



# A Survey on Deep Stereo Matching in the Twenties

Fabio Tosi<sup>1</sup> · Luca Bartolomei<sup>1</sup> · Matteo Poggi<sup>1</sup>

Received: 1 August 2024 / Accepted: 11 December 2024 / Published online: 26 February 2025  
© The Author(s) 2025

## Abstract

Stereo matching is close to hitting a half-century of history, yet witnessed a rapid evolution in the last decade thanks to deep learning. While previous surveys in the late 2010s covered the first stage of this revolution, the last five years of research brought further ground-breaking advancements to the field. This paper aims to fill this gap in a two-fold manner: first, we offer an in-depth examination of the latest developments in deep stereo matching, focusing on the pioneering architectural designs and groundbreaking paradigms that have redefined the field in the 2020s; second, we present a thorough analysis of the critical challenges that have emerged alongside these advances, providing a comprehensive taxonomy of these issues and exploring the state-of-the-art techniques proposed to address them. By reviewing both the architectural innovations and the key challenges, we offer a holistic view of deep stereo matching and highlight the specific areas that require further investigation. To accompany this survey, we maintain a regularly updated project page that catalogs papers on deep stereo matching in our Awesome-Deep-Stereo-Matching repository.

**Keywords** Stereo matching · Machine learning · Deep learning

## 1 Introduction

Stereo matching—the task of estimating dense disparity maps from a pair of rectified images—has been a fundamental problem in computer vision for nearly half a century, playing a crucial role in a wide range of applications, such as autonomous driving, robotics, and augmented reality. After about twenty-five years of hand-designed stereo algorithms Scharstein and Szeliski (2002), the use of end-to-end deep neural networks has become the dominant paradigm in the late 2010s.

Existing surveys Poggi et al. (2021); Laga et al. (2020) have offered valuable insights on this rapid revolution, categorizing end-to-end architectures into 2D and 3D classes

according to their cost-volume computation and optimization strategies, while also emphasizing the challenges remaining open.

Since then, however, the domain has progressed rapidly, with the emergence of novel methods and paradigms inspired by breakthroughs in other areas of deep learning. Transformer-based Li et al. (2021) and iterative refinement Lipson et al. (2021) architectures are prime examples of how the field has evolved, demonstrating the potential for further improvements in accuracy and efficiency.

Despite the remarkable achievements, multiple challenges have emerged as deep stereo matching has advanced. One of the most critical issues, already highlighted in previous surveys, is the lack of generalization, particularly when facing domain shifts between synthetic and real data. Although synthetic datasets have been crucial for pre-training deep stereo networks, these models often perform poorly when applied to real scenes without fine-tuning. Recognizing the importance of this issue, several techniques have been proposed in recent years to improve zero-shot generalization and to develop domain adaptation methods for seamless adaptation to unknown target domains.

Alongside the domain-shift problem, deep stereo matching has overcome several other limitations. The tendency of deep networks to over-smooth depth at object boundaries, leading to inaccurate 3D reconstructions, has been

---

Communicated by D. Scharstein.

---

✉ Fabio Tosi  
fabio.tosi5@unibo.it  
https://fabiotosi92.github.io/

Luca Bartolomei  
luca.bartolomei5@unibo.it  
https://bartn8.github.io/

Matteo Poggi  
m.poggi@unibo.it  
https://mattpoggi.github.io/

<sup>1</sup> Department of Computer Science and Engineering,  
University of Bologna, Bologna, Italy

a persistent challenge. The increasing diversity of camera setups and the need to process images with different resolutions has underscored the importance of developing flexible algorithms. The demand for high-resolution and highly detailed disparity estimation, combined with the necessity for real-time performance on resource-constrained devices, has further complicated the task. These challenges have led to the development of more robust, versatile, and efficient deep stereo models that effectively address these limitations and improve their practical application.

Furthermore, the integration of complementary imaging modalities, including depth sensors, non-visible spectrum cameras, event cameras, and gated cameras, has opened new avenues for enhancing the robustness and accuracy of stereo matching in challenging environments. By leveraging the strengths and complementary nature of these modalities, multimodal stereo matching techniques aim to extend the applicability and reliability of stereo vision to a broader range of real-world scenarios, overcoming the limitations of relying solely on visible-spectrum cameras.

Driven by these significant developments, this survey provides a thorough and up-to-date review of the latest advancements in deep stereo matching that have emerged during the 2020s. We cover a wide range of topics explored since the last surveys Poggi et al. (2021); Laga et al. (2020), including novel architectures, multi-modal approaches, and cutting-edge techniques designed to address the aforementioned critical challenges. Our overview comprehensively analyzes over 100 distinct contributions to deep stereo matching, all of which have been presented at top conferences and published in prestigious journals. Our goal is to offer a thorough overview of the state-of-the-art, discussing the progress made since previous surveys and highlighting the current trends and future directions in the field.

The remaining sections are organized as follows:

- Sect. 2 covers the latest developments in deep stereo matching frameworks, highlighting novel architectural paradigms, computational efficiency techniques, and specialized multi-modality designs.
- Sect. 3 explores key challenges in deep stereo matching and investigates recent approaches to address them.
- Sect. 4 reports quantitative results on the main online benchmarks, both established and recently introduced, demonstrating the progress made in the field over the last few years.
- Sects. 5 and 6 focus on potential future research directions and provide a summary of the survey.

## 1.1 Background

To better understand the latest trends in the field, we introduce the fundamentals of deep stereo matching that have

been driving advancements until the late 2010s. In the interest of space, we recommend referring to the appendix for a detailed overview of the background information. This additional resource also provides detailed descriptions of the datasets commonly used in this context, as well as the primary evaluation metrics employed in this domain. Finally, for a more comprehensive overview and detailed descriptions of the body of research that arose before 2020, readers can refer to existing surveys related to deep stereo matching, such as Poggi et al. (2021); Laga et al. (2020); Poggi et al. (2021, 2017).

## 2 Deep Stereo Frameworks

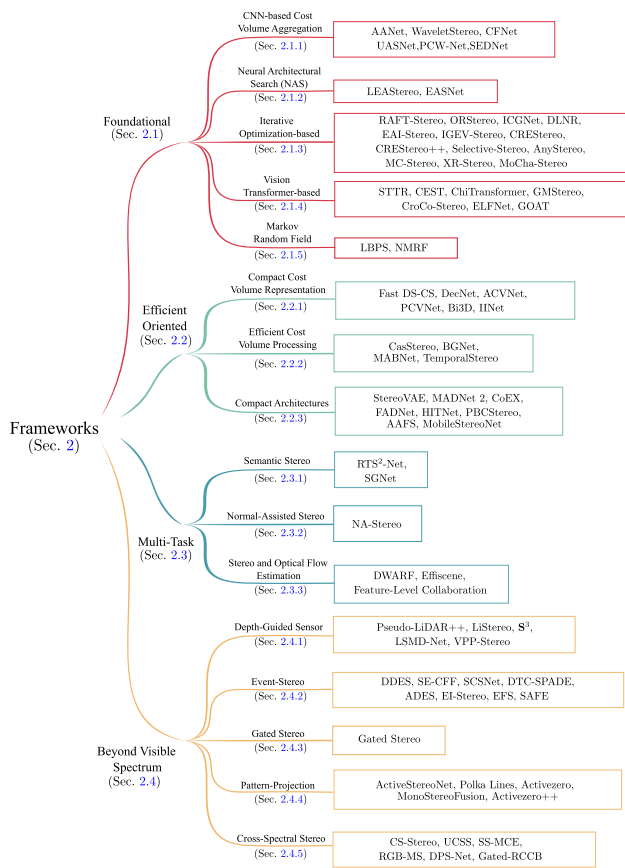
### 2.1 Foundational Deep Stereo Architectures

This section categorizes and analyzes classical deep stereo models that emerged in recent years. These are grouped into five categories: CNN-based cost volume aggregation (Sec. 2.1.1), Neural Architecture Search (NAS) for stereo matching (Sec. 2.1.2), iterative optimization-based (Sec. 2.1.3), Vision Transformer-based (Sec. 2.1.4), and Markov random field-based architectures (Sec. 2.1.5). These represent the primary trends and techniques that have significantly advanced the state-of-the-art of stereo depth estimation. The taxonomy depicted in Fig. 1 summarizes these categories.

#### 2.1.1 CNN-Based Cost Volume Aggregation

Here, we cover recent deep stereo methods using established CNN-based cost volume aggregation techniques Poggi et al. (2021); Laga et al. (2020), which broadly fall into two categories: 2D and 3D architectures, distinguished by their strategy for encoding features and geometry. Both construct a cost volume from the left and right input images - 2D architectures typically employ a correlation layer on the extracted features, while 3D architectures concatenate or compute the feature difference over the full disparity range. This cost volume is then processed via plain 2D convolutions in 2D encoder-decoder architectures, or through 3D convolutions in 3D architectures that explicitly encode geometry.

Among them, **AA<sub>Net</sub>**[Code] Xu and Zhang (2020) aims to replace computationally expensive 3D convolutions while maintaining high accuracy by introducing a sparse points based intra-scale cost aggregation method using deformable convolutions to address the edge-fattening issue at disparity discontinuities, and an approximation of traditional cross-scale cost aggregation using neural network layers to handle large textureless regions. **WaveletStereo** Yang et al. (2020) takes a different approach by learning wavelet coefficients of the disparity map instead of directly estimating



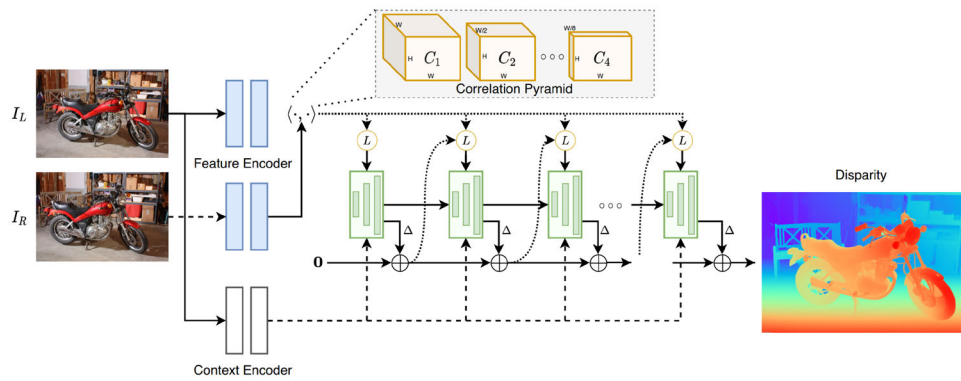
**Fig. 1** A taxonomy of deep stereo frameworks in the 2020s. We categorize the reviewed methods based on their key designs and paradigms (Color figure online)

disparity. It comprises low-frequency and high-frequency sub-modules to handle smooth and detailed regions, respectively, and reconstructs the disparity map iteratively by mapping multi-resolution cost volumes to wavelet coefficients through convolutional networks and inverse wavelet transforms. **CFNet** [Code] Shen et al. (2021), on the other hand, combines a fused cost volume representation, which fuses multiple low-resolution cost volumes to capture global and structural information, and a cascade cost volume representation, which adaptively adjusts the disparity search range at each stage using variance-based uncertainty estimation. **UASNet** Mao et al. (2021) constructs cascade 3D cost volumes with improved disparity range prediction and effective sampling by introducing an uncertainty distribution-guided range prediction (URP) module to handle ambiguities and an uncertainty-based disparity sampler (UDS) module that discretizes the per-pixel predicted disparity range based on the matching uncertainty. **PCW-Net** [Code] Shen et al. (2022) employs a multi-scale cost volume fusion module to construct combination volumes on the upper levels of the pyramid and integrate them for initial disparity estimation, covering multi-scale receptive fields and extracting

domain-invariant structural cues. It also introduces a warping volume-based disparity refinement module to narrow down the residue searching range from the initial disparity searching range to a fine-grained one. **SEDNet** [Code] Chen et al. (2023) focuses on joint disparity and uncertainty estimation by extending the stereo backbone GWCNet Guo et al. (2019) with a lightweight uncertainty estimation subnetwork. It matches the distribution of disparity errors with the distribution of uncertainty estimates using a KL divergence-based loss function enabled by a differentiable histogramming scheme, ensuring that the uncertainty estimates follow the same distribution as the true errors of the disparity estimator.

### 2.1.2 Neural Architecture Search for Stereo Matching

Neural architecture search (NAS) aims to automate the design of neural architectures, alleviating the manual effort required by human experts. While NAS has achieved success in high-level vision tasks like classification and detection, applying it to the dense prediction problem of stereo matching is challenging. State-of-the-art deep stereo networks are computationally intensive, making the naïve application of traditional NAS methods prohibitively expensive due to the vast search space. To overcome this limitation, recent works have proposed tailored NAS frameworks that incorporate task-specific priors from the stereo matching pipeline. The key motivation is to enable efficient architecture search for this domain while maintaining high accuracy, without solely relying on manually designed architectures. **LEAStereo** [Code] Cheng et al. (2020) is the first end-to-end hierarchical NAS framework for stereo that incorporates task-specific human knowledge. It consists of a 2D feature net for local image feature extraction and a 3D matching net for computing and aggregating matching costs from a 4D feature volume. The cell-level structure of these sub-networks is searched using a novel residual cell design and tailored candidate operations. At the network level, the search explores the arrangement of cells and the size of intermediate feature maps and volumes. This hierarchical approach, which embeds geometric priors, leads to accurate stereo matching models without relying on human-designed architectures. Building upon **LEAStereo**, **EASNet** [Code] Wang et al. (2022) addresses its limitations, such as high computational costs and limited scalability, by introducing a stereo network that can be deployed on devices with varying computing resources. The architecture search space covers depth, width, kernel size, and scale, allowing the network to adapt to different resource constraints. A progressive shrinking training strategy enables efficient optimization, allowing sub-networks with diverse configurations to be extracted without additional training while maintaining high accuracy. This is



**Fig. 2** RAFT-Stereo Lipson et al. (2021) architecture. It constructs a correlation pyramid from correlation features (blue) extracted from each image. A context encoder extracts “context” image features (white) and an initial hidden state. The disparity field starts at zero. In each iteration, the GRU(s) (green) sample from the correlation pyramid using the

current disparity estimate. Sampled correlation features, initial image features, and current hidden state(s) are processed by the GRU(s) to update the hidden state and disparity estimate. Picture from Lipson et al. (2021)

a significant advantage over LEAStereo, which requires re-searching and re-training for new deployment scenarios.

### 2.1.3 Iterative Optimization-Based Architectures

Iterative optimization-based approaches have emerged as a promising paradigm for stereo matching, offering significant advantages over existing deep cost aggregation methods. Inspired by the success of iterative refinement in optical flow estimation, particularly the RAFT architecture Teed and Deng (2020), these methods have been adapted to stereo matching with impressive results. By bypassing explicit cost aggregation and instead iteratively updating a disparity estimate using a high-resolution cost volume as guidance – i.e., the cost volume is computed only once, without further processing it thereafter– these methods address key limitations of conventional approaches. Specifically, iterative stereo enables the direct use of high-resolution cost volumes without the computational burden of 3D convolutions, making it applicable to high-resolution images. Moreover, by avoiding a predefined disparity range, it can handle a wide range of disparities without sacrificing accuracy or efficiency. In fact, the recurrent nature of the update operator allows for a flexible trade-off between accuracy and speed through a variable number of iterations.

**RAFT-Stereo** [Code] Lipson et al. (2021), a deep stereo architecture inspired by the optical flow network RAFT Teed and Deng (2020), has been a game-changer for recent stereo solutions. By introducing novel principles, illustrated in Fig. 2, such as lightweight cost volumes, and iterative refinement, RAFT-Stereo has paved the way for more efficient and effective processing of high-resolution stereo pairs. The architecture consists of three main components: a feature extractor, a correlation pyramid, and a Convolutional Gated Recurrent Units (ConvGRU)-based update operator.

