing plant phenotyping easier [69], and others yet to imagine.

While the number of traditional cameras on high-end phones have been growing steadily in recent years, MS sensors have not been ported yet on consumer devices like phones or action cameras due to several limitations [22]. In particular, MS sensors resolution is significantly smaller than standard RGB cameras which feature several Megapixels of resolution, because the most suitable technology to realize them extends the Bayer pattern used for color imaging into multi-spectral filter arrays [41], with each native pixel of the imaging sensor detecting one band by placing in front of it the corresponding optical filter, i.e. one MS "pixel" for a camera sensing 16 bands uses a 4×4 grid of native pixels. Thus, MS sensors tend also to be larger and bulkier than RGB cameras; and they are orders of magnitude more expensive, *i.e.* MS cameras cost at least tens of thousands of dollars/euros. There exist linear MS cameras [62] or MS cameras realized with filter wheel technology [8] which may feature high resolution but can sense only static scenes and are not appropriate for deployment on mobile devices.

As a result of these technological limitations, MS cameras compatible with cost and size requirements of mobile devices feature very small resolution, insufficient for the applications listed above. Moreover, beside up-sampling the MS image to usable resolutions, it is usually important to align MS information with RGB streams coming from traditional cameras on the device, where objects or areas of interest can be easily identified with effective algorithms. Due to the challenges of matching images across spectra, the mostly explored setup to acquire MS images and RGB images simultaneously has been to physically align their optical centers by using beam splitters [11,31], which are however unfeasible to deploy in mobile devices, where instead the dense registration between the images in the pair must be computed by computer vision algorithms.

Although establishing dense image correspondences is one of the fundamental and most studied problems in computer vision, solutions that address the cross-spectral images and/or unbalanced resolutions are rare. They only investigated case is the special incarnation of the problem where an RGB image is matched to a Near Infra-Red (NIR) or Infra-Red (IR) one at the same resolution [14, 36, 37, 59, 78]. To the best of our knowledge, the general MS-RGB case, both balanced and unbalanced in terms of resolution, is yet unexplored in literature. Research on this topic has also been hindered by the lack of publicly available datasets: existing methods tackling the NIR/IR-RGB case have been tested on datasets acquired with custom hardware setups targeting specific use cases, like autonomous driving or object detection, and never made publicly available [14, 78] or are tested on small datasets with sparse ground-truth: e.g. the dataset used by [36, 37] and proposed in [59] has 7 image pairs where 50-100 object cor-

ners on average per pair have been manually annotated, i.e. less 700 ground-truth correspondences.

In this work, we propose the first large-scale and publicly-available dataset to study RGB-MS registration. In particular, we cast the registration problem as a *stereo matching* one due to the two cameras being synchronized. Our dataset features more than 11K stereo pairs composed of a low-res 510×254 MS image and a high-res 3222×1605 RGB image which can be used to compute a high-res MS image registered at pixel-level with the RGB image. Examples of pairs are shown in Fig. 1a-b. Another key feature of our dataset is that 34 stereo pairs coming from 13 different scenes are densely annotated (Fig. 1c), thanks to an original acquisition methodology whereby a second RGB image and several projectors are used to create a very accurate active space-time stereo setup [16], which result in more than 125 millions of ground-truth correspondences. We also propose a deep learning architecture to tackle the challenging cross-spectral and resolution-unbalanced problem, that can be used to compute registered MS images at arbitrary resolutions and serves as a baseline for the dataset, whose results are shown in Fig. 1d. When training the network, we leverage the large body of unlabelled images in our dataset by sourcing *proxy*-labels, which are obtained by exploiting again the second RGB camera to run passive stereo matching. Our dataset enables the community to study the challenging problem of cross-spectral and resolution-unbalanced matching of images, which is key to enable porting of existing MS applications in the consumer space as well as to unlock new applications of MS imaging specific to the mobile device setup. Our contributions are:

• we propose the first investigation into the challenging problem of cross-spectral and resolution-unbalanced dense matching, which is a key enabling technology to unlock MS applications on mobile devices;

• we present the first large-scale publicly available dataset for the problem;

• thanks to a peculiar acquisition methodology exploiting two registered high-res RGB cameras, we also make available the first densely labelled set of images for this problem and we propose a training methodology which can leverage unlabelled images via proxy supervision;

• we propose a deep architecture to compute correspondences between images at different resolutions and with different spectral content, which can be used to generate MS images registered at pixel-level with the RGB stream.

The project page is available at https://cvlab-unibo.github.io/rgb-ms-web/.

2. Related Work

Deep Stereo. First attempts to learn in stereo matching focused on the design of robust matching functions between image patches, implemented with shallow CNNs

[13, 45, 75]. Then, along the established pipeline [56] other sub-tasks were cast in the form of a neural network, such as optimization [57, 58] and disparity refinement [1, 5, 19, 23, 33]. However, the development of end-to-end architectures [46] represented a real turning point in the field, with more and more works focusing nowadays on the design of new architectures rather than on classical algorithms. According to [53], two main families of deep stereo networks exist, respectively 2D [32, 42, 44, 46, 49, 61, 72] and 3D architectures [9, 12, 17, 34, 60, 71, 76, 77].

Self/Proxy-Supervised Stereo. In parallel to the spread of end-to-end stereo, some first strategies were developed to train such models without ground-truth labels [63, 80]. Several works followed, exploiting image reprojection across the two views as a form of supervision [40, 65, 70] or, in alternative, the guidance given by traditional stereo algorithms [2, 52, 63, 64] was used to distill *proxy* labels to supervise the stereo network.

