A Survey on Deep Learning Techniques for Stereo-Based Depth Estimation

Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun, *Senior Member, IEEE*

Abstract—Estimating depth from RGB images is a long-standing ill-posed problem, which has been explored for decades by the computer vision, graphics, and machine learning communities. Among the existing techniques, stereo matching remains one of the most widely used in the literature due to its strong connection to the human binocular system. Traditionally, stereo-based depth estimation has been addressed through matching hand-crafted features across multiple images. Despite the extensive amount of research, these traditional techniques still suffer in the presence of highly textured areas, large uniform regions, and occlusions. Motivated by their growing success in solving various 2D and 3D vision problems, deep learning for stereo-based depth estimation has attracted a growing interest from the community, with more than 150 papers published in this area between 2014 and 2019. This new generation of methods has demonstrated a significant leap in performance, enabling applications such as autonomous driving and augmented reality. In this paper, we provide a comprehensive survey of this new and continuously growing field of research, summarize the most commonly used pipelines, and discuss their benefits and limitations. In retrospect of what has been achieved so far, we also conjecture what the future may hold for deep learning-based stereo for depth estimation research.

Index Terms—CNN, deep learning, 3D reconstruction, stereo matching, multi-view stereo, disparity estimation, feature leaning, feature matching

1 Introduction

Depth estimation from one or multiple RGB images is a long standing ill-posed problem, with applications in various domains such as robotics, autonomous driving, object recognition and scene understanding, 3D modeling and animation, augmented reality, industrial control, and medical diagnosis. This problem has been extensively investigated for many decades. Among all the techniques that have been proposed in the literature, stereo matching is traditionally the most explored one due to its strong connection to the human binocular system.

The *first* generation of stereo-based depth estimation methods relied typically on matching pixels across multiple images captured using accurately calibrated cameras. Although these techniques can achieve good results, they are still limited in many aspects. For instance, they are not suitable when dealing with occlusions, featureless regions, or highly textured regions with repetitive patterns. Interestingly, we, as humans, are good at solving such ill-posed inverse problems by leveraging prior knowledge. For example, we can easily infer the approximate sizes of objects, their relative locations, and even their approximate relative distance to our eye(s). We can do this because all the previously seen objects and scenes have enabled us to build prior knowledge and develop mental models of how the 3D world looks like. The *second* generation of methods tries to leverage this prior knowledge by formulating the problem as a learning task. The advent of deep learning techniques in computer vision [1] coupled with the increasing availability of large training datasets, have led to a *third* generation of methods that are able to recover the lost dimension. Despite being recent, these methods have demonstrated exciting and promising results on various tasks related to computer vision and graphics.

In this paper, we provide a comprehensive and structured review of the recent advances in stereo image-based depth estimation using deep learning techniques. These methods use two or more images captured with spatially-distributed RGB cameras.[1] We have gathered more than 150 papers, which appeared between January 2014 and December 2019 in leading computer vision, computer graphics, and machine learning conferences and journals.[2] The goal is to help the reader navigate in this emerging field, which has gained a significant momentum in the past few years.

The major contributions of this paper are as follows;

- To the best of our knowledge, this is the first paper that surveys stereo-based depth estimation using deep learning techniques. We present a

- *Hamid Laga is with the Information Technology Discipline, Murdoch University, Murdoch 6150, Australia, and with the University of South Australia, The Phenomics and Bioinformatics Research Centre, SA 5000, Australia. E-mail: H.Laga@murdoch.edu.au.*
- *Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun are with the University of Western Australia, Perth, WA 6009, Australia. E-mail: laurent.jospin@research.uwa.edu.au, {farid.boussaid, mohammed.bennamoun}@uwa.edu.au.*

Manuscript received 16 May 2020; revised 6 Oct. 2020; accepted 14 Oct. 2020. Date of publication 20 Oct. 2020; date of current version 4 Mar. 2022. (Corresponding author: Hamid Laga.) Recommended for acceptance by O. Veksler. Digital Object Identifier no. 10.1109/TPAMI.2020.3032602

1. Deep learning-based depth estimation from monocular images and videos is an emerging field and requires a separate survey.
2. At the time of writing this paper.

comprehensive review of more than 150 papers, which appeared in the past six years in leading conferences and journals.
- We provide a comprehensive taxonomy of the state-of-the-art. We first describe the common pipelines and then discuss the similarities and differences between methods within each pipeline.
- We provide a comprehensive review and an insightful analysis on all the aspects of the problem, including the training data, the network architectures and their effect on the reconstruction performance, the training strategies, and the generalization ability.
- We provide a comparative summary of the properties and performances of some key methods using publicly available datasets and in-house images. The latter have been chosen to test how these methods would perform on completely new scenarios.

The rest of this paper is organized as follows; Section 2 formulates the problem and lays down the taxonomy. Section 3 surveys the various datasets which have been used to train and test stereo-based depth reconstruction algorithms. Section 4 focuses on the works that use deep learning architectures to learn how to match pixels across images. Section 5 reviews the end-to-end methods for stereo matching, while Section 6 discusses how these methods have been extended to the multi-view stereo case. Section 7 focuses on the training procedures including the choice of the loss functions and the degree of supervision. Section 8 discusses the performance of key methods. Finally, Section 9 discusses the potential future research directions, while Section 10 summarizes the main contributions of this paper.

2 Scope and Taxonomy

Let $\mathbf{I} = \{I_k, k = 1, \ldots, n\}$ be a set of $n \geq 1$ RGB images of the same 3D scene, captured using cameras whose intrinsic and extrinsic parameters can be *known* or *unknown*. The goal is to estimate one or multiple depth maps, which can be from the same viewpoint as the input [2], [3], [4], [5], or from a new arbitrary viewpoint [6], [7], [8], [9], [10]. This paper focuses on deep learning methods for stereo-based depth estimation, i.e., $n = 2$ in the case of stereo matching, and $n > 2$ for the case of Multi-View Stereo (MVS). Monocular and video-based depth estimation methods are beyond the scope of this paper and require a separate survey.

Learning-based depth reconstruction can be summarized as the process of learning a predictor f_θ that can infer from the set of images $\mathbf{I}$, a depth map $\hat{D}$ that is as close as possible to the unknown depth map D. In other words, we seek to find a function f_θ such that $\mathcal{L}(\mathbf{I}) = d(f_\theta(\mathbf{I}), D)$ is minimized. Here, θ is a set of parameters, and $d(\cdot, \cdot)$ is a certain measure of distance between the real depth map D and the reconstructed depth map $f_\theta(\mathbf{I})$. The reconstruction objective $\mathcal{L}$ is also known as the *loss function*.

We can distinguish two main categories of methods. Methods in the *first* class (Section 4) mimic the traditional stereo-matching techniques [11] by explicitly learning how to match, or put in correspondence, pixels across the input images. Such correspondences can then be converted into an optical flow or a disparity map, which in turn can be converted into depth at each pixel in the reference image. The predictor f is composed of three modules: a feature extraction module, a feature matching and cost aggregation module, and a disparity/depth estimation module. Each module is trained independently from the others.

The *second* class of methods (Section 5) solves the stereo matching problem using a pipeline that is trainable end-to-end. Two main classes of methods have been proposed. Early methods formulated the depth estimation as a regression problem. In other words, the depth map is directly regressed from the input without explicitly matching features across the views. While these methods are simple and fast at runtime, they require a large amount of training data, which is hard to obtain. Methods in the second class mimic the traditional stereo matching pipeline by breaking the problem into stages composed of differentiable blocks and thus allowing end-to-end training. While a large body of the literature focused on pairwise stereo methods, several papers have also addressed the multi-view stereo case and these will be reviewed in Section 6.

In all methods, the estimated depth maps can be further refined using refinement modules [2], [3], [12], [13] and/or progressive reconstruction strategies where the reconstruction is refined every time new images become available.

Finally, the performance of deep learning-based stereo methods depends not only on the network architecture but also on the datasets on which they have been trained (Section 3) and on the training procedure used to optimise their parameters (Section 7). The latter includes the choice of the loss functions and the supervision mode, which can be fully supervised with 3D annotations, weakly supervised, or self-supervised. We will discuss all these aspects in the subsequent sections.

3 Datasets

Table 1 summarizes some of the datasets that have been used to train and test deep learning-based depth estimation algorithms. Below, we discuss these datasets based on their sizes, their spatial and depth resolution, the type of depth annotation they provide, and the domain gap (or shift) issue faced by many deep learning-based algorithms.

1) Dataset Size. The first datasets, which appeared prior to 2016, are of small scale due to the difficulty of creating ground-truth 3D annotations. An example is the two KITTI datasets [15], [21], which contain 200 stereo pairs with their corresponding disparity ground-truth. They have been extensively used to train and test patch-based CNNs for stereo matching algorithms (see Section 4), which have a small receptive field. As such a single stereo pair can result in thousands of training samples. However, in end-to-end architectures (Sections 5 and 6), a stereo pair corresponds to only one sample. End-to-end networks have a large number of parameters, and thus require large datasets for efficient training. While collecting large image datasets is very easy, e.g., by using video sequences as in e.g., NYU2 [17], ETH3D [25], SUN3D [19], and ETH3D [25], annotating them with 3D labels is time consuming. Recent works, e.g., the ApolloScape [34] and A2D2 [35], use LIDAR to acquire dense 3D annotations.

Data augmentation strategies, e.g., by applying geometric and photometric transformations to the images that are

TABLE 1
Datasets for Depth/Disparity Estimation

	Year	Type	Purpose	Images					Depth					Cam. params.	
				Resolution	# Scenes	# Views per scene	# Tr. scenes	# Ts. scenes	Resolution	#GT frames	Type	Depth range	Disparity range	Int.	Ext.
Make3D [14]	2009	Real	Monocular depth	2272×1704	534	monocular	400	134	78×51	534	Dense	–	–		
KITTI2012 [15]	2012	Real	Stereo	1240×376	389	2	194	195	1226×370	–	Sparse	–	–	Y	Y
MPI Sintel [16]	2012	Synthetic	Optical flow	1024×436	35 videos	50	23 videos	12 videos	–	–	Dense	–	–		
NYU2 [17]	2012	Real - indoor	Monocular depth, object segmentation	640×480	464 videos, 100+ fr. per video	monocular	–	–	–	1,449	Kinect depth	–	–	N	N
RGB-D SLAM [18]	2012	Real	SLAM	640×480	19 videos		15 videos	4 videos	–	–	Dense	–	–	Y	Y
SUN3D [19]	2013	Real - rooms	Monocular video	640×480	415 videos, 10–1000+ fr. per video	–	–	–	–	–	Dense, SfM	–	–		Y
Middleburry [20]	2014	Indoor	Stereo	2948×1988	30	2	15	15	2948×1988	30	Dense	–	260	Y	Y
KITTI 2015 [21]	2015	Real	Stereo	1242×375	400	4	200	200	1242×375	–	Sparse	–	–	Y	Y
KITTI-MVS2015 [21]	2015	Real	MVS	1242×375	400	20	200	200	–	–	Sparse	–	–	Y	Y
FlyingThings3D, Monkaa, Driving [22]	2016	Synthetic	Stereo, Video, Optical flow	960×540	39K frames	2	21,818	4,248	384×192	–	Dense	–	160px	Y	Y
CityScapes [23]	2016	Street scenes	Semantic seg., dense labels	2048×1024	5K	2	2975	1525	–	–	NA	–	–	Ego-motion	
			Semantic seg.,coarse labels	2048×1024	20K	2			NA	NA	NA	NA	NA	Ego-motion	
DTU [24]	2016	Real, small objects	MVS	1200×1600	80	$49 - 64$	–	–	–	–	Structured light scans	–	–	Y	Y
ETH3D [25]	2017	Real, in/outdoor	Low-res, Stereo	940×490	47	2	27	20	–	47	Dense	–	–	Y	Y
			Low-res, MVS on video	940×490	10 videos	4	5 videos	5 videos	–	–	Dense	–	–	Y	Y
			High-res, MVS on images from DSLR camera	940×490	25	$14 - 76$	13	12	–	25	Dense	–	–	Y	Y
SUNCG [26]	2017	Synthetic, indoor	Scene completion	–	45K	–	–	–	640×480	–	Depth and Vol. GT	–	–		
MVS-Synth [27]	2018	Synth - urban	MVS	1920×1080	120	100			–	–	Dense	–	–	Y	Y
MegaDepth [28]	2018	Real (Internet images)	Monocular, Eucl. and ord. depth	1600×1600	130K	monocular	–	–	–	100K (Eucl.), 30K (Ord.)	Dense, Eucl., Ord.	–	–		
Jeon and Lee [29]	2018	Real	Depth enhancement	–	4K images	–	–	–	640×480	4,000	Dense	$0.01 - 30$m	–	Y	Y
OmniThings [30], [31]	2019	Synthetic, fish-eye images	Omnidirectional MVS	800×768	10240	4	9216	1024	640×320	–	Dense	–	$\leq$ 192px		
OmniHouse [30], [31]	2019	Synthetic, fish-eye images	Omnidirectional MVS	800×768	2,560	4	2048	512	640×320	–	Dense	–	$\leq$ 192px		
HR-VS [32]	2019	Synthetic, outdoor	High res. stereo	2056×2464	780	2	–	–	1918×2424	780	Dense, Eucl.	2.52 to 200m	9.66 to 768px		
		Real, outdoor	High res. stereo	1918×2424	33	2	–	–	1918×2424	33	Dense, Eucl.		5.41 to 182.3px		
DrivingStereo [33]	2019	Driving	High res. stereo	1762×800	182,188	2	174,437	7,751	1762×800	182,188	Sparse	up to 80m			
ApolloScape [34]	2019	Auto. driving	High res. stereo	3130×960	5,165	2	4,156	1,009	–	5165	LIDAR	$-$ to $-$m	–	Y	–
A2D2 [35]	2020	Auto. driving	High res. stereo	2.3M pixel	41,277	6	–	–	–	–	LIDAR	up to 100m	–	Y	Y

"GT": ground-truth, "Tr.": training, "Ts.": testing, "fr.": frames, "Vol.": volumetric, "Eucl": euclidean, "Ord": ordinal, "Int.": intrinsic, "Ext.": extrinsic.

available, have been extensively used in the literature. There are, however, a few other strategies that are specific to depth estimation. This includes artificially synthesizing and rendering from 3D CAD models 2D and 2.5D views from various (random) viewpoints, poses, and lighting conditions. One can also overlay rendered 3D models on the top of real images. This approach has been used to generate the FlyingThings3D, Monkaa, and Driving datasets of [22], and the OmniThings and OmniHouse datasets for benchmarking MVS for omnidirectional images [30], [31]. Huang *et al.* [27] followed a similar idea but used scenes from video games to generate MVS-Synth, a photo-realistic synthetic dataset prepared for learning-based Multi-View Stereo algorithms.

The main challenge is that generating large amounts of synthetic data containing varied real-world appearance and motion is not trivial [36]. As a result, a number of works overcome the need for ground-truth depth information by training their deep networks without 3D supervision, see Section 7.1. Others used traditional depth estimation and structure-from-motion (SfM) techniques to generate 3D annotations. For example, Li *et al.* [28] used modern structure-from-motion and multiview stereo (MVS) methods together with multiview Internet photo collections to create the large-scale MegaDepth dataset providing improved depth estimation accuracy via bigger training dataset sizes. This dataset has also been automatically augmented with ordinal depth relations generated using semantic segmentation.

2) Spatial and Depth Resolutions. The disparity/depth information can be either in the form of maps of the same or lower resolution than the input images, or in the form of sparse depth values at some locations in the reference image. Most of the existing datasets are of low spatial resolution. In recent years, however, there has been a growing focus on stereo matching with high-resolution images. An example of a high-resolution dataset is the HR-VS and HR-RS of Yang *et al.* [32], where each RGB pair of resolution 1918×2424 is annotated with a depth map of the same resolution. However, the dataset only contains 800 pairs of stereo images, which is relatively small for end-to-end training. Other datasets such as the ApolloScape [34] and A2D2 [35] contain very high resolution images, of the order of 3130×960, with more that $100+$ hours of stereo driving videos, in the case of ApolloScape, have been specifically designed to test autonomous driving algorithms.

3) euclidean versus Ordinal Depth. Instead of manually annotating images with exact, i.e., euclidean, depth values, some papers, e.g., MegaDepth [28], provide ordinal annotations, i.e., pixel x_1 is closer, farther, or at the same depth, as pixel x_2. Ordinal annotation is simpler and faster to achieve than euclidean annotation. In fact, it can be accurately obtained using traditional stereo matching algorithms, since ordinal depth is less sensitive to innacuracies in depth estimation

4) Domain Gap. While artificially augmenting training datasets allows enriching existing ones, the domain shift caused by the very different conditions between real and synthetic data can result in a lower accuracy when applied to real-world environments. We will discuss, in Section 7.3, how this domain shift issue has been addressed in the literature.

4 Depth by Stereo Matching

Stereo-based depth reconstruction methods take $n = 2$ RGB images and produce a disparity map D that minimizes an energy function of the form

$$E(D) = \sum_x C(x, d_x) + \sum_x \sum_{y \in \mathcal{N}_x} E_s(d_x, d_y). \quad (1)$$

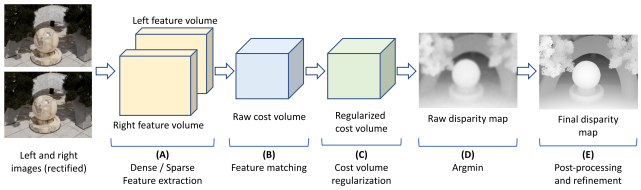


Fig. 1. The building blocks of a stereo matching pipeline.

Here, x and y are image pixels, and $\mathcal{N}_x$ is the set of pixels that are within the neighborhood of x. The first term of Eqn. (1) is the matching cost. When using rectified stereo pairs, $C(x, d_x)$ measures the cost of matching the pixel $x = (i, j)$ of the left image with the pixel $y = (i, j - d_x)$ of the right image. In this case, $d_x = D(x) \in [d_{min}, d_{max}]$ is the disparity at pixel x. Depth can then be inferred by triangulation. When the disparity range is discritized into n_d disparity levels, C becomes a 3D cost volume of size $W \times H \times n_d$. In the more general multiview stereo case, i.e., $n \geq 2$, the cost $C(x, d_x)$ measures the inverse likelihood of x on the reference image having depth d_x. The second term of Eqn. (1) is a regularization term used to impose constraints such as smoothness and left-right consistency.

Traditionally, this problem has been solved using a pipeline of four building blocks [11], see Fig. 1: (1) feature extraction, (2) feature matching across images, (3) disparity computation, and (4) disparity refinement and post-processing. The first two blocks construct the cost volume C. The third block regularizes the cost volume and then finds, by minimizing Eqn. (1), an initial estimate of the disparity map. The last block refines and post-processes the initial disparity map.

This section focuses on how these individual blocks have been implemented using deep learning-based methods. Table 2 summarises the state-of-the-art methods.