The feature extractor comprises a feature encoder, applied to both images to generate dense feature maps, and a context encoder, applied only to the left image to initialize the update operator and inject context features. The correlation pyramid is constructed without predefining the range of disparities by computing the visual similarity between feature vectors, restricting the computation to pixels sharing the same y-coordinate, resulting in a 3D cost volume:

$$\mathbf{C}_{ijk} = \sum_h \mathbf{f}_{ijh} \cdot \mathbf{g}_{ikh}, \quad \mathbf{C} \in \mathbb{R}^{H \times W \times W} \quad (1)$$

Here,  $\mathbf{C}$  represents the cost volume,  $\mathbf{f}$  and  $\mathbf{g}$  are the dense feature maps extracted from the left and right images, respectively. This lightweight cost volume is computed using a single matrix multiplication, making it computationally efficient. The update operator iteratively retrieves local cost values from All-Pairs Correlations (APC) and updates the disparity field using multi-resolution ConvGRUs with cross-connections. The multi-level GRUs efficiently propagate information across the image, improving the global consistency of the disparity field. This iterative refinement process allows for a trade-off between accuracy and efficiency through early stopping. The final disparity is upsampled to the original resolution using a convex upsampling method. Building upon the principles introduced by RAFT-Stereo, several methods have proposed enhancements and adaptations to address specific challenges. Among them, **ORStereo** Hu et al. (2021) addresses the challenge of training stereo matching models on high-resolution images with large disparity ranges by proposing a two-phase approach. It estimates a down-sampled disparity map and occlusion mask in the first phase, followed by patch-wise refinement using a recurrent residual updater (RRU) and a normalized local refinement (NLR) module in the second phase.

The RRU recurrently updates the disparity and occlusion predictions, while the NLR employs normalization techniques to generalize to unseen disparity ranges. Building upon the principles of iterative refinement, **CREStereo** [Code] Li et al. (2022) introduces a hierarchical network with recurrent refinement and an adaptive group correlation layer (AGCL) for robust matching. The AGCL computes correlations in local search windows, reducing memory and computation requirements, and incorporates deformable search windows and group-wise correlations. In contrast, **EAI-Stereo** [Code] Zhao et al. (2022) focuses on error correction by incorporating an error-aware refinement module that combines left-right warping with learning-based upsampling, allowing the network to learn error correction capabilities. **IGEV-Stereo** [Code] Xu et al. (2023) takes a different approach by building a Combined Geometry Encoding Volume (CGEV) that encodes geometry, context information, and local matching details. It combines a Geometry Encoding Volume (GEV), obtained through lightweight 3D regularization of the cost volume, with All-Pairs Correlations to provide both non-local geometry and local matching details. **DLNR** [Code] Zhao et al. (2023), on the other hand, focuses on preserving high-frequency information during the iterative process by incorporating a Decouple LSTM module, a Normalization Refinement module, and a multi-scale, multi-stage Channel-Attention Transformer feature extractor. **CREStereo++** Jing et al. (2023) extends CREStereo by introducing an uncertainty guided adaptive correlation (UGAC) module to enhance robustness and accuracy. It employs a content-aware warping layer with learnable offsets guided by variance-based uncertainty estimation, adjusting the sampling range when building the cost volume to capture more potential correspondences in ill-posed regions. **Selective-Stereo** [Code] Wang et al. (2024) addresses the fixed receptive field limitation by proposing a Selective Recurrent Unit (SRU) with multiple GRU branches to adaptively capture and fuse multi-frequency information, and a Contextual Spatial Attention (CSA) module to guide the adaptive fusion based on image regions. **Any-Stereo** [Code] Liang and Li (2024) focuses on maintaining the regularized disparity at a higher resolution by modeling the disparity as a continuous representation over 2D spatial coordinates using an Implicit Neighbor Mask Function (INMF). It proposes Intra-scale Similarity Unfolding (ISU) and Cross-scale Feature Alignment (CFA) strategies to complement missing information and details in the latent code. **XR-Stereo** [Code] Cheng et al. (2024), instead, presents an architecture that leverages temporal information in video streams. It employs a RAFT-style scheme, where the key idea is to warp the previous frame's disparity and hidden state to the current frame using camera poses, serving as a warm-start for the GRU.

**MC-Stereo** [Code] Feng et al. (2024) addresses the limitations of existing approaches in handling multi-peak

distribution and fixed search range by introducing a multi-peak lookup strategy and a cascade search range during the iterative process. **MoCha-Stereo** [Code] Chen et al. (2024) tackles the loss of geometric edge details by capturing repeated geometric contours across channels as motif channels that preserve common geometric structures. It employs a Motif Channel Attention (MCA) mechanism, a Motif Channel Correlation Volume (MCCV), and a Reconstruction Error Motif Penalty (REMP) module to recover lost geometric edge details through an iterative process. Finally, **ICGNet** [Code] Gong et al. (2020) builds over existing architectures such as IGEV-Stereo and introduces additional intra-view and cross-view geometry constraints, by extracting keypoints with an off-the-shelf detector and enforcing the image features extractor to predict the very same keypoints (intra-view), as well as by forcing the very correspondences between left and right image features as those occurring between left and right keypoints processed by an off-the-shelf matcher (cross-view).

#### 2.1.4 Vision Transformer-Based Architectures

In recent years, Vision Transformers (ViT) Dosovitskiy et al. (2020) have emerged as a promising alternative to CNNs for various computer vision tasks, including stereo matching. Transformers, originally developed for natural language processing Vaswani et al. (2017), have demonstrated remarkable performance in capturing long-range dependencies and global context information. These methods move away from traditional cost-volume construction and instead formulate stereo matching as a sequence-to-sequence problem, utilizing attention mechanisms to establish correspondences between pixels in the left and right images. Key components of vision transformer-based stereo matching include self-attention, cross-attention, and positional encoding schemes that provide spatial cues.

One of the pioneering works in this area is the STereo TRansformer (**STTR**) [Code] Li et al. (2021), which introduces a Transformer architecture that, unlike traditional cost-volume construction, relaxes the fixed disparity range constraint and alternates between self-attention along epipolar lines and cross-attention between left-right pairs. This approach allows the network to capture long-range pixel associations and resolve ambiguities in matching regions. Moreover, the architecture incorporates a relative positional encoding scheme to provide discriminative spatial cues and enforces a uniqueness constraint during matching, ensuring one-to-one correspondence between pixels across the stereo pair.

Building upon STTR, subsequent works have focused on enhancing the contextual information and refining the disparity estimation process. [Code] Guo et al. (2022) introduces the Context Enhanced Path (CEP), a plug-in module that

extracts long-range context information from low-resolution features. By following a coarse-to-fine approach and employing attention masking, optimal transport for uniqueness constraints, and post-processing modules like upsampling and context adjustment, CEST achieves improved disparity map refinement. Similarly, **ChiTransformer** [Code] Su and Ji (2022) proposes a self-supervised method with an encoder-decoder transformer architecture, utilizing parallel ViT streams and depth cue rectification blocks to fuse binocular cues and produce depth estimates. While these methods focus on static stereo matching, **Dynamic-Stereo** [Code] Karaev et al. (2023) tackles the challenge of temporally consistent disparity estimation from stereo videos. By introducing a hybrid encoder-decoder network that combines transformer-based attention across space, view, and time with an iterative refinement approach similar to RAFT, Dynamic-Stereo achieves good results in capturing dynamic scenarios.

Some works have also explored transformers for unifying various dense correspondence tasks. For instance, **GMStereo** [Code] Xu et al. (2023) presents a unified model for optical flow, rectified and unrectified stereo matching tasks, formulating them as a dense correspondence matching problem. By exploiting cross-attention in a Transformer-based model and employing parameter-free task-specific matching layers (2D, 1D, or another form) to obtain dense correspondences, GMStereo demonstrates the versatility and effectiveness of transformers in handling multiple related tasks.

Pre-training has also emerged as a promising approach to further improve the performance of ViT-based stereo matching models. **CroCo-Stereo** [Code] Weinzaepfel et al. (2023) introduces a large-scale pre-training approach that leverages a cross-view completion pretext task to learn robust representations. The pre-trained model, CroCo v2, utilizes a ViT encoder and decoder architecture with self-attention and cross-attention mechanisms, and is then finetuned for stereo matching using a Dense Prediction Transformer (DPT) head, which directly processes the encoder and decoder features to predict the final disparity, in contrast to existing methods relying on cost volumes or iterative refinement.

Recognizing the complementary strengths of cost-volume-based and transformer-based approaches, **ELFNet** [Code] Lou et al. (2023) proposes a novel framework that fuses the two paradigms. By employing heads in each branch to estimate aleatoric and epistemic uncertainties using deep evidential learning, and introducing a two-stage fusion strategy, ELFNet effectively integrates local and global information based on the estimated uncertainties.

The progressive refinement of disparity estimates in occluded regions is another crucial aspect addressed by ViT-based models. **GOAT** [Code] Liu et al. (2024) introduces a parallel disparity and occlusion estimation module (PDO) and an iterative occlusion-aware global aggregation module (OGA). By exploiting restricted global spatial correlation

within a focus scope guided by the occlusion mask, GOAT achieves robust disparity refinement in occluded regions.

### 2.1.5 Markov Random Field-Based Architectures

Markov Random Fields (MRFs) have been widely used in traditional stereo matching methods before the advent of deep learning, thanks to their ability to reduce matching ambiguities in challenging regions by enforcing spatial coherence and smoothness constraints on the disparity map. However, traditional MRF models rely on hand-crafted potential functions and message-passing procedures, which often lead to suboptimal results. Recently, a new line of research has emerged that aims to leverage the power of deep learning to create fully data-driven MRF models for stereo. The first attempt in this direction is made by **LBPS** [Code] Knobelreiter et al. (2020). Specifically, LBPS adapted the Belief Propagation (BP) framework, in particular the truncated max-product Belief Propagation, by adding the necessary components to formulate it as a custom layer inside a deep learning model—a BP-layer. Purposely, differentiable losses on marginals were introduced, allowing for performing backpropagation through the BP-layer. Then, the authors built a stereo model using two UNets: the former processes the stereo pair to obtain dense features, used to compute a cost volume, and the latter processes the left image to retrieve pixel-wise scores. These two cues are forwarded to the BP-layer, in charge of refining the cost volume before computing the final disparity map. Lately, the Neural Markov Random Field (**NMRF**) model [Code] Guan et al. (2024) followed this path as well, by introducing a novel fully data-driven approach that consists of a local feature CNN, a Disparity Proposal Network (DPN), a Neural MRF inference stage, and a disparity refinement stage. The local feature CNN extracts multi-level features from the stereo pair, while the DPN prunes the search space by identifying the top  $k$  disparity modals for each pixel through neural message passing. The inference stage performs coarse-level disparity estimation using a probabilistic graphical model with self and neighbor edges, where the potential functions and message passing are learned through neural networks. Finally, the disparity refinement stage further improves the coarse estimates. The model is trained using a combination of proposal loss, initialization loss, and disparity loss, which measures the discrepancy between candidate labels, cost volume, and the ground truth disparity. By learning complex pixel relationships and potential functions directly from data while retaining the graph inductive bias of MRFs, NMRF opens up new possibilities for combining the strengths of both traditional MRFs and deep learning in stereo matching.

## 2.2 Efficiency-Oriented Architectures

While deep learning has enabled dramatic accuracy improvements over traditional stereo methodologies by training convolutional neural networks to learn powerful feature representations and cost volume filtering for better matching, many deep stereo networks are too computationally intensive for real-time operation, especially on embedded or mobile devices with strict power, memory, and computational constraints. In such cases, there is an important need to develop efficient deep neural architectures that can provide the accuracy benefits of learned approaches while operating at real-time speeds with low latency on low-power edge devices.

### 2.2.1 Compact Cost Volume Representations

Efficient stereo matching often requires compact cost volume representations to reduce memory and computational demands. Various approaches have been proposed to extract essential matching information while suppressing redundant data, enabling faster processing without significant accuracy degradation.

Among these frameworks, **Fast DS-CS** [Code] Yee and Chakrabarti (2020), unlike most existing stereo networks using expensive learned matching costs and 3D convolutions, uses traditional efficient matching costs to construct initial cost volumes and learns a mapping to convert the cost information at each pixel into a low-dimensional “cost signature” feature vector, which is then processed using an encoder-decoder network with 2D convolutions. In contrast, **DecNet** [Code] Yao et al. (2021) proposes a decomposition model that reduces the computational cost growth as resolution increases by decomposing stereo matching into a dense matching at very low resolution and a series of sparse matchings at higher resolutions, enabling logarithmic complexity growth. **ACVNet** [Code] Xu et al. (2022) introduces an attention-based filtering approach, leveraging correlation clues and a multi-level adaptive patch matching mechanism to generate the Attention Concatenation Volume (ACV). **PCVNet** [Code] Zeng et al. (2023), instead, proposes a parameterized cost volume representation that encodes the entire disparity space using a multi-Gaussian distribution, allowing for a global view of the disparity space and progressively focusing on local regions for fine-grained matching. In contrast, **Bi3D** [Code] Badki et al. (2020) formulates stereo matching as a collection of binary classification tasks, training a neural network to classify whether an object is closer or farther than a given depth plane, enabling selective depth estimation within a specific range of interest and offering a flexible trade-off between accuracy and computational efficiency. **IINet** Li et al. (2024) aims to address the redundancy in explicit 3D cost volumes by proposing a compact 2D implicit net-

work, incorporating intra-image context information, Fast Multi-scale Score Volume (FMSV), Confidence Based Filtering (CBF), and Intra-Inter Fusing (IIF) network with Residual Context-aware Upsamplers (RCUs) for enhanced information transmission. These methods employ different strategies, such as cost signatures, dense-sparse decomposition, attention-based filtering, parameterized representations, and implicit networks, to achieve compact yet informative cost volume representations.

### 2.2.2 Efficient Cost Volume Processing

Processing high-dimensional cost volumes can be a major bottleneck in real-time stereo matching. To address this, various techniques have been developed to efficiently process cost volumes, such as cascaded architectures, edge-aware upsampling, and multi-scale feature extraction.

Accordingly, **CasStereo** [Code] Gu et al. (2020) addresses this by proposing a cascade cost volume formulation, constructing a cost volume using a feature pyramid and iteratively narrowing down the depth range at each stage, recovering the output in a coarse-to-fine manner. In contrast, **BGNet** [Code] Xu et al. (2021) focuses on edge-preserving cost volume upsampling using a learned bilateral grid, enabling efficient upsampling while preserving sharp edges and allowing computationally expensive operations to be performed at lower resolutions without compromising accuracy. **MABNet** [Code] Xing et al. (2020) introduces the Multi-branch Adjustable Bottleneck (MAB) module for multi-level feature extraction, with the 2D MAB module having three branches with different dilation rates and the 3D MAB module factorizing 3D convolutions into disparity-wise and spatial convolutions, aiming for high accuracy with a small model size. **TemporalStereo** [Code] Zhang et al. (2023), instead, employs a coarse-to-fine architecture with sparse cost volumes, enriching them with multi-level context, statistical fusion for robust, global cost aggregation, and an adaptive shifting strategy that adjusts the disparity candidates based on the aggregated costs. Moreover, it can seamlessly operate in both single stereo pair and temporal modes, leveraging past information to boost matching accuracy with a negligible runtime increase.

### 2.2.3 Compact Architectures

Computing the cost volume is the primary contributor to computational complexity in end-to-end networks, but not the only one. Thus, designing an architecture that is compact in any part can further decrease complexity.

For instance, **StereoVAE** Chang et al. (2023) proposes a computationally efficient and lightweight framework combining a traditional matching algorithm and a variational autoencoder (VAE) based neural network. Initially, the for-

mer generates a coarse disparity map from a low-resolution stereo image pair, leveraging its low computational complexity. Subsequently, a tiny VAE-based network upscales and refines the coarse disparity map, taking the coarse map and the left image as inputs.