Cross-spectral Matching. Finding correspondences between images sensing different parts of the wavelength spectrum represents an additional challenge. This task has been investigated in different settings, in most cases by matching RGB-IR [14, 47], RGB-thermal [50] and RGB-NIR [36, 37, 59, 78] modalities, as well as between stereo images with very different radiometric variations [28, 29]. Traditionally, most approaches aimed at designing handcrafted functions [14, 27, 28, 59] or descriptors [29, 36, 50]. More modern approaches deployed deep-learning to learn a deep correlation function [37] starting from a self-similarity measure while Zhi *et al.* [78] designed an end-to-end network trained in semi-supervised manner on RGB-NIR, by explicitly leveraging knowledge about materials after manual annotation of the training set.

Only a few recent works [10, 18, 73] on remote sensing focus on affine trasformations to tackle the RGB-MS registration task. However, in contrast to our method, they do not explicitly reason on searching dense correspondences between the two input signals. Among published papers, [78] is the most relevant to ours, as it proposes an end-to-end model trained without ground-truth, although in the RGB-NIR setup. Conversely to [78], our novel *proxy*-supervised training paradigm does not require any manual annotation, conveniently exploiting an additional RGB camera during training data acquisition. Moreover, we propose a novel dataset with semi-dense ground-truth disparity maps, counting a total of more than 125M annotated pixels, whereas in [78] were limited to a few, sparse pixels, *i.e.* 5K.

3. RGB-MS Dataset

In this section, we present our novel RGB-MS dataset. We start by describing the acquisition setup and then we dig into the procedure to annotate the RGB-MS pairs with semi-dense ground-truth labels.

3.1. Camera Setup

Our acquisition setup consists of two synchronized Ximea cameras: an RGB camera equipped with a Sony IMX253LQR-C 12.4 Mpx sensor and an MS camera based on a IM-SM4X4-VIS2 2.2 Mpx sensor with 10 bands. Both cameras are mounted on a common support with a distance of about 4cm between their optical centers.

The two cameras are calibrated, so as to form a stereo system where the RGB device (left) is used as the reference camera. This is achieved by the OpenCV tools for camera calibration [7], after having acquired a set of images with both cameras framing a planar chessboard. Then, corner detection is performed on both images after conversion to grayscale. As for the MS camera, we found it sufficient to average the 10 bands pixel-wise in order to obtain a pseudo-grayscale image in which corners are easily detectable. Thus, we estimate intrinsic and lens distortion parameters for both cameras and then, after undistortion, calibrate the RGB-MS stereo camera system and estimate the transformations to rectify the RGB and MS frames. Given the dramatic difference between the two images in terms of resolution, we cannot rely on the standard procedure to estimate the rectification parameters, since it would lead to rectified images resized to a resolution in between the two. Thus, we define the concept of unbalance rectification in which we consider two images as *unbalance rectified* if, after resizing them to same resolution, they are rectified. Additional details in the supplementary material.

3.2. Annotation Pipeline

We now introduce our annotation pipeline, aimed at enriching our RGB-MS dataset with semi-dense ground-truth disparity maps. A variety of active depth sensors would fit this purpose, although, on the one hand, significantly limiting the resolution at which our ground-truth could be collected, on the other introducing the need to carry out a registration between the RGB camera and the active sensor itself, a non-trivial task itself.

To overcome these issues, we leverage a space-time stereo pipeline [16] by adding a second RGB sensor to our camera setup. Specifically, we mount this additional device on the right of the MS sensor, i.e. with a baseline of about 8cm with respect to the other RGB camera, thus implementing a 12.2 Mpx RGB stereo setup. This configuration allows to fully exploit the high resolution of the RGB cameras to provide much more accurate ground-truth labels compared to the deployment of any ancillary active sensor (e.g. Lidar or ToF) while removing the need of complex registration across sensors. Moreover, the larger baseline increases depth resolution and produces finer-grained labels.

Image Acquisition. To collect a single scene for which we aim at providing an RGB-MS pair together with ground-truth disparities, we use our trinocular cameras setup to ac-

quire a set of *passive* RGB-MS images, *i.e.* the static scene is acquired in absence of any external perturbation. For each scene, we make several acquisition with different illuminations. These frames represent the actual images that will be distributed with our dataset. Then, we use a set of portable projectors to perturb the sensed environment with random black-and-white banded patterns generated through Non-recurring De Bruijn sequence [43], as shown in Fig. 2a-b. We acquire several dozens of *active* images for each scene to be processed by the next step in the pipeline.

Space-Time Semi-Global-Matching. We implement a robust space-time stereo pipeline [16] to process the active stereo pairs acquired on the same static scene and obtain a highly accurate disparity map (Fig. 2d). Such a framework leverages on the variety of patterns projected during each acquisition, greatly increasing the distinctiveness of any single pixel and thus easing the matching process. In particular, our algorithm implements four steps:

i) *Single-pair cost computation.* For each active RGB-RGB pair t, we compute an initial Disparity Space Image (DSI), storing d_{max} matching costs for any pixel in the reference image at coordinates (x, y). Such costs are obtained by applying a Census transform [74] on 9×7 windows to both frames and then computing the Hamming distance between 63-bit strings:

$$\text{DSI}_t(x,y,d) = \sum_i \mathcal{C}_i^{L(t)}(x,y) \oplus \mathcal{C}_i^{R(t)}(x-d,y) \quad (1)$$

with $\mathcal{C}^{L(t)}, \mathcal{C}^{R(t)}$ being the t-th left and right census transformed images and i any single bit in the 63-bit strings.

ii) *Space-time integration and SGM optimization.* Once single DSI_t volumes have been initialized, we integrate them over time to obtain more robust matching costs thanks to the variety of different patterns projected onto the scene:

$$\text{DSI}(x,y,d) = \sum_t \text{DSI}_t(x,y,d) \quad (2)$$

Then, we further optimize the matching cost distributions by means of the Semi-Global Matching algorithm [30], before selecting the final disparity $\hat{d}(x,y)$ by means of a Winner-Takes-All strategy.