4.1 Learning Feature Extraction and Matching

Early deep learning techniques for stereo matching replace the hand-crafted features (block A of Fig. 1) with learned features [37], [38], [39], [42]. They take two patches, one centered at a pixel $x = (i, j)$ on the left image and another one centered at pixel $y = (i, j - d)$ on the right image (with $d \in \{0, \ldots, n_d\}$), compute their corresponding feature vectors using a CNN, and then match them (block B of Fig. 1), to produce a similarity score $C(x, d)$, using either standard similarity metrics such as the L_1, the L_2, and the correlation metric, or metrics learned using a top network. The two components can be trained either separately or jointly.

4.1.1 The Basic Network Architecture

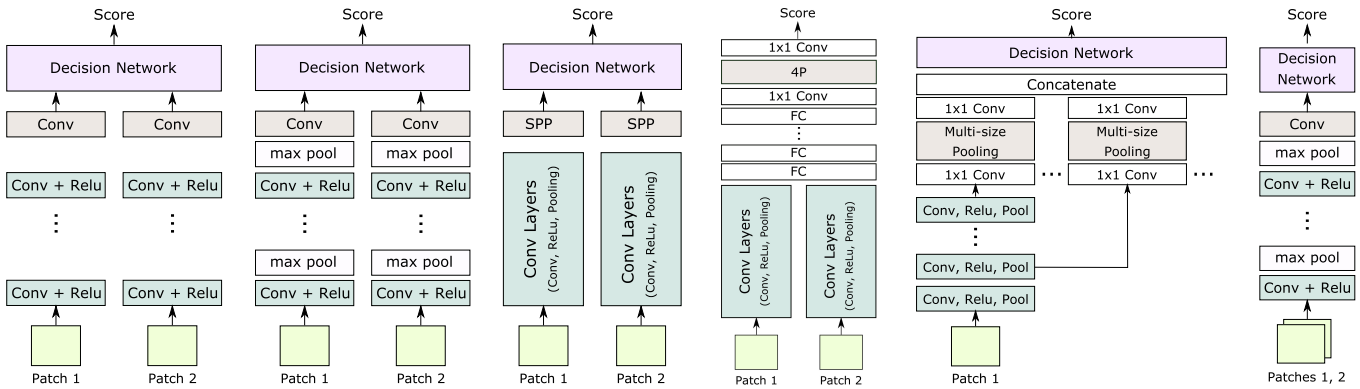
The basic network architecture, introduced in [37], [38], [39], [42] and shown in Fig. 2a, is composed of two CNN encoding branches, which act as descriptor computation modules. The first branch takes a patch around a pixel $x = (i, j)$ on the left image and outputs a feature vector that characterizes that patch. The second branch takes a patch around the pixel $y = (i, j - d)$, where $d \in [d_{min}, d_{max}]$ is a candidate disparity. Zbontar and LeCun [39] and later Zbontar *et al.* [42] use an encoder composed of four convolutional layers, see Fig. 2a. Each layer, except the last one, is followed by a ReLU unit. Zagoruyko and Komodakis [37] and Han *et al.* [38] use a similar architecture but add:

- max-pooling and subsampling after each layer, except the last one, see Fig. 2b. As such, the network is able to account for larger patch sizes and a larger variation in the viewpoint compared to [39], [42].
- a Spatial Pyramid Pooling (SPP) module at the end of each feature extraction branch [37] so that the network can process patches of arbitrary sizes while producing features of a fixed size, see Fig. 2c. Its role is to aggregate the features of the last convolutional layer, through spatial pooling, into a feature grid of a fixed size. The module is designed in such a way that the size of the pooling regions varies with the size of the input to ensure that the output feature grid has a fixed size independently of the size of the input patch or image. Thus, the network is able to process patches/images of arbitrary sizes and compute feature vectors of the same dimension without changing its structure or retraining.

The learned features are then fed to a top module, which returns a similarity score. It can be implemented as a standard similarity metric, e.g., the L_2 distance, the cosine distance, and the (normalized) correlation distance (or inner

TABLE 2
Taxonomy and Comparison of Deep Learning-Based Stereo Matching Techniques

Method	Year	Feature computation		Similarity	Training		Regularization
		Architectures	Dimension		Degree of supervision	Loss	
Zagoruyko [37]	2015	ConvNet	Multiscale	FCN	Supervised with positive/negative samples	Hinge and squared L_2	NA
Han [38]	2015	ConvNet	Fixed scale	FCN	Supervised	Cross-entropy	NA
Zbontar [39]	2015	ConvNet	Fixed scale	Hand-crafted	Triplet contrastive learning	L_1	MRF
Chen [40]	2015	ConvNet	Multiscale	Correlation + voting	Supervised with positive/negative samples	L_1	MRF
Simo [41]	2015	ConvNet	Fixed scale	L_2	Supervised with positive/negative samples	L_2	NA
Zbontar [42]	2016	ConvNet	Fixed scale	Hand-crafted, FCN	Supervised with known disparity	Hinge	Classic stereo
Balantas [43]	2016	ConvNet	Fixe scale	L_2	Supervised, triplet contrastive learning	Soft-Positive-Negative (Soft-PN)	–
Mayer [22]	2016	ConvNet	Fixed-scale	Hand-crafted	Supervised	–	Encoder-decoder
Luo [44]	2016	ConvNet	Fixed scale	Correlation	Supervised	Cross-entropy	MRF
Kumar [45]	2016	ConvNet	Fixed scale	ConvNet	Supervised, triplet contrastive learning	Maximise inter-class distance, minimize inter-class distance.	–
Shaked [46]	2017	Highway network with multilevel skip connections	Fixed scale	FCN	Supervised	Hinge+cross-entropy	Classic+4Conv+5FC
Hartmann [47]	2017	ConvNet	Fixed scale	ConvNet	Supervised	Croos-entropy	Encoder
Park [48]	2017	ConvNet	Fixed scale	1×1 Convs, ReLU, SPP	Supervised	–	NA
Ye [49]	2017	ConvNet Multisize pooling	Fixed scale	FCN (1×1 convs)	Supervised	L_1	SGM
Tulyakov [50]	2017	Generic - independent of the network architecture			Weakly supervised	MIL, Contrastive, Contrastive-DP	–



(a) MC-CNN [39], [42]. (b) [37] and [38]. (c) [37]. (d) LW-CNN [48]. (e) FED-D2DRR [49]. (f) [37].

Fig. 2. Feature learning and matching architectures. The basic architecture in (a) has been extended in many ways. First, by adding pooling and subsampling layers in (b), the network can process larger patches and thus allowing for larger variations in the viewpoints between the two patches. The architectures in (c), (d) and (e) add Spatial Pyramid Pooling (SPP) modules to process patches of arbitrary sizes while producing features of fixed size. The architectures in (c) and (e) place the SPP modules right after the feature computation module, which is computationally more efficient than placing them after the feature matching/decision module as in (d). The architecture in (f), which unifies feature learning and metric learning, is easier to train but is computationally more expensive at runtime.

product) as in the MC-CNN-fast (MC-CNN-fst) architecture of [39], [42]. The main advantage of the correlation over the L_2 distance is that it can be implemented using a layer of 2D [51] or 1D [22] convolutional operations, called *correlation layer*. A correlation layer does not require training since the filters are in fact the features computed by the second branch of the network. As such, correlation layers have been extensively used in the literature [22], [39], [41], [42], [44].

Instead of using hand-crafted similarity measures, recent works use a decision network composed of (1) fully-connected (FC) layers [37], [38], [42], [46], [49], which can be implemented as 1×1 convolutions, (2) fully convolutional layers [47], or (3) convolutional layers followed by fully-connected layers. The decision network is trained jointly with the feature extraction module to assess the similarity between two image patches. Han *et al.* [38] use a top network composed of three fully-connected layers followed by a softmax. Zagoruyko and Komodakis [37] use two linear fully connected layers (each with 512 hidden units) that are separated by a ReLU activation layer while the MC-CNN-acrt network of Zbontar *et al.* [42] use up to five fully-connected layers. In all cases, the features computed by the two branches of the feature encoding module are first concatenated and then fed to the top network. Hartmann *et al.* [47], on the other hand, aggregate the features coming from multiple patches using mean pooling before feeding them to a decision network. The main advantage of aggregation by pooling over concatenation is that the former can handle any arbitrary number of patches without changing the architecture of the network or re-training it. As such, it is suitable for computing multi-patch similarity.

Using a decision network instead of hand-crafted similarity measures enables learning, from data, the appropriate similarity measure instead of imposing one at the outset. It is more accurate than using a correlation layer but is significantly slower.

4.1.2 Network Architecture Variants

Since its introduction, the baseline architecture has been extended in several ways in order to: (1) improve training using residual networks (ResNet) [46], (2) enlarge the receptive field of the network without losing in resolution or in computation efficiency [48], [49], [52], (3) handling multi-scale features [37], [40], (4) reducing the number of forward passes [37], [44], and (5) easing the training procedure by learning similarity without explicitly learning features [37].

ConvNet versus ResNet: While Zbontar *et al.* [39], [42] and Han *et al.* [38] use standard convolutional layers in the feature extraction block, Shaked and Wolf [46] add residual blocks with multilevel weighted residual connections to facilitate the training of very deep networks. Its particularity is that the network learns by itself how to adjust the contribution of the added skip connections. It was demonstrated that this architecture outperforms the base network of Zbontar *et al.* [39].

Enlarging the receptive field of the network: The scale of the learned features is defined by (1) the size of the input patches, (2) the receptive field of the network, and (3) the kernel size of the convolutional filters and pooling operations used in each layer. While increasing the kernel sizes allows the capture of more global interactions between the image pixels, it induces a high computational cost. Also, the conventional pooling, as used in [39], [42], reduces resolution and could cause the loss of fine details, which is not suitable for dense correspondence estimation.

To enlarge the receptive field without losing resolution or increasing the computation time, some techniques, e.g., [52], use dilated convolutions, i.e., large convolutional filters but with holes and thus they are computationally efficient. Other techniques, e.g., [48], [49], use Spatial Pyramid Pooling (SPP) modules placed at different locations in the network, see Figs. 2c, 2d, and 2e. For instance, Park *et al.* [48], who introduced FW-CNN for stereo matching, append an SPP module at the end of the decision network, see Fig. 2d. As a result, the receptive field can be enlarged. However, for each pixel in the reference image, both the fully-connected layers and the pooling operations need to be computed n_d times where n_d is the number of disparity levels. To avoid this, Ye *et al.* [49] place the SPP module at the end of each feature computation branch, see Figs. 2c and 2e. In this way, it is only computed once for each patch. Also,

Ye *et al.* [49] employ multiple one-stride poolings, with different window sizes, to different layers and then concatenate their outputs to generate the feature maps, see Fig. 2e.

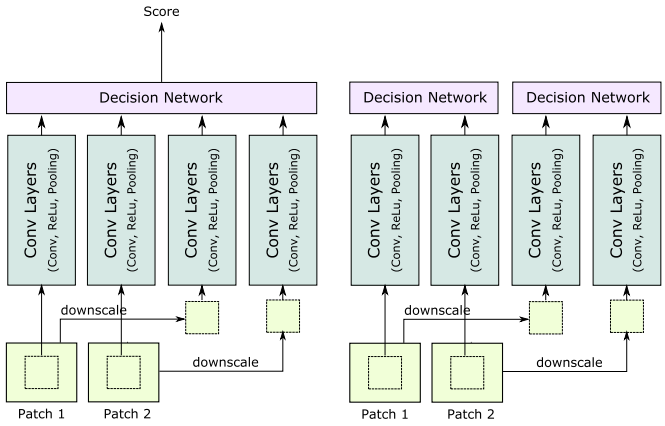
Learning multiscale features: The methods described so far can be extended to learn features at multiple scales by using multi-stream networks, one stream per patch size [37], [40], see Fig. 3. Zagoruyko and Komodakis [37] propose a two-stream network, which is essentially a network composed of two siamese networks combined at the output by a top network, see Fig. 3a. The first siamese network, called central high-resolution stream, receives as input two 32×32 patches centered around the pixel of interest. The second network, called surround low-resolution stream, receives as input two 64×64 patches but down-sampled to 32×32. The output of the two streams are then concatenated and fed to a top decision network, which returns a matching score. Chen *et al.* [40] use a similar approach but instead of aggregating the features computed by the two streams prior to feeding them to the top decision network, it appends a top network on each stream to produce a matching score. The two scores are then aggregated by voting, see Fig. 3b.

The main advantage of the multi-stream architecture is that it can compute features at multiple scales in a single forward pass. It, however, requires one stream per scale, which is not practical if more than two scales are needed.

Reducing the number of forward passes: Using the approaches described so far, inferring the raw cost volume from a pair of stereo images is performed using a moving window-like approach, which would require multiple forward passes, n_d forward passes per pixel where n_d is the number of disparity levels. However, since correlations are highly parallelizable, the number of forward passes can be significantly reduced. For instance, Luo *et al.* [44] reduce the number of forward passes to one pass per pixel by using a siamese network, whose first branch takes a patch around a pixel while the second branch takes a larger patch that expands over all possible disparities. The output is a single 64D representation for the left branch, and $n_d \times 64$ for the right branch. A correlation layer then computes a vector of length n_d, where its $d-$th element is the cost of matching the pixel x on the left image with the pixel $x - d$ on the rectified right image.

Zagoruyko and Komodakis [37] showed that the outputs of the two feature extraction sub-networks need to be computed only once per pixel, and do not need to be recomputed for every disparity under consideration. This can be done in a single forward pass, for the entire image, by propagating full-resolution images instead of small patches. Also, the output of the top network composed of fully-connected layers in the accurate architecture (i.e., MC-CNN-Accr) can be computed in a single forward pass by replacing the fully-connected layers with convolutional layers of 1×1 kernels. However, it still requires one forward pass for each disparity under consideration.

Learning similarity without feature learning: Joint training of feature extraction and similarity computation networks unifies the feature learning and the metric learning steps. Zagoruyko and Komodakis [37] propose another architecture that does not have a direct notion of features, see Fig. 2f. In this architecture, the left and right patches are packed together and fed jointly into a two-channel network composed of convolution and ReLU layers followed by a set of fully connected layers. Instead of computing features, the network directly outputs the similarity between the input pair of patches. Zagoruyko and Komodakis [37] showed that this architecture is easy to train. However, it is expensive at runtime since the whole network needs to be run n_d times per pixel.



(a) Central-surround [37] (b) Voting-based [40].

Fig. 3. Multiscale feature learning architectures.

4.1.3 Training Procedures

The networks described in this section are composed of a feature extraction block and a feature matching block. Since the goal is to learn how to match patches, these two modules are jointly trained either in a supervised (Section 4.1.3.1) or in a weakly supervised manner (Section 4.1.3.2).

Supervised training: Existing methods for supervised training use a training set composed of positive and negative examples. Each positive (respectively negative) example is a pair composed of a reference patch and its matching patch (respectively a non-matching one) from another image. Training either takes one example at a time, positive or negative, and adapts the similarity [37], [38], [40], [41], or takes at each step both a positive and a negative example, and maximizes the difference between the similarities, hence aiming at making the two patches from the positive pair *more similar* than the two patches from the negative pair [39], [43], [45]. This latter scheme is known as *Triplet Contrastive learning*.

Zbontar *et al.* [39], [42] use the ground-truth disparities of the KITTI2012 [15] or Middlebury [20] datasets. For each known disparity, the method extracts one negative pair and one positive pair as training examples. As such, the method is able to extract more than 25 million training samples from KITTI2012 [15] and more than 38 million from the Middlebury dataset [20]. This method has been also used by Chen *et al.* [40], Zagoruyku and Komodakis [37], and Han *et al.* [38]. The amount of training data can be further augmented by using data augmentation techniques, e.g., flipping patches and rotating them in various directions.

Although the supervised learning works very well, the complexity of the neural network models requires very large labeled training sets, which are hard or costly to collect for real applications (e.g., consider the stereo reconstruction

of the Mars landscape). Even when such large sets are available, the ground truth is usually produced from depth sensors and often contains noise that reduces the effectiveness of the supervised learning [53]. This can be mitigated by augmenting the training set with random perturbations [39] or synthetic data [22], [54]. However, synthesis procedures are hand-crafted and do not account for the regularities specific to the stereo system and target scene at hand.

Loss Functions. Supervised stereo matching networks are trained to minimize a matching loss, which is a function that measures the discrepancy between the ground-truth and the predicted matching scores for each training sample. It can be defined using (1) the L_1 distance [40], [42], [46], (2) the hinge loss [42], [46], or (3) the cross-entropy loss [44].

Weakly supervised learning: Weakly supervised techniques exploit one or more stereo constraints to reduce the amount of manual labelling. Tulyakov *et al.* [50] consider Multi-Instance Learning (MIL) in conjunction with stereo constraints and coarse information about the scene to train stereo matching networks with datasets for which ground truth is not available. Unlike supervised techniques, which require pairs of matching and non-matching patches, the training set is composed of N triplets. Each triplet is composed of: (1) W reference patches extracted on a horizontal line of the reference image, (2) W positive patches extracted from the corresponding horizontal line on the right image, and (3) W negative patches, i.e., patches that do not match the reference patches, extracted from another horizontal line on the right image. As such, the training set can automatically be constructed from stereo pairs without manual labelling.

The method is trained by exploiting five constraints: the epipolar constraint, the disparity range constraint, the uniqueness constraint, the continuity (smoothness) constraint, and the ordering constraint. They then define three losses that use different subsets of these constraints, namely:

- The Multi Instance Learning loss, which uses the epipolar and the disparity range constraints. From these two constraints, we know that every non-occluded reference patch has a matching positive patch in a known index interval, but does not have a matching negative patch. Therefore, for every reference patch, the similarity of the best reference-positive match should be greater than the similarity of the best reference-negative match.
- The constractive loss, which adds to the MIL method the uniqueness constraint. It tells us that the matching positive patch is unique. Thus, for every patch, the similarity of the best match should be greater than the similarity of the second best match.
- The constractive-DP uses all the constraints but finds the best match using Dynamic Programming (DP).

The method has been applied to train a deep siamese neural-network that takes two patches as an input and predicts a similarity measure. Benchmarking on standard datasets shows that the performance is as good as or better than the published results on MC-CNN-fst [39], which uses the same network architecture but trained using fully labeled data.