**MobileStereoNet** [Code] Shamsafar et al. (2022) proposes a 3D and a 2D lightweight stereo models to pursue efficiency. The former builds upon GwcNet-g Guo et al. (2019), while the latter uses 2D encoder-decoder to process a 3D cost volume, which is constructed efficiently using a new "Interlacing Cost Volume" module. Both models replace various components with 2D and 3D versions of MobileNet-V1 and MobileNet-V2 convolution blocks to further reduce complexity.

Bringing lightweight design to the extreme, **PBCStereo** Cai et al. (2022) is a fully binarized deep stereo network. To mitigate accuracy degradation due to quantization, PBCStereo introduces the Interpolation+Binary Convolution (IBC) module to replace binary deconvolutions and fuse shallow/deep features, and proposes the Binary Input Layer (BIL) coding method to binarize inputs while preserving pixel precision. PBCStereo incorporates IBC modules for any of feature extraction, cost aggregation, and disparity refinement stages, with all operations binarized, following a typical stereo pipeline optimized for efficiency. **FADNet** [Code] Wang et al. (2020) focuses on efficient 2D-based correlation layers with stacked residual blocks. It employs a DispNetC-based backbone extensively reformed with residual blocks, point-wise correlation layers, and multi-scale residual learning, adopting a coarse-to-fine training strategy with a loss weight scheduling technique. **HITNet** [Code] Tankovich et al. (2021) avoids building a full 3D cost volume and uses fast multi-resolution initialization. The disparity map is represented as planar tiles with learnable feature descriptors at multiple resolutions, and the network iteratively refines disparity hypotheses via differentiable 2D geometric propagation and warping, reasoning about slanted plane hypotheses for accurate warping and upsampling in coarse-to-fine manner. **CoEX** [Code] Bangunharcana et al. (2021), instead, pursues efficiency and accuracy through the Guided Cost volume Excitation (GCE) module, which utilizes extracted image features from the MobileNetV2 Sandler et al. (2018) backbone to excite relevant channels in the cost volume, improving feature extraction without significant computational overhead. CoEX also introduces top-k selection before soft-argmin disparity regression, computing the final disparity estimate using only the top-k matching cost values. **AAFS** [Code] Chang et al. (2020) combines an efficient backbone made of depthwise separable convolutions, an attention-aware feature aggregation (ACCV) module, and a cascaded 3D CNN architecture to achieve real-time stereo on edge devices. The ACCV module adaptively aggregates information from different scales of the feature

maps, enhancing their representational capacity, while the cascaded 3D CNN architecture regularizes cost volumes at multiple in coarse-to-fine manner. Lastly, **MADNet 2** [Code] Poggi and Tosi (2024) revises MADNet Tonioni et al. (2019) with the all-pairs correlation module from RAFT-Stereo, while removing its original context network. Combined with refined augmentation techniques, these modifications enable MADNet 2 to significantly outperform its predecessor while preserving computational efficiency.

## 2.3 Multi-task Deep Architectures

In this section, we focus on models combining stereo matching with other dense tasks within a multi-task framework.

### 2.3.1 Semantic Stereo Matching

Integrating semantic information into stereo matching has the potential to significantly improve the performance of depth estimation algorithms. By exploiting high-level semantic cues, these methods can better handle challenging scenarios such as textureless regions, occlusions, and depth discontinuities. Despite the many attempts Zhan et al. (2019); Yang et al. (2018); Jiang et al. (2019); Wu et al. (2019) reported in previous surveys Poggi et al. (2021), only few further works have emerged in the 20s, specifically targeting real-time stereo matching with semantic guidance. Among them, **RTS<sup>2</sup>Net** Dovesi et al. (2020) proposes a lightweight, real-time architecture for joint semantic segmentation and disparity estimation. Key aspects include a multi-stage, coarse-to-fine pyramidal decoder design for trading accuracy and speed, a shared encoder extracting common features for both tasks, separate semantic and disparity decoders operating on the shared features, and a synergy refinement module that exploits the complementarity between semantics and depth. **SGNet** Chen et al. (2020), instead, built upon PSM-Net Chang and Chen (2018), introduces three modules to embed semantic constraints: a confidence module computing the consistency between disparity and semantic features, a residual module optimizing the initial disparity map based on semantic categories, and a loss module supervising disparity smoothness based on semantic boundaries and regions.

### 2.3.2 Normal-Assisted Stereo Matching

These methods aim to exploit the geometric coherence between depth and surface normals to regularize the depth estimation process and enhance the overall performance. Indeed, the latter cue can serve as a strong prior to any depth estimation model, for instance by enforcing smoothness on the predictions in correspondence of large surfaces, either planar or not. This very same constraint can be imposed also on disparity regression. One notable example of this

approach is HITNet Tankovich et al. (2021), which achieves high accuracy by geometrically reasoning about disparities and inferring slanted plane hypotheses. Building upon a similar concept, **NA-Stereo** Kusupati et al. (2020) was proposed for generic multi-view stereo that can also be applied to the classic binocular stereo matching task. The key idea is to incorporate a normal estimation network (NNet) into the depth estimation framework, jointly optimizing for both depth and normals through a consistency loss. The method constructs a 3D cost volume by plane sweeping and accumulates multi-view image information, which is then regularized by NNet that predicts surface normals from the accumulated features.

### 2.3.3 Joint Stereo And Optical Flow

This family of frameworks processes stereo videos by jointly estimating three elements for each pair of consecutive left images: the disparity map for that image pair, the optical flow between the two left images, and optionally, the change in disparity between them. Predicting these three cues together allows for estimating the 3D scene flow of the observed scene Vedula et al. (1999). It is worth mentioning that, among the literature, several works directly estimate 3D scene flow starting from point clouds or pre-computed disparity maps Teed and Deng (2021); Liu et al. (2022). Here, we survey only those approaches estimating disparity maps as well. Among these, **DWARF** [Code] Aleotti et al. (2020) proposed a compact coarse-to-fine architecture estimating disparity, flow, and disparity change at multiple scales, by building compact correlation volumes exploiting features warping across the scales. Specifically, a 3D correlation layer is deployed across the correlation scores computed between two consecutive stereo pairs. On the contrary, **Effiscene** Jiao et al. (2021) tackles scene flow estimation in an unsupervised manner, by decomposing the task into predicting stereo depth, optical flow, camera poses, and motion segmentation. The four ingredients combined together allow for estimating the 3D scene flow of the scene in the absence of any annotation. Finally, **Feature-Level Collaboration** Chi et al. (2021) follows the same track and proposes an architecture for joint estimation of stereo depth and optical flow, followed by a pose decoder estimating the camera ego-motion from the previous two predictions.

## 2.4 Beyond Visible Spectrum Deep Stereo Matching

In recent years, there has been a significant increase in the use of modalities beyond conventional color cameras for various tasks. This section examines frameworks that perform stereo estimation by combining color imagery with data from external depth sensors or different types of cameras, or by operating entirely outside the visible RGB spectrum.

### 2.4.1 Depth-Guided Sensor Stereo Matching

Guided stereo matching aims to enhance the accuracy and robustness of stereo networks by leveraging sparse depth cues from external sensors, such as LiDAR. These depth hints, independent of visual appearance, help to overcome limitations in challenging scenarios like textureless regions and mitigate the impact of domain shift.

**Pseudo-LiDAR++** You et al. (2020) adapts a stereo network to directly estimate depth by constructing a depth cost volume and refines the estimates using a graph-based depth correction algorithm that propagates sparse LiDAR measurements. In contrast, **LiStereo** Zhang et al. (2020) employs a two-branch architecture, with a color image branch for stereo feature extraction and a LiDAR branch for processing sparse depth maps, which are then fused to output dense depth maps.

The Sparse Signal Superdensity ( $S^3$ ) Huang et al. (2021) framework, instead, addresses the challenges of low density and imbalanced distribution of sparse depth cues by expanding the depth estimates around sparse signals based on the RGB image and controlling their influence through confidence weighting. **LSMD-Net** Yin et al. (2022) fuses LiDAR and stereo information using a dual-branch disparity predictor, where the outputs from these branches are fused at each pixel using a mixture density module that models the estimated disparity at each pixel as a probability distribution using Laplacian distributions. The final disparity map is then determined by the expectation of the more confident branch.

Differently from the previous strategies, **VPP-Stereo** [Code] Bartolomei et al. (2023) uses depth measurements from an active sensor to hallucinate patterns onto the stereo images, thus simplifying visual correspondence. Various patterning strategies are proposed to improve performance. The hallucinated stereo pair is processed by any stereo matching algorithm (either learned or hand-crafted) to obtain dense depth estimations. Lastly, **SDG-Depth** [Code] Li et al. (2024) incorporates a sparse LiDAR deformable propagation module to generate a Semi-Dense hint Guidance map and a confidence map, which guide cost aggregation in stereo matching. Additionally, SDG-Depth introduces a learned disparity-depth conversion module to reduce triangulation errors.

### 2.4.2 Event-Camera-Based Stereo Matching

Event cameras Gallego et al. (2022), also known as neuro-morphic cameras, are peculiar sensors designed to mitigate the shortcomings of conventional imaging devices, such as limited dynamic range, sensor noise, and motion blur caused by rapid movements. In contrast to color cameras, which capture frames at fixed intervals, event cameras operate asynchronously. As will be discussed, the majority of these methods adapt successful models from traditional stereo lit-

erature Poggi et al. (2021) to this novel paradigm, primarily by developing appropriate techniques for handling asynchronous event streams.

Specifically, given a timestamp  $t_d$  at which disparity map estimation is desired, events are usually sampled backward in time from the stream, either based on a time interval (SBT) or a maximum number of events (SBN), and processed to obtain tensors ready to be processed by standard CNNs.

The pioneering effort dates back to the 2010s: **DDES** [Code] Tulyakov et al. (2019), inspired by the previous successes in the classical stereo literature Kendall et al. (2017). Events are organized in queues and processed by temporal fully connected layers, implemented as MLPs to aggregate the information encoded along time by the events, and used to build a 4D cost volume, optimized with a 3D UNet inspired by GC-Net Kendall et al. (2017).

More recent works delved into strategies for codifying the events streams: **SE-CFF** [Code] Nam et al. (2022) aims at solving the loss of meaningful events due to overriding, by proposing a ConcentrationNet for aggregating the events streams into a compact representation preserving fine details. Furthermore, SE-CFF distills knowledge from future events during training to compensate for the missing information in past events. SE-CFF builds a correlation-based cost volume and processes it with dilated 2D convolutions Xu and Zhang (2020) and a 2D UNet to get the final prediction. Given the many events sampled from streams, **SCSNet** Cho and Yoon (2022) implements a differentiable mechanism for retaining only those relevant for estimating disparity at the desired timeframe. Furthermore, a Neighbor Cross Similarity Feature (NCSF) extraction module is designed to encode the similarity across the different modalities—images and events. SCSNet uses group-wise correlation layers Guo et al. (2019) and 3D convolutions in its backbone. In contrast, **DTC-SPADE** Zhang et al. (2022) samples events according to SBT and stores them in voxel grids, that are processed by Discrete-Time Convolution (DTC) modules—recurrent operators designed to embed data in a variety of temporal scales. Furthermore, spatially adaptive denormalization modules (SPADE Park et al. (2019)) are used to complement the sparse voxel grid embeddings with edge information from the original raw events. Following DDES Tulyakov et al. (2019), DTC-SPADE deploys a GC-Net-like backbone Kendall et al. (2017) to predict a dense disparity map. Peculiarly, **ADES** Cho et al. (2023) focuses on domain adaptation of an event-stereo model, either based on AANet Xu and Zhang (2020) or PSMNet Chang and Chen (2018), trained on the pseudo-events frame generated by EventGAN Zhu et al. (2021) from color images. During adaptation on real events streams, these are used to generate pseudo-color images Rebecq et al. (2019) and compute a smudge-aware self-supervised loss. Furthermore, a Motion-invariant Consistency module is designed to perturb events and simulate

slower or faster motion and then impose a consistency loss between the disparity maps obtained from perturbed and original inputs.

Eventually, some frameworks emerged to process both color and events. **EI-Stereo** [Code] Mostafavi et al. (2021) has been proposed to match pairs of event and color images to exploit the best of the two worlds. An event-intensity recycling network iteratively processes each image with a single channel of the corresponding event stack—from the most recent to the oldest—while keeping a hidden state propagated through the channels. Then, an AANet-like architecture Xu and Zhang (2020) is deployed for estimating the disparity map. On the same track, **EFS** Cho and Yoon (2022) deploys distinct feature extractors for images and events. The features are fused across the modalities and used to build a correlation-based, multi-scale cost volume, that passes through a 3D aggregation network to estimate a dense disparity map. An additional cost volume between event features is built at training time, and processed to obtain a sparse disparity map.

Lastly, **SAFE** Chen et al. (2024) estimates disparity from asymmetric inputs—a color image and an event stream—by dividing the task into three sub-problems: (i) asymmetric color-event matching, (ii) color Structure-from-Motion (SfM), and (iii) event SfM. Two feature extractors process images and events, and then three cost volumes are built, (i) between image and events, (ii) time-adjacent images, and (iii) time-adjacent events, and combined in a single volume, to be processed by a fusion module deploying 3D ConvLSTMs.

### 2.4.3 Gated Stereo Matching

Adverse weather conditions such as fog and snow can significantly affect the performance of conventional color cameras. Gated cameras Andersson (2006) are designed to be robust against these conditions. During acquisition, an illuminator emits light to flood a specific range of the scene in front of it. Multiple acquisitions are performed by illuminating predetermined distances (slices)—each one not interfering with the others—and then accumulated in a single frame. **Gated Stereo** Walz et al. (2023) utilizes a synchronized wide-baseline gated stereo camera and deploys a network made of monocular and stereo branches, exploiting both time-of-flight intensity cues and multi-view cues, respectively. These branches are combined through a fusion network to produce the final depth map. The framework is trained in a semi-supervised manner, with a novel ambient-aware and illuminator-aware self-supervised loss on gated images.

### 2.4.4 Pattern Projection-Based

Pattern projection-based approaches for active stereo vision have gained significant attention in recent years due to their ability to enhance depth estimation accuracy and robustness

in challenging scenarios. These methods employ a projector to cast a structured light pattern onto the scene, providing additional cues for stereo matching. By exploiting the known pattern geometry and the correspondence between the projected pattern and the captured images, pattern projection-based techniques can effectively handle textureless regions, and other ambiguities that often pose difficulties for traditional passive matching algorithms. The projected patterns are typically designed to have high spatial frequency and good uniqueness properties to facilitate accurate and efficient stereo matching.

In the 10s, **ActiveStereoNet** Zhang et al. (2018) appeared as the pioneering work in end-to-end learning for active stereo systems, introducing a novel reconstruction loss based on local contrast normalization (LCN) to address the challenges of illumination variations and textureless regions. Furthermore, ActiveStereoNet proposes an invalidation network to explicitly handle occlusions, which is trained end-to-end without ground-truth data. Building upon this foundation, In the 20s, **Polka Lines** Baek and Heide (2021) takes a step further by jointly learning the structured illumination pattern and the reconstruction algorithm, parameterized by a diffractive optical element (DOE) and a neural network, respectively. This joint optimization approach enables the learned "Polka Lines" patterns to be tailored to the reconstruction network, achieving high-quality depth estimates across various conditions. **Activezero** Liu et al. (2022) addresses the lack of real-world depth annotations by proposing a mixed-domain learning framework that combines supervised learning on synthetic data with self-supervised learning on real data. It introduces a novel temporal IR reprojection loss that is more robust to noise and textureless patches and invariant to illumination changes. Despite the similarities with Polka Lines, it focuses on improving generalization. **MonoStereoFusion** Xu et al. (2022) takes a different path by integrating a monocular structured light subsystem and a binocular stereo subsystem to leverage their complementary advantages. The monocular subsystem handles textureless regions, while the binocular subsystem is used for distant objects and outdoor scenes. In contrast to other methods, MonoStereoFusion introduces an IR camera without a narrow-band filter, allowing it to receive both visible and IR light simultaneously. Lastly, **ActiveZero++** Chen et al. (2023) extends the ActiveZero framework by incorporating an illumination-invariant feature matching module and a confidence-based depth completion module. The illumination-invariant feature matching module uses a self-attention mechanism to learn robust features that are insensitive to illumination variations. The confidence-based depth completion module uses the confidence from the stereo network to identify and improve erroneous areas in depth prediction through depth-normal consistency.