iii) *Outliers suppression.* Although SGM dramatically regularizes the initial cost volume and leads to smooth disparity maps, several outliers are still present. To remove them, we apply a confidence-based approach [68] to filter out the least confident pixels. Given a pool of conventional confidence measures, we discard all pixels marked as not reliable according to each measure. To this aim, we select two binary measures from [51] – ACC, LRC – together with a Weighted Median Disparity Deviation over a 41×41 window, considering pixels as not confident if larger than 1.

iv) *Sub-pixel interpolation.* Finally, we estimate sub-pixel disparities to improve depth resolution. Starting from

Active Stereo *Pre-Warping* *Warped*

(a) (c) (e)

(b) (d) (f)

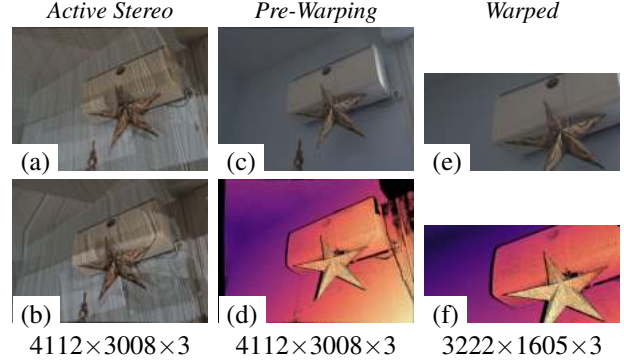$4112\times3008\times3$ $4112\times3008\times3$ $3222\times1605\times3$

Figure 2. **Ground-truth Acquisition.** Given a set of active RGB-RGB stereo pairs (a,b), we compute the ground-truth disparity (d) aligned with the left image of the RGB-RGB stereo system (c) with our space-time stereo algorithm. Then, we warp it (f) to be aligned with the left image of the RGB-MS stereo system (e).

the DSI output of SGM, an interpolation algorithm based on [48] is applied to pixels that have not been filtered out previously. This is traditionally achieved by fitting, for any pixel (x, y) in the image, a function between its minimum cost $c_{\hat{d}} = DSI(x, y, \hat{d})$ and its disparity neighbours $c_{\hat{d}-1}$, $c_{\hat{d}+1}$. The interpolation function can be assumed as monotonically increasing in $[0, 1]$, then defining sub-pixel precise disparity $\hat{d}_{\text{sub}}$ as

$$\hat{d}_{\text{sub}} = \begin{cases} \hat{d} - 0.5 + \frac{c_{\hat{d}-1} - c_{\hat{d}}}{c_{\hat{d}+1} - c_{\hat{d}}} & \text{if } c_{\hat{d}-1} > c_{\hat{d}+1} \\ \hat{d} - 0.5 + \frac{c_{\hat{d}+1} - c_{\hat{d}}}{c_{\hat{d}-1} - c_{\hat{d}}} & \text{otherwise} \end{cases} \quad (3)$$

Following [48], we fit a low-order polynomial function, *i.e.* a parabola $ax^2 + bx + c$, setting $a = 1$, $b = 1$ and $c = 0$.

Manual Cleaning on Point Clouds. Despite the aforementioned filtering strategies, the disparity maps obtained so far may still present some outliers. Thus, we manually clean the ground-truth maps of any noisy label by projecting them into 3D point clouds and selecting all points resulting isolated from the main structures in the scene. For pixels corresponding to points that have been filtered out in the point cloud, we remove them from the ground-truth maps.

Disparity Warping. The ground-truth disparity maps produced so far are aligned to the reference image of the RGB-RGB setup, i.e. the left image rectified for the RGB-RGB setup (Fig. 2c). The reference image for the RGB-MS setup is the same, but it is rectified differently due to the different mechanical alignment and the smaller field of view of the MS camera (see supplementary for more details). Hence, we use the estimated rectification homographies to perform backward warping of the disparity map and align it with the RGB image rectified for the RGB-MS setup (Fig. 2e), while considering the rotation of the camera reference frame and the different baselines in the RGB-MS and RGB-RGB setups, thus producing the final ground-truth disparity
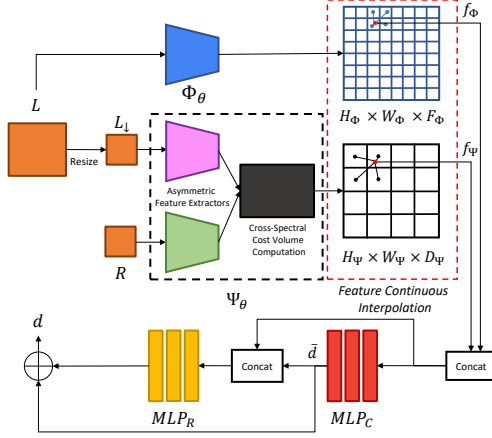
Figure 3. **Architecture Overview.** Given an unbalanced stereo pair composed of a reference high-resolution image L and a target multi-spectral low-resolution image R, our network estimates a disparity map aligned with L by combining cross-spectral cost probabilities computed by a stereo backbone Ψ_θ and deep features from L obtained by the feature extractor Φ_θ.

maps for the RGB-MS pair (Fig. 2f).

4. Cross-Spectral Matching

Fig. 3 illustrates our deep architecture specifically designed to address the RGB-MS matching problem. Given a rectified RGB-MS pair, we extract deep features from the high-resolution RGB image and compute the matching cost volume between the MS image and the RGB image resized at the same spatial resolution. Then, the extracted features and the cost volume are used to estimate a continuous dense correspondence field, which can be used to synthesize a high-resolution multi-spectral image aligned with the RGB one. In this section, we describe in detail the network architecture, the loss function and the training protocol.

4.1. Problem Statement

Let $L \in \mathbb{R}^{W_L \times H_L \times 3}$ denote the RGB image acquired by the high-resolution camera and $R \in \mathbb{R}^{W_R \times H_R \times N}$ ($N = 10$ in our experiments) the image captured by the low-resolution MS sensor after unbalanced rectification. As highlighted in Fig. 1, the two cameras feature very different properties in terms of resolution and number of channels. Our goal is to estimate a disparity map $\mathcal{D}$ aligned with the reference image L at an arbitrary spatial resolution.