4.2 Regularization and Disparity Estimation

Once the raw cost volume is estimated, one can estimate the disparity by dropping the regularization term of Eqn. (1), or equivalently block C of Fig. 1, and taking the argmin, the softargmin, or the subpixel MAP approximation (block D of Fig. 1). However, the raw cost volume computed from image features could be noise-contaminated, e.g., due to the existence of non-Lambertian surfaces, object occlusions, or repetitive patterns. Thus, the estimated depth maps can be noisy. Several methods overcome this problem by using traditional MRF-based stereo framework for cost volume regularization [39], [40], [44]. In these methods, the initial cost volume C is fed to a global [11] or a semi-global [55] matcher to compute the disparity map. Semi-global matching provides a good tradeoff between accuracy and computation requirements. In this method, the smoothness term of Eqn. (1) is defined as

$$E_s(d_x, d_y) = \alpha_1 \delta(|d_{xy} = 1) + \alpha_2 \delta(|d_{xy} > 1), \tag{2}$$

where $d_{xy} = d_x - d_y$, α_1 and α_2 are positive weights chosen such that $\alpha_2 > \alpha_1$, and δ is the Kronecker delta function, which gives 1 when the condition in the bracket is satisfied, otherwise 0. To solve this optimisation problem, the SGM energy is broken down into multiple energies E_s, each one defined along a path s. The energies are minimised independently and then aggregated. The disparity at x is computed using the winner-takes-all strategy of the aggregated costs of all directions

$$d_x = \arg\min_d \sum_s E_s(x, d). \tag{3}$$

This method requires setting the two parameters α_1 and α_2 of Eqn. (2). Instead of manually setting them, Seki *et al.* [56] proposed SGM-Net, a neural network trained to provide these parameters at each image pixel. They obtained better penalties than hand-tuned methods as in [39].

The SGM method, which uses an aggregated scheme to combine costs from multiple 1D scanline optimizations, suffers from two major issues: (1) streaking artifacts caused by the scanline optimization approach, at the core of this algorithm, may lead to inaccurate results, and (2) the high memory footprint that may become prohibitive with high resolution images or devices with constrained resources. As such Schonberger *et al.* [57] reformulate the fusion step as the task of selecting the best amongst all the scanline optimization proposals at each pixel in the image. They solve this task using a per-pixel random forest classifier.

Poggi *et al.* [58] learn a weighted aggregation where the weight of each 1D scanline optimisation is defined using a confidence map computed using either traditional techniques [59] or deep neural networks, see Section 5.5.

5 End-to-End Depth From Stereo

Recent works solve the stereo matching problem using a pipeline that is trained end-to-end. Two main classes of methods have been proposed. Early methods, e.g., FlowNet-Simple [51] and DispNetS [22], use a single encoder-decoder, which stacks together the left and right images into a 6D volume, and regresses the disparity map. These

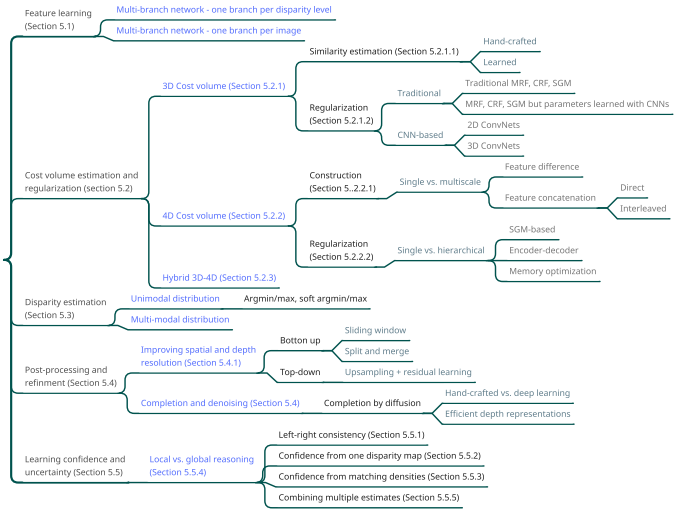


Fig. 4. Taxonomy of the network architectures for stereo-based disparity estimation using end-to-end deep learning.

methods, which do not require an explicit feature matching module, are fast at runtime. They, however, require a large amount of training data, which is hard to obtain. Methods in the second class mimic the traditional stereo matching pipeline by breaking the problem into stages, each stage is composed of differentiable blocks and thus allowing end-to-end training. Below, we review in details these techniques. Fig. 4 provides a taxonomy of the state-of-the-art architectures, while Table 3 compares 28 key methods based on this taxonomy.

5.1 Feature Learning

Feature learning networks follow the same architectures as the ones described in Figs. 2 and 3. However, instead of processing individual patches, the entire images are processed in a single forward pass producing feature maps of the same or lower resolution as the input images. Two strategies have been used to enable matching features across the images:

1) Multi-Branch Networks Composed of n Branches Where n is the Number of Input Images. Each branch produces a feature map that characterizes its input image [22], [60], [61], [62], [63], [64], [65]. These techniques assume that the input images have been rectified so that the search for correspodnences is restricted to be along the horizontal scanlines.

2) Multi-Branch Networks Composed of n_d Branches Where n_d is the Number of Disparity Levels. The dth branch, $1 \leq d \leq n_d$, processes a stack of two images, as in Fig. 2f; the first one is the reference. The second one is the right image but re-projected to the dth depth plane [66]. Each branch produces a feature map that characterizes the similarity between the reference image and the right image re-projected onto a given depth plane. While these techniques do not rectify the images, they assume that the intrinsic and extrinsic camera parameters are known. Also, the number of disparity levels cannot be varied without updating the network architecture and retraining it.

In both methods, the feature extraction module uses either fully convolutional (ConvNet) networks such as VGG, or residual networks such as ResNets [67]. The latter facilitates the training of very deep networks [68]. They can also capture and incorporate more global context in the unary features by using either dilated convolutions (Section 4.1.2.2) or multi-scale approaches. For instance, the PSM-Net of Chang and Chen [64] append a Spatial Pyramid Pooling module in order to extract and aggregate features at multiple scales. Chabra *et al.* [69] used Vortex Pooling [70], an extension of the Atrous Spatial Pyramid Pooling (ASPP) module [67], to aggregate the local and contextual information. Unlike ASPP modules, which treat the aggregated features equally, vortex pooling gives more attention to features near the central pixel, which usually provides more related semantic information. Nie *et al.* [65] extend PSM-Net using a multi-level context aggregation pattern termed *Multi-Level Context Ultra-Aggregation (MLCUA)*. It encapsulates all convolutional features into a more discriminative representation by intra and inter-level features combination. It combines the features at the shallowest, smallest scale with features at deeper, larger scales using just shallow skip connections. This results in an improved performance, compared to PSM-Net [64], without significantly increasing the number of parameters in the network.

5.2 Cost Volume Construction

Once the features have been computed, the next step is to compute the matching scores, which will be fed, in the form of a cost volume, to a top network for regularization and disparity estimation. The cost volume can be three dimensional (3D) where the third dimension is the disparity level (Section 5.2.1), four dimensional (4D) where the third dimension is the feature dimension and the fourth one is the disparity level (Section 5.2.2), or hybrid to benefit from the properties of the 3D and 4D cost volumes (Section 5.2.3). In general, the cost volume is constructed at a lower resolution, e.g., at $1/8$-th, than the input [72], [75]. It is then either subsequently upscaled and refined, or used as is to estimate a low resolution disparity map, which is then upscaled and refined using a refinement module.

5.2.1 3D Cost Volumes

Construction: A 3D cost volume can be built by taking the L_1, L_2, or correlation distance between the features of the left image and those of the right image that are within a predefined disparity range, see [22], [69], [72], [73], [75], [78], [82], [88], and the FlowNetCorr of [51]. Correlation-based dissimilarities can be implemented using a convolutional layer that does not require training (its filters are the features computed by the second branch of the network). Flow estimation networks such as FlowNetCorr [51] use 2D correlations. Disparity estimation networks, such as [22], [68], iResNet [63], DispNet3 [83], EdgeStereo [80], HD^3 [88], and [76], [82], use 1D correlations.

Regularization of 3D cost volumes: Once a cost volume is computed, an initial disparity map can be estimated using the argmin, the softargmin, or the subpixel MAP approximation over the depth dimension of the cost volume, see for example [72] and Fig. 5a. This is equivalent to dropping the regularization term of Eqn. (1). In general, however, the raw cost volume is noise-contaminated (e.g., due to the existence of non-Lambertian surfaces, object occlusions, and

TABLE 3
Taxonomy and Comparison, on the KITTI2015 Test Dataset, of 27 End-to-End Disparity Estimation Techniques

Method	Year	Feature computation		Cost volume			Disparity	Refinement/post processing		Supervision	Performance	
		Architecture	Dimension	Type	Construction	Regularization		Spatial/depth resolution	Completion/ denoising		D1-est	D1-fg
FlowNetCorr [51]	2015	ConvNet	Single scale	3D	Correlation	2D ConvNet	–	Up-convolutions	Ad-hoc, variational	Supervised	–	–
Zhang *et al.* [72] (Active stereo)	2018	ConvNet	Single scale	3D	Hand-crafted	NA	Soft argmin	Upsampling and residual learning		Self-supervised	–	–
Wang *et al.* [73] (stage 4)	2019	ConvNet	Multires. maps	3D	L_1	Progressive refinement (3D Conv)	Soft argmin	Upsampling, residual learning	Spatial propagation network	Supervised	6.2	–
Knobelreiter *et al.* [74]	2017	ConvNet	Single scale	3D	Correlation	Hybrid CNN-CRF		No post-processing		Supervised	5.50	–
Khamis *et al.* [75]	2018	ResNet	Single scale	3D	L_2	3D ConvNet	Soft argmin	Hierarchical, Upsamling and residual learning		Supervised	4.83	7.45
Tonioni *et al.* [76]	2019	ConvNet	Multiscale	3D	Correlation		Encoder	Recrusively upsampling and residual learning		Online self-adaptive	4.66	–
DispNetC [22]	2016	ConvNet	Single scale	3D	Correlation	2D ConvNet	–	–	–	Supervised	4.34	4.32
Zhong *et al.* [77]	2017	ConvNet with skip conn.	Single scale	4D	Interleaved feature concat.	3D Conv, encoder-decoder	Soft argmin	Self-improvement at runtime		Self-supervised	3.57	7.12
Jie *et al.* [78]	2018	Constant highway net	Single scale	3D	FCN	–	RNN-based LRCR	– –	– –	Supervised	3.03	5.42
Kendall *et al.* [61]	2017	ConvNet with skip conn.	Single scale	4D	Feature concat.	3D Conv, encoder-decoder, hierarchical	Soft argmax	–	–	Supervised	2.87	6.16
Yu *et al.* [79]	2018	ResNet	Single scale	3D	Feature concatenation Encoder-decoder	3D Conv + SGM	Soft argmin	–	–	Supervised	2.79	5.46
Pang *et al.* [62]	2017	ConvNet	Single scale	3D	Correlation	2D ConvNet		Upsampling and residual learning		Supervised	2.67	3.59
Liang *et al.* [63]	2018	ConvNet	Multiscale	3D	Correlation	2D ConvNet	Encoder-decoder	Iterative upsampling and residual learning		Supervised	2.67	3.59
Song *et al.* [80]	2018	Sallow ConvNet	Single scale	3D	Correlation Encoder	Edge-guided, Context Pyramid	Residual pyramid	–	–	Supervised	2.59	4.18
Tulyakov *et al.* [81]	2018	–	Single scale	4D	Compressed matching features	3D Conv	Multimodal - Sub-pixel MAP	–	–	Supervised	2.58	4.05
Duggal *et al.* [82]	2019	ResNet, SPP	Multiscale	3D, sparse	Correlation, Adaptive pruning with PatchMatch	Encoder-decoder	Soft argmax	Encoder	–	Supervised	2.35	**3.43**
Chang *et al.* [64]	2018	SPP	Multiscale	4D	Feature concat.	3D Conv, Stacked encoder-decoders	Soft argmin	Progressive refinement		Supervised	2.32	4.62
Chabra *et al.* [69]	2019	ConvNet + Vortex pooling	Multiscale	3D	L_1	Dilated 3D ConvNet	Soft argmax	Upsampling+residual learning		Supervised	2.26	4.95
Yang *et al.* [68]	2018	Shallow ResNet	Single scale	3D	Correlation, Left features, segmentation mask		Regression with Encoder-decoder	–	–	Self-supervised	2.25	4.07
Ilg *et al.* [83]	2018	ConvNet	Single scale	3D	Correlation	2D Conv, Encoder-decoder, joint disparity and occlusion		Cascade of encoder-decoders, residual learning		Supervised	2.19	–
Yang *et al.* [32]	2019	ConvNet, SPP	Multiscale	Pyramid, 4D	Feature difference	Conv3D, Volumetric Pyramid Pooling	Softargmax	No refinement or postprocessing		Supervised	2.14	3.85
Chen *et al.* [84]	2019						Single-modal weighted avg			Supervised	2.14	4.33
Guo *et al.* [85]	2019	SPP	Multiscale	Hybrid 3D-4D	Group-wise correlation	Stacked hourglass nets	Soft argmin	–	–	Supervised	2.11	3.93
Wu *et al.* [86]	2019	ResNet50, SPP	Multiscale	Pyramid 4D	Feature concat.	Encoder-decoder + Feature fusion	3D Conv, soft argmin	–	–	Supervised	2.11	3.89
EMCUA *et al.* [65]	2019	SPP	Multiscale	4D	Feature concat.	3D Conv, MCUA	Arg softmin	–	–	Supervised	2.09	4.27
Zhang *et al.* [87]	2019	Stacked hourglass	Single scale	4D	Concatenation	Semi-global aggregation layers, Local-guided aggregation layers	Soft argmax	–	–	Supervised	2.03	3.91
Yin *et al.* [88]	2019	DLA net	Multiscale	3D	Correlation	Density decoder	Discrete matching distribution	–	–	Supervised	**2.02**	3.63

"FCN": Fully-Connected Network, "SPN": Spatial Propagation Network. "LRCR": Left-Right Comparative Recurrent model, "MCUA": Multi-Level Context Ultra-Aggregation for Stereo Matching. "DLA": Deep layer aggregation [71], "VPP": Volumetric Pyramid Pooling, "D1-est" and "D1-fg": D1 error in all estimated pixels and on foreground pixels, respectively. D1 error is defined as the percentage of pixels with disparity error larger than 3 pixels.

repetitive patterns). The goal of the regularization module is to leverage context along the spatial and/or disparity dimensions to refine the cost volume before estimating the initial disparity map.

1) Regularization Using Traditional Methods. Early papers use traditional techniques, e.g., Markov Random Fields (MRF), Conditional Random Fields (CRF), and Semi-Global Matching (SGM), to regularize the cost volume by explicitly incorporating spatial constraints, e.g., smoothness, of the depth maps. Recent papers showed that deep learning networks can be used to fine-tune the parameters of these methods. For example, Knöbelreiter *et al.* [74] proposed a hybrid CNN-CRF. The CNN computes the matching term of Eqn. (1), which becomes the unary term of a CRF module. The pairwise term of the CRF is parameterized by edge weights computed using another CNN. The end-to-end trained CNN-CRF pipeline could achieve a competitive performance using much fewer parameters (thus a better utilization of the training data) than the earlier methods.

Zheng *et al.* [89] provide a way to model CRFs as Recurrent Neural Networks (RNN) for segmentation tasks so that the entire pipeline can be trained end-to-end. Unlike segmentation, in depth estimation, the number of depth samples, whose counterparts are the semantic labels in segmentation tasks, is expected to vary for different scenarios. As such, Xue *et al.* [90] re-designed the RNN-formed CRF module so that the model parameters are independent of the number of depth samples. Paschalidou *et al.* [91]

formulate the inference in a MRF as a differentiable function, hence allowing end-to-end training using back propagation. Note that Zheng *et al.* [89] and Paschalidou *et al.* [91] focus on multi-view stereo (Section 6). Their approaches, however, are generic and can be used to regularize 3D cost volumes obtained using pairwise stereo networks.

2) Regularization Using 2D Convolutions (2DConvNet), Figs. 5b and 5c. Another approach is to process the 3D cost volume using a series of 2D convolutional layers producing another 3D cost volume [22], [51], [62], [63]. 2D convolutions are computationally efficient. However, they only capture and aggregate context along the spatial dimensions, see Fig. 5b, and ignore context along the disparity dimension. Yao *et al.* [92] sequentially regularize the 2D cost maps along the depth direction via a Gated Recurrent Unit (GRU), see Fig. 5c. This reduces drastically the memory consumption, e.g., from 15.4 GB in [93] to around 5 GB, making high-resolution reconstruction feasible, while capturing context along both the spatial and the disparity dimensions.

3) Regularization Using 3D Convolutions (3DConvNet), Fig. 5d. Khamis *et al.* [75] use the L_2 distance to compute an initial 3D cost volume and 3D convolutions to regularize it across both the spatial and disparity dimensions, see Fig. 5d. Due to its memory requirements, the approach first estimates a low-resolution disparity map, which is then progressively improved using residual learning. Zhang *et al.* [72] follow the same approach but the refinement block starts with separate convolution layers running on the upsampled disparity and input image respectively, and merge the features later to produce the residual. Chabra *et al.* [69] observe that the cost volume regularization step is the one that uses most of the computational resources. They propose a regularization module that uses 3D dilated convolutions in the width, height, and disparity dimensions, to reduce the computation time while capturing a wider context.

Wang *et al.* [73], on the other hand, build a 3D cost volume, and regularizes it in a coarse-to-fine manner. Their network takes a low resolution feature map, builds a low resolution cost volume and then uses 3D convolutions to estimate a low resolution disparity map by searching on a small disparity range. Each subsequent disparity estimation block takes the disparity predicted by the previous block, but upsampled to a higher resolution, and the learned features at a higher resolution, and then estimates the disparity residuals. The advantage is two-fold; *first*, at higher resolutions, the network only learns to predict residuals, which reduces the computation cost. *Second*, the approach is progressive and one can select to return the intermediate disparities, trading accuracy for speed.

5.2.2 4D Cost Volumes

Construction: 4D cost volumes preserve the dimension of the features [32], [61], [64], [65], [77], [86]. The rational behind 4D cost volumes is to let the top network learn the appropriate similarity measure for comparing the features instead of using hand-crafted ones as in Section 5.2.1.

4D cost volumes can be constructed by feature differences across a pre-defined disparity range [32], which results in cost volume of size $H \times W \times 2n_d \times c$, or by concatenating the features computed by the different branches of the network [61], [64], [65], [77], [86]. Using this method, Kendall *et al.* [61] build a 4D volume of size $H \times W \times (n_d + 1) \times c$ (c here is the dimension of the features). Zhong *et al.* [77] follow the same approach but concatenate the features in an interleaved manner. That is, if $\mathbf{f}_L$ is the feature map of the left image and $\mathbf{f}_R$ the feature map of the right image, then the final feature volume is assembled in such a way that its $2i-$th slice holds the left feature map while the $(2i+1)-$th slice holds the right feature map but at disparity $d = i$. This results in a 4D cost volume that is twice larger than the cost volume of Kendall *et al.* [61]. To capture multi-scale context in the cost volume, Chang and Chen [64] generate for each input image a pyramid of features, upsamples them to the same dimension, and then builds a single 4D cost volume by concatenation. Wu *et al.* [86] build from the multiscale features (four scales) multiscale 4D cost volumes.

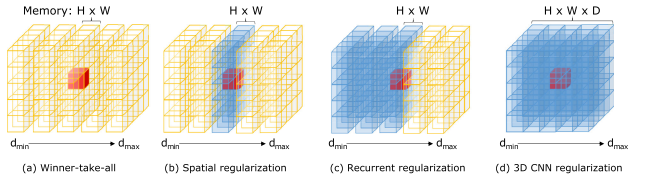


Fig. 5. Cost volume regularization schemes [92]: (a) does not consider context, (b) captures context along the spatial dimensions using 2D convolutions, (c) captures context along the spatial and disparity dimensions by recurrent regularization using 2D convolutions, and (d) captures context in all dimensions by using 3D convolutions.