## 2.4.5 Cross-Spectral Stereo Networks

Cross-spectral stereo matching has emerged as a promising approach for estimating depth by finding correspondences between images captured in different spectral bands, such as visible (RGB) and near-infrared (NIR), short-wave infrared (SWIR), or thermal infrared (TIR). Indeed, in many real-world scenarios, RGB stereo matching often struggles to find reliable correspondences—e.g., in low-light conditions, fog, etc. By leveraging the distinct appearance characteristics of materials and objects across different spectral bands, cross-spectral stereo aims to overcome these limitations. This comes with unique challenges, such as the significant appearance differences between images from different spectral bands or the scarcity of annotated, cross-spectral datasets.

Before the 20s, **CS-Stereo** Zhi et al. (2018) processed RGB and NIR images without depth supervision. It simultaneously estimates disparity and translates the RGB image to a pseudo-NIR image using a disparity prediction network and a spectral translation network, incorporating a material recognition network to handle unreliable matching regions. Similarly, **UCSS** Liang et al. (2019) employed a spectral translation network based on F-cycleGAN to minimize appearance variations between cross-spectral images and a stereo matching network to estimate disparity, enabling iterative end-to-end unsupervised learning. **SS-MCE** Walters et al. (2021) takes a different approach, estimating dense flow fields between images of different spectra without requiring ground truth. It employs a dual-spectrum siamese-like structure with two flow estimation modules, each estimating the flow field to align one image with the other and then warp it back, demonstrating its versatility across different spectral combinations. **RGB-MS** Tosi et al. (2022) focuses on registering RGB and MS images with different resolutions using a self-supervised deep architecture consisting of coarse and fine-grained sub-modules, employing a proxy label distillation strategy for supervision. **DPS-Net** Tian et al. (2023), instead, is an end-to-end network for polarimetric stereo depth estimation that leverages multi-domain similarity and geometric constraints, constructing RGB and polarization correlation volumes, introducing an iso-depth cost to handle polarization ambiguities, and employing a cascaded dual-GRU architecture to recurrently update the disparity. Finally, **Gated-RCCB** Brucker et al. (2024) is an approach that fuses high-resolution RCCB and gated NIR cameras to estimate dense, accurate depth maps. It exploits active and passive signals across visible and NIR spectra, proposing a cross-spectral stereo network with a fusion module to effectively integrate features from both modalities.

### 3 Challenges and Solutions

The development of more accurate stereo architectures in the 2020s allowed researchers to tackle many challenges that were unresolved in the late 2010s Poggi et al. (2021). As illustrated in Fig. 1, these challenges can be categorized into four main groups, discussed in the following sections.

#### 3.1 Domain Shift

Existing deep stereo networks achieve remarkable performance when trained and tested on data from the same domain. However, when these models are deployed to the target scenarios, they often encounter a significant drop in performance due to the domain shift between the training and testing data. This shift can be attributed to:

- Variations in color, illumination, contrast, and texture.
- Differences in stereo camera setups, such as baseline distance, focal length, and sensor properties.
- Variations in the disparity range due to different depths of the scenes (*e.g.*, indoor vs. outdoor environments).

These differences can make stereo networks learn domain-specific features or shortcuts that do not transfer well to novel scenarios. To address this challenge, several techniques have been explored lately.

##### 3.1.1 Zero-Shot Generalization

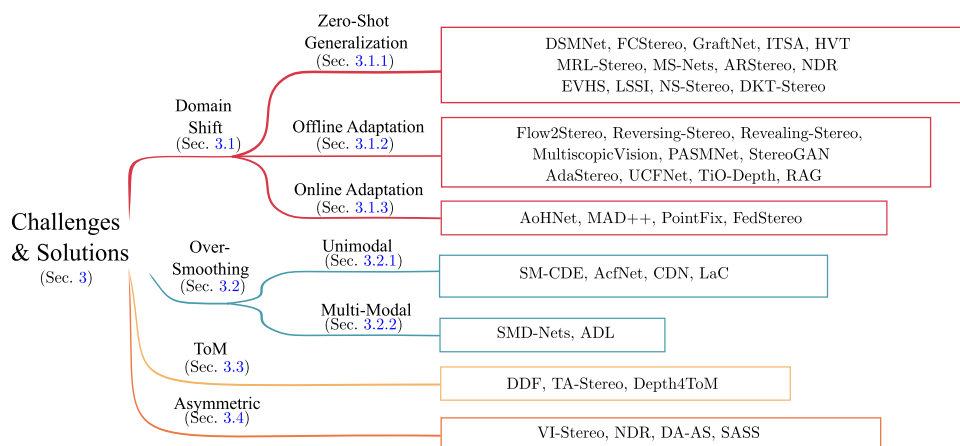
Zero-shot generalization refers to the capability of a stereo network to generalize from one domain (*e.g.*, synthetic data) to a completely different domain (*e.g.*, real-world scenes) without the need for fine-tuning or adaptation. This is particularly desirable when collecting stereo images or ground truth data for the target domain is expensive or infeasible. It is worth mentioning that the recent iterative models reviewed in Sec. 2.1.3 excel at this Tosi et al. (2023), despite not implementing any strategy specific for this purpose.

**Domain-Agnostic Feature Modeling** These methods focus on various aspects, such as feature regularization, feature consistency, shortcut avoidance, and representation learning, each tackling the problem from a unique perspective while sharing the common goal of learning robust, domain-invariant representations. **DSMNet** [Code] Zhang et al. (2020) and **FCStereo** [Codes] Zhang et al. (2022) both aim to learn domain-invariant features, but they differ in their approach. DSMNet introduces a Domain Normalization (DN) layer to regularize the distribution of learned features across spatial and channel dimensions, reducing sensitivity to image-level style variations and local contrast differences between domains. In contrast, FCStereo focuses on

explicitly encouraging feature consistency between matching pixels from left and right views through two loss functions: the stereo contrastive feature (SCF) loss and the stereo selective whitening (SSW) loss. While DSMNet emphasizes feature regularization, FCStereo prioritizes feature consistency across domains. **GraftNet** [Code] Liu et al. (2022) takes a different path by leveraging broad-spectrum features from a model pre-trained on large-scale datasets. It grafts these features to the cost aggregation module of an existing stereo network and uses a shallow network to restore task-related information. In comparison to DSMNet and FCStereo, which learn domain-invariant features from scratch, GraftNet exploits existing knowledge from pre-trained models to improve generalization. **ITSA** [Code] Chuah et al. (2022) and **HVT** [Code] Chang et al. (2023) both address the issue of shortcut learning, where networks exploit spurious correlations or superficial cues in synthetic training data rather than learning transferable representations. On the one hand, ITSA uses information-theoretic losses to automatically restrict the encoding of shortcut-related information into feature representations by combining the task loss with an approximation of the Fisher information loss. On the other hand, HVT emphasizes data augmentation, by transforming synthetic training images hierarchically at global, local, and pixel levels to diversify the training domain and prevent the model from learning dataset-dependent shortcuts. **MRL-Stereo** Rao et al. (2023), instead, introduces an approach to learning domain-invariant representations using masked representation learning. By feeding a masked left image and a complete right image into the model and reconstructing the original left image, MRL-Stereo encourages the learning of structural and domain-invariant features (Fig. 3).

**Non-parametric Cost Volumes** Non-parametric cost volume construction methods build cost volumes using conventional, domain-agnostic matching functions rather than relying on learned features that may be sensitive to domain-specific characteristics. Matching-Space Networks (**MS-Nets**) [Code] Cai et al. (2020), for example, move the learning process from the color space to the Matching Space by replacing learning-based feature extraction with four conventional matching functions (NCC, ZSAD, CENSUS, and SOBEL) and their associated confidence scores. These functions and scores are combined to generate a 4D matching volume, which is then regularized using adapted versions of popular architectures like GCNet Kendall et al. (2017) and PSMNet Chang and Chen (2018). Similarly, **ARStereo** [Code] Cheng et al. (2022) leverages Census Transform binary patterns as the matching features for building cost volumes, together with a backbone applied to extract high-level semantic contextual features from the reference image alone, avoiding the vulnerabilities of matching at features level. Then, this hybrid cost volume is regularized through the remaining layers.

**Fig. 3** A taxonomy of the main challenges (and solutions) in deep stereo matching. For each, we highlight the key problem areas and novel techniques developed.



**Integration of Additional Geometric Cues** Incorporating complementary geometric information, such as sparse depth hints or refined disparities from traditional algorithms, can guide stereo networks toward more robust and generalizable predictions. Two notable methods in this direction are Neural Disparity Refinement (NDR) [Code] Aleotti et al. (2021); Tosi et al. (2024) and Expansion of Visual Hints for Stereo (EVHS) Pilzer et al. (2023). On the one hand, NDR combines traditional stereo algorithms with deep learning to obtain refined, high-resolution disparity maps with sharp edges. Trained solely on synthetic data, the network can refine and super-resolve disparity maps from any source, including classical stereo methods or deep stereo networks, generalizing well to real images in a zero-shot manner. On the other hand, EVHS exploits visual hints from visual-inertial odometry. The method expands sparse and unevenly distributed 3D cues using a 3D random geometric graph, connecting hints that are close in the 3D world to improve the learning and inference process, filtered by confidence. Expanded hints are integrated within DeepPruner Duggal et al. (2019), guiding the differentiable patchmatch algorithm within a narrow disparity range.

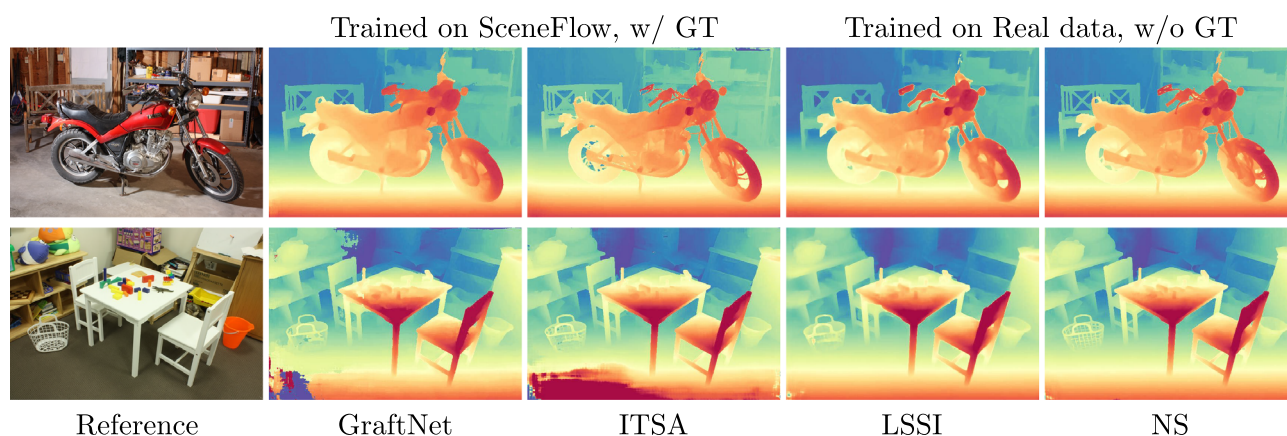
**Real-World Monocular to Synthetic Stereo Data** Generating diverse, realistic training data is crucial for improving generalization. To address this, approaches like rendering synthetic data or using real-world single/sparse images have been proposed. Methods like Learning Stereo from Single Images (LSSI) [Code] Watson et al. (2020) and NeRF-Supervised Deep Stereo (NS-Stereo) [Code] Tosi et al. (2023) create diverse stereo training data directly from easily acquired real-world monocular images. This exposes stereo networks to natural textures/appearances, promoting zero-shot generalization, while enabling training on numerous scenes leveraging intrinsic real-world visual properties. Specifically, LSSI leverages pre-trained monocular depth networks (e.g., MiDaS Ranftl et al. (2022)) to predict depth maps for single images, which are then converted to disparity maps and used both to synthesize stereo pairs, as well as pseudo-labels

for training a stereo network. Similarly, NS-Stereo generates stereo training data from sparse real-world image sequences captured with a single handheld camera. However, instead of using monocular depth estimation, NS-Stereo fits a NeRF model Müller et al. (2022) to each scene and uses it to render arbitrary stereo pairs by synthesizing a reference view and a target view displaced by a virtual baseline. Interestingly, NS-Stereo takes a step further by generating stereo triplets to handle occlusions in the photometric loss and exploiting the depth maps rendered by NeRF as proxy supervision. The qualitative improvement in zero-shot generalization compared to networks trained on synthetic or real stereo image synthesis can be observed in Fig. 4.

**Knowledge Transfer** Fine-tuning a pre-trained network often harms generalization itself. This happens, in particular, when the original training dataset is much larger than the few samples used for fine-tuning, however, few works focused on overcoming this issue. On this path, DKT-Stereo [Code] Zhang et al. (2024) aims at preserving the original, “Dark Knowledge” of a model during fine-tuning, with a frozen teacher network, an exponential moving average (EMA) teacher network, and a student network, all initialized with the same pre-trained weights. A further Filter and Ensemble (F&E) module discards the region being inconsistent across the pseudo-labels by the frozen and the EMA teachers to avoid insufficient regularization, and ensembles the labels in the consistent region to prevent overfitting ground-truth details at the expense of generalization.

### 3.1.2 Offline Adaptation

When a set of stereo pairs is available from an unseen domain, the domain shift can be compensated for by carrying out offline domain adaptation. In the absence of ground-truth annotations, a common approach consists of using self-supervised learning techniques, such as photometric losses,



**Fig. 4** Qualitative comparison – PSMNet variants. From left to right: reference image, disparity maps predicted by networks trained on synthetic data with ground-truth (Graft-PSMNet, ITSA-PSMNet) or on real data without any ground-truth (Mfs-PSMNet, NS-PSMNet)

to either train from scratch a network specialized for the target domain or to fine-tune a pre-trained one.

**Flow2Stereo** [Code] Liu et al. (2020) pursues the former strategy by jointly learning optical flow and stereo matching by exploiting the 3D geometry of stereoscopic videos in two stages. In the first stage, a teacher network predicts confident optical flow using photometric consistency and geometric constraints. In the second stage, the student network is refined using a self-supervised loss and proxy learning tasks, with the teacher network’s confident predictions serving as pseudo-labels. On the same track, **Reversing-Stereo** [Code] Aleotti et al. (2020) reverses the typical relationship between monocular and stereo depth estimation Godard et al. (2017). At the core, a monocular completion network (MCN) leverages sparse disparity points computed by any traditional stereo algorithm and single-image cues to predict dense and accurate disparity maps, by aggregating multiple MCN predictions with a randomized subset of the input points. These high-quality proxy labels are then used to supervise the training of deep stereo networks. Similarly, **Revealing-Stereo** Chen et al. (2021) proposes a framework to improve both stereo and monocular depth estimation by leveraging their reciprocal relations, introducing an occlusion-aware distillation strategy to train a monocular depth network with reliable predictions from a stereo network, and an occlusion-aware fusion module to combine their advantages. **TiO-Depth** [Code] Zhou and Dong (2023), instead, is a two-in-one self-supervised model handling monocular and binocular tasks via a Siamese architecture with monocular feature matching and multi-stage joint training. **PASMNet** [Code] Shen et al. (2023) deploys a parallax-attention mechanism (PAM) integrating epipolar constraints with an attention mechanism to calculate feature similarities along the epipolar line. PASMnet, instead, employs this mechanism within a feature extractor, a cascaded parallax-attention module for coarse-to-fine matching cost regression, and a disparity refinement

module, and is trained with a combination of photometric loss, smoothness loss, and PAM-specific losses that enforce left-right consistency, smoothness, and cycle consistency of the parallax-attention maps at multiple scales.