4.2. Deep Cross-Spectral Network

Our proposed architecture composed of 1) a coarse sub-module responsible for integrating geometric information between the RGB image and the MS signal at low-resolution and 2) a fine-grained sub-module aimed at recovering details by leveraging on the original high-resolution

RGB image. More specifically, let

$$\Psi_\theta : \mathbb{R}^{W_R \times H_R \times (N+3)} \to \mathbb{R}^{W_\Psi \times H_\Psi \times D_\Psi} \qquad (4)$$

denote a deep stereo backbone with parameters θ and maximum disparity D_Ψ. Examples for such backbones are stereo architectures that process a standard stereo pair by means of siamese towers and perform 3D convolutions on 4D feature volumes [12, 26, 71, 76] to produce a 3D cost volume without performing the last softargmax step. For our specific setup, Ψ_θ takes as input $L_\downarrow$ and R, where $L_\downarrow$ represents the reference RGB image L resized to match the resolution of R, and outputs a feature vector D_Ψ. By doing so, the two images can be considered rectified based on the definition of unbalanced rectification [1]. However, differently from the standard RGB-RGB setting where the stereo pair is processed by means of siamese towers, we modify Ψ_θ to compute deep features from $L_\downarrow$ and R using two feature extractors – that are independent, because of the different information sensed in the two images and encoded in a different number of input channels ($L_\downarrow$:3, R: N). Once high-dimensional features have been extracted from the two images, we keep unchanged the remaining part of the stereo backbone Ψ_θ in charge of building the cross-spectral cost volume and regularizing the aggregated costs. Our intuition is that, by doing so, the network effectively learns to match deep features computed from two completely different signals at low-resolution and, thus, to recover the final 3D structure of the scene. Yet, Ψ_θ does not make full use of the available high resolution RGB image, where important high-frequency details may be extracted. Thus, we deploy also image features learned from the high-resolution image by means of a fully convolutional module that has the same architecture of the low resolution feature extractor adopted in Ψ_θ but does not share weights with it and defined as:

$$\Phi_\theta : \mathbb{R}^{W_L \times H_L \times 3} \to \mathbb{R}^{W_\Phi \times H_\Phi \times F_\Phi} \qquad (5)$$

where F_Φ represents the dimension of extracted features.

Following recent advances in continuous function representations [1, 67], we adopt a Multilayer Perceptron (MLP) to compute the final disparity map $d \in \mathbb{R}$ at arbitrary spatial location $\mathbf{x}_L \in \mathbb{R}^2$ where the input features for the MLP are obtained by concatenating the bilinear interpolated features f_Ψ and f_Φ computed on the output of Ψ_θ and Φ_θ, respectively. We train our model by adopting the recent loss function proposed in [1] that allows estimating accurate disparity maps as well as sharp depth boundaries. More specifically, we use two MLPs: MLP_C, in charge of computing a categorical distribution over disparity values $[0, d_{\max}]$ from which an initial integer disparity value is selected, and MLP_R, which estimates a real-valued offset to recover sub-pixel information. The final disparity value predicted by our network can be expressed as:

$$d = \bar{d} + \mathrm{MLP}_R(f_\Psi(\mathbf{x}_{L_\downarrow}), f_\Phi(\mathbf{x}_L), \bar{d}) \qquad (6)$$

RGB   MS   *Proxies*

(a)   (b)   (c)

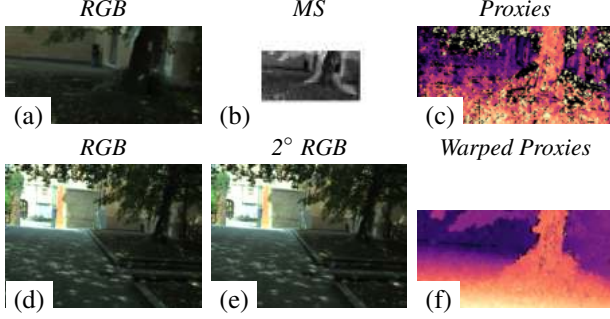RGB $2°$ RGB   *Warped Proxies*

(d)   (e)   (f)

Figure 4. **Proxy Labels Distillation.** A classical algorithm [30] struggles at dealing with RGB-MS images (a-c). Using a second RGB camera yields much more accurate proxy labels (d-f).

where $\bar{d}$ is the integer disparity computed by MLP_C, i.e.

$$\bar{d} = \text{argmax}(\text{MLP}_C(f_\Psi(\mathbf{x}_{L_\downarrow}), f_\Phi(\mathbf{x}_L))) \qquad (7)$$

where $\mathbf{x}_{L_\downarrow} \in \mathbb{R}^2$ is the 2D continuous location in $L_\downarrow$, while $\mathbf{x}_L \in \mathbb{R}^2$ is the corresponding location at high resolution, defined as $\mathbf{x}_L = \frac{W_L}{W_{L_\downarrow}} \mathbf{x}_{L_\downarrow}$.

Thus, by indicating the disparity labels as d^*, our network can be trained by minimizing the following loss function:

$$\mathcal{L} = - \mathcal{N}(d^*, \sigma) * \log\Big(\text{MLP}_C(f_\Psi(\mathbf{x}_{L_\downarrow}), f_\Phi(\mathbf{x}_L))\Big) \\ + \Big| MLP_R(f_\Psi(\mathbf{x}_{L_\downarrow}), f_\Phi(\mathbf{x}_L)) - d_R^* \Big| \qquad (8)$$

where $\mathcal{N}(d^*, \sigma)$ represents the Gaussian distribution centered at d^* and sampled at integer values in $[0, d_{\max}]$, while $d_R^* = d^* - \bar{d}$.