4D cost volumes carry richer information compared to 3D cost volumes. Note, however, that volumes obtained by concatenation contain no information about the feature similarities, so more parameters are required in the subsequent modules to learn the similarity function.

Regularization of 4D cost volumes: 4D cost volumes are regularized with 3D convolutions, which exploit the correlation in height, width and disparity dimensions, to produce a 3D cost volume. Kendall *et al.* [61] use a U-net encoder-decoder with 3D convolutions and skip connections. Zhong *et al.* [77] use a similar approach but add residual connections from the contracting to the expanding parts of the regularization network. To take into account a large context without a significant additional computational burden, Kendall *et al.* [61] regularize the cost volume hierarchically, with four levels of subsampling, allowing to explicitly leverage context with a wide field of view. Muliscale 4D cost volumes [86] are aggregated into a single 3D cost volume using a 3D multi-cost aggregation module, which operates in a pairwise manner starting with the smallest volume. Each volume is processed with an encoder-decoder, upsampled to the next resolution in the pyramid, and then fused using a 3D feature fusion module.

Also, semi-global matching techniques have been used to regularize 4D cost volumes where their parameters are estimated using convolutional networks. In particular, Yu *et al.* [79] process the initial 4D cost volume with an encoder-decoder composed of 3D convolutions and upconvolutions, and produces another 3D cost volume. The subsequent aggregation step is performed using an end-to-end two-stream network: the *first* stream generates three cost aggregation proposals C_i, one along each of the tree dimensions, i.e., the height, width, and disparity. The *second* stream is a

guidance stream used to select the best proposals. It uses 2D convolutions to produce three guidance (confidence) maps W_i. The final 3D cost volume is produced as a weighted sum of the three proposals, i.e., $\max_i(C_i * W_i)$.

3D convolutions are expensive in terms of memory requirements and computation time. As such, subsequent works that followed the seminal work of Kendall *et al.* [61] focused on (1) reducing the number of 3D convolutional layers [87], (2) progressively refining the cost volume and the disparity map [64], and (3) compressing the 4D cost volume [81]. Below, we discuss these approaches.

1) Reducing the Number of 3D Convolutional Layers. Zhang *et al.* [87] introduced GANet, which replaces a large number of the 3D convolutional layers in the regularization block with (1) two 3D convolutional layers, (2) a semi-global aggregation layer (SGA), and (3) a local guided aggregation layer (LGA). SGA is a differentiable approximation of the semi-global matching. Unlike SGM, in SGA the user-defined parameters are learnable. Moreover, they are added as penalty coefficients/weights of the matching cost terms. Thus, they are adaptive and more flexible at different locations for different situations. The LGA layer, on the other hand, is appended at the end and aims to refine the thin structures and object edges. The SGA and LGA layers, which are used to replace the costly 3D convolutions, capture local and whole-image cost dependencies. They significantly improve the accuracy of the disparity estimation in challenging regions such as occlusions, large textureless/reflective regions, and thin structures.

2) Progressive Approaches. Some techniques avoid directly regularizing high resolution 4D cost volumes using the expensive 3D convolutions. Instead, they operate in a progressive manner. For instance, Chang and Chen [64] introduced PSM-Net, which first estimates a low resolution 4D cost volume, and then regularizes it using stacked hourglass 3D encoder-decoder blocks. Each block returns a 3D cost volume, which is then upsampled and used to regress a high resolution disparity map using additional 3D convolutional layers followed by a softmax operator. As such, the stacked hourglass blocks can be seen as refinement modules.

3) 4D Cost Volume Compression. Tulyakov *et al.* [81] reduce the memory usage, without having to sacrify accuracy, by compressing the features into compact matching signatures. As such, the memory footprint is significantly reduced. More importantly, it allows the network to handle an arbitrary number of multiview images and to vary the number of inputs at runtime without having to re-train the network.

5.2.3 Hybrid 3D-4D Cost Volumes

The correlation layer provides an efficient way to measure feature similarities, but it loses much information because it produces only a single-channel map for each disparity level. On the other hand, 4D cost volumes obtained by feature concatenation carry more information but are resource-demanding. They also require more parameters in the subsequent aggregation network to learn the similarity function. To benefit from both, Guo *et al.* [85] propose a hybrid approach, which constructs two cost volumes; one by feature concatenation but compressed into 12 channels using two convolutions. The second one is built by dividing the high-dimension feature maps into N_g groups along the feature channel, computing correlations within each group at all disparity levels, and finally concatenating the correlation maps forming another 4D volume. The two volumes are then combined together and passed to a 3D regularization module composed of four 3D convolutional layers followed by three stacked 3D hourglass networks. This approach results in a significant reduction of parameters compared to 4D cost volumes built by only feature concatenation, without losing too much information like full correlations.

5.3 Disparity Computation

The simplest way to estimate the disparity map from the regularized cost volume C is by using the pixel-wise argmin, i.e., $d_x = \arg\min_d C(x,d)$ (or equivalently $\arg\max$ if the volume C encodes the likelihood). However, the argmin/argmax operator is unable to produce sub-pixel accuracy and cannot be trained with back-propagation due to its non-differentiability. Another approach is the differentiable soft argmin/max over disparity [61], [66], [72], [75]

$$d^* = \frac{1}{\sum_{j=0}^{n_d} e^{-C(x,j)}} \sum_{d=0}^{n_d} d \times e^{-C(x,d)}. \tag{4}$$

The soft argmin operator approximates the sub-pixel MAP solution when the distribution is unimodal and symmetric [81]. When this assumption is not fulfilled, the softargmin blends the modes and may produce a solution that is far from all the modes and may result in over smoothing. Chen *et al.* [84] observe that this is particularly the case at boundary pixels where the estimated disparities follow multimodal distributions. To address these issues, Chen *et al.* [84] only apply a weighted average operation on a window centered around the modal with the maximum probability, instead of using a full-band weighted average on the entire disparity range.

Tulyakov *et al.* [81] introduced the sub-pixel MAP approximation, which computes a weighted mean around the disparity with the maximum posterior probability as

$$d^* = \sum_{d:|\hat{d}-d|\leq\delta} d \cdot \sigma(C(x,d)), \tag{5}$$

where δ is a meta parameter set to 4 in [81], $\sigma(C(x,d))$ is the probability of the pixel x having a disparity d, and $\hat{d} = \arg\max_d C(x,d)$. The sub-pixel MAP is only used for inference. Tulyakov *et al.* [81] also showed that, unlike the softargmin/max, this approach allows changing the disparity range at runtime without re-training the network.

5.4 Variants

The pipeline described so far infers disparity maps that can be of low-resolution (along the width, height, and disparity dimensions), incomplete, noisy, missing fine details, and suffering from over-smoothing especially at object boundaries. As such, many variants have been introduced to (1) improve their resolution (Section 5.4.1), (2) improve the

processing time, especially at runtime (Section 5.4.3), and (3) perform disparity completion and denoising (Section 5.4.2).

5.4.1 Learning to Infer High Resolution Disparity Maps

Directly regressing high-resolution depth maps that contain fine details, e.g., by adding further upconvolutional layers to upscale the cost volume, would require a large number of parameters and thus are computationally expensive and difficult to train. As such, state-of-the-art methods struggle to process high resolution imagery because of memory constraints or speed limitations. This has been addressed by using either bottom-up or top-down techniques.

Bottom-up techniques operate in a sliding window-like approach. They take small patches and estimate the refined disparity either for the entire patch or for the pixel at the center of the patch. Lee *et al.* [94] follow a split-and-merge approach. The input image is split into regions, and a depth is estimated for each region. The estimates are then merged using a fusion network, which operates in the Fourier domain so that depth maps with different cropping ratios can be handled. While both sliding window and split-and-merge approaches reduce memory requirements, they require multiple forward passes, and thus are not suitable for realtime applications. Also, these methods do not capture the global context, which can limit their performance.

Top-down techniques, on the other hand, operate on the disparity map estimates in a hierarchical manner. They first estimate a low-resolution disparity map and then upsample them to the desired resolution, e.g., using bilinear upsampling, and further process them using residual learning to recover small details and thin structures [69], [72], [75]. This process can also be run progressively by cascading many of such refinement blocks, each block refines the estimate of the previous block [62], [75]. Unlike upsampling cost volumes, refining disparity maps is computationally efficient since it only requires 2D convolutions. Existing methods mainly differ in the type of additional information that is appended to the upsampled disparity map for refinement. For instance:

- Khamis *et al.* [75] concatenate the upsampled disparity map with the original reference image.
- Liang *et al.* [63] append to the initial disparity map the cost volume and the reconstruction error, defined as the difference between the left image and the right image but warped to the left image using the estimated disparity map.
- Chabra *et al.* [69] take the left image and the reconstruction error on one side, and the left disparity and the geometric error map, defined as the difference between the estimated left disparity and right disparity but warped onto the left view. These are independently filtered using one layer of convolutions followed by batch normalization. The results of the two streams are concatenated and then further processed using a series of convolutional layers to produce the refined disparity map.

These methods improve the spatial resolution but not the disparity resolution. To refine both the spatial and depth resolution, while operating on high resolution images, Yang *et al.* [32] propose to search for correspondences incrementally over a coarse-to-fine hierarchy. The approach constructs a pyramid of four 4D cost volumes, each with increasing spatial and depth resolutions. Each volume is filtered by six 3D convolution blocks, and further processed with a Volumetric Pyramid Pooling block, an extension of Spatial Pyramid Pooling to feature volumes, to generate features that capture sufficient global context for high resolution inputs. The output is then either (1) processed with another conv3D block to generate a 3D cost volume from which disparity can be directly regressed. This allows to report on-demand disparities computed from the current scale, or (2) tri-linearly-upsampled to a higher spatial and disparity resolution so that it can be fused with the next 4D volume in the pyramid. To minimise memory requirements, the approach uses striding along the disparity dimensions in the last and second last volumes of the pyramid. The network is trained end-to-end using a multi-scale loss. This hierarchical design also allows for anytime on-demand reports of disparity by capping intermediate coarse results, allowing accurate predictions for near-range structures with low latency (30 ms).

This approach shares some similarities with the approach of Kendall *et al.* [61], which constructs hierarchical 4D feature volumes and processes them from coarse to fine using 3D convolutions. Kendall *et al.*'s approach [61], however, has been used to leverage context with a wide field of view while Yang *et al.* [32] apply coarse-to-fine principles for high-resolution inputs and anytime, on-demand processing.

5.4.2 Learning for Completion and Denoising

Raw disparities can be noisy and incomplete, especially near object boundaries where depth smearing between objects remains a challenge. Several techniques have been developed for denoising and completion. Some of them are ad-hoc, i.e., post-process the noisy and uncomplete initial estimates to generate clean and complete depth maps. Other methods addressed the issue of the lack of training data for completion and denoising. Others proposed novel depth representations that are more suitable for this task, especially for solving the depth smearing between objects.

Ad-hoc methods process the initially estimated disparities using variational approaches [51], [95], Fully-Connected CRFs (DenseCRF) [27], [96], hierarchical CRFs [2], and diffusion processes [40] guided by confidence maps [97]. They encourage pixels that are spatially close and with similar colors to have closer disparity predictions. They have been also explored by Liu *et al.* [5]. However, unlike Li *et al.* [2], Liu *et al.* [5] used a CNN to minimize the CRF energy. Convolutional Spatial Propagation Networks (CSPN) [98], [99], which implement an anisotropic diffusion process, are particularly suitable for depth completion since they predict the diffusion tensor using a deep CNN. This is then applied to the initial map to obtain the refined one.

One of the main challenges of deep learning-based depth completion and denoising is the lack of labelled training data, i.e., pairs of noisy, incomplete depth maps and their corresponding clean depth maps. To address this issue, Jeon and Lee [29] propose a pairwise depth image dataset

generation method using dense 3D surface reconstruction with a filtering method to remove low quality pairs. They also present a multi-scale Laplacian pyramid based neural network and structure preserving loss functions to progressively reduce the noise and holes from coarse to fine scales. The approach first predicts the clean complete depth image at the coarsest scale, which has a quarter of the original resolution. The predicted depth map is then progressively upsampled through the pyramid to predict the half and original-sized image. At the coarse level, the approach captures global context while at finer scales it captures local information. In addition, the features extracted during the downsampling are passed to the upsampling pyramid with skip connections to prevent the loss of the original details in the input depth image during the upsampling.

Instead of operating on the network architecture, the loss function, or the training datasets, Imran *et al.* [100] propose a new representation for depth called Depth Coefficients (DC) to address the problem of depth smearing between objects. The representation enables convolutions to more easily avoid inter-object depth mixing. The representation uses a multi-channel image of the same size as the target depth map, with each channel representing a fixed depth. The depth values increase in even steps of size b. (The approach uses 80 bins.) The choice of the number of bins trades-off memory versus precision. The vector composed of all these values at a given pixel defines the depth coefficients for that pixel. For each pixel, these coefficients are constrained to be non-negative and sum to 1. This representation of depth provides a much simpler way for CNNs to avoid depth mixing. First, CNNs can learn to avoid mixing depths in different channels as needed. Second, since convolutions apply to all channels simultaneously, depth dependencies, like occlusion effects, can be modelled and learned by neural networks. The main limitation, however, is that the depth range needs to be set in advance and cannot be changed at runtime without re-training the network. Imran *et al.* [100] also show that the standard Mean Squared Error (MSE) loss function can promote depth mixing, and thus propose to use cross-entropy loss for estimating the depth coefficients.

5.4.3 Learning for Realtime Processing

The goal is to design efficient stereo algorithms that not only produce reliable and accurate estimations, but also run in realtime. For instance, in the PSMNet [64], the cost volume construction and aggregation takes more than 250ms (on nNvidia Titan-Xp GPU). This renders realtime applications infeasible. To speed the process, Khamis *et al.* [75] first estimate a low resolution disparity map and then hierarchically refine it. Yin *et al.* [88] employ a fixed, coarse-to-fine procedure to iteratively find the match. Chabra *et al.* [69] use 3D dilated convolutions in the width, height, and disparity channels when filtering the cost volume. This halves the computational cost in comparison to state of the art cost filtering architectures. Duggal *et al.* [82] combine deep learning with PatchMatch [101] to adaptively prune out the potentially large search space and significantly speed up inference. PatchMatch-based pruner module is able to predict a confidence range for each pixel, and construct a sparse cost volume that requires significantly less operations. This also allows the model to focus only on regions with high likelihood and save computation and memory. To enable end-to-end training, Duggal *et al.* [82] unroll PatchMatch as an RNN where each unrolling step is equivalent to an iteration of the algorithm. This approach achieved a performance that is comparable to the state-of-the-art, e.g., [64], [68], while reducing the computation time from 600 ms for [68] to 60 ms per image in the SceneFlow dataset, as reported in [82].

5.5 Learning Confidence Maps

The ability to detect, and subsequently remedy to, failure cases is important for applications such as autonomous driving and medical imaging. Thus, a lot of research has been dedicated to estimating confidence or uncertainty maps, which are then used to sparsify the estimated disparities by removing potential errors and then replacing them from the reliable neighboring pixels. Disparity maps can also be incorporated in a disparity refinement pipeline to guide the refinement process [78], [102], [103]. Seki *et al.* [102], for example, incorporate the confidence map into a Semi-Global Matching module for dense disparity estimation. Gidaris *et al.* [103] use confidence maps to detect the incorrect estimates, replace them with disparities from neighbouring regions, and then refine the disparity using a refinement network. Jie *et al.* [78], on the other hand, estimate two confidence maps, one for each of the input images, concatenate them with their associated cost volumes, and use them as input to a 3D convolutional LSTM to selectively focus in the subsequent step on the left-right mismatched regions.

Conventional confidence estimation methods are mostly based on assumptions and heuristics on the matching cost volume analysis, see [59] for a review and evaluation of the early methods. Recent techniques are based on supervised learning [104], [105], [106], [107], [108], [109]. They estimate confidence maps directly from the disparity space either in an ad-hoc manner, or in an integrated fashion so that they can be trained end-to-end along with the disparity/depth estimation. Poggi *et al.* [110] provide a quantitative evaluation. Below, we discuss some of these techniques.

5.5.1 Confidence From Left-Right Consistency Check

Left-right consistency is one of the most commonly-used criteria for measuring confidence in disparity estimates. The idea is to estimate two disparity maps, one from the left image (D_{left}), and another from the right image (D_{right}). An error map can then be computed by taking a pixel-wise difference between D_{left} and D_{right}, but warped back onto the left image, and converting them into probabilities [63]. This measure is suitable for detecting occlusions, i.e., regions that are visible in one view but not in the other.

Left-right consistency can also be learned using deep or shallow networks composed of fully convolutional layers [78], [102]. Seki *et al.* [102] propose a patch-based confidence prediction (PBCP) network, which requires two disparity maps, one estimated from the left image and the other one from the right image. PBCP uses a two-channel network. The first channel enforces left-right consistency while the

second one enforces local consistency. The network is trained in a classifier manner. It outputs a label per pixel indicating whether the estimated disparity is correct.

Instead of treating left-right consistency check as an isolated post-processing step, Jie *et al.* [78] perform it jointly with disparity estimation, using a Left-Right Comparative Recurrent (LRCR) model. It consists of two parallel convolutional LSTM networks [111], which produce two error maps; one for the left disparity and another for the right disparity. The two error maps are then concatenated with their associated cost volumes and used as input to a 3D convolutional LSTM to selectively focus in the next step on the left-right mismatched regions.

5.5.2 Confidence From a Single Raw Disparity Map

Left-right consistency checks estimate two disparity maps and thus are expensive at runtime. Shaked and Wolf [46] train, via the binary cross entropy loss, a network, composed of two fully-connected layers, to predict the correctness of an estimated disparity from only the reference image. Poggi and Mattoccia [107] pose the confidence estimation as a regression problem and solve it using a CNN trained on small patches. For each pixel, the approach extracts a square patch around the pixel and forwards it to a CNN trained to distinguish between patterns corresponding to correct and erroneous disparity assignments. It is a single channel network, designed for 9×9 image patches. Zhang *et al.* [72] use a similar confidence map estimation network, called *invalidation network*. The key idea is to train the network to predict confidence using a pixel-wise error between the left disparity and the right disparity. At runtime, the network only requires the left disparity. Finally, Poggi and Mattoccia [112] show that one can improve the confidence maps estimated using previous algorithms by enforcing local consistency in the confidence estimates.

5.5.3 Confidence Map From Matching Densities

Traditional deep networks represent activations and outputs as deterministic point estimates. Gast and Roth [113] explore the possibility of replacing the deterministic outputs by probabilistic output layers. To go one step further, they replace all intermediate activations by distributions. As such, the network can be used to estimate the matching probability densities, hereinafter referred to as *matching densities*, which can then be converted into uncertainties (or confidence) at runtime. The main challenge of estimating matching densities is the computation time. To make it tractable, Gast and Roth [113] assume parametric distributions. Yin *et al.* [88] relax this assumption and propose a pyramidal architecture to make the computation cost sustainable and allow for the estimation of confidence at run time.