**MultiscopicVision** Yuan et al. (2021) proposes a self-supervised framework for stereo matching utilizing multiple images captured at aligned camera positions, introducing losses to optimize the network without ground-truth. **StereoGAN** [Code] Liu et al. (2020) tackles the domain gap between synthetic and real data by jointly optimizing a domain translation network and a stereo matching network. The domain translation component utilizes a GAN-based adversarial loss to generate realistic images from synthetic data while maintaining stereo consistency and epipolar geometry through novel constraints. By leveraging knowledge of the target domain, StereoGAN effectively adapts synthetic data to more closely resemble the characteristics of the real-world domain.

Assuming a pre-trained model, **AdaStereo** Song et al. (2021) aligns input representations and improves cross-domain adaptation ability, by deploying (1) a non-adversarial progressive color transfer algorithm for input image-level alignment, (2) an efficient, parameter-free cost normalization layer for channel-level feature alignment, to regulate the norm distribution of pixel-wise feature vectors; and (3) a self-supervised auxiliary task that reconstructs target-domain images using predicted disparities and occlusion masks.

In contrast, **UCFNet** [Code] Shen et al. (2023) extends CFNet Shen et al. (2021) and adapts a pre-trained model to the target domain using the pseudo-labels predicted by the model itself, and filtered according to uncertainty. Lastly, **RAG** [Code] Zhang et al. (2022) exploits NAS for discovering novel cells and adapting a pre-trained network to a specific domain, by expanding its architecture. Accordingly, RAG discovers a different structure for each domain: at test time, the proper one is selected depending on the domain

faced, by a Scene Router network ensembling a set of auto-encoders trained on the single domains during adaptation.

### 3.1.3 Online Stereo Adaptation

When stereo images from the target domain are not available beforehand, a stereo network can be adapted in an online manner during deployment Tonioni et al. (2019). In this case, stereo image pairs from the target domain are collected continuously, and the network adapts by relying on self-supervised learning objectives, such as photometric consistency loss. The main challenges of this setting consist of avoiding degradation of the model and preserving efficiency.

The first works extending Tonioni et al. (2019) are **AoHNet** Wang et al. (2020) and **MAD++** [Code] Poggi et al. (2021). Both enhance the effectiveness and speed of the adaptation procedure by replacing photometric losses with pseudo-labels from traditional stereo algorithms. In addition, AoHNet deploys an Adapt-or-Hold (AoH) mechanism to figure out whether to adapt or not, thus reducing the overall computational overhead. In contrast, **PointFix** Kim et al. (2022) exploits Model Agnostic Meta-Learning (MAML) to improve online adaptation. In the inner loop, wrong pixels are selected from the prediction by the base network and fixed by a PointFixNet, which predicts residuals used to adapt the base network itself. In the outer loop, both the base network and the PointFixNet are optimized based on the performance of the former after adaptation. At test time, only the inner loop is performed. Lastly, **FedStereo** [Code] Poggi and Tosi (2024) casts online adaptation as a distributed process, in which a server receives and aggregates updated weights from a set of actively adapting clients. These weights are then sent to a listening client not carrying out adaptation. To minimize communication overhead, a strategy inspired by MAD Tonioni et al. (2019) is used to transfer only a subset of the weights.

## 3.2 Over-Smoothing

A common limitation of most stereo networks is their tendency to over-smooth depth discontinuities. During inference, 3D stereo architectures typically employ a *soft argmax* operation to obtain the final disparity estimate by calculating the mean of the often multimodal predicted distribution. This leads to disparity estimates falling between the foreground and background modes, resulting in erroneous predictions and over-smoothed depth discontinuities. This approach leads to disparity estimates that fall between the foreground and background modes, resulting in erroneous predictions and over-smoothed depth discontinuities. 2D stereo architectures, despite not explicitly using soft argmax, are still affected by this issue. This yields inconsistent “bleeding” artifacts in the reconstructed 3D geometry around object

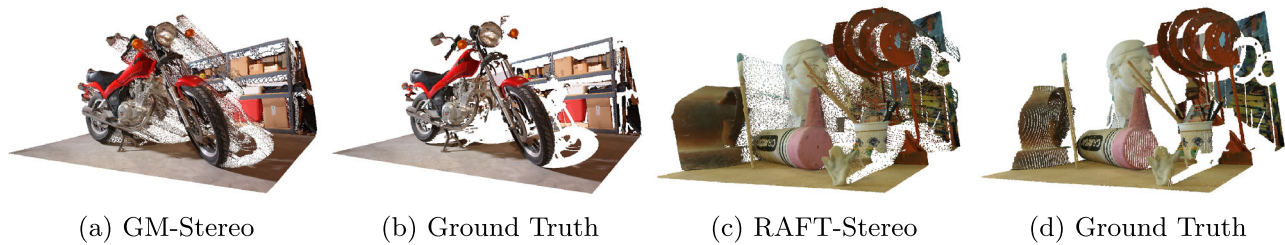
boundaries, highly undesirable for applications requiring accurate reconstruction with precise contours, as those shown in Fig. 5b and d. To mitigate this issue, some methodologies proposed to alleviate the “smoothing bias” and improve depth boundary reconstruction. These methods can be broadly categorized into two main approaches: *unimodal distribution modeling* and *multi-modal distribution modeling*.

### 3.2.1 Unimodal Distribution Modeling

Unimodal distribution modeling aims to alleviate the over-smoothing issue by constraining the disparity estimation to a single dominant mode. **SM-CDE** Chen et al. (2019) introduces a single-modal weighted average operation during inference. It considers the locality of the estimated disparity distribution and applies weighted averaging only to the dominant mode. Additionally, SM-CDE analyzes different loss functions and demonstrates that using cross-entropy loss with Gaussian distribution during training provides more stable and fine-grained supervision. Similarly, **AcfNet** [Code] Zhang et al. (2020) aims to improve the sharpness of disparity maps by directly supervising the cost volume with adaptive unimodal ground truth distributions. It introduces a confidence estimation network to modulate the variance of the unimodal distribution based on the confidence of finding a unique match. This ensures that pixels with high confidence have sharp peaks, while those with low confidence have flatter peaks. AcfNet also proposes a stereo focal loss to address the sample imbalance problem in the cost volume. Another approach, **CDN** [Code] Garg et al. (2020), addresses the issue by introducing an architecture that outputs a continuous distribution over arbitrary disparity values. It predicts probabilities and real-valued offsets for each disparity value in a pre-defined discrete set and takes the mode of this distribution as the final prediction. CDN also uses a novel loss function based on the Wasserstein distance between the true and predicted distributions. While the above methods focus on modeling unimodal distributions, **LaC** [Code] Liu et al. (2022) introduces a Local Similarity Pattern (LSP) to capture local structural information by explicitly revealing relationships between a point and its neighbors. To address the over-smoothing problem caused by static convolutional filters, LaC proposes Cost Self-Reassembling (CSR) and Disparity Self-Reassembling (DSR) strategies to adaptively propagate reliable disparity values based on image content.

### 3.2.2 Multi-Modal Distribution Modeling

Moving beyond unimodal distributions, multi-modal distribution modeling has emerged as a promising approach to achieve sharp depth discontinuities. **SMD-Nets** [Code] Tosi et al. (2021) is a pioneering work in this direction, utilizing bimodal mixture densities as output representa-



**Fig. 5** Bleeding Artifacts. The smooth disparities predicted between foreground and background objects project into flying points in 3D space (a,c), whereas precise 3D reconstructions demand sharp discontinuities (b,d)

tion. It encodes the input into a feature map, from which a multi-layer perceptron estimates the parameters of a bimodal Laplacian mixture distribution at any continuous 2D location. The bimodal distribution effectively models both the foreground and background disparities, allowing for precise depth estimation near object boundaries. The final disparity is obtained by selecting the mode with the highest density value. Building upon this concept, **ADL** [Code] Xu et al. (2024) proposes an adaptive multi-modal cross-entropy loss that models the ground-truth disparity distribution as a mixture of Laplacians. The number of modes and their weights are determined by local clustering and statistics within a local window. ADL also introduces a dominant-modal disparity estimator (DME) to robustly locate the dominant mode from the multi-modal outputs.

### 3.3 Transparent and Reflective Objects

Non-Lambertian materials have always posed significant challenges to matching algorithms, as they introduce misleading visual information about scene geometry. When a transparent object is present, a matching algorithm computes correspondences between points behind it, failing to perceive the object's distance from the camera. Alternatively, when dealing with a reflective surface, the algorithm might fail to triangulate depth properly. However, driven by data, deep learning has the potential to address these difficult challenges as well. In this regard, **DDF** Chai et al. (2020) proposes a framework for fusing the disparity predicted from stereo images with the depth acquired by a structured-light camera. The fusion is performed as a per-pixel weighted average of the two, with the weights predicted by a 2D UNet. An additional UNet is deployed for refining the fused depth map. An alternative strategy consists of segmenting the non-Lambertian materials and use the semantic masks to assist the stereo network. **TA-Stereo** Wu et al. (2023) and **Depth4ToM** [Code] Costanzino et al. (2023) implement this approach from two different perspectives. The former applies the segmentation masks to the stereo images to enforce similar appearance across the two and directly ease matching at testing time. In contrast, Depth4Tom in-paints non-Lambertian

objects according to the segmentation masks and processes them to obtain pseudo-labels for fine-tuning the stereo model. Pseudo-labels are predicted by a pre-trained monocular depth model and are fused with the predictions by the stereo network itself, replacing these latter in correspondence with non-Lambertian materials.

### 3.4 Asymmetric Stereo

Most stereo frameworks assume the stereo image pair is captured by cameras with identical properties. However, asymmetries between the images are common, due to differences in the cameras' intrinsic parameters, resolutions, or noise levels. These asymmetries can affect the matching process between the two images. Consequently, a new research direction has emerged to design stereo solutions that are robust to such asymmetries between the stereo pairs.

Visually-Imbalanced Stereo (**VI-Stereo**) Liu et al. (2020) is the first work in this direction. Given an HR left and LR right image pair, it deploys a UNet to upsample the latter and restore a balanced stereo pair, which is processed by a DispNet Mayer et al. (2016) to predict an HR disparity map. On the contrary, **NDR** [Code] Aleotti et al. (2021); Tosi et al. (2024) downsamples the left image to the resolution of the right one, runs a traditional stereo algorithm to obtain an LR and then upsample it, guided by the HR features, according to a continuous formulation. Differently from the previous works, **DA-AS** Chen et al. (2022) studies a self-supervised setting and proposes feature-metric consistency to replace the photometric loss, which yields sub-optimal results with asymmetric images, yet can be used an initial stereo model to learn good features for matching. Then, these features can be used to compute feature-metric consistency loss and train a new model starting from it. Lastly, **SASS** Song et al. (2023) proposes a novel spatially-adaptive self-similarity measure in a self-supervised manner, by extending the concept of self-similarity to generate deep features that are robust to the asymmetries by leveraging contrastive learning.

## 4 Experiments and Analysis

In this section, we compare the stereo frameworks surveyed so far on the standard benchmarks in the field. Specifically, we report the leaderboards of the KITTI 2015 and the Middlebury v3 online benchmarks—the former being the most popular dataset used in the literature, while the latter still represented a challenging benchmark in the late 2010s. Additionally, we include results from the Robust Vision Challenge (ROB), which demonstrates the recent efforts to achieve robustness to domain shifts. Finally, we present the leaderboard of the Booster online benchmark, representing the challenges that remain open in the 2020s. In each table, we highlight the **first**, **second**, and **third** best results in single categories, with the absolute best being reported in **purple**.

We decide to report only results being available from online benchmarks to avoid any ambiguities in the evaluation criteria that may arise across different papers, thus ensuring the fairest evaluation among any method. Unfortunately, online benchmarks to evaluate the impact of specific challenges such as over-smoothing or synthetic-to-real generalization do not exist. For an evaluation focusing on those aspects, we refer the reader to our prior works on such topics Tosi et al. (2021, 2023).

### 4.1 KITTI 2015

Table 1 collects entries from the KITTI 2015 online leaderboard concerning the stereo frameworks reviewed in our survey, with methods ranked according to the D1-all metric. The table reports three main sub-categories: (1) foundational stereo networks published in the 2020s, (2) efficient architectures published in 2020–2024, and (3) representative models from the 2010s. Although the KITTI benchmark was already saturated before the 2020s, the most recent advances in the field allowed for further improvements, with the latest proposals from 2023 to 2024 establishing consistently at the very top of the leaderboard, outperforming LEAStereo Yin et al. (2019) after nearly four years. Concerning efficient architectures, we can appreciate how the gap with state-of-the-art models has been significantly reduced, making them a viable alternative for practical applications.

### 4.2 Middlebury v3

Table 2 reports the results sourced from the Middlebury v3 online benchmark, with methods being ranked according to the bad2.0 metric. We divide it into two sub-tables, grouping on top the models among those covered in this survey, and at the bottom those from previous years Poggi et al. (2021). At first glance, we can appreciate a large improvement over the previous state-of-the-art, which got better and better through the five years that passed from the previous sur-

**Table 1** KITTI 2015 online benchmark

Method	Venue	KITTI 2015			
		D1-bg ↓	D1-fg ↓	D1-all ↓	Time (s)
Deep Stereo Networks (2020-2024)					
MoCha-Stereo	CVPR 2024	<b>1.36</b>	<b>2.43</b>	<b>1.53</b>	0.27
GANet+ADL	CVPR 2024	1.38	<b>2.38</b>	<b>1.55</b>	0.67
Selective-IGEV	CVPR 2024	<b>1.33</b>	2.61	<b>1.55</b>	0.24
MC-Stereo	3DV 2024	<b>1.36</b>	2.51	<b>1.55</b>	0.40
Any-IGEV	AAAI 2024	1.43	<b>2.35</b>	1.58	0.32
NMRF-Stereo	CVPR 2024	<b>1.28</b>	3.13	1.59	0.09
CroCo-RVC	ICCV 2023	1.38	2.65	1.59	0.93
IGEV-Stereo	CVPR 2023	1.38	2.67	1.59	0.18
LEAStereo	NeurIPS 2020	1.40	2.91	1.65	0.30
ACVNet	CVPR 2022	1.37	3.07	1.65	0.2
PCW-Net	ECCV 2022	1.37	3.16	1.67	0.44
LaC+GANet	AAAI 2022	1.44	2.83	1.67	1.8
CREStereo	CVPR 2022	1.45	2.86	1.69	0.41
Any-RAFT	AAAI 2024	1.44	3.04	1.70	0.34
DKT-IGEV	CVPR 2024	1.46	3.05	1.72	0.18
GMStereo	TPAMI 2023	1.49	3.14	1.77	0.17
GOAT	WACV 2024	1.71	2.51	1.84	0.29
CFNet	CVPR 2021	1.54	3.56	1.88	0.18
AcfNet	AAAI 2020	1.51	3.80	1.89	0.48
CDN	NeurIPS 2020	1.66	3.20	1.92	0.4
AANet+	CVPR 2020	1.65	3.96	2.03	0.06
PSM + SMD-Nets	CVPR 2021	1.69	4.01	2.08	0.41
MSNet3D	WACV 2022	1.75	3.87	2.10	1.5
Bi3D	CVPR 2020	1.95	3.48	2.21	0.48
AANet	CVPR 2020	1.99	5.39	2.55	0.062
WaveletStereo	CVPR 2020	2.24	4.62	2.63	0.27
AdaStereo	CVPR 2021	2.59	5.55	3.08	0.41
DSMNet-synthetic	ECCV 2020	3.11	6.72	3.71	1.6
Efficiency-Oriented Deep Stereo Networks (2020-2024)					
TemporalStereo	IROS 2023	<b>1.61</b>	<b>2.78</b>	<b>1.81</b>	0.04
PCVNet	ICCV 2023	<b>1.68</b>	<b>3.19</b>	<b>1.93</b>	0.05
HITNet	CVPR 2021	1.74	<b>3.20</b>	<b>1.98</b>	0.02
CasStereo	CVPR 2020	<b>1.59</b>	4.03	2.00	0.6
CoEX	IROS 2021	1.74	3.41	2.02	0.027
BGNet+	CVPR 2021	1.81	4.09	2.19	0.03
MABNet_origin	ECCV 2020	1.89	5.02	2.41	0.38
Fast_DS-CS	WACV 2020	2.83	4.31	3.08	0.02
MABNet_tiny	ECCV 2020	3.04	8.07	3.88	0.11
StereoVAE	ICRA 2023	4.25	10.18	5.23	0.03
AAFS	ACCV 2020	6.27	13.95	7.54	0.01
Deep Stereo Networks (2016-2019)					
CSPN	TPAMI 2019	<b>1.51</b>	<b>2.88</b>	<b>1.74</b>	1.0
GANet	CVPR 2019	<b>1.48</b>	<b>3.46</b>	<b>1.81</b>	1.8
GwcNet	CVPR 2019	<b>1.74</b>	3.93	<b>2.11</b>	0.32
HSMNet	CVPR 2019	1.80	<b>3.85</b>	2.14	0.14
PSMNet	CVPR 2018	1.86	4.62	2.32	0.41
GC-Net	ICCV 2017	2.21	6.16	2.87	0.9
DispNetC	CVPR 2016	4.32	4.41	4.34	0.06

veys. By looking at the top 6 positions of the leaderboard, we find RAFT-Stereo Lipson et al. (2021) and five more architectures derived from it, suggesting how the former has been a game changer for a benchmark challenging as Middlebury v3. Indeed, it is worth noticing how RAFT-Stereo has been one of the very first architectures, after HSMNet, capable of processing Middlebury images at full resolution, unleashing the possibility of maintaining many more details in the final disparity maps, as well as its outstanding generalization performance contributed to its success on this benchmark. Selective-IGEV currently represents the state-of-the-art: on the one hand, it achieves an average error on non-occluded pixels below the pixel for the first time. On the other hand, the average error raises over 1.5 on all pixels, in particular those large ones caused by very close objects, suggesting that the occlusions may still represent an open challenge.