4.3. Proxy Label Distillation

We rely on proxy label distillation [2,52,63,66] to obtain a strong and reliable source of supervision for our architecture from the unlabeled images. Thus, following [66], we adopt the popular Semi-Global Matching (SGM) algorithm to obtain proxy-labels from single-channel frames encoding the average of the image channels for the multi-spectral images and the converted grayscale for the RGB images. However, since SGM aims at computing matches between the two images, in our RGB-MS setup it is intrinsically tampered by the different modalities of the two, as shown in Fig. 4 (top). This leads to less accurate proxy-labels and, thus, to a less effective weakly-supervised training.

Thus, to source a reliable supervision also on the challenging RGB-MS setup, we propose to enhance our proxy distillation pipeline according to a novel strategy. Specifically, we have already discussed that a second high-resolution RGB image allows us to obtain accurate semi-dense ground-truth maps by means of the space-time stereo

framework. We argue that, in a similar manner, we can deploy the second RGB camera during unlabeled data acquisition as well to obtain better proxy supervision. This allows for complementing any RGB-MS pair with a corresponding RGB-RGB pair over which classical stereo algorithms can produce much more accurate proxy labels to be deployed at training time. Hence, we exploit the very same procedure described before to warp the proxy-labels from the rectified RGB-RGB left image to the rectified RGB-MS left image. This second strategy, illustrated in Fig. 4 (bottom), allows us to supervise our network and teach it how to match RGB-MS modalities supervised by established knowledge concerning the matching between RGB-RGB, at the cost of requiring an additional RGB camera during the *offline* self-training. This kind of strategy as already been exploited to pursue training without ground-truth depth labels, e.g. in the case of monocular depth estimation by means of image reconstruction losses [24,25] or to distill proxy labels [66]. Despite representing an additional requirement for the acquisition setup – yet cheap and effortless compared to any annotation pipeline, such as the one used to collect our ground-truth data – we will show in our experiments how this strategy, when such an additional sensor is available, leads to a more effective training of our RGB-MS and, thus, to more accurate disparity estimations.

5. Experimental Results

5.1. Additional Datasets

UnrealStereo4K: The UnrealStereo4K dataset [67] is a synthetic RGB-RGB stereo dataset containing around $8K$ high-resolution (3840×2160) stereo pairs with dense ground-truth disparities. Although it does not include multi-spectral images, we employ UnrealStereo4K as an additional source of training data to improve the overall final results yielded by our cross-modal networks. In particular, as it is a standard RGB-RGB dataset, we simulate a RGB-MS setup by simply converting the right image of the stereo pair from RGB to grayscale and, then, stacking it to form a volume of 10 channels. Moreover, we resize input images fed to Ψ using a downsample factor of 6 to mimic the unbalanced factor featured by our real RGB-MS setup.

5.2. Implementation Details

Our proposed architecture is implemented with PyTorch. We adopt Adam [38] as optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. In our implementation, we adopt different stereo backbones for our Ψ_θ submodule. In particular, we conduct experiments on two different 3D architectures such as PSM [12] and GWC [26], showing that our proposal can be effectively combined with diverse stereo networks. For both the selected backbones, we use the official code provided by the authors and modify the input channels of the

Method	Input	2° RGB	Synth	D-AEPE	ADE (m)
SGM	Single-Channel			14.66	0.215
Zhi et al. [78] (photo)	Single-Channel			89.86	1.010
Zhi et al. [78]	Single-Channel	✓		19.45	0.116
PSM	All-Channels			14.51	0.103
PSM	All-Channels	✓		10.04	0.084
PSM + MLPs	All-Channels	✓		19.12	0.182
PSM + MLPs + Φ	All-Channels	✓		**8.36**	**0.065**
PSM + MLPs + Φ	Single-Channel	✓		24.13	0.737
PSM + MLPs + Φ	All-Channels	✓	✓	**7.31**	**0.057**
GWC + MLPs + Φ	All-Channels	✓	✓	8.67	0.100

Table 1. **Disparity/Depth Errors**. Quantitative results on the proposed RGB-MS dataset. We report error metrics for the disparity/depth predictions. Best results in bold.

Method	Input	2° RGB	Synth	F-AEPE	bad_1	bad_2	bad_3
SGM	Single-Channel			2.32	52.40	25.73	14.28
Zhi et al. [78] (photo)	Single-Channel			14.22	85.05	73.87	66.42
Zhi et al. [78]	Single-Channel	✓		3.08	70.51	47.81	32.65
PSM	All-Channels			2.30	44.49	21.42	13.07
PSM	All-Channels	✓		1.59	50.90	21.41	11.20
PSM + MLPs	All-Channels	✓		3.03	67.42	36.86	21.20
PSM + MLPs + Φ	All-Channels	✓		**1.32**	**34.89**	**13.32**	**7.05**
PSM + MLPs + Φ	Single-Channel	✓		3.82	73.20	43.08	22.20
PSM + MLPs + Φ	All-Channels	✓	✓	**1.16**	**32.15**	**11.54**	**6.39**
GWC + MLPs + Φ	All-Channels	✓	✓	1.37	37.91	15.12	7.64

Table 2. **Flow Errors**. Quantitative results on the proposed RGB-MS dataset. We report error metrics for the flow maps employed during registration. Best results in bold.

feature extractor associated with the right image to match the number of bands of our multi-spectral camera (i.e., 10). For MLP_C and MLP_R, instead, we follow the implementation of [1]. We train each network for 70 epochs on a single NVIDIA 3090 GPU with a learning rate of 10^{-4}. As RGB-MS training set we use the whole set of 11K unlabeled images, computing SGM proxy labels on each stereo pair. We discard proxy disparity maps with less than 70% of valid pixels to avoid too sparse supervision, obtaining approximately 9K valid training images. When training also on UnrealStereo4K, we first pre-train for 30 epochs solely on synthetic data and then we train for other 70 epochs with both real and synthetic data. We empirically noticed that this training procedure produces sharper disparity maps as final results compared to training on RGB-MS pairs only. During training we use a batch size of 2, random crops of size 1152×1152 from L and select the corresponding crop of size 192×192 on the multi-spectral image R. Moreover, we randomly sample $P = 30000$ points from each crop at continuous spatial locations in the image domain to pursue training based on our continuous formulation. Additional training details are reported in the supplementary material.