5.5.4 Local versus Global Reasoning

Some techniques, e.g., Seki *et al.* [102]'s, reason locally by enforcing local consistency. Tosi *et al.* [114] introduced LGC-Net to move beyond local reasoning. The input reference image and its disparity map are forwarded to a local network, e.g., C-CNN [107], and a global network, e.g., an encoder/decoder architecture with a large receptive field.

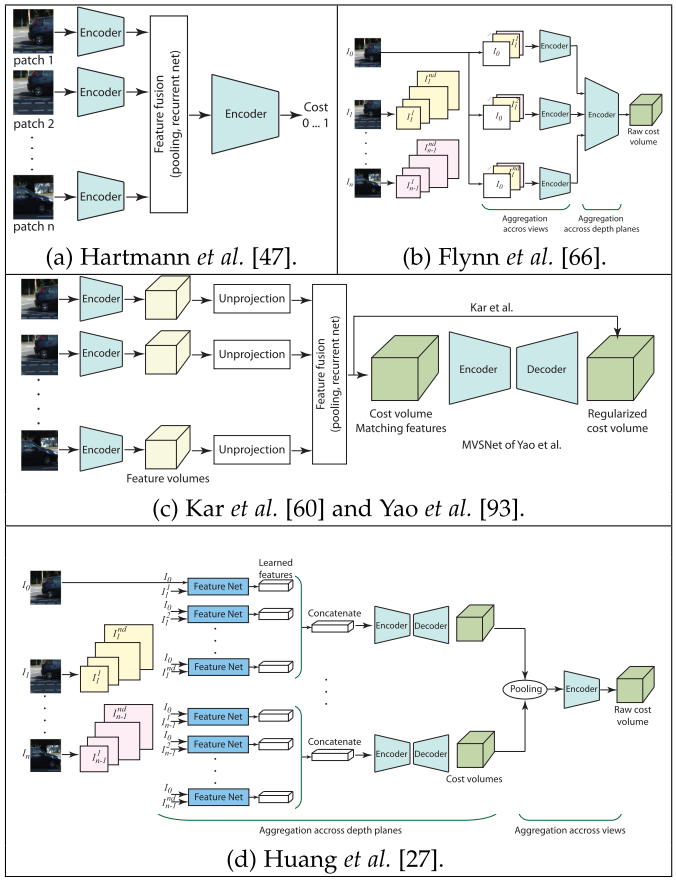


Fig. 6. Taxonomy of multivew stereo methods. (a), (b), and (c) perform early fusion, while (d) performs early fusion by aggregating features across depth plans, and late fusion by aggregating cost volumes across views.

The output of the two networks and the initial disparity, concatenated with the reference image, are further processed with three independent convolutional towers whose outputs are concatenated and processed with three 1×1 convolutional layers to finally infer the confidence map.

5.5.5 Combining Multiple Estimators

Some papers combine the estimates of multiple algorithms to achieve a better accuracy. Haeusler *et al.* [104] fed a random forest with a pool of 23 confidence maps, estimated using conventional techniques, yielding a much better accuracy compared to any confidence map in the pool. Batsos *et al.* [109] followed a similar idea but combine the strengths and mitigate the weaknesses of four basic stereo matchers in order to generate a robust matching volume for the subsequent optimization and regularization steps. Poggi and Mattoccia [58] train an ensemble regression trees classifier. These methods are independent of the disparity estimation module, and rely on the availability of the cost volume.

6 Learning Multiview Stereo

Multiview Stereo methods follow the same pipeline as of depth-from-stereo. Early works focused on computing the similarity between multiple patches. For instance, Hartmann *et al.* [47] (Fig. 6a) replace the pairwise correlation layer used in stereo matching by an average pooling layer

TABLE 4
Taxonomy and Comparison of 13 Deep Learning-Based MVS Techniques

Method	Year	Representation	Fusion	Training	Peformance on (DTU, SUN3D, ETH3D)				Complexity			
					#images	Error (mm)	% $< 1mm$	% $< 2mm$	#Params	Memory	Complexity	Time
Kar *et al.* [60]	2017	Volumetric	Recurrent fusion of 3D feature grids	Supervised	Variable	–	–	–	–	–	–	–
Hartmann *et al.* [47]	2017	Replace correlation by pooling		Supervised	5 (can vary)	$(1.356, -, -)$	–	–	–	–	–	–
Ji *et al.* [116]	2017	Volumetric	Reconstructed surfaces	Supervised	5	$(0.745, -, -)$	69.95	74.4	–	–	–	4 hrs
Choi *et al.* [117]	2018	Volumetric	Pairwise cost volumes	Supervised	5	$(0.6511, -, -)$	–	–	–	–	–	–
Huang *et al.* [27]	2018	PSV	Encoder-decoder for intra-volume, Max pooling for inter-volume	Supervised	Variable	$(-, 0.419, 0.412)$	–	–	–	–	–	–
Leroy *et al.* [118]	2018	PSV	Depth fusion	Supervised	Variable	$(0.599, -, -)$	–	–	72K	–	–	–
Paschalidou *et al.* [91]	2018	Depth-based	Avg. pooling over pairwsie correlations	Supervised	Variable	$(-, -, -)$	–	–	–	7GB	–	25 mins
Yao *et al.* [93]	2018	PSV	Feature pooling by variance	Supervised	5	$(0.462, 0.397, 0.470)$	75.69	80.25	363K	5.28GB	$O(HWn_d)$	0.9s
Wang *et al.* [120]	2018	PSV and abs. difference	Concatenation of pairwise cost volumes and ref. image	Supervised	Variable	$(-, 0.114, 0.257)$	–	–	33.9M for $n_d = 64$	–	–	0.04s
Hou *et al.* [115]	2019	–	Temporal fusion of the latent rep.	Supervised	Variable (video sequence)	$(-, \mathbf{0.101}, 0.229)$	–	–	–	–	–	–
Luo *et al.* [119]	2019	PSV	Feature pooling by variance	Supervised	Variable	$(0.406, -, -)$	–	–	–	–	–	–
Xue *et al.* [90]	2019	PSV	Cost volume pooling by variance	Supervised	5 (can vary)	$(\mathbf{0.398}, -, -)$	80.02	83.84	571K	5.43GB	$O(HWn_d)$	1.8s
Won *et al.* [30]	2019	Spherical PSV	Concatenation	Supervised	–	–	–	–	–	–	–	–

to aggregate the learned features of $n \geq 2$ input patches, and then feed the output to a top network, which returns a matching score. With this method, computing the best match for a pixel on the reference image requires n_d^{n-1} forward passes. (n_d is the number of depth levels and n is the number of images.) This is computationally very expensive especially when dealing with high resolution images.

Techniques that compute depth maps in a single forward pass differ in the way the information from the multiple views is fed to the network and aggregated. We classify them into whether they are volumetric (Section 6.1) or Plane-Sweep Volume (PSV)-based (Section 6.2). The latter does not rely on intermediate volumetric representations of the 3D geometry. The only exception is the approach of Hou *et al.* [115], which performs temporal fusion of the latent representations of the input images. The approach, however, requires temporally-ordered images. Table 4 provides a taxonomy and compares 13 state-of-the-art MVS techniques.

6.1 Volumetric Representations

One of the main issues for MVS reconstruction is how to match, in an efficient way, features across multiple images. Pairwise stereo methods rectify the images so that the search for correspondences is restricted to the horizontal epipolar lines. This is not possible with MVS due to the large view angle differences between the images. This has bee addressed using volumetric representations of the scene geometry [60], [116]. Depth maps are then generated by projection from the desired viewpoint. For a given input image, with known camera parameters, a ray from the viewpoint is cast through each image pixel. The voxels intersected by that ray are assigned the color [116] or the learned feature [60] of that pixel. Existing methods differ in the way information from multiple views are fused:

1) Fusing Feature Grids. Kar *et al.* [60] (Fig. 6c) fuse, recursively, the back-projected 3D feature grids using a recurrent neural network. The produced 3D grid is regularized using an encoder-decoder. To avoid dependency on the order of the images, Kar *et al.* [60] randomly permute the input images during training while constraining the output to be the same.

2) Fusing Pairwise Cost Volumes. Choi *et al.* [117] fuse the cost volumes, computed from each pair of images, using a weighted sum where the weight of each volume is the confidence map computed from that cost volume.

3) Fusing the Reconstructed Surfaces. Ji *et al.* [116] process each pair of volumetric grids using a 3D CNN, which classifies whether a voxel is a surface point or not. To avoid the exhaustive combination of every possible image pairs, Ji *et al.* [116] learn their relative importance, using a network composed of fully-connected layers, automatically select a few view pairs based on their relative importance to reconstruct multiple volumetric grids, and take their weighted sum to produce the final 3D reconstruction.

To handle high resolution volumetric grids, Ji *et al.* [116] split the whole space into small Colored Voxel Cubes (CVCs) and regress the surface cube-by-cube. While this reduces the memory requirements, it requires multiple forward passes and thus increases the computation time. Paschalidou *et al.* [91] avoid the explicit use of the volumetric representation. Instead, each voxel of the grid is projected onto each of the input views, before computing the pairwise correlation between the corresponding learned features on each pair of views, and then averaging them over all pairs of views. Repeating this process for each depth value will result in the depth distribution on each pixel. This depth distribution is regularized using an MRF formulated as a differentiable function to enable end-to-end training.

In terms of performance, the volumetric approach of Ji *et al.* [116] requires 4 hours to obtain a full reconstruction of a typical scene in DTU dataset [24]. The approach of Paschalidou *et al.* [91] takes approximately 25 mins, on an Intel i7 computer with an Nvidia GTX Titan X GPU, for the same task. Finally, methods that perform fusion post-reconstruction have higher reconstruction errors compared to those that perform early fusion.

6.2 Plane-Sweep Volume Representations

These methods directly estimate depth maps from the input without using intermediate volumetric representations of the 3D geometry. As such, they are computationally more efficient. The main challenge to address is how to efficiently

match features across multiple views in a single forward pass. This is done by using the Plane-Sweep Volumes [27], [66], [90], [93], [118], [119], i.e., they back project the input images [27], [66], [118] or their learned features [90], [93], [119] into planes at different depth values, forming PSVs from which the depth map is estimated. Existing methods differ in the way the PSVs are processed with the feature extraction and feature matching blocks.

Flynn *et al.*'s network [66] (Fig. 6b) is composed of n_d branches, one for each depth plane. The $d-$th branch of the network takes as input the reference image and the planes of the PSVs of the other images which are located at depth d. These are packed together and fed to a two-stage network. The first stage computes matching features between the reference image and all the PSV planes located at depth d. The second stage models interactions across depth planes using convolutional layers. The final block of the network is a per-pixel softmax over depth, which returns the most probable depth value per pixel. The approach requires that the number of views and the camera parameters of each view to be known.

Huang *et al.* [27]'s approach (Fig. 6d) starts with a pairwise matching step where a cost volume is computed between the reference image and each of the input images. For a given pair $(I_1, I_i), i = 2, \ldots, n$, I_i is first back-projected into a PSV. A siamese network then computes a matching cost volume between I_1 and each of the PSV planes. These volumes are aggregated into a single cost volume using an encoder-decoder network. This is referred to as intra-volume aggregation. Finally a max-pooling layer is used to aggregate the multi intra-volumes into a single inter-volume, which is then used to predict the depth map. Unlike Flynn *et al.* [66], Huang *et al.* [27]'s approach does not require a fixed number of input views since aggregation is performed using pooling. In fact, the number of views can vary between training and at runtime.

Unlike [27], [66], which back-project the input images, the MVSNet of Yao *et al.* [93] use the camera parameters to back-project the learned features into a 3D frustum of a reference camera sliced into parallel frontal planes, one for each depth value. The approach then generates the matching cost volume upon a pixel-wise variance-based metric, and finally a generic 3D U-Net is used to regularize the matching cost volume to estimate the depth maps. Luo *et al.* [119] extend MVSNet [93] to P-MVSNet in two ways. *First*, a raw cost volume is processed with a learnable patch-wise aggregation function before feeding it to the regularization network. This improves the matching robustness and accuracy for noisy data. *Second*, instead of using a generic 3D-UNet network for regularization, P-MVSNet uses a hybrid isotropic-anisotropic 3D-UNet. The plane-sweep volumes are essentially anisotropic in depth and spatial directions, but they are often approximated by isotropic cost volumes, which could be detrimental. In fact, one can infer the corresponding depth map along the depth direction of the matching cost volume, but cannot get the same information along other directions. Luo *et al.* [119] exploit this fact, through the proposed hybrid 3D U-Net with isotropic and anisotropic 3D convolutions, to guide the regularization of matching confidence volume.

The main advantage of using PSVs is that they eliminate the need to supply rectified images. In other words, the camera parameters are implicitly encoded. However, in order to compute the PSVs, the intrinsic and extrinsic camera parameters need to be either provided in advance or estimated using, for example, Structure-from-Motion techniques as in [27]. Also, these methods require setting in advance the disparity range and its discretisation. Moreover, they often result in a complex network architecture. Wang *et al.* [120] propose a light-weight architecture. It stacks together the reference image and the cost volume, computed using the absolute difference between the reference image and each other image but at different depth planes, and feeds them to an encoder-decoder network, with skip connections, to estimate the inverse depth at three different resolutions. Wang *et al.* [120] use a view selection rule, which selects the frames that have enough angle or translation difference and then use the selected frames to compute the cost volume.

Finally, note that feature back-projection has been also used by Won *et al.* [30] for omnidirectional depth estimation from a wide-baseline multi-view stereo setup. The approach uses spherical maps and spherical cost volumes.

7 Training End-to-End Stereo Methods

The training process aims to find the network parameters W that minimize a loss function $\mathcal{L}(W; \hat{D}, \Theta)$ where $\hat{D}$ is the estimated disparity, and Θ are the supervisory cues. The loss function is defined as the sum of a data term $\mathcal{L}_1(\hat{D}, \Theta, W)$, which measures the discrepancy between the ground-truth and the estimated disparity, and a regularization or smoothness term $\mathcal{L}_2(\hat{D}, W)$, which imposes local or global constraints on the solution. The type of supervisory cues defines the degree of supervision (Section 7.1), which can be supervised with 3D groundtruth (Section 7.1.1), self-supervised using auxiliary cues (Section 7.1.2), or weakly supervised (Section 7.1.3). Some methods use additional cues, in the form of constraints on the solution, to boost the accuracy and performance (Section 7.2). One of the main challenges of deep learning-based techniques is their ability to generalize to new domains. Section 7.3 reviews methods that addressed this issue. Finally, Section 7.4 reviews methods that learn network architectures.

7.1 Supervision Methods

7.1.1 3D Supervision Methods

Supervised methods are trained to minimise a loss function that measures the error between the ground truth disparity and the estimated disparity. It is of the form

$$\mathcal{L} = \frac{1}{N} \sum C(x) H(C(x) - \epsilon) \mathcal{D}\Big(\Phi(d_x), \Phi(\hat{d}_x)\Big), \tag{6}$$

where: d_x and $\hat{d}_x$ are, respectively, the groundtruth and the estimated disparity at pixel x. $\mathcal{D}$ is a measure of distance, which can be the L_2, the L_1 [61], [62], [99], [121], the smooth L_1 [64], or the smooth L_1 but approximated using the two-parameter robust function $\rho(\cdot)$ [75], [122]. $C(x) \in [0, 1]$ is the confidence of the estimated disparity at x. Setting $C(x) = 1$ and $\epsilon = 0, \forall x$ is equivalent to ignoring the confidence map.

$H(x)$ is the heavyside function, with $H(x) = 1$ if $x \geq 0$, and $H(x) = 0$ otherwise. $\Phi(\cdot)$ is either the identify or the log function. The latter avoids overfitting to large disparities.

Some papers restrict the sum in Eqn. (6) to be over only the valid pixels or regions of interest, e.g., foreground or visible pixels [123], to avoid outliers. Others, e.g., Yao *et al.* [93], divide the loss into two parts, one over the initial disparity and the other one over the refined disparity. The overall loss is defined as the weighted sum of the two losses.

7.1.2 Self-Supervised Methods

Self-supervised methods, originally used in optical flow estimation [124], [125], have been proposed as a possible solution in the absence of sufficient ground-truth training data. These methods mainly rely on image reconstruction losses, taking advantage of the projective geometry, and the spatial and temporal coherence when multiple images of the same scene are available. The rationale is that if the estimated disparity map is as close as possible to the ground truth, then the discrepancy between the reference image and any of the other images but unprojected using the estimated depth map onto the reference image, is also minimized. The general loss function is of the form

$$\mathcal{L} = \frac{1}{N} \sum_{x} \mathcal{D}\big(\Phi\big(I_{ref}\big)(x) - \Phi\big(\tilde{I}_{ref}\big)(x)\big), \tag{7}$$

where $\tilde{I}_{ref}$, which is I_{right} but unwarped onto I_{ref} using the estimated disparity, and $\mathcal{D}$ is a measure of distance. The mapping function Φ can be:

- The identity [68], [77], [126], [127]. In this case, the loss of Eqn. (7) is called a photometric or image reconstruction loss.
- A mapping to the feature space [68], i.e., $\Phi\big(I_{ref}\big) = \mathbf{f}$ where $\mathbf{f}$ is the learned feature map.
- The gradient of the image, i.e., $\Phi\big(I_{ref}\big) = \nabla I_{ref}$, which is less sensitive to variations in lighting and acquisition conditions than the photometric loss.

The distance $\mathcal{D}$ can be the L_1 or L_2 distance. Some papers [77] also use more complex metrics such as the structural dissimilarity [128] between patches in I_{ref} and in $\tilde{I}_{ref}$.

While stereo-based supervision methods do not require ground-truth 3D labels, they rely on the availability of calibrated stereo pairs during training.

7.1.3 Weakly Supervised Methods

Supervised methods for disparity estimation can achieve promising results if trained on large quantities of ground truth depth data. However, manually obtaining ground-truth depth data is extremely difficult and expensive, and is prone to noise and inaccuracies. Weakly supervised methods rely on auxiliary signals to reduce the amount of manual labelling. In particular, Tonioni *et al.*[129] used as a supervisory signal the depth estimated using traditional stereo matching techniques to fine-tune depth estimation networks. Since such depth data can be sparse, noisy, and prone to errors, they propose a confidence-guided loss that penalizes ground-truth depth values that are deemed not reliable. It is defined using Eqn. (6) by setting $\mathcal{D}(\cdot)$ to be the L_1 distance, and $\epsilon > 0$. Kuznietsov *et al.* [130] use sparse ground-truth depth for supervised learning, while enforcing the deep network to produce photo-consistent dense depth maps in a stereo setup using a direct image alignment/reprojection loss. These two methods rely on an ad-hoc disparity estimator. To avoid that, Zhou *et al.* [131] propose an iterative approach, which starts with a randomly initialized network. At each iteration, it computes matches from the left to the right images, and matches from the right to the left images. It then selects the high confidence matches and adds them as labelled data for further training in the subsequent iterations. The confidence is computed using the left-right consistency of Eqn. (12).