**Table 2** Middlebury-v3 online benchmark

Method	Venue	Res	Non-Occ		All	
			bad2.0 ↓	EPE ↓	bad2.0 ↓	EPE ↓
Deep Stereo Networks (2020-2024)						
Selective-IGEV	CVPR 2024	F	2.51	0.91	6.04	1.54
DLNR	CVPR 2023	F	3.20	1.06	6.98	1.91
EAI-Stereo	ACCV 2022	F	3.68	1.09	7.53	1.92
CREStereo	CVPR 2022	F	3.71	1.15	8.13	2.10
RAFT-Stereo	3DV 2021	F	4.74	1.27	9.37	2.71
IGEV-Stereo	CVPR 2023	F	4.83	2.89	8.16	3.64
HIT-Net	CVPR 2021	H	6.46	1.71	12.8	3.29
GMSTereo	TPAMI 2023	H	7.14	1.31	11.7	1.89
LEAStereo	NeurIPS 2020	H	7.15	1.43	12.1	2.89
CroCo-Stereo	ICCV 2023	F	7.29	1.76	11.1	2.36
PCVNet	ICCV 2023	H	8.19	1.53	13.6	2.71
GOAT	WACV 2024	H	8.73	1.64	13.8	2.71
ACVNet	CVPR 2022	H	13.6	8.24	19.5	12.1
AdaStereo	CVPR 2021	F	13.7	2.22	19.8	3.39
AAANet++	CVPR 2020	H	15.4	6.37	22.0	9.77
CasStereo	CVPR 2020	H	18.8	4.50	26.0	8.98
Deep Stereo Networks (2016-2019)						
HSM-Smooth-Occ	CVPR 2019	F	10.8	2.15	17.5	3.44
MC-CNN-acrt	JMLR 2016	H	8.08	3.82	19.1	17.9
GANetREF_RVC	CVPR 2019	H	18.9	12.2	24.9	15.8
iResNet	CVPR 2018	H	22.9	3.31	29.5	4.67
DeepPruner_ROB	ICCV 2019	Q	30.1	4.80	36.4	6.56
PSMNet_ROB	CVPR 2018	Q	42.1	6.68	47.2	8.78
Hand-Crafted Stereo Algorithms (pre-2016)						
SGM	TPAMI 2007	H	18.4	5.32	25.7	9.27

### 4.3 Robust Vision Challenge (RVC)

Table 3 collects the leaderboards from the three editions of the Robust Vision Challenge that took place in 2018, 2020, and 2022, reported respectively from the bottom to the top. All methods are ranked on the individual benchmarks respectively according to D1-all, bad1, and avg error metrics on KITTI, Middlebury, and ETH3D, while the overall rank is computed according to the Schulze Proportional Ranking (PR) Schulze (2011). We highlight that the challenge allowed for training on any existing dataset—both synthetic and real

ones; for a synthetic-to-real benchmark, we refer the readers to Tosi et al. (2023). The trend highlighted through the three editions is consistent with the evolution of the field we have reported in this survey. Indeed, end-to-end deep stereo was still young in 2018, as we can notice from the bottom-most section of the table, where iResNet Liang et al. (2018) and PSMNet Chang and Chen (2018) are the only frameworks belonging to this category. Then, the field developed quickly in the following two years and CFNet Shen et al. (2021) ranked first in 2020, surpassing any of the networks from the 10s, with HSM-Net being the best among these latter models. Lastly, in 2022 the first three positions were taken by models extending the optimization paradigm introduced by RAFT-Stereo Lipson et al. (2021), confirming how the design strategy emerged from this latter has been a game-changer in the field. The rapid improvements observed through the three editions can be appreciated, in particular, by focusing on the error rates achieved on the Middlebury dataset, more specifically on the bad1 metric: indeed, this latter was higher than 45% for the winner of the first edition in 2018, dropping to about 26% for the winner of the 2020 edition and, finally, getting down to 16.5% in the latest edition.

### 4.4 Booster

Table 4 reports entries from the Booster online benchmark. Specifically, we show error metrics computed at full resolution (i.e., approximately 12 Megapixels) across different categories: *All*, which covers the whole regions of the images, as well as classes 0 to 3, which represents pixels belonging to materials from opaque to highly transparent/reflective. All methods are ranked, for each category, according to the bad2 score. We can appreciate how the whole benchmark remains

**Table 3** Robust Vision Challenge

Method	Venue	KITTI 2015				Middlebury				ETH3D				Overall Rank
		D1-bg ↓	D1-fg ↓	D1-all ↓	Rank	bad1 ↓	bad2 ↓	EPE ↓	Rank	bad1 ↓	bad2 ↓	EPE ↓	Rank	
Robust Vision Challenge 2022														
CREStereo++_RVC	ICCV 2023	1.55	3.53	1.88	2	16.5	9.46	2.2	1	1.70	0.37	0.16	1	1
iRaftStereo_RVC	3DV 2021	1.88	3.03	2.07	4	24.0	13.3	2.9	2	1.88	0.55	0.17	2	2
raft+_RVC	3DV 2021	1.60	2.98	1.83	1	22.6	14.4	3.86	3	2.18	0.71	0.21	4	3
CroCo_RVC	ICCV 2023	2.04	3.75	2.33	4	32.9	19.7	5.1	4	1.54	0.50	0.21	3	3
MaskLacGwcNet_RVC	AAAI 2022	1.65	3.68	1.99	3	31.3	15.8	13.5	5	6.42	1.88	0.38	5	5
Robust Vision Challenge 2020														
CFNet_RVC	CVPR 2021	1.65	3.53	1.96	2	26.2	16.1	5.07	2	3.7	0.97	0.26	1	1
NLCA.NET_v2_RVC	TNNLS 2022	1.51	3.97	1.92	1	29.4	16.4	5.6	3	4.11	1.2	0.29	2	2
HSM-Net_RVC	CVPR 2019	2.74	8.73	3.74	6	31.2	16.5	3.44	1	4.4	1.51	0.28	3	3
CVANet_RVC	ICCVW 2019	1.76	4.91	2.28	3	58.5	38.5	8.64	4	4.58	1.32	0.32	4	4
AAANet_RVC	CVPR 2020	2.23	4.89	2.67	5	42.9	31.8	12.8	5	5.41	1.95	0.33	5	5
GANetREF_RVC	CVPR 2019	1.88	4.58	2.33	4	43.1	24.9	15.8	6	6.97	1.25	0.45	6	6
Robust Vision Challenge 2018														
iResNet_ROB	CVPR 2018	2.27	4.89	2.71	2	45.9	31.7	6.56	3	4.67	1.22	0.27	2	1
DN-CSS_ROB	ECCV 2018	2.39	5.71	2.94	3	41.3	28.3	5.48	1	3.0	0.96	0.24	1	2
PSMNet_ROB	CVPR 2018	1.79	4.92	2.31	1	67.3	47.2	8.78	6	5.41	1.31	0.36	3	3
CBMV_ROB	CVPR 2018	3.55	12.09	4.97	4	21.6	13.3	6.65	2	4.66	2.06	0.36	4	4
SGM_ROB	TPAMI 2007	5.06	13.00	6.38	5	38.6	26.4	14.2	4	10.77	4.67	0.57	5	5
ELAS_ROB	ACCV 2010	7.38	21.15	9.67	6	51.7	34.6	13.4	5	17.82	8.75	0.8	6	6

**Table 4** Booster Dataset

Method	Venue	All			Class 0			Class 1			Class 2			Class 3		
		bad2 ↓	EPE ↓	Rank	bad2↓	EPE ↓	Rank	bad2↓	EPE ↓	Rank	bad2↓	EPE ↓	Rank	bad2↓	EPE ↓	Rank
RAFT-Stereo	3DV 2021	35.67	16.29	7	28.08	<b>2.49</b>	3	39.24	9.25	7	71.49	39.26	9	80.13	48.71	8
RAFT-Stereo (ft)	3DV 2021	34.67	5.26	6	30.87	2.70	5	37.17	4.54	5	56.32	7.15	5	64.82	<b>12.32</b>	1
CFNet	CVPR 2021	61.35	27.61	8	60.62	22.05	9	56.38	11.64	9	74.55	23.44	7	81.68	43.51	7
CFNet (ft)	CVPR 2021	66.85	19.65	9	62.93	9.72	8	66.37	11.29	8	80.66	38.80	8	87.28	42.46	6
PCVNet	ICCV 2023	<b>27.41</b>	<b>6.21</b>	1	<b>24.67</b>	<b>2.49</b>	2	<b>26.36</b>	<b>3.04</b>	2	<b>36.76</b>	<b>3.93</b>	1	<b>53.11</b>	<b>20.96</b>	3
DKT-RAFT	CVPR 2024	<b>28.60</b>	<b>6.82</b>	2	<b>24.81</b>	2.50	4	<b>25.67</b>	<b>3.03</b>	1	<b>42.81</b>	<b>4.77</b>	2	<b>60.73</b>	23.90	4
DKT-IGEV	CVPR 2024	34.48	9.30	5	31.01	5.44	7	34.90	3.91	4	53.41	5.86	4	<b>61.52</b>	28.69	5
CREStereo (ft)	CVPR 2022	<b>29.53</b>	<b>5.10</b>	3	<b>24.40</b>	3.09	6	<b>30.28</b>	<b>3.31</b>	3	<b>53.14</b>	<b>4.79</b>	3	67.63	<b>18.94</b>	2
CREStereo	CVPR 2022	33.07	12.56	4	25.55	<b>2.12</b>	1	36.31	5.62	6	62.77	20.89	6	79.50	53.34	9

**Table 5** DSEC Online Benchmark

Method	Venue	DSEC			
		1PE ↓	2PE ↓	MAE ↓	RMSE ↓
SCSNet	ECCV 2022	<b>5.673</b>	<b>0.992</b>	<b>0.387</b>	<b>0.852</b>
EI-Stereo	ICCV 2021	<b>5.814</b>	<b>1.055</b>	<b>0.396</b>	<b>0.906</b>
SE-CFF	CVPR 2022	9.583	2.620	<b>0.519</b>	<b>1.231</b>
DTC-SPADE	CVPR 2022	<b>9.277</b>	<b>2.416</b>	0.527	1.291

very difficult for current state-of-the-art stereo networks, with the very high resolution and the presence of non-Lambertian objects being the two main challenges. PCVNet Zeng et al. (2023) is the absolute winner, outperforming any other framework on classes 2 and 3, while remaining competitive on 0 and 1. By looking at RAFT-Stereo and CREStereo, for which results are reported by using both the original weights as well as those obtained after running a fine-tuning on the Booster training set, we highlight how the error rates on the most challenges classes 2 and 3 are largely reduced after fine-tuning, proving that state-of-the-art models have the potential to learn how to deal with non-Lambertian objects from context. Nonetheless, we point out how significant efforts are still necessary to properly deal with the two, aforementioned challenges.

#### 4.5 DSEC

Table 5 collects the entries from the DSEC online benchmark. We can appreciate how, at first glance, the error rates are much lower compared to those observed on the previous benchmarks. We ascribe this to the VGA resolution of the event cameras used to collect it, as well as to the relatively sparser ground truth and low diversity of the scenes. Unsurprisingly, models processing both RGB images and event data—SCSNet Cho and Yoon (2022) and EI-Stereo Mostafavi et al. (2021)—outperform those processing only the latter by a notable margin, hinting at the need for further developments in this field to bridge the gap with the standard color-based stereo literature.

## 5 Discussion

In this section, we highlight the key messages from our survey, highlighting the key advances made in the 2020s and suggesting potential avenues for future developments in deep stereo matching.

**Architecture Design.** As the benchmark results show, the new design strategy introduced by RAFT-Stereo has been a game changer, bringing much higher robustness to domain shifts. Most of the latest frameworks released a few months before this survey followed this new paradigm, and we expect more to come. Nonetheless, the quest for finding novel and effective architectures is not over, as witnessed by the very latest proposals Guan et al. (2024) achieving increasingly better results.

**Stereo Beyond RGB.** A trend that has been consolidating over the past five years is the use of other modalities, such as images from thermal, multispectral, or event cameras, as input to stereo matching networks. This brings fresh perspectives to a longstanding yet vivid field. However, online benchmarks for these new tasks are still rare, and more would help to consolidate this track.

**Open Challenges.** Despite the many successes in addressing some of the challenges anticipated by previous surveys Poggi et al. (2021), some persist. Indeed, the Booster dataset Ramirez et al. (2022, 2023) highlighted how images at very high resolution remain difficult to deal with, as well as non-Lambertian objects are critical, mainly due to the lack of training data or sub-optimal approaches to handle them. Similarly, challenging weather conditions still represent potential obstacles.

**Foundational Models.** Finally, in the wake of the emergence of visual foundational models for various computer vision tasks, we argue that a foundational model for stereo matching is still missing. While some attempts have been made recently for single-image depth estimation Yang et al. (2024a, b) and, more in general, for developing foundational models for 3D vision Wang et al. (2024); Leroy et al. (2024); Zhang et al. (2024), an effort in this direction has not yet been made focused on stereo. Although pure stereo matching is a low-level task based on patch-matching and may not require

the same scaling as general-purpose foundational models for 3D vision Wang et al. (2024); Leroy et al. (2024); Zhang et al. (2024)—which aim to tackle a broader range of tasks such as depth estimation, camera parameter/pose estimation, and more—we argue that casting stereo beyond the simple pattern-matching perspective could play a crucial role in addressing most of the challenges highlighted in this paper—such as domain shifts, matching in the presence of non-Lambertian surfaces and so on.

## 6 Conclusion

In this paper, we have surveyed the advances in deep stereo matching that have emerged in the 2020s. Building upon existing surveys Poggi et al. (2021); Laga et al. (2020) whose time horizon was limited to the end of 2019, we have deeply investigated the new architectures and design trends that have appeared recently and have become the standard implementation patterns today. We also looked at more advanced applications of stereo matching, ranging from sensor fusion to cross-spectral matching across different modalities. Finally, we have analyzed the performance of various methods on popular online benchmarks. This analysis not only identifies the current leading approaches but also sheds light on the remaining challenges and potential future research directions. We believe that this survey can serve as a practical guide for both novices and seasoned experts, providing inspiration for their work.