5.3. Evaluation Metrics

To evaluate our RGB-MS model, we use several metrics for optical flow, disparity and depth estimation, computing all the metrics at high resolution. In case of the optical flow metrics, we consider the registration error from the RGB to the MS camera, which, indeed, is the performance metric most relevant in our scenario which, in essence, deals with enriching the RGB image with MS information taken from the available lower resolution sensor. Hence, the metrics used to evaluate the performance of our models are: 1) *Flow Average End Point Error (F-AEPE)*, defined as the norm of the difference between the ground-truth and predicted flow vectors expressed in pixels, 2) *Disparity Average End-Point-Error (D-AEPE)*, measured in pixels between the predicted disparity map and the ground-truth disparity map, averaged across all pixels, 3) *Absolute depth error* (ADE) measured in meters (m), between the predicted depth map

and the ground-truth depth map and 4) *Percentage of bad pixels (bad$_\tau$)*, where τ is a tolerance on the optical flow error to accept an estimated flow as correct.

5.4. Results

We report experimental results and ablation studies dealing with disparity/depth and flow error metrics in Tab. 1 and Tab. 2, respectively. As a reference, in the first row of both tables, we show the results yielded by SGM on the RGB-MS test set when both stereo frames are converted to grayscale, averaging all channels in case of the MS image. Moreover, we compare against [78] by using their network on our RGB-MS stereo setup.

Ablation on Supervision and Architecture. In rows from 4 to 7 of both tables, we ablate the contributions of the proposed supervision strategy and architectural components. First of all, comparing the basic configuration of PSM trained on SGM proxy labels obtained with or without the adoption of the second additional RGB camera (rows 4 vs 5) we note an improvement for almost all metrics, with a sensible decrease of 4 pixels and almost 1 pixel in the D-AEPE and F-AEPE error metrics, respectively. These results support our idea that a better supervision obtained by two cameras with the same modality can improve the matching performance across different modalities significantly. In the sixth row, we add MLP_C and MLP_R to the PSM backbone, so as to introduce our continuous formulation potentially yielding much more accurate and sharp high-resolution disparity maps. However, this causes a substantial deterioration of all performance metrics. We ascribe this issue to the task of recovering an accurate high-resolution disparity map through our continuous formulation being too difficult in absence of any source of high-resolution information. Indeed, in the following rows, we add the Φ high-resolution feature extractor (Fig. 3) to the architecture comprising the PSM backbone and the MLPs, achieving the best results in all metrics with a huge boost compared to all previous configurations.

We can appreciate the effect of each component in the first 4 columns of Fig. 5. For instance, we can perceive how

PSM	*+ 2° RGB*	*+ MLPs*	*+ Φ*	*+ Synth*
(a)	(b)	(c)	(d)	(e)

Figure 5. **Ablation Qualitative Results.** Disparity maps for the stereo pair in Fig. 1 computed by: PSM (a), PSM with the proposed 2°
RGB supervision (b), PSM with the continuous and sharp formulation leveraging the two MLPs (c). Finally, the full architecture with the
high-resolution feature extractor Φ without (d) and with synthetic supervision (e).

the continuous formulation based on the MLPs (third col-
umn) can yield disparity maps exhibiting sharp edges, but
only with the introduction of the encoder Φ (fourth column)
we are able to recover correctly the high-resolution details.

Comparison with Existing Baselines. In the second
and third rows, we report the results obtained by [78] by
training it on our RGB-MS dataset, using the original pho-
tometric loss between the RGB image and the MS image
averaged across channels, as well as on proxy labels ex-
tracted using the second RGB image. We can observe how,
in both tables, the photometric loss (row 2) is in-effective as
the original formulation relies on photometrically calibrated
RGB-NIR images and a mapping from RGB to pseudo-
NIR, while the equivalent RGB to pseudo-MS mapping
does not exist according to our MS camera manufacturer.
On the contrary, by training it on proxy labels, it notably
ameliorates the results – although our network still repre-
sents the best solution for this task, since the architecture
by [78] is not designed to handle our unbalanced setup.

Comparison with Alternative Input Modality. In row
8, we perform an additional test comparing our proposal
to a baseline strategy to handle different input modalities
with a different number of channels, – i.e., collapsing both
the left and right image into a single-channel representation.
We achieve this by converting RGB images into grayscale
and averaging the 10 channels of MS images. Thus, we can
use the original formulation of the PSM architecture, which
relies on a classical shared feature extractor. From Tab.1
and 2 we notice how the network struggles with this concise
input representation, suggesting that we better rely on the
information from all bands to learn a reliable matching.

Auxiliary Synthetic Data. In both tables, we also report
the results obtained by training the best architecture (rows
7) with additional synthetic data (Sec. 5.1 and Sec. 5.2).
We note that we can further push its performances (rows
9 vs rows 7) achieving the best results in all metrics and
with an average flow estimation error (F-AEPE) as small as
almost 1 pixel (1.16), a remarkable result for such a chal-
lenging registration task dealing with images featuring di-
verse modalities and highly unbalanced resolutions. In the
last column of Fig. 5 we can perceive qualitatively how
leveraging on auxiliary synthetic training data can help to
further improve the predicted disparity maps, in particular

in terms of sharpness of the finer details thanks to the per-
fect ground-truth supervision in challenging regions such as
depth discontinuities.

Eventually, in the last rows we report also the results ob-
tained using a different backbone, such as GWC [26], prov-
ing to be effective as well.