7.2 Incorporating Additional Cues

Several works incorporate additional cues and constraints to improve the quality of the disparity estimates. Examples include smoothness [77], left-right consistency [77], maximum depth [77], and scale-invariant gradient loss [121]. Such cues can also be in the form of auxiliary information such as semantic cues used to guide the disparity estimation network. Below, we discuss a number of these works.

1) Smoothness. In general, one can assume that neighboring pixels have similar disparity values. Such smoothness constraint can be enforced by minimizing:

- The absolute difference between the disparity predicted at x and those predicted at each pixel y within a certain predefined neighborhood $\mathcal{N}_x$ around x

$$\mathcal{L} = \frac{1}{N} \sum_{x} \sum_{y \in \mathcal{N}_x} |d_x - d_y|. \tag{8}$$

  Here, N is the total number of pixels.
- The magnitude of the first-order gradient ∇ of the estimated disparity map [68]

$$\mathcal{L} = \frac{1}{N} \sum_{x} \{(\nabla_u d_x) + (\nabla_v d_x)\}, x = (u, v). \tag{9}$$

- The magnitude of the 2nd-order gradient of the estimated disparity [127], [132]

$$\mathcal{L} = \frac{1}{N} \sum_{x} \Big\{(\nabla_u^2 d_x)^2 + (\nabla_v^2 d_x)^2\Big\}. \tag{10}$$

- The 2nd-order gradient of the estimated disparity weighted by the image's 2nd-order gradients [77]

$$\mathcal{L} = \frac{1}{N} \sum_{x} \Big\{|\nabla_u^2 d_x| e^{-|\nabla_u^2 I(x)|} + |\nabla_v^2 d_x| e^{-|\nabla_v^2 I(x)|}\Big\}. \tag{11}$$

2) Consistency. Zhong *et al.* [77] introduced the loop-consistency loss, which is constructed as follows. Consider the left image I_{left} and the synthesized image $\tilde{I}_{left}$ obtained by warping the right image to the left image coordinate using the disparity map defined on the right image. A second synthesized left image $\tilde{\tilde{I}}_{left}$ can also be generated by warping the left image to the right image coordinates, by using the disparities at the left image, and then warping it back to the left image using the disparity at the right image. The three versions of the left image provide two constraints: $I_{left} =$

$\tilde{I}_{left}$ and $I_{left} = \tilde{\tilde{I}}_{left}$, which can be used to regularize the disparity maps. Godard *et al.* [133] introduce the left-right consistency term, which is a linear approximation of the loop consistency. The loss attempts to make the left-view disparity map equal to the projected right-view disparity map. It is defined as

$$\mathcal{L} = \frac{1}{N}\sum_{x} |d_x - \tilde{d}_x|, \tag{12}$$

where $\tilde{d}$ is the disparity at the right image but reprojected onto the coordinates of the left image.

3) Maximum-Depth Heuristic. There may be multiple warping functions that achieve a similar warping loss, especially for textureless areas. To provide strong regularization in these areas, Zhong *et al.* [77] use the Maximum-Depth Heuristic (MDH) [134] defined as the sum of all depths/disparities

$$\mathcal{L} = \frac{1}{N}\sum_{x} |d_x|. \tag{13}$$

4) Scale-Invariant Gradient Loss [121]. It is defined as

$$\mathcal{L} = \sum_{h \in A}\sum_{x} \|g_h[D](x) - g_h[\hat{D}](x)\|_2, \tag{14}$$

where $A = \{1, 2, 4, 8, 16\}$, $x = (i, j)$, $f_{i,j} \equiv f(i, j)$, and

$$g_h[f](i,j) = \left(\frac{f_{i+h,j} - f_{i,j}}{|f_{i+h,j} - f_{i,j}|}, \frac{f_{i,j+h} - f_{i,j}}{|f_{i,j+h} - f_{i,j}|}\right)^{\top}. \tag{15}$$

This loss penalizes relative depth errors between neighbouring pixels. This loss stimulates the network to compare depth values within a local neighbourhood for each pixel. It emphasizes depth discontinuities, stimulates sharp edges, and increases smoothness within homogeneous regions.

5) Incorporating Semantic Cues. Some papers incorporate additional cues such as normal [135], segmentation [68], and edge [80] maps, to guide the disparity estimation. These can be either provided at the outset, e.g., estimated with a separate method as in [80], or estimated jointly with the disparity map. Qi *et al.* [135] propose a mechanism that uses the depth map to refine the quality of the normal estimates, and the normal map to refine the quality of the depth estimates. This is done using a two-stream network: a depth-to-normal network for normal map refinement using the initial depth estimates, and a normal-to-depth network for depth refinement using the estimated normal map.

Yang *et al.* [68] and Song *et al.* [80] incorporate semantics by stacking semantic maps (segmentation masks in the case of [68] and edge features in the case of [80]) with the 3D cost volume. Yang *et al.* [68] train jointly a disparity estimation network and a segmentation network by using a loss function defined as a weighted sum of the reconstruction error, a smoothness term, and a segmentation error. Song *et al.* [80] further incorporate edge cues in the edge-aware smoothness loss to penalize drastic depth changes in flat regions. Also, to allow for depth discontinuities at object boundaries, the edge-aware smoothness loss is defined based on the gradient map obtained from the edge detection sub-network, which is more semantically meaningful than the variation in raw pixel intensities.

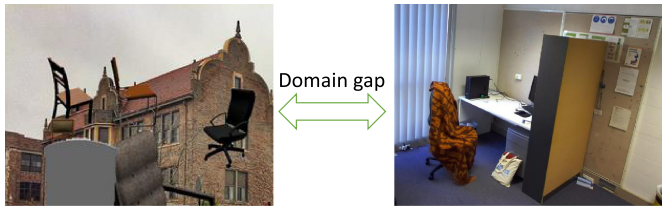


Fig. 7. Illustration of the domain gap between synthetic (left) and real (right) images. The left image is from the FlyingThings synthetic dataset [22].

Wu *et al.* [86] introduced an approach that fuses multi-scale 4D cost volumes with semantic features obtained using a segmentation sub-network. The approach uses the features of the left and the right images as input to a semantic segmentation network similar to PSPNet [136]. Semantic features for each image are then obtained from the output of the classification layer of the segmentation network. A 4D semantic cost volume is obtained by concatenating each unary semantic feature with their corresponding unary from the opposite stereo image across each disparity level. Both the spatial pyramid cost volumes and the semantic cost volume are fed into a 3D multi-cost aggregation module, which aggregates them, using an encoder-decoder followed by a 3D feature fusion module, into a single 3D cost volume in a pairwise manner starting with the smallest volume.

In summary, appending semantic features to the cost volume improves the reconstruction of fine details, especially near object boundaries.

7.3 Domain Adaptation and Transfer Learning

Deep architectures for depth estimation are severely affected by the domain shift issue, which hinders their effectiveness when performing inference on images significantly diverse from those used during the training stage. This can be observed, for instance, when moving from indoor to outdoor environments, from synthetic to real data, see Fig. 7, and when changing the camera model/parameters. As such, deep learning networks trained on one domain, e.g., by using synthetic data, suffer when applied to another domain, e.g., real data, resulting in blurry object boundaries and errors in ill-posed regions such as object occlusions, repeated patterns, and textureless regions. These are referred to as *generalization glitches* [137].

Several strategies have been proposed to adress this domain bias issue. They can be classified into two categories: adaptation by fine-tuning (Section 7.3.1) and adaptation by data transformation (Section 7.3.2). In both cases, the adaptation can be offline or online.

7.3.1 Adaptation by Fine-Tuning

Methods in this category perform domain adaptation by first training a network on images from a certain domain, e.g., synthetic images as in [22], and then fine-tuning it on images from a target domain. A major difficulty is to collect accurate ground-truth depth for stereo or multiview images from the target domain. Relying on active sensors (e.g., LiDAR) to obtain such supervised labeled data is not feasible in practical applications. As such, recent works, e.g.,

[129], [137], [138] rely on off-the-shelf stereo algorithms to obtain ground-truth disparity/depth labels in an unsupervised manner, together with state-of-the-art confidence measures to ascertain the correctness of the measurements of the off-the-shelf stereo algorithms. The latter is used in [129], [138] to discriminate between reliable and unreliable disparity measurements, to select the former and fine tune a pre-trained model, e.g., DispNet [22], using such smaller and sparse set of points as if they were ground-truth labels.

Pang *et al.* [137] also use a similar approach as in [129], [138] to address the generalization glitches. The approach, however, exploits the scale diversity, i.e., up-sampling the stereo pairs enables the model to perform stereo matching in a localized manner with subpixel accuracy, by performing iterative optimisation of predictions obtained at multiple resolutions of the input.

Note that self-supervised and weakly supervised techniques for disparity estimation, e.g., [133], [139], [140], [141], can also be used for offline domain adaptation. In particular, if stereo pairs of the target domain are available, these techniques can be fine-tuned, in an unsupervised manner, using reprojection losses, see Sections 7.1.2 and 7.1.3.

Although effective, these offline adaptation techniques reduce the usability of the methods since users are required to train the models every time they are exposed to a new domain. As a result, several recent papers developed online adaptation techniques. For example, Tonioni *et al.* [76] address the domain shift issue by casting adaptation as a continuous learning process whereby a stereo network can evolve online based on the images gathered by the camera during its real deployment. This is achieved in an unsupervised manner by computing error signals on the current frames, updating the whole network by a single back-propagation iteration, and moving to the next pair of input frames. To keep a high enough frame rate, Tonioni *et al.* [76] propose a lightweight, fast, and modular architecture, called MADNet, which allows training sub-portions of the whole network independently from each other. This allows adapting disparity estimation networks to unseen environments without supervision at approximately 25 fps, while achieving an accuracy comparable to DispNetC [22]. Similarly, Zhong *et al.* [142] use video sequences to train a deep network online from a random initialization. They employ an LSTM in their model to leverage the temporal information during the prediction.

Zhong *et al.* [142] and Tonioni *et al.* [76] consider online adaptation separately from the initial training. Tonioni *et al.* [143], on the other hand, incorporate the adaptation procedure to the learning objective to obtain a set of initial parameters that are suitable for online adaptation, i.e., they can be adapted quickly to unseen environments. This is implemented using the model agnostic meta-learning framework of [144], an explicit *learn-to-adapt* framework that enables stereo methods to adapt quickly and continuously to new target domains in an unsupervised manner.

7.3.2 Adaptation by Data Transformation

Methods in this category transform the data of one domain to look similar in style to the data of the other domain. For example, Atapour-Abarghoue *et al.* [145] proposed a two-staged approach. The first stage trains a depth estimation model using synthetic data. The second stage is trained to transfer the style of synthetic images to real-world images. By doing so, the style of real images is first transformed to match the style of synthetic data and then fed into the depth estimation network, which has been trained on synthetic data. Zheng *et al.* [146] perform the opposite by transforming the synthetic images to become more realistic and using them to train the depth estimation network. Zhao *et al.* [147] consider both synthetic-to-real [146] and real-to-synthetic [145], [148] translations. The two translators are trained in an adversarial manner using an adversarial loss and a cycle-consistency loss. That is, a synthetic image when converted to a real image and converted back to the synthetic domain should look similar to the original one.

Although these methods have been used for monocular depth estimation, they are applicable to (multi-view) stereo matching methods.

7.4 Learning the Network Architecture

Much research work in depth estimation is being spent on manually optimizing network architectures, but what about if the optimal network architecture, along with its parameters, could be also learnt from data? Saika *et al.* [149] show how to use and extend existing AutoML techniques [150] to efficiently optimize large-scale U-Net-like encoder-decoder architectures for stereo-based depth estimation. Traditional AutoML techniques have extreme computational demand limiting their usage to small-scale classification tasks. Saika *et al.* [149] apply Differentiable Architecture Search (DARTs) [151] to encoder-decoder architectures. Its main idea is to have a large network that includes all architectural choices and to select the best parts of this network by optimization. This can be relaxed to a continuous optimization problem, which, together with the regular network training, leads to a bilevel optimization problem. Experiments conducted on DispNet of [83], an improved version of [22], show that the automatically optimized DispNet (AutoDispNet) yields better performance compared to the baseline DispNet, with about the same number of parameters. The paper also shows that these benefits carry over to large stacked networks.

8 Discussion and Comparison

Tables 3 and 4, respectively, compare the performance of the methods surveyed in this paper on standard datasets such as KITTI2015 for pairwise stereo methods, and DTU, SUN3D and ETH3D for multiview stereo methods. Most of these methods have been trained on subsets of these publicly available datasets. A good disparity estimation method, once properly trained, should achieve good performance not only on publicly available benchmarks but on arbitrary novel images. They should not require re-training or fine-tuning every time the domain of usage changes. In this section, we will look at how some of these methods perform on novel unseen images. We will first describe in Section 8.1 the evaluation protocol, the images that will be used, and the evaluation metrics. We then discuss the performance of these methods in Sections 8.2 and 8.3.

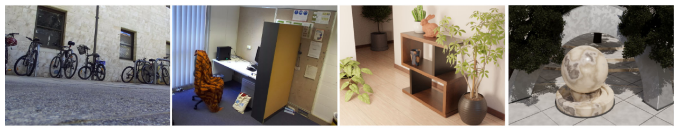
(a) Baseline: images with good lighting conditions.

(b) Challenge: images with challenging lighting conditions.

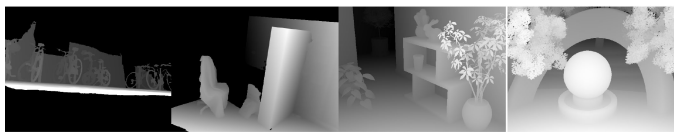
Fig. 8. Examples of stereo pairs and their ground-truth disparity maps from the ApolloScape dataset [34].

(a) Left image.

(b) Highlights of regions of interest where ground-truth disparity is estimated with high confidence.



(c) Right image.



$d \in [9.3, 34.0]$ $d \in [18.7, 29.9]$ $d \in [5.6, 14.5]$ $d \in [5.5, 13.2]$
Depth $\in [2.1, 7.8]$ Depth $\in [2.4, 3.3]$ Depth $\in [7.8, 25.0]$ Depth $\in [10.8, 25.6]$

(d) Ground-truth disparity maps.

Fig. 9. Four images, collected in-house and used to test 16 state-of-the-art methods. The green masks on some of the left images highlight the pixels where the ground-truth disparity is available. The disparity range is shown in pixels while the depth range is in meters. d refers to disparity.

8.1 Evaluation Protocol

We consider several key methods, trained independently on different datasets, and evaluate their performance on the stereo subset of the ApolloScape dataset [34], and on an in-house collected set of four images. The motivation behind this choice is two-fold. *First*, the ApolloScape dataset is composed of stereo images taken outdoor in autonomous driving setups. Thus, it exhibits several challenges related to uncontrolled complex and varying lighting conditions, and heavy occlusions. *Second*, the dataset is novel and existing methods have not been trained or exposed to this dataset. Thus, it can be used to assess how these methods generalize to novel scenarios. In this dataset, ground truth disparities have been acquired by accumulating 3D point clouds from Lidar and fitting 3D CAD models to individually moving cars. We also use four *in-house* images of size $W = 640$ and $H = 480$, see Fig. 9, specifically designed to challenge these methods. Two of the images are of real scenes: a Bicycles scene composed of bicycles in a parking, and an indoor Desk scene composed of office furnitures. We use a moving stereo camera to capture multiple stereo pairs, and Structure-from-Motion to build a 3D model of the scenes. We then render depth maps from the real cameras' viewpoints. Regions where depth is estimated with high confidence will be used as ground-truth. The remaining two images are synthetic, but real-looking. They include objects with complex structures, e.g., thin structures such as plants, large surfaces with either uniform colors or textures and repetitive patterns, presenting several challenges to stereo-based depth estimation algorithms.

We have tested 16 stereo-based methods published in 9 papers (between 2018 and 2019), see below. We use the network weights as provided by the authors.

1) AnyNet [73]: It is a four-stages network, which builds 3D cost volumes in a coarse-to-fine manner. The first stage estimates a low resolution disparity map by searching on a small disparity range. The subsequent stages estimate refined disparity maps using residual learning.

2) DeepPruner [82]: It combines deep learning with PatchMatch [101] to speed up inference by adaptively pruning out the potentially large search space for correspondences. Two variants have been proposed: DeepPruner (Best), which downsamples the cost volume by a factor of 4, and DeepPruner (Fast), which downsamples it by a factor of 8.

3) DispNet3 [83], an improved version of DispNet [22] where occlusions and disparity maps are jointly estimated.

4) GANet [87]: It replaces a large number of the 3D convolutional layers in the regularization block with (1) two 3D convolutional layers, (2) a semi-global aggregation layer, and (3) a local guided aggregation layer. SGA and LGA layers capture local and whole-image cost dependencies. They are meant to improve the accuracy in challenging regions such as occlusions, large textureless/reflective regions, and thin structures.

5) HighResNet [32]: To refine both the spatial and the depth resolutions while operating on high resolution images, this method searches for correspondences incrementally using a coarse-to-fine hierarchy. Its hierarchical design also allows for anytime on-demand reports of disparity.

6) PSMNet [64]: It progressively regularizes a low resolution 4D cost volume, estimated from a pyramid of features.

7) iResNet [63]: The initial disparity and the learned features are used to calculate a feature constancy map, which measures the correctness of the stereo matching. The initial disparity map and the feature constancy map are then fed into a sub-network for disparity refinement.

8) UnsupAdpt [129]: It is a unsupervised adaptation approach that enables fine-tuning without any ground-truth information. It first trains DispNet-Corr1D [22] using the KITTI2012 training dataset and then adapts the network to KITTI2015 and Middlebury 2014.