## Appendix A: Background

We introduce the fundamentals of deep stereo matching that have been driving advancements until the late 2010s. For a more comprehensive overview and detailed descriptions of the body of research that arose before 2020, readers can refer to existing surveys related to stereo matching, such as Poggi et al. (2021); Laga et al. (2020); Poggi et al. (2017).

### A.1 Theory

#### A.1.1 Learning for the Stereo Pipeline

In the early days, deep learning was studied for improving individual components of the traditional stereo pipeline Scharstein and Szeliski (2002). Specifically, CNN-based approaches were explored to learn more robust and discriminative matching cost functions, optimize the cost volume, and refine the final disparity map.

One of the primary areas of interest was the development of learned matching cost functions to replace hand-crafted ones, such as the sum of absolute differences (SAD) or the

census transform (CT) Zabih and Woodfill (1994). These learned cost functions, typically implemented using Siamese CNN architectures, were trained to predict the similarity between image patches extracted from stereo pairs, resulting in more accurate and robust matching costs Han et al. (2015); Žbontar and LeCun (2016); Chen et al. (2015); Luo et al. (2016). The resulting cost volumes were then processed using conventional optimization techniques, such as SGM Hirschmuller (2007), and refined using traditional post-processing steps including bilateral and/or median filtering.

Some efforts were also made to improve the cost volume optimization and refinement stages using learning-based approaches. Several methods have been proposed to learn how to modulate the cost volume based on the reliability of matching costs Park and Yoon (2015), select highly confident pixels as constraints for optimization Spyropoulos et al. (2014), and adapt the aggregation step in SGM to reduce streaking artifacts Poggi and Mattoccia (2016). CNNs were also employed to refine the final disparity map Shaked and Wolf (2017); Gidaris and Komodakis (2017); Batsos and Mordohai (2018); Aleotti et al. (2021), replacing conventional filtering techniques.

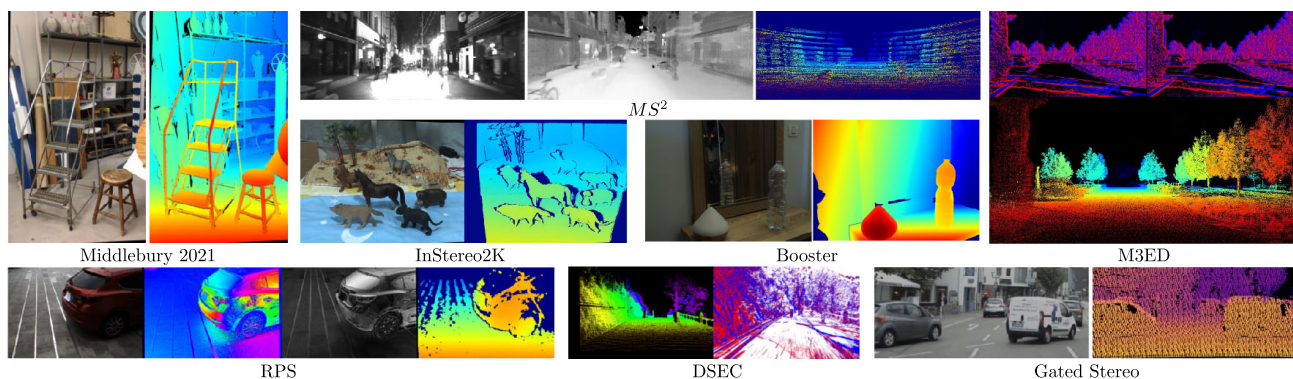
These early learning-based approaches demonstrated that deep learning could improve stereo matching by replacing certain hand-crafted components with learned methods, while still relying on the traditional pipeline structure.

The success of these methods in improving individual components of the stereo pipeline paved the way for the development of end-to-end deep stereo networks, which would come to dominate the field in the following years. By showing that learned components could outperform their hand-crafted counterparts, these early approaches laid the groundwork for the paradigm shift towards fully learnable models that directly estimate disparity maps from stereo image pairs.

#### A.1.2. End-to-End Deep Stereo

The advent of deep learning has revolutionized the field of stereo matching, enabling the development of end-to-end models that directly estimate disparity maps from stereo image pairs. These models have largely replaced traditional stereo matching pipelines, which typically consist of multiple hand-crafted steps Scharstein and Szeliski (2002). According to Poggi et al. (2021); Laga et al. (2020), end-to-end deep stereo networks directly estimating disparity maps from stereo image pairs can be broadly categorized into two main classes, based on their architecture: 2D networks and 3D networks.

Both 2D and 3D stereo networks begin by extracting deep features from the left and right input images using shared-weight CNNs. The key difference lies in how they construct



**Fig. 6** New stereo datasets in the 20s (Real-World)

and process the cost volume, which encodes the similarity between features at different disparity levels.

2D networks, such as DispNet Mayer et al. (2016), usually build a 3D cost volume by computing the correlation between features at corresponding pixels across a range of disparities, encoding the similarity between patches centered on each pixel:

$$c(x_1, x_2) = \sum_{o \in [-k, k] \times [-k, k]} \langle f_1(x_1 + o), f_2(x_2 + o) \rangle \quad (2)$$

where  $f_1$  and  $f_2$  are the features from the left and right images, and  $k$  determines the neighborhood size. The resulting 3D cost volume is processed using 2D convolutions using an encoder-decoder design, inspired by the U-Net model Ronneberger et al. (2015), to directly regress disparity values for each pixel. The use of plain 2D convolutions allows these networks to achieve real-time performance.

In contrast, 3D networks, first introduced by GC-Net Kendall et al. (2017) and later followed by more advanced architectures such as PSMNet Chang and Chen (2018) and GA-Net Zhang et al. (2019), construct a 4D cost volume by concatenating or computing the difference between features at all possible disparities. This 4D cost volume is then processed using 3D convolutions, which explicitly encode the geometry of the scene and capture the relationships between pixels at different disparity levels. The final disparity map is obtained using a fully differentiable soft argmin operation, which allows for sub-pixel disparity estimates:

$$\text{soft-argmin} = \sum_{d=0}^D d \cdot \sigma(-c_d) \quad (3)$$

where  $\sigma$  is the softmax operator applied along the disparity dimension  $D$ , and  $c_d$  are the cost values at each disparity level  $d$ . First, the predicted costs  $c_d$  from the cost volume are converted to a probability volume by taking the negative of each value. The probability volume is then normalized

across the disparity dimension using the softmax operation,  $\sigma(\cdot)$ . Finally, the disparity is computed as the sum of each disparity  $d$  weighted by its normalized probability. While 3D networks generally achieve higher accuracy than their 2D counterparts, they have significantly higher computational and memory requirements due to the use of 3D convolutions.

While 3D networks generally achieve higher accuracy than their 2D counterparts, they have significantly higher computational and memory requirements due to the additional dimension. Various techniques have been proposed to mitigate this computational burden, such as coarse-to-fine strategies Khamis et al. (2018); Wang et al. (2019); Yin et al. (2019), adaptive search space pruning Duggal et al. (2019), and hierarchical architectures Yang et al. (2019), and multi-scale feature extraction.

Additionally, multi-task learning approaches have been explored to leverage the complementary nature of tasks like semantic segmentation Yang et al. (2018) and edge detection Song et al. (2019). For a more comprehensive overview and detailed descriptions of these methods, readers can refer to existing surveys, such as Poggi et al. (2021); Laga et al. (2020).

## Appendix B: Datasets

In this section, we introduce the most relevant datasets being released in the 20ies. For details about datasets available before, we refer the reader to previous surveys Poggi et al. (2021); Laga et al. (2020).

### B.1 Real-World

We start by listing those datasets collecting real stereo images and, optionally, ground truth disparity maps acquired by means of active sensors. Figure 6 shows an overview of some samples taken from these datasets.

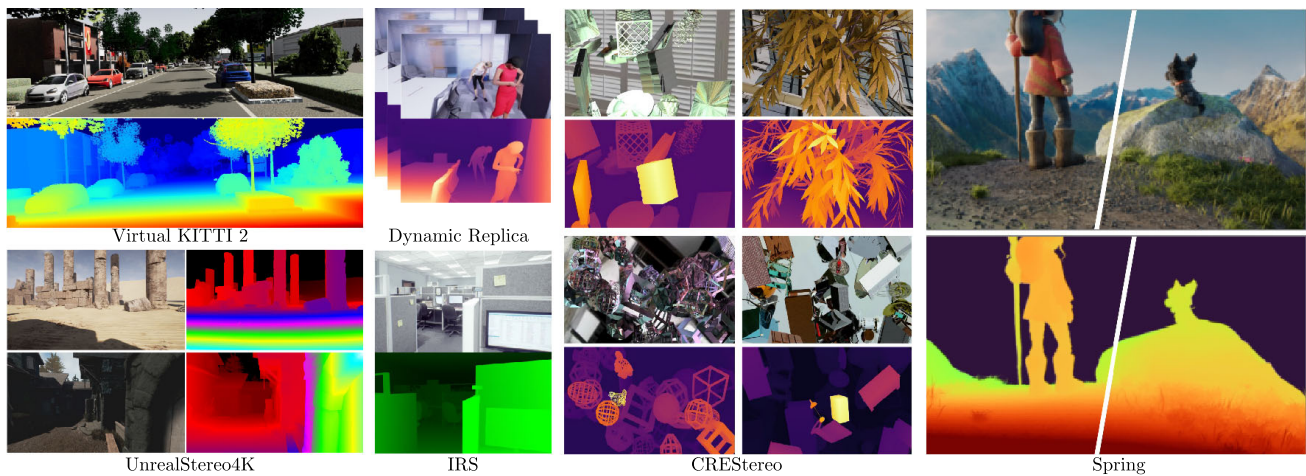


Fig. 7 New stereo datasets in the 20s (Synthetic)

### B.1.1 Passive Stereo Datasets

**Middlebury 2021** Scharstein et al. (2014).<sup>1</sup> This data collection consists of 24 indoor stereo datasets captured using a mobile device (Apple iPod touch 6G) mounted on a robotic arm, enabling acquisition of ground truth disparities via structured lighting. Spanning 11 distinct scenes imaged from 1-3 viewpoints under varying illuminations and exposures, including flash, ambient light, and device torch lighting, each dataset provides two  $1920 \times 1080$  resolution views organized into directories containing multiple illumination/exposure variants alongside calibration data and ground truth disparity maps.

**Booster** Ramirez et al. (2022).<sup>2</sup> This work introduces a novel high-resolution stereo dataset targeting the open challenges of non-Lambertian surfaces and high-resolution stereo matching. The dataset consists of 419 high-resolution (12 Mpx) stereo image pairs, as well as 419 unbalanced pairs with a 12 Mpx left image and 1.1 Mpx right image, collected across 64 different indoor scenes. The scenes contain a variety of specular and transparent surfaces, which are carefully annotated with dense ground-truth disparities using a novel deep space-time stereo framework. This framework leverages a pre-trained deep stereo network to accumulate cost volumes from multiple textured stereo pairs, enabling accurate sub-pixel disparity labels even for the challenging non-Lambertian regions. In addition to the ground-truth disparities, it provides manually annotated material segmentation masks to facilitate analysis of network performance on different surface types. The dataset is divided into 228 training and 191 test samples, where the test ground truth disparities are withheld, providing a challenging benchmark to

encourage further research on these open problems in stereo matching. The dataset is accompanied by an online evaluation benchmark for assessing stereo methods.

**Holopix50k** Hua et al. (2020).<sup>3</sup> Holopix50k is an in-the-wild stereo dataset comprising 49,368 image pairs captured by users of the Holopix mobile social platform, covering a wide variety of realistic scenarios in mobile photography. The images have an average resolution of  $0.74 (\pm 0.30)$  Mpx and were collected from the first Lightfield-enabled social media application. The dataset underwent post-processing steps to ensure high-quality stereo pairs, including removing vertical disparity and disparity-based filtering. However, Holopix50k contains only stereo RGB images without any ground truth disparity, making it suitable for self-supervised training of stereo networks. Moreover, due to the lack of ground truth disparity, Holopix50k is not typically used as a benchmark for evaluating stereo vision algorithms.

**InStereo2K** Bao et al. (2020).<sup>4</sup> The InStereo2K dataset is a large-scale stereo dataset designed for indoor scenes, containing 2050 pairs of stereo images with highly accurate disparity maps obtained using a structured light system. The dataset is split into a training set of 2000 image pairs and a test set of 50 image pairs, with a resolution of  $1080 \times 860$  pixels. InStereo2K covers a wide range of indoor scenes, including offices, classrooms, bedrooms, living rooms, and dormitories. The structured light system used to generate the dataset consists of two color cameras with a resolution of  $1280 \times 960$  pixels and a projector with a resolution of  $1024 \times 768$  pixels.

<sup>1</sup> <https://vision.middlebury.edu/stereo/data/scenes2021/>

<sup>2</sup> <https://cvlab-unibo.github.io/booster-web/>

<sup>3</sup> <https://github.com/leiainc/holopix50k>

<sup>4</sup> <https://github.com/YuhuaXu/StereoDataset>

### B.1.2 Multimodal Stereo Datasets

**CATS** Treible et al. (2017).<sup>5</sup> The Color and Thermal Stereo (CATS) dataset is a large-scale benchmark consisting of approximately 1400 images captured by a sensor platform with two visible-band cameras at a resolution of  $1280 \times 960$  pixels, two long wave infrared (LWIR) thermal cameras at a resolution of  $640 \times 480$  pixels, and a LiDAR for high-accuracy ( $<2\text{mm}$ ) ground truth. The dataset features 100 indoor and 80 outdoor cluttered scenes with diverse objects under various lighting and environmental conditions. CATS provides rectified stereo pairs in color, thermal, and cross-modality configurations, along with corresponding ground truth disparity maps generated by projecting the LiDAR data onto the images using a novel semi-automatic calibration method.

**MVSEC** Zhu et al. (2018).<sup>6</sup> The Multi Vehicle Stereo Event Camera dataset (MVSEC) collects data from both indoor and outdoor environments with event cameras. Specifically, a 10 cm-baseline stereo camera made with DAVIS 346B sensors collects event frames at  $346 \times 260$  resolution, together with a grayscale stereo camera mounting Aptina MT9V034 sensors working at  $752 \times 480$  resolution, a Velodyne Puck LITE (16-lines) to collect sparse depth data, an IMU, motion capture, and GPS. This sensor suite allows for extracting accurate ground truth pose and disparity maps—the latter obtained by accumulating LiDAR point clouds according to a LiDAR odometry pipeline (LOAM Zhang and Singh (2014)). Sequences are collected in a mixture of indoor and outdoor environments from different vehicles: 6 from a hexacopter, 5 from a car, one from a motorcycle, and 2 when the camera is handheld, both during day and night.

**DSEC** Gehrig et al. (2021).<sup>7</sup> The DSEC (Driving Stereo Event Camera) dataset is designed to address the challenges faced by standard cameras in adverse lighting conditions such as night, sunrise and sunset. It provides a rich set of sensory data, including a wide-baseline stereo setup with two color frame cameras ( $1440 \times 1080$  resolution, 51 cm baseline) and two high-resolution monochrome event cameras ( $640 \times 480$  resolution, 60 cm baseline), which offer unique advantages like high temporal resolution, high dynamic range, reduced motion blur, and low latency. The dataset contains 53 sequences collected in various illumination conditions and environments in Switzerland, totaling over 4 TB of raw data, and also includes LiDAR data, RTK GPS measurements, accurate time synchronization, and calibration between sensors. DSEC provides ground truth disparity and incorporates sequences with challenging illumination conditions.