6. Conclusion, Limitations and Future work

In this paper, we have explored RGB-MS image regis-
tration for the first time. Purposely, we have collected a
dataset of 34 synchronized RGB-MS pairs annotated with
accurate semi-dense ground-truth disparity maps. We have
also proposed a deep network architecture to perform cross-
modal matching between the two image modalities. More-
over, to avoid the need of hard-to-source ground-truth la-
bels, we have proposed a novel supervision strategy aimed
at distilling knowledge from an easier RGB-RGB matching
problem. This strategy has been realized by training our net-
work on 11K unlabeled RGB-MS/RGB-RGB image pairs
acquired as part of our dataset. Testing the trained model
on the annotated RGB-MS pairs shows promising results,
with an average registration error close to 1 pixel despite
the dramatic diversity in image modality and resolution.

Although our dataset is the first to address RGB-MS
matching problem, it has several limitations. In particu-
lar, although comparable in terms of number of samples to
some popular stereo datasets [55], more images could be
collected to make our benchmark more complete and chal-
lenging. Moreover, at the moment, ground-truths are lim-
ited to indoor scenes only. Finally, images are acquired with
only one specific 10-bands MS camera. As future work, we
plan to increase the size and variety of the dataset by acquir-
ing more environments and employing diverse MS devices.

Many other ideas should definitely be explored to bet-
ter tackle the novel and challenging registration task ad-
dressed in this paper. In this respect, we believe that our
proposed architecture and training methodology set forth a
strong baseline to steer further investigation towards more
effective and more general solutions.

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. 3, 5, 7

[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. 3, 6

[3] Sangwook Baek, Euisun Choi, Yoonkil Baek, and Chulhee Lee. Detection of counterfeit banknotes using multispectral images. *Digital Signal Processing*, 78:294–304, 2018. 1

[4] W. G. C. Bandara, G. W. K. Prabhath, D. W. S. C. B. Dissanayake, H. M. V. R. Herath, G. M. R. I. Godaliyadda, M. P. B. Ekanayake, S. S. P. Vithana, S. M. D. Demini, and T. Madhujith. A multispectral imaging system to assess meat quality. In *2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pages 1–6, 2018. 1

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

[6] B. Bayer. Color imaging array. U.S. America Patent US 3 971 065A, Jul. 20, 1975. 1

[7] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 3

[8] Johannes Brauers, Nils Schulte, and Til Aach. Multispectral filter-wheel cameras: Geometric distortion model and compensation algorithms. *IEEE Transactions on Image Processing*, 17(12):2368–2380, 2008. 2

[9] 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. 3

[10] Si-Yuan Cao, Hui-Liang Shen, Shu-Jie Chen, and Chunguang Li. Boosting structure consistency for multispectral and multimodal image registration. *IEEE Transactions on Image Processing*, 29:5147–5162, 2020. 3

[11] Xun Cao, Xin Tong, Qionghai Dai, and Stephen Lin. High resolution multispectral video capture with a hybrid camera system. In *CVPR 2011*, pages 297–304, 2011. 2

[12] 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. 3, 5, 6

[13] 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. 3

[14] Walon Wei-Chen Chiu, Ulf Blanke, and Mario Fritz. Improving the kinect by cross-modal stereo. In *BMVC*, 2011. 2, 3

[15] Philippe Colantoni, Ruven Pillay, Christian Lahanier, and Denis Pitzalis. Analysis of multispectral images of paint-

ings. In *2006 14th European Signal Processing Conference*, pages 1–5, 2006. 1

[16] 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, 4

[17] 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. 3

[18] Claudio I. Fernández, Ata Haddadi, Brigitte Leblon, Jinfei Wang, and Keri Wang. Comparison between three registration methods in the case of non-georeferenced close-range multispectral images. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 8205–8208, 2021. 3

[19] Maxime Ferrera, Alexandre Boulch, and Julien Moras. Fast stereo disparity maps refinement by fusion of data-based and model-based estimations. In *2019 International Conference on 3D Vision (3DV)*, pages 9–17. IEEE, 2019. 3

[20] Bruno Garavelli, Alessandra Mencarelli, and Luca Zanotti. Xspectra¡sup¿®¡/sup¿: The most advanced real time food contaminants detector. In *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4, 2017. 1

[21] Jingjing Ge, Danying Fu, Xiaoming Zhong, and Xiaojie Yu. A new target identification measurement based on multispectral imaging. In Gary W. Kamerman and Ove Steinvall, editors, *Electro-Optical Remote Sensing XIII*, volume 11160, pages 154 – 160. International Society for Optics and Photonics, SPIE, 2019. 1

[22] Nils Genser, Jürgen Seiler, and André Kaup. Camera array for multi-spectral imaging. *IEEE Transactions on Image Processing*, 29:9234–9249, 2020. 2

[23] 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. 3

[24] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 6

[25] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 6

[26] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *CVPR*, 2019. 5, 6, 8

[27] Nina Heide, Christian Frese, Thomas Emter, and Janko Petereit. Real-time hyperspectral stereo processing for the generation of 3d depth information. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3299–3303. IEEE, 2018. 3

[28] Yong Seok Heo, Kyong Mu Lee, and Sang Uk Lee. Robust stereo matching using adaptive normalized cross-correlation. *IEEE Transactions on pattern analysis and machine intelligence*, 33(4):807–822, 2010. 3

[29] Yong Seok Heo, Kyoung Mu Lee, and Sang Uk Lee. Joint depth map and color consistency estimation for stereo images with different illuminations and cameras. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1094–1106, 2012. 3

[30] 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. 4, 6

[31] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045, 2015. 1, 2

[32] Eddy Ilg, Tonmoy Saikia, Margret Keuper, and Thomas Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *The European Conference on Computer Vision (ECCV)*, September 2018. 3

[33] Zequn Jie, Pengfei Wang, Yonggen Ling, Bo Zhao, Yunchao Wei, Jiashi Feng, and Wei Liu. Left-right comparative recurrent model for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[34] 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. 3