9) SegStereo [68]: It is a self-supervised disparity estimation method, which uses segmentation masks to guide the disparity estimation. Both segmentation and

TABLE 5
Computation Time, Memory Consumption, at Runtime, and Reconstruction Accuracy, in Terms of RMSE (the lower the better) and Bad-2 (the lower the better), on Images of Size 640×480

Method	Supervision mode	Cost (vol.)	Time (sec)	Memory (GB)	Training set	RMSE - Baseline			RMSE - Challenge			Overall Bad-2 (%)	
						Bkg	Fg	Overall	Bkg	Fg	Overall	Baseline	Challenge
iResNet [63]	Supervised	3D	0.939	7.656	KITTI2015	60.04	61.72	60.54	45.87	46.85	47.86	97.15	96.96
					ROB [152]	22.08	17.16	18.08	23.01	16.51	18.83	73.75	74.08
PSMNet [64]	Supervised	4D	1.314	1.900	KITTI2015	9.88	9.81	9.80	10.10	9.42	9.93	77.25	76.01
					KITTI2012	10.17	10.24	10.29	10.66	10.33	11.00	71.07	68.60
HighResNet [32]	Supervised	4D	**0.037**	0.474	Middleburry [20], KITTI2015 [21], ETH3D [25], HR-VS [32]	9.47	9.91	9.94	8.58	9.64	**9.78**	63.26	62.52
DeepPruner (Fast) [82]	Supervised	3D	3.930	6.166	KITTI2012+2015	9.56	9.90	9.94	8.74	9.75	9.86	61.49	61.57
AnyNet [73]	Supervised	3D	0.285	**0.232**	KITTI2015	9.46	10.74	10.34	9.83	11.60	11.15	63.02	59.50
					KITTI2012	9.80	10.29	10.20	9.34	10.62	10.61	61.63	58.09
UnsupAdpt [129]	Self-supervised	3D	–	–	Shadow-on-Truck	**8.52**	10.08	9.58	10.66	10.88	10.27	59.56	59.30
					KITTI2012 adapted	9.44	10.39	10.19	10.10	10.42	10.78	59.65	59.26
					to KITTI2015	9.44	1						
SegStereo [68]	Self-supervised	3D	0.195	~ 12.00	CityScapes [23]	9.26	10.30	10.17	9.03	10.49	10.54	57.56	56.77
DispNet3 [83]	Supervised	3D	–	10.953	css-FlyingThings3D [22]	9.29	9.98	9.87	9.66	10.34	10.61	58.76	58.64
					CSS-FlyingThings3D [22]	9.11	9.64	9.54	8.97	9.91	10.19	54.44	55.81
					CSS-ft-KITTI	9.68	9.62	9.70	**8.38**	11.00	11.11	53.36	52.41
DeepPruner (Best) [82]	Supervised	3D	8.430	8.845	KITTI2012+2015	9.64	9.43	9.46	12.38	**8.74**	10.48	56.17	55.03
GANet [87]	Supervised	4D	8.336	3.017	KITTI2012	9.98	10.29	10.25	10.69	10.95	11.55	57.93	56.13
					KITTI2015	9.55	**9.38**	**9.39**	9.37	9.50	9.89	53.07	52.67

SegStereo [68] has been tested on a PC equipped with an Nvidia GeForce RTX 2080. The other methods have been tested on a PC equipped with an Nvidia Tesla K40 GPU with a 12 GB graphic memory. See the Supplementary Material for a visual representation, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2020.3032602.

disparity maps are jointly estimated with an end-to-end network.

The methods (1) to (7) are supervised with ground-truth depth maps while the methods (8) and (9) are self-supervised. We compare their accuracy at runtime using the overall Root Mean Square Error (RMSE) defined as

$$\mathrm{RMSE}^2_{\mathrm{linear}} = \frac{1}{N}\sum_{N} |d_i - \hat{d}_i|^2, \tag{16}$$

and the Bad-n error defined as the percentage of pixels whose estimated disparity deviates with more than n pixels from the ground truth. We use $n \in \{0.5, 1, 2, 3, 4, 5\}$. The Bad-n error considers the distribution and spread of the error and thus provides a better insight on the accuracy of the methods. In addition to accuracy, we also report the computation time and memory footprint at runtime.

8.2 Computation Time and Memory Footprint

From Table 5, we can distinguish three types of methods; slow methods, e.g., PSMNet [64], DeepPruner (Best) and (Fast) [82], and GANet [87], require more than 1 second to estimate one disparity map. They also require between 3 GB and 10 GB (for DispNet3 [83]) of memory at runtime. As such, these methods are very hard to deploy on mobile platforms. Average-speed methods, e.g., AnyNet [73] and iResNet [63], produce a disparity map in around one second. Finally, fast methods, e.g., HighResNet [32], require less than 0.1 seconds. In general, methods that use 3D cost volumes are faster and less memory demanding than those that use 4D cost volumes. There are, however, two exceptions: iResNet [63] and DeepPruner [82], which use 3D cost volumes but require a large amount of memory at runtime. While iResNet requires less than a second to process images of size $W = 640, H = 480$, since it uses 2D convolutions to regularize the cost volume, DeepPruner [82] requires more than 3 seconds. We also observe that HighResNet [32], which uses 4D cost volumes but adopts a hierarchical approach to produce disparity on demand, is very efficient in terms of computation time as it only requires 37 ms, which is almost 8 times faster than AnyNet [73], which uses 3D cost volumes. Note that AnyNet [73] can run on mobile devices due to its memory efficiency.

8.3 Reconstruction Accuracy

Table 5 shows the average RMSE and the Bad-2 error of each of the methods described in Section 8.1. We report the results on a baseline subset composed of 141 images that look more or less like KITTI2012 images, hereinafter referred to as *baseline*, and on another subset composed of 33 images with challenging lighting conditions, hereinafter referred to as *challenge*. Here, we focus on the relative comparison across methods since some of the high errors observed might be attributed to the way the ground-truth has been acquired in ApolloScape [34] dataset, rather than to the methods themselves.

We observe that these methods behave almost equally on the two subsets. However, the reconstruction error, is significantly important, larger than 8 pixels, compared to the errors reported on standard datasets such as KITTI2012 and KITTI2015. This suggests that, when there is a significant domain gap between training and testing then the reconstruction accuracy can be significantly affected.

We also observe the same trend on the Bad-n curves of Fig. 10 where, in all methods, more than 25 percent of the pixels had a reconstruction error that is larger than 5 pixels. The Bad-n curves show that the errors are large on the foreground pixels, i.e., pixels that correspond to cars, with more than 55 percent of the pixels having an error that is larger than 3 pixels (against 35 percent on the background pixels). Interestingly, Table 5 and Fig. 10 show that most of the methods achieve similar reconstruction accuracies. The only exception is iResNet [63] trained on Kitti2015 and on ROB [152], which had more than 90 percent, respectively 55 percent, of pixels with an error that is larger than 5 pixels. In all methods, less than 25 percent of the pixels had an error that is less than 1 pixel. This

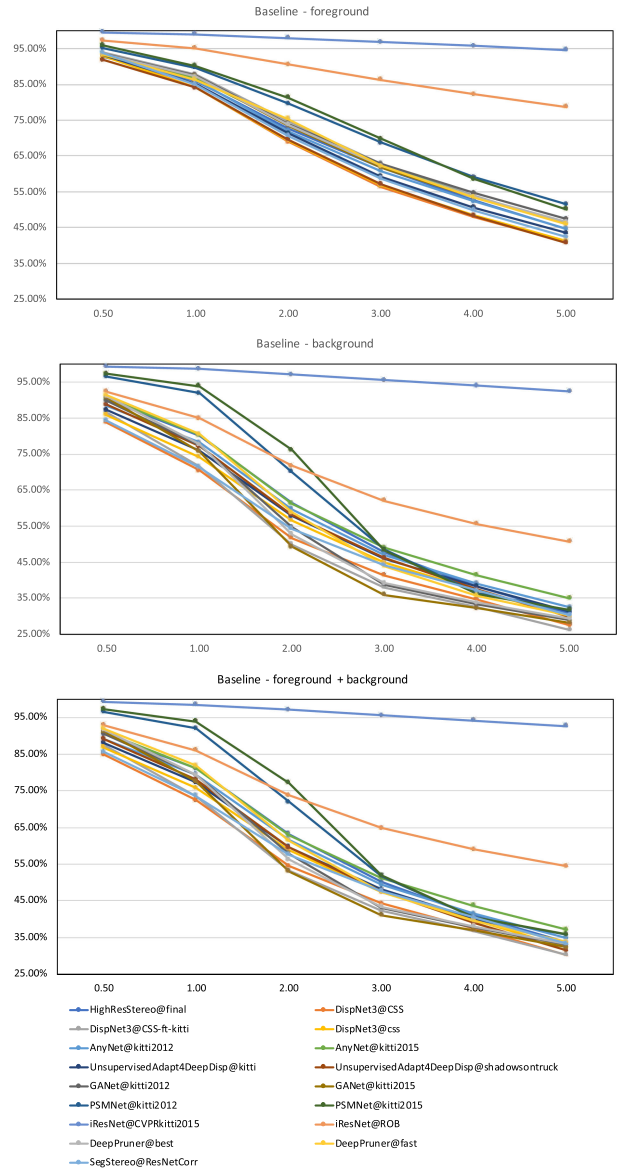


Fig. 10. Overall Bad-n error, $n \in [0.5, 5.0]$ on a selection of 141 (baseline) images from the stereo vision challenge of ApolloScape dataset [34]. A similar behaviour is observed on the challenge subset, see the supplementary material, available online. The horizontal axis is the error n while the vertical axis is the percentage of pixels whose estimated disparity deviates with more than n pixels from the ground truth.

suggests that achieving sub-pixel accuracy remains an important challenge for future research.

Using the Bad-2 error as a mean of comparison (see Table 5), we observe that methods that incorporate and jointly estimate additional semantic cues, e.g., segmentation masks as in SegStereo [68] and occlusions as in [83], achieve a better performance than those that estimate disparity alone. SegStereo [68], which is self-supervised, achieves a similar or better performance than many of the supervised methods. Also, GANet [87], which uses semi-global aggregation layers and local-guided aggregation layers during the cost volume regularization stage, achieves one of the best performances. This suggests that the SGA, a differentiable approximation of the semi-global matching, which, unlike the standard Semi-Global Matching, learns the SGM coefficients that are adapted to different locations for different situations, helps in improving the performance. Also, the LGA layers help refine the disparity estimation at the thin structures and object edges. This suggests that accuracy can be significantly improved by effectively guiding the disparity estimation algorithms.

While most of the methods use a refinement stage to improve the resolution and the overall accuracy, the HighResNet of Yang *et al.* [32] uses multiscale 4D volumes, aggregated using a Volume Pyramid Pooling block, to estimate high resolution disparity maps achieving 8 percent improvement in Bad-2 comapred to PSMNet [64].

Finally, the unsupervised self-adaptation method of Tonioni *et al.* [129], which takes the baseline DispNet-Corr1D network [22] trained on KITTI2012 and adapts it to KITTI2015 and Middlebury 2014, achieves one of the best performances on the foreground regions.

In terms of the visual quality of the estimated disparities, see Fig. 11, we observe that most of the methods were able to recover the overall shape of trees but fail to reconstruct the details especially the leaves. The reconstruction errors are high in flat areas and around object boundaries. Also, highly reflective materials and poor lighting conditions remain a big challenge to these methods as shown in Fig. 11b. The supplementary material, available online, provides more results on the four stereo pairs of Fig. 9.

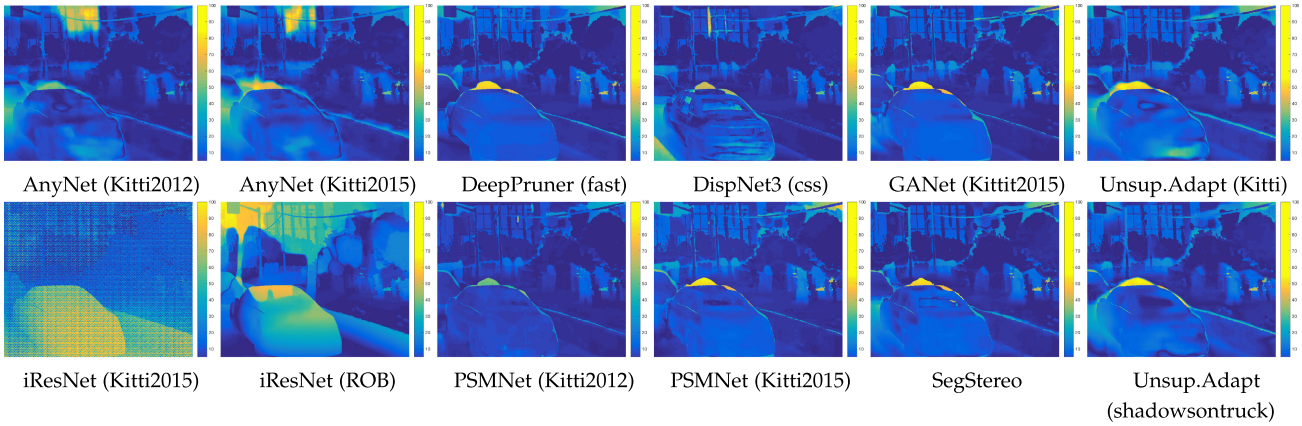
9 Future Research Directions

Deep learning methods for stereo-based depth estimation have achieved promising results. The topic, however, is still in its infancy and further developments are yet to be expected. In this section, we present some of the current issues and highlight directions for future research.

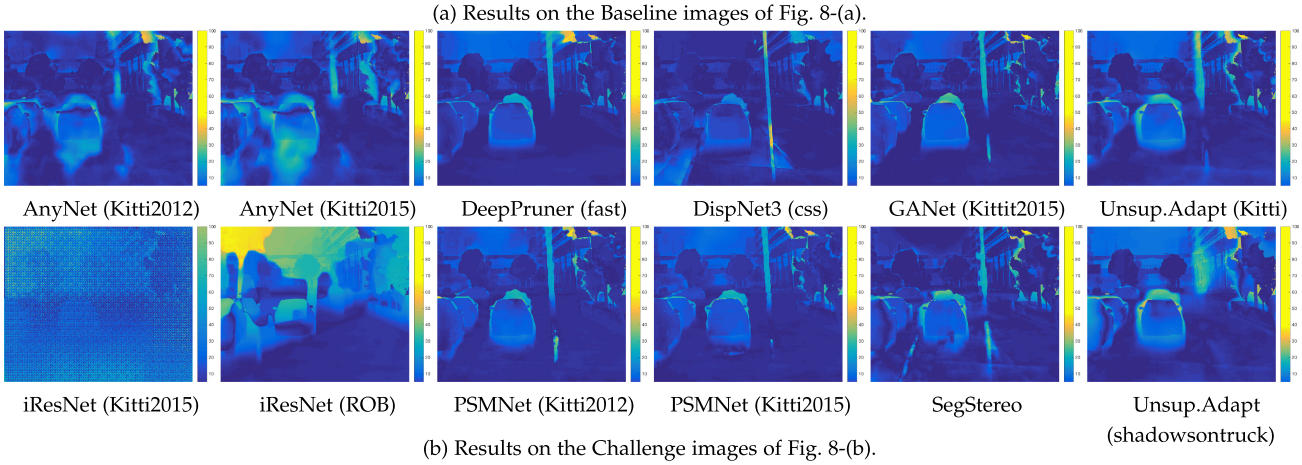
1) Camera Parameters. Most of the stereo-based techniques surveyed in this paper require rectified images. Multi-view stereo techniques use Plane-Sweep Volumes or back-projected images/features. Both image rectification and PSVs require known camera parameters, which are challenging to estimate in the wild. Many papers attempted to address this problem for monocular depth estimation and for 3D shape reconstruction by jointly optimising for the camera parameters and the geometry of the 3D scene [153].

2) Lighting Conditions and Complex Material Properties. Poor lighting conditions and complex materials properties remain a challenge to most of the current methods, see for example Fig. 11b. Combining object recognition, high-level scene understanding, and low-level feature learning can be one promising avenue to address these issues.

3) Spatial and Depth Resolution. Most of the current techniques do not handle high resolution input images and generally produce depth maps of low spatial and depth resolution. Depth resolution is particularly limited, making the methods unable to reconstruct thin structures, e.g., vegetation and hair, and structures located at a far distance from the camera. Although refinement modules can improve the resolution of the estimated depth maps, the gain is still small compared to the resolution of the input images. This has recently been addressed using hierarchical techniques, which allow on-demand reports of disparity by capping the resolution of the intermediate results [32]. In these methods, low resolution depth maps can be produced in realtime,



(a) Results on the Baseline images of Fig. 8-(a).



(b) Results on the Challenge images of Fig. 8-(b).

Fig. 11. Pixel-wise errors between the ground-truth disparities and the disparities estimated from the images of Fig. 8. We also refer the reader to Table 5, which provides the average reconstruction errors of these methods.

and thus can be used on mobile platforms, while high resolution maps would require more computation time. Producing, in realtime, accurate maps of high spatial and depth resolutions remains a challenge for future research.

4) Realtime Processing. Most deep learning methods for disparity estimation use 3D and 4D cost volumes, which are processed and regularized using 2D and 3D convolutions. They are expensive in terms of memory requirements and processing time. We expect to see in the future more research on novel lightweight, and subsequently fast, end-to-end deep networks that can run on edge devices.

5) Disparity Range. Existing techniques uniformly discretize the disparity range. This results in multiple issues. In particular, although the reconstruction error can be small in the disparity space, it can result in an error of meters in the depth space, especially at far ranges. One way to mitigate this is by discritizing disparity and depth uniformly in the log space. Also, changing the disparity range requires retraining the networks. Treating depth as a continuum could be one promising avenue for future research.

6) Training. Deep networks heavily rely on the availability of training images annotated with ground-truth labels. This is very expensive and labor intensive for depth/disparity reconstruction. As such, the performance of the methods and their generalization ability can significantly be affected including the risk of overfitting the models to specific domains. Existing techniques mitigate this problem by either designing loss functions that do not require 3D annotations, or by using domain adaptation and transfer learning strategies. The former, however, requires calibrated cameras. Domain adaptation techniques, especially unsupervised ones [138], are recently attracting more attention since, with these techniques, one can train with both synthetic data, which are easy to obtain, and real-world data. They also adapt, in an unsupervised manner and at run-time to ever-changing environments as soon as new images are gathered. Early results are very encouraging and thus expect in the future to see the emergence of large datasets, similar to ImageNet but for 3D reconstruction.

7) Automatically Learning the Network Architecture, its Activation Functions, and its Parameters From Data. Most existing research has focused on designing novel network architectures and novel training methods for optimizing their parameters. It is only recently that some papers started to focus on automatically learning optimal architectures. Early attempts, e.g., [149], focus on simple architectures. We expect in the future to see more research on automatically learning complex disparity estimation architectures and their activation functions, using, for example, the neuro-evolution theory [154], [155], which will free the need for manual network design.

10 Conclusion

Since 2014, we have entered a new era where data-driven and machine learning techniques play a central role in stereo image-based depth reconstruction. We have seen that, from 2014 to 2019, more than 150 papers on the topic have

been published in the major computer vision, computer graphics, and machine learning conferences and journals. Even during the final stages of this submission, more new papers are being published making it difficult to keep track of the recent developments, and more importantly, understanding their differences and similarities, especially for new comers to the field. This timely paper provides a comprehensive survey of these recent developments and can thus serve as a guide to the reader to navigate this fast-growing field of research. Although this paper compares the performance of some key disparity estimation methods on arbitrary novel images, evaluating and understanding how well these methods would perform on object boundaries and thin structures, and on scene reconstruction tasks can be useful for many applications, e.g., Augmented Reality and 3D modelling.

Finally, there are several related topics that have not been covered in this survey. Examples include image-based 3D object reconstruction using deep learning, which has been recently surveyed by Han *et al*. [153], and monocular and video-based depth estimation, which requires a separate survey paper given the large amount of papers that have been published on the topic in the past 5 to 6 years. Other topics include photometric stereo and active stereo [156], [157], which are outside the scope of this paper.