<sup>5</sup> <https://bigdatavision.org/CAT/download.html>

<sup>6</sup> <https://daniilidis-group.github.io/mvsec/>

<sup>7</sup> <https://dsec.ifi.uzh.ch/>

**M3ED** Chaney et al. (2023).<sup>8</sup> This dataset provides 57 indoor/outdoor scenes collected with a compact multi-sensor block mounted on three different vehicles—a car, a UAV, and a quadruped robot. The event stereo camera mounts Prophesee Gen 4 sensors, capturing frames at  $1280 \times 720$  resolution with a 12 cm baseline, accompanied by a color stereo camera mounding OVC 3b sensors working at  $1280 \times 800$  resolution with an equivalent baseline. A 64-line Ouster OS1-64U LiDAR is used to generate semi-dense ground-truth disparity maps, by accumulating single scans through LiDAR odometry (FasterLIO Bai et al. (2022)). In total, M3ED provides about 3TB of raw and processed data.

**Gated Stereo** Walz et al. (2023).<sup>9</sup> Gated Stereo is a long-range dataset that was captured to facilitate training and evaluation of depth estimation methods that leverage gated imaging and stereo cues. Gated imaging enables robust long-range depth measurement by using active illumination and measuring time-of-flight, overcoming limitations of passive techniques. The dataset was acquired by driving over 1000 km in Germany with a sensor suite comprising a long-range LiDAR, an RGB stereo camera, and a synchronized NIR gated stereo camera. The gated camera captures  $1280 \times 720$  resolution images at 120 Hz, providing 3 gated slices and 2 ambient exposures enabled by active VCSEL illumination. The RGB stereo camera provides  $1920 \times 1080$  images at 30 Hz. The dataset contains over 107,000 samples captured across day, night, and varying weather conditions. After being curated for scenario diversity, it is divided into 54,320 samples for training, 728 for validation, and 2,463 for testing.

**RGB-MS** Tosi et al. (2022).<sup>10</sup> The RGB-MS dataset tackles the novel challenge of registering synchronized RGB and multi-spectral (MS) images with vastly different resolutions via stereo matching. It comprises 34 indoor RGB-MS image pairs from 13 scenes, with 12.4MP RGB images paired with 2.2MP 10-band MS images and annotated with high-resolution semi-dense ground truth disparities. These were generated using an additional 12.2MP RGB camera during acquisition - static scenes perturbed with patterns enabled robust disparity estimation from active RGB stereo pairs via cost volume integration, semi-global matching, outlier removal, and warping to the RGB-MS views. 11K unlabeled RGB-RGB-MS triplets are also provided in both indoor and outdoor environments.

**MS<sup>2</sup>** Shin et al. (2023).<sup>11</sup> The Multi-Spectral Stereo (MS<sup>2</sup>) dataset is a large-scale collection of synchronized multi-modal data captured for autonomous driving, comprising around 195K data pairs of stereo RGB images

<sup>8</sup> <https://m3ed.io/>

<sup>9</sup> <https://light.princeton.edu/gatedstereo/>

<sup>10</sup> <https://cvlab-unibo.github.io/rgb-ms-web/>

<sup>11</sup> <https://sites.google.com/view/multi-spectral-stereo-dataset>

(2448 × 2048 resolution, 15fps), stereo near-infrared (NIR) images (1280 × 720 resolution, 15fps), stereo long-wave infrared (thermal) images (640 × 512 resolution, 30fps), stereo LiDAR point clouds (10fps), and GNSS/IMU information. The data was acquired across diverse environments like cities, residential areas, roads, and campuses, under varying conditions including daytime, nighttime, cloudy, and rainy weather. The thermal stereo images are provided along with semi-dense ground truth depth maps generated by accumulating successive LiDAR sweeps and utilizing interpolated odometry.

**RPS** Tian et al. (2023).<sup>12</sup> The Real Polarimetric Stereo dataset (RPS) is a collection of 1,300 stereo pairs—1,000 of which have been collected indoors, while the remaining 300 are in outdoor environments. Two Lucid PHX050S-Q polarization cameras have been used to capture stereo polarization images at 1280×960 resolution, paired with an RGB-D camera for collecting ground-truth depth. Each polarization camera in the pair captures four polarization images, with different polarizer angles in a single shot.

## B.2 Synthetic

We now introduce more datasets obtained through graphic engines, allowing for generating vast amounts of stereo images and dense ground truth disparities with little effort. Figure 7 shows an overview of some samples taken from these datasets.

**HR-VS** Yang et al. (2019).<sup>13</sup> The High-res Virtual Stereo (HR-VS) dataset, presented in Yang et al. (2019) along with the HSMNet architecture, is a synthetic high-resolution stereo dataset generated using the Carla simulator Dosovitskiy et al. (2017). It contains 780 pairs of stereo images at 2056 × 2464 pixel resolution collected under 4 weather conditions in Town01, with a camera baseline of 0.54m. The dataset mimics real-world driving scenarios by limiting the disparity range to [9.66, 768] pixels and the depth range to [2.52, 200] meters, making it suitable for training high-resolution stereo matching algorithms for autonomous driving and urban scene understanding.

**Virtual KITTI 2** Cabon et al. (2020).<sup>14</sup> Building upon its predecessor, Virtual KITTI Gaidon et al. (2016), this dataset offers photo-realistic renderings of driving scenarios generated using the Unity game engine. The dataset comprises 5 image sequences, each a clone of a real-world sequence from the KITTI tracking benchmark, with a resolution of 1242 × 375 pixels. What sets Virtual KITTI 2 apart is its

diverse set of renditions for each sequence, including variations in weather conditions like fog and rain, as well as modifications to the camera configuration, such as rotations of 15° and 30°. In addition to RGB images, the dataset provides ground truth data for depth, optical flow, scene flow, instance segmentation, and semantic segmentation.

**TartanAir** Wang et al. (2020).<sup>15</sup> The TartanAir dataset is a large-scale, challenging dataset designed for robot navigation tasks, with a special focus on stereo vision and SLAM-related problems. It provides synchronized stereo RGB images across 1037 long motion sequences in 30 diverse environments, covering structured urban and indoor scenes as well as unstructured natural environments, resulting in over 1 million frames. Collected in photo-realistic simulation using the Unreal Engine and AirSim plugin, the dataset also includes depth images, segmentation labels, camera poses, occupancy grid maps, optical flow, stereo disparity, and simulated LiDAR measurements. TartanAir covers a wide range of motion patterns and includes challenging scenes with dynamic lighting, low illumination, adverse weather, and dynamic objects, aiming to bridge the gap between simulation and real-world performance.

**UnrealStereo4K** Tosi et al. (2021).<sup>16</sup> The UnrealStereo4K dataset is a large-scale synthetic stereo dataset created using the Unreal Engine and the open-source plugin UnrealCV. It consists of 8 scenes, including both indoor and outdoor environments, with a total of 8,000 stereo pairs at 3840 × 2160px resolution. The dataset is divided into 7,720 training pairs, 80 validation pairs, and 200 in-domain test pairs. An additional 200 out-of-domain test pairs from an unseen scene are provided to evaluate the generalization ability of stereo matching methods. The baseline is set to 20 cm for indoor environments and 50 cm for outdoor environments. The camera poses are randomly sampled, and invalid camera poses and stereo pairs are filtered out based on minimum depth, image intensity, and disparity map gradient criteria to ensure non-trivial geometry structure.

**IRS** Wang et al. (2021).<sup>17</sup> The IRS (Indoor Robotics Stereo) dataset is a large-scale synthetic stereo dataset designed for indoor robotics applications, containing over 100,000 stereo image pairs (84,946 for training and 15,079 for testing) with a resolution of 960 × 540 pixels. The dataset covers four indoor scene types (home, office, restaurant, and store) and 70 different scene layouts, with more than 2,091 identical furniture objects of various types placed within the constructed spaces. IRS provides high-quality and dense ground truth labels for surface normal and disparity maps, which are crucial for training deep learning models. The dataset is generated using a customized version of Unreal

<sup>12</sup> <https://github.com/Ethereal-Tian/DPS-Net>

<sup>13</sup> <https://github.com/gengshan-y/high-res-stereo?tab=readme-ov-file>

<sup>14</sup> <https://europe.naverlabs.com/Research/Computer-Vision/Proxy-Virtual-Worlds/>

<sup>15</sup> <https://theairlab.org/tartanair-dataset/>

<sup>16</sup> <https://github.com/fabiotosi92/SMD-Nets>

<sup>17</sup> <https://github.com/HKBU-HPML/IRS>

Engine 4 with originally implemented plug-ins, ensuring that the rendered images closely resemble real-world scenes in terms of material texture, light reflection and transmission, shadows, bloom, and lens flare.

**CREStereo** Li et al. (2022).<sup>18</sup> In the paper introducing the CREStereo architecture, a synthetic dataset for training stereo matching algorithms is also proposed. This dataset, generated using Blender, includes left-right image pairs and corresponding dense disparity maps. The dataset scenes feature a variety of object shapes sourced from ShapeNet and custom models with complex characteristics such as holes and open-work structures. Additionally, the dataset incorporates complex lighting environments with random colors and luminance, using real-world images as textures for objects and backgrounds. To ensure a wide range of disparity, objects are randomly positioned within the virtual cameras' field of view.

**SimStereo** Jospin et al. (2022).<sup>19</sup> The Active-Passive SimStereo dataset contains 515 computer-generated image pairs with a resolution of  $640 \times 480$  pixels, rendered using a physically-based rendering engine. This ensures that the images closely resemble real-world scenes in terms of material texture, light reflection, transmission, shadows, bloom, and lens flare. The dataset is split into a training set of 412 image pairs and a test set of 103 image pairs, covering various shapes, depth ranges, and styles. The test set includes realistic scenes and abstract compositions to effectively benchmark the performance of stereo vision models. SimStereo provides both active and passive frames for each scene, allowing for a direct comparison of the relative performance of algorithms in both scenarios.

**SimSIN** Wu et al. (2022).<sup>20</sup> The SimSIN dataset is a synthetic dataset counting about 500K stereo images with rendered ground truth depth. Images are generated adopting a stereo baseline of 13 cm following the camera, at  $512 \times 512$  resolution, and are derived from existing, synthetic datasets: specifically, about 80K, 205K, and 215K images are rendered from Replica, MP3D, and HM3D respectively. Despite being synthetic, it represents one of the largest sources of annotated stereo images involving indoor scenes.

**Spring** Mehl et al. (2023).<sup>21</sup> The Spring dataset introduces a large, high-resolution, high-detail, synthetic benchmark for scene flow, optical flow, and stereo estimation. Based on rendered scenes from the open-source Blender movie "Spring", it provides 6,000 stereo image pairs from 47 sequences at  $1920 \times 1080$  resolution, along with 12,000 ground truth frames for stereo and 23,812 ground truth frames for motion - 60x more than KITTI 2015 and 15x more than

MPI Sintel. To adequately assess fine details, the dataset uses a novel evaluation methodology based on super-resolved ( $3840 \times 2160$ px) ground truth, and provides focused evaluations on different image regions including high-detail, unmatched, non-rigid, sky, and large-displacement areas. The dataset is accompanied by a public benchmark website to upload, analyze and compare novel methods.

**Dynamic Replica** Karaev et al. (2023).<sup>22</sup> Dynamic Replica is a synthetic stereo video dataset featuring 524 videos of animated humans and animals in everyday scenes, designed for training temporally consistent models. The dataset includes 484 training videos, 20 validation videos (both 10s long with 300 frames each), and 20 test videos (30s long), all rendered at  $1280 \times 720$  resolution. The dataset is based on Facebook Replica reconstructions, using 375 3D human scans and 13 animal categories, with randomized camera baselines between 4 cm and 30 cm in the training subset. All samples contain ground-truth depth maps, optical flow, and segmentation masks, and camera parameters for both stereo views.

## Appendix C: Evaluation Metrics

The evaluation of stereo methods is typically performed by comparing the estimated disparity maps against ground truth data using several metrics. A widely adopted metric is the End-Point Error (EPE) $\downarrow$ , defined as:

$$\text{EPE} = \frac{1}{N} \sum_p |D_p - D_p^{gt}| \quad (4)$$

where  $D_p$  and  $D_p^{gt}$  are the predicted and ground truth disparity values for pixel  $p$ , respectively, and  $N$  is the total number of pixels. The EPE calculates the average absolute difference between the predicted and ground truth disparity values across all pixels. Optionally, the Root Mean Squared Error (RMSE) is also taken into account in some benchmarks, such as Middlebury v3 and Booster:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_p (|D_p - D_p^{gt}|)^2} \quad (5)$$

Moreover, existing benchmarks commonly report the bad- $\tau$  error $\downarrow$ , defined as the percentage of pixels whose disparity estimation deviates by more than  $\tau$  pixels from the ground truth:

$$\text{bad-}\tau = \frac{1}{N} \sum_p \delta(|D_p - D_p^{gt}| > \tau) \quad (6)$$

<sup>18</sup> <https://github.com/megvii-research/CREStereo>

<sup>19</sup> <https://iee-dataport.org/open-access/active-passive-simstereo>

<sup>20</sup> <https://distdepth.github.io/>

<sup>21</sup> <http://spring-benchmark.org>

<sup>22</sup> <https://dynamic-stereo.github.io/>

where  $\delta$  is the indicator function that returns 1 if the condition is true, and 0 otherwise. On the KITTI 2015 benchmark, a variant of the bad metric, the D1 error, is adopted:

$$D1 = \frac{1}{N} \sum_p \delta(|D_p - D_p^{gt}| > 3 \wedge |D_p - D_p^{gt}| > 0.05 \cdot D_p^{gt}) \quad (7)$$

considering as outliers those disparities whose error surpasses 3 pixels and 5% of the ground truth value.

**Acknowledgements** This study was funded by the European Union - Next Generation EU within the framework of the National Recovery and Resilience Plan NRRP - Mission 4 “Education and Research” - Component 2 - Investment 1.1 “National Research Program and Projects of Significant National Interest Fund (PRIN)” (Call D.D. MUR n. 104/2022) - PRIN2022 - Project reference: RIVERWATCH [Codice MUR: 2022MMBA8X\_002] [CUP: J53D23002270001]

**Data Availability** The datasets and benchmark results supporting the findings of this study are publicly available. The KITTI dataset and its benchmark results can be accessed at [https://www.cvlibs.net/datasets/kitti/eval\\_scene\\_flow.php?benchmark=stereo](https://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo). The Middlebury v3 dataset and its benchmark results are available at <https://vision.middlebury.edu/stereo/eval3/>. The Booster dataset and its benchmark results can be accessed at <https://cvlab-unibo.github.io/booster-web/benchmark.html>. The Robust Vision Challenge datasets and benchmark results can be found at <http://www.robustvision.net/>. The DSEC dataset and its benchmark results can be accessed at <https://dsec.ifi.uzh.ch/>

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47, 7–42.
- Poggi, M., Tosi, F., Batsos, K., Mordohai, P., & Mattoccia, S. (2021). On the synergies between machine learning and binocular stereo for depth estimation from images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9), 5314–5334.
- Laga, H., Jospin, L. V., Boussaid, F., & Bennamoun, M. (2020). A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4), 1738–1764.
- Li, Z., Liu, X., Drenkow, N., Ding, A., Creighton, F.X., Taylor, R.H., Unberath, M. (2021). Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceed-*

- ings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6197–6206
- Lipson, L., Teed, Z., Deng, J. (2021). Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*
- Poggi, M., Kim, S., Tosi, F., Kim, S., Aleotti, F., Min, D., Sohn, K., & Mattoccia, S. (2021). On the confidence of stereo matching in a deep-learning era: A quantitative evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9), 5293–5313.
- Poggi, M., Tosi, F., Mattoccia, S. (2017). Quantitative evaluation of confidence measures in a machine learning world. In *Proceedings of the IEEE International Conference on Computer Vision*, 5228–5237
- Xu, H., Zhang, J. (2020) Aanet: Adaptive aggregation network for efficient stereo matching. In *proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 1959–1968
- Yang, M., Wu, F., Li, W. (2020). Waveletstereo: Learning wavelet coefficients of disparity map in stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Shen, Z., Dai, Y., Rao, Z. (2021). Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13906–13915
- Mao, Y., Liu, Z., Li, W., Dai, Y., Wang