[35] Sewoong Kim, Dongrae Cho, Jihun Kim, Manjae Kim, Sangyeon Youn, Jae Eun Jang, Minkyu Je, Dong Hun Lee, Boreom Lee, Daniel L. Farkas, and Jae Youn Hwang. Smartphone-based multispectral imaging: system development and potential for mobile skin diagnosis. *Biomed. Opt. Express*, 7(12):5294–5307, Dec 2016. 1

[36] Seungryong Kim, Dongbo Min, Bumsub Ham, Minh N Do, and Kwanghoon Sohn. Dasc: Robust dense descriptor for multi-modal and multi-spectral correspondence estimation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1712–1729, 2016. 2, 3

[37] Seungryong Kim, Dongbo Min, Stephen Lin, and Kwanghoon Sohn. Deep self-correlation descriptor for dense cross-modal correspondence. In *European Conference on Computer Vision*, pages 679–695. Springer, 2016. 2, 3

[38] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[39] Ilona Kuzmina, Ilze Diebele, Dainis Jakovels, Janis Spigulis, Lauma Valeine, Janis Kapostinsh, and Anna Berzina. Towards noncontact skin melanoma selection by multispectral imaging analysis. *Journal of Biomedical Optics*, 16(6):1 – 4, 2011. 1

[40] 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. 3

[41] Pierre-Jean Lapray, Xingbo Wang, Jean-Baptiste Thomas, and Pierre Gouton. Multispectral filter arrays: Recent advances and practical implementation. *Sensors*, 14(11):21626–21659, 2014. 2

[42] 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. 3

[43] Jongwoo Lim. Optimized projection pattern supplementing stereo systems. In *2009 IEEE International Conference on Robotics and Automation*, pages 2823–2829. IEEE, 2009. 4

[44] Lahav Lipson, Zachary Teed, and Jia Deng. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. *arXiv preprint arXiv:2109.07547*, 2021. 3

[45] 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. 3

[46] 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. 3

[47] Max Mehltretter, Sebastian P Kleinschmidt, Bernardo Wagner, and Christian Heipke. Multimodal dense stereo matching. In *German Conference on Pattern Recognition*, pages 407–421. Springer, 2018. 3

[48] Vlad-Cristian Miclea, Cristian-Cosmin Vancea, and Sergiu Nedevschi. New sub-pixel interpolation functions for accurate real-time stereo-matching algorithms. In *2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 173–178. IEEE, 2015. 4

[49] 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. 3

[50] Peter Pinggera12, Toby Breckon, and Horst Bischof. On cross-spectral stereo matching using dense gradient features. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, volume 2, 2012. 3

[51] Matteo Poggi, Seungryong Kim, Fabio Tosi, Sunok Kim, Filippo Aleotti, Dongbo Min, Kwanghoon Sohn, and Stefano Mattoccia. On the confidence of stereo matching in a deep-learning era: a quantitative evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 4

[52] 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. 3, 6

[53] 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. 3

[54] Jianwei Qin, Kuanglin Chao, Moon S. Kim, Renfu Lu, and Thomas F. Burks. Hyperspectral and multispectral imaging for evaluating food safety and quality. *Journal of Food Engineering*, 118(2):157–171, 2013. 1

[55] 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. 8

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

[57] Akihito Seki and Marc Pollefeys. Patch based confidence prediction for dense disparity map. In *BMVC*, pages 23.1–23.13, 2016. 3

[58] Akihito Seki and Marc Pollefeys. SGM-Nets: Semi-global matching with neural networks. In *CVPR*, pages 231–240, 2017. 3

[59] Xiaoyong Shen, Li Xu, Qi Zhang, and Jiaya Jia. Multi-modal and multi-spectral registration for natural images. In *ECCV*, 2014. 2, 3

[60] 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. 3

[61] Xiao Song, Xu Zhao, Hanwen Hu, and Liangji Fang. Edgestereo: A context integrated residual pyramid network for stereo matching. In *ACCV*, 2018. 3

[62] Nicolaas Tack, Andy Lambrechts, Philippe Soussan, and Luc Haspeslagh. A compact, high-speed, and low-cost hyperspectral imager. In Joel Kubby and Graham Trevor Reed, editors, *Silicon Photonics VII*, volume 8266, pages 126 – 138. International Society for Optics and Photonics, SPIE, 2012. 2

[63] 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. 3, 6

[64] 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. 3

[65] 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. 3

[66] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[67] 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. 5, 6

[68] Fabio Tosi, Matteo Poggi, Stefano Mattoccia, Alessio Tonioni, and Luigi di Stefano. Learning confidence measures in the wild. In *BMVC*, 2017. 4

[69] Liangju Wang, Jian Jin, Zhihang Song, Jialei Wang, Libo Zhang, Tanzeel U. Rehman, Dongdong Ma, Neal R. Carpenter, and Mitchell R. Tuinstra. Leafspec: An accurate and portable hyperspectral corn leaf imager. *Computers and Electronics in Agriculture*, 169:105209, 2020. 2

[70] 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. 3

[71] 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. 3, 5

[72] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *ECCV*, pages 636–651, 2018. 3

[73] Rahat Yasir, Mark Eramian, Ian Stavness, Steve Shirtliffe, and Hema Duddu. Data-driven multispectral image registration. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 230–237, 2018. 3

[74] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *European conference on computer vision*, pages 151–158. Springer, 1994. 4

[75] 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. 3

[76] 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. 3, 5

[77] 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. 3

[78] Tiancheng Zhi, Bernardo R Pires, Martial Hebert, and Srinivasa G Narasimhan. Deep material-aware cross-spectral stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1916–1925, 2018. 2, 3, 7, 8

[79] Zilong Zhong, Jonathan Li, Zhiming Luo, and Michael Chapman. Spectral–spatial residual network for hyperspectral image classification: A 3-d deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):847–858, 2018. 1

[80] 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. 3

[81] Zhuang Zhou, Shengyang Li, and Yuyang Shao. Crops classification from sentinel-2a multi-spectral remote sensing images based on convolutional neural networks. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 5300–5303, 2018. 1