Acknowledgments

We would like to thank all the authors of the reference papers who have made their codes and datasets publicly available. This work was supported in part by Murdoch University's Vice Chancellor's Small Steps of Innovation Funding Program, and by the Australian Research Council (Grants DP150100294 and DP150104251).

References

[1] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, "A guide to convolutional neural networks for computer vision," *Synthesis Lectures Comput. Vis.*, vol. 8, no. 1, pp. 1–207, 2018.

[2] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1119–1127.

[3] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2650–2658.

[4] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 740–756.

[5] F. Liu, C. Shen, G. Lin, and I. D. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.

[6] J. Yang, S. E. Reed, M.-H. Yang, and H. Lee, "Weakly-supervised disentangling with recurrent transformations for 3D view synthesis," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1099–1107.

[7] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep convolutional inverse graphics network," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2539–2547.

[8] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Multi-view 3D models from single images with a convolutional network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 322–337.

[9] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 286–301.

[10] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg, "Transformation-grounded image generation network for novel 3D view synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 702–711.

[11] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1/3, pp. 7–42, 2002.

[12] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2366–2374.

[13] X. Wang, D. Fouhey, and A. Gupta, "Designing deep networks for surface normal estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 539–547.

[14] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009.

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

[16] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 611–625.

[17] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.

[18] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 573–580.

[19] J. Xiao, A. Owens, and A. Torralba, "SUN3D: A database of big spaces reconstructed using SfM and object labels," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1625–1632.

[20] D. Scharstein *et al.*"High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.*, 2014, pp. 31–42.

[21] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3061–3070.

[22] N. Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4040–4048.

[23] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[24] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 153–168, 2016.

[25] T. Schöps *et al.*, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3260–3269.

[26] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1746–1754.

[27] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "DeepMVS: Learning multi-view Stereopsis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2821–2830.

[28] Z. Li and N. Snavely, "MegaDepth: Learning single-view depth prediction from internet photos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2041–2050.

[29] J. Jeon and S. Lee, "Reconstruction-based pairwise depth dataset for depth image enhancement using CNN," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 422–438.

[30] C. Won, J. Ryu, and J. Lim, "OmniMVS: End-to-end learning for omnidirectional stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8987–8996.

[31] C. Won, J. Ryu, and J. Lim, "End-to-end learning for omnidirectional stereo matching with uncertainty prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 06, 2020, doi: 10.1109/TPAMI.2020.2992497.

[32] G. Yang, J. Manela, M. Happold, and D. Ramanan, "Hierarchical deep stereo matching on high-resolution images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5515–5524.

[33] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, "DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 899–908.

[34] P. Wang, X. Huang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The apolloScape open dataset for autonomous driving and its application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2702–2719, Oct. 2020.

[35] J. Geyer *et al.*, "A2D2: Audi autonomous driving dataset," 2020, *arXiv: 2004.06320*.

[36] N. Mayer *et al.*, "What makes good synthetic training data for learning disparity and optical flow estimation?" *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 942–960, 2018.

[37] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4353–4361.

[38] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3279–3286.

[39] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1592–1599.

[40] W. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang, "A deep visual correspondence embedding model for stereo matching costs," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 972–980.

[41] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 118–126.

[42] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, no. 1/32, 2016, Art. no. 2.

[43] V. Balntas, E. Johns, L. Tang, and K. Mikolajczyk, "PN-Net: Conjoined triple deep network for learning local image descriptors," 2016, *arXiv:1601.05030*.

[44] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5695–5703.

[45] B. G. V. Kumar, G. Carneiro, and I. Reid, "Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5385–5394.

[46] A. Shaked and L. Wolf, "Improved stereo matching with constant highway networks and reflective confidence learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4641–4650.

[47] W. Hartmann, S. Galliani, M. Havlena, L. Van Gool, and K. Schindler, "Learned multi-patch similarity," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1595–1603.

[48] H. Park and K. M. Lee, "Look wider to match image patches with convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1788–1792, Dec. 2017.

[49] X. Ye, J. Li, H. Wang, H. Huang, and X. Zhang, "Efficient stereo matching leveraging deep local and context information," *IEEE Access*, vol. 5, pp. 18 745–18 755, 2017.

[50] S. Tulyakov, A. Ivanov, and F. Fleuret, "Weakly supervised learning of deep metrics for stereo reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1339–1348.

[51] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2758–2766.

[52] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2002–2011.

[53] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[54] P. Fischer, A. Dosovitskiy, and T. Brox, "Descriptor matching with convolutional neural networks: A comparison to sift," 2014, *arXiv:1405.5769*.

[55] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[56] A. Seki and M. Pollefeys, "SGM-Nets: Semi-global matching with neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 21–26.

[57] J. L. Schönberger, S. N. Sinha, and M. Pollefeys, "Learning to fuse proposals from multiple scanline optimizations in semi-global matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 739–755.

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

[59] X. Hu and P. Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2121–2133, Nov. 2012.

[60] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 364–375.

[61] A. Kendall *et al.*, "End-to-end learning of geometry and context for deep stereo regression," *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 66–75.

[62] J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, vol. 7, pp. 878–886.

[63] Z. Liang *et al.*, "Learning for disparity estimation through feature constancy," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2811–2820.

[64] J. Chang and Y. Chen, "Pyramid stereo matching network," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5410–5418.

[65] G.-Y. Nie *et al.*, "Multi-level context ultra-aggregation for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3283–3291.

[66] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "DeepStereo: Learning to predict new views from the world's imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5515–5524.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[68] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "SegStereo: Exploiting semantic information for disparity estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 636–651.

[69] R. Chabra, J. Straub, C. Sweeney, R. Newcombe, and H. Fuchs, "StereoDRNet: Dilated residual StereoNet," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11 786–11 795.

[70] C.-W. Xie, H.-Y. Zhou, and J. Wu, "Vortex pooling: Improving context representation in semantic segmentation," 2018, *arXiv: 1804.06242*.

[71] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2403–2412.

[72] Y. Zhang *et al.*, "ActiveStereoNet: End-to-end self-supervised learning for active stereo systems," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 784–801.

[73] Y. Wang *et al.*, "Anytime stereo image depth estimation on mobile devices," in *Proc. Int. Conf. Robot. Autom.*, 2019, pp. 5893–5900.

[74] P. Knöbelreiter, C. Reinbacher, A. Shekhovtsov, and T. Pock, "End-to-end training of hybrid CNN-CRF models for stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1456–1465.

[75] S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi, "StereoNet: Guided hierarchical refinement for real-time edge-aware depth prediction," *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 596–613.

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

[77] Y. Zhong, Y. Dai, and H. Li, "Self-supervised learning for stereo matching with self-improving ability," 2017, *arXiv: 1709.00930*.

[78] Z. Jie *et al.*, "Left-right comparative recurrent model for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3838–3846.

[79] L. Yu, Y. Wang, Y. Wu, and Y. Jia, "Deep Stereo Matching with Explicit Cost Aaggregation Sub-architecture," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7517–7524.

[80] X. Song, X. Zhao, H. Hu, and L. Fang, "EdgeStereo: A context integrated residual pyramid network for stereo matching," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 20–35.

[81] S. Tulyakov, A. Ivanov, and F. Fleuret, "Practical deep stereo (PDS): Toward applications-friendly deep stereo matching," *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 5871–5881.

[82] S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtasun, "DeepPruner: Learning efficient stereo matching via differentiable PatchMatch," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4384–4393.

[83] E. Ilg, T. Saikia, M. Keuper, and T. Brox, "Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 614–630.
[84] C. Chen, X. Chen, and H. Cheng, "On the over-smoothing problem of CNN based disparity estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8997–9005.
[85] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3273–3282.
[86] Z. Wu, X. Wu, X. Zhang, S. Wang, and L. Ju, "Semantic stereo matching with pyramid cost volumes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7484–7493.
[87] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "GA-net: Guided aggregation net for end-to-end stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 185–194.
[88] Z. Yin, T. Darrell, and F. Yu, "Hierarchical discrete distribution decomposition for match density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6044–6053.
[89] S. Zheng *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1529–1537.
[90] Y. Xue *et al.*, "MVSCRF: Learning multi-view stereo with conditional random fields," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4312–4321.
[91] D. Paschalidou, O. Ulusoy, C. Schmitt, L. Van Gool, and A. Geiger, "RayNet: Learning volumetric 3D reconstruction with ray potentials," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3897–3906.
[92] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5525–5534.
[93] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 767–783.
[94] J.-H. Lee, M. Heo, K.-R. Kim, and C.-S. Kim, "Single-image depth estimation based on fourier domain analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 330–339.
[95] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.
[96] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 109–117.
[97] X. Sun, X. Mei, S. Jiao, M. Zhou, Z. Liu, and H. Wang, "Real-time local stereo via edge-aware disparity propagation," *Pattern Recognit. Lett.*, vol. 49, pp. 201–206, 2014.
[98] S. Liu, S. De Mello, J. Gu, G. Zhong, M.-H. Yang, and J. Kautz, "Learning affinity via spatial propagation networks," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1520–1530.
[99] X. Cheng, P. Wang, and R. Yang, "Depth estimation via affinity learned with convolutional spatial propagation network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 103–119.
[100] S. Imran, Y. Long, X. Liu, and D. Morris, "Depth coefficients for depth completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12438–12447.
[101] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, 2009, Art. no. 24.
[102] A. Seki and M. Pollefeys, "Patch based confidence prediction for dense disparity map," in *Proc. Brit. Mach. Vis. Conf.*, 2016, vol. 2, Art. no. 4.
[103] S. Gidaris and N. Komodakis, "Detect, replace, refine: Deep structured prediction for pixel wise labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5248–5257.
[104] R. Haeusler, R. Nair, and D. Kondermann, "Ensemble learning for confidence measures in stereo vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 305–312.
[105] A. Spyropoulos, N. Komodakis, and P. Mordohai, "Learning to detect ground control points for improving the accuracy of stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1621–1628.
[106] M.-G. Park and K.-J. Yoon, "Leveraging stereo matching with learning-based confidence measures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 101–109.
[107] M. Poggi and S. Mattoccia, "Learning from scratch a confidence measure," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 46.1–46.13.
[108] A. S. Wannenwetsch, M. Keuper, and S. Roth, "ProbFlow: Joint optical flow and uncertainty estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1173–1182.
[109] K. Batsos, C. Cai, and P. Mordohai, "CBMV: A coalesced bidirectional matching volume for disparity estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2060–2069.
[110] M. Poggi, F. Tosi, and S. Mattoccia, "Quantitative evaluation of confidence measures in a machine learning world," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5228–5237.
[111] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 802–810.
[112] M. Poggi and S. Mattoccia, "Learning to predict stereo reliability enforcing local consistency of confidence maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2452–2461.
[113] J. Gast and S. Roth, "Lightweight probabilistic deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3369–3378.
[114] F. Tosi, M. Poggi, A. Benincasa, and S. Mattoccia, "Beyond local reasoning for stereo confidence estimation with deep learning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 319–334.
[115] Y. Hou, J. Kannala, and A. Solin, "Multi-view stereo by temporal nonparametric fusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2651–2660.
[116] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang, "SurfaceNet: An end-to-end 3D neural network for multiview stereopsis," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2307–2315.
[117] S. Choi, S. Kim, K. Park, and K. Sohn, "Learning descriptor, confidence, and depth estimation in multi-view stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 276–282.
[118] V. Leroy, J.-S. Franco, and E. Boyer, "Shape reconstruction using volume sweeping and learned photoconsistency," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 781–796.
[119] K. Luo, T. Guan, L. Ju, H. Huang, and Y. Luo, "P-MVSNet: Learning patch-wise matching confidence aggregation for multi-view stereo," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 10 452–10 461.
[120] K. Wang and S. Shen, "MVDepthNet: Real-time multiview depth estimation neural network," in *Proc. Int. Conf. 3D Vis.*, 2018, pp. 248–257.
[121] B. Ummenhofer *et al.*, "DeMoN: Depth and motion network for learning monocular stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, vol. 5, pp. 5622–5631.
[122] J. T. Barron, "A general and adaptive robust loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4331–4339.
[123] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros, "Learning dense correspondence via 3D-guided cycle consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 117–126.
[124] A. Ahmadi and I. Patras, "Unsupervised convolutional neural networks for motion estimation," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 1629–1633.
[125] J. Y. Jason, A. W. Harley, and K. G. Derpanis, "Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 3–10.
[126] M. Bai, W. Luo, K. Kundu, and R. Urtasun, "Exploiting semantic information and deep matching for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 154–170.
[127] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, vol. 2, pp. 6612–6619.
[128] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[129] A. Tonioni, M. Poggi, S. Mattoccia, and L. Di Stefano, "Unsupervised adaptation for deep stereo," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1614–1622.
[130] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6647–6655.
[131] C. Zhou, H. Zhang, X. Shen, and J. Jia, "Unsupervised learning of stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1567–1575.
[132] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki, "SfM-net: Learning of structure and motion from video," 2017, *arXiv: 1704.07804*.

[133] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, vol. 2, pp. 6602–6611.
[134] M. Perriollat, R. Hartley, and A. Bartoli, "Monocular template-based reconstruction of inextensible surfaces," *Int. J. Comput. Vis.*, vol. 95, no. 2, pp. 124–137, 2011.
[135] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "GeoNet: Geometric neural network for joint depth and surface normal estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 283–291.
[136] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.
[137] J. Pang *et al.*, "Zoom and learn: Generalizing deep stereo matching to novel domains," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2070–2079.
[138] A. Tonioni, M. Poggi, S. Mattoccia, and L. Di Stefano, "Unsupervised domain adaptation for depth prediction from images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2396–2409, Oct. 2020.
[139] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Weakly-supervised transfer for 3D human pose estimation in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, p. 7.
[140] Z. Zhang, C. Xu, J. Yang, Y. Tai, and L. Chen, "Deep hierarchical guidance and regularization learning for end-to-end depth estimation," *Pattern Recognit.*, vol. 83, pp. 430–442, 2018.
[141] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "Towards real-time unsupervised monocular depth estimation on CPU," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 5848–5854.
[142] Y. Zhong, H. Li, and Y. Dai, "Open-world stereo video matching with deep RNN," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 101–116.
[143] A. Tonioni, O. Rahnama, T. Joy, L. D. Stefano, T. Ajanthan, and P. H. Torr, "Learning to adapt for stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9661–9670.
[144] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.
[145] A. Atapour-Abarghouei and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2800–2810.
[146] C. Zheng, T.-J. Cham, and J. Cai, "T2Net: Synthetic-to-realistic translation for solving single-image depth estimation tasks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 767–783.
[147] S. Zhao, H. Fu, M. Gong, and D. Tao, "Geometry-aware symmetric domain adaptation for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9788–9798.
[148] J. Nath Kundu, P. Krishna Uppala, A. Pahuja, and R. Venkatesh Babu, "AdaDepth: Unsupervised content congruent adaptation for depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2656–2665.
[149] T. Saikia, Y. Marrakchi, A. Zela, F. Hutter, and T. Brox, "AutoDispNet: Improving disparity estimation with AutoML," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1812–1823.
[150] F. Hutter, L. Kotthoff, and J. Vanschoren, in *Automated Machine Learning-Methods, Systems, Challenges*. Berlin, Germany: Springer, 2019.
[151] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable architecture search," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 3–16.
[152] Robust vision challenge, 2020. Accessed: May 6, 2020. [Online]. Available: http://www.robustvision.net/
[153] X. Han, H. Laga, and M. Bennamoun, "Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2019.2954885.
[154] K. O. Stanley, J. Clune, J. Lehman, and R. Miikkulainen, "Designing neural networks through neuroevolution," *Nat. Mach. Intell.*, vol. 1, no. 1, pp. 24–35, 2019.
[155] G. Bingham, W. Macke, and R. Miikkulainen, "Evolutionary optimization of deep learning activation functions," in *Proc. Genetic Evol. Comput. Conf.*, 2020, pp. 289–296. [Online]. Available: https://doi.org/10.1145/3377930.3389841
[156] B. Haefner, Z. Ye, M. Gao, T. Wu, Y. Queau, and D. Cremers, "Variational uncalibrated photometric stereo under general lighting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8538–8547.
[157] Q. Zheng, Y. Jia, B. Shi, X. Jiang, L.-Y. Duan, and A. C. Kot, "SPLINE-net: Sparse photometric stereo through lighting interpolation and normal estimation networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8548–8557.

Hamid Laga received the MSc and PhD degrees in computer science from the Tokyo Institute of Technology, Japan, in 2003 and 2006, respectively. He is currently an associate professor at Murdoch University, Australia, and an adjunct associate professor with the Phenomics and Bioinformatics Research Centre (PBRC) of the University of South Australia, Australia. His research interests span various fields of machine learning, computer vision, computer graphics, and pattern recognition, with a special focus on the 3D reconstruction, modeling and analysis of static and deformable 3D objects, and on image analysis and big data in agriculture and health. He is the recipient of the best paper awards at SGP2017, DICTA2012, and SMI2006.

Laurent Valentin Jospin received the bachelor's and master's degrees at EPFL, Switzerland. He is currently working toward the PhD degree in computer science at the University of Western Australia, Australia, in 2019. He is a PhD research student in the field of deep learning for computer vision. His main research interests include 3D reconstruction, sampling and image acquisition strategies, computer vision applied to robotic navigation, computer vision applied to environmental sciences, and Bayesian statistics applied to computer vision. His research career started as an intern in two EPFL labs, publishing his first three papers in the process. His thesis project focus on real time 3D reconstruction with different computer vision techniques.

Faird Boussaid received the MS and PhD degrees in microelectronics from the National Institute of Applied Science (INSA), Toulouse, France, in 1996 and 1999 respectively. He joined Edith Cowan University, Perth, Australia, as a postdoctoral research fellow, and a member of the Visual Information Processing Research Group, in 2000. He joined the University of Western Australia, Crawley, Australia, in 2005, where he is currently a professor. His current research interests include neuromorphic engineering, smart sensors, and machine learning.

Mohammed Bennamoun (Senior Member, IEEE) is Winthrop professor with the Department of Computer Science and Software Engineering, University of Western Australia (UWA), Australia and is a researcher in computer vision, machine/deep learning, robotics, and signal/speech processing. He has published four books (available on Amazon), one edited book, one Encyclopedia article, 14 book chapters, more than 150 journal papers, more than 260 conference publications, 16 invited and keynote publications. His h-index is 77 and his number of citations is more than 13,500 (Google Scholar). He was awarded more than 70 competitive research grants, from the Australian Research Council, and numerous other Government, UWA and industry Research Grants. He successfully supervised more than 26 PhD students to completion. He won the Best Supervisor of the Year Award at Queensland University of Technology (1998), and received award for research supervision at UWA (2008 and 2016) and Vice-Chancellor Award for mentorship (2016). He delivered conference tutorials at major conferences, including: IEEE CVPR 2016, Interspeech 2014, IEEE ICASSP, and ECCV. He was also invited to give a Tutorial at an International Summer School on Deep Learning (DeepLearn 2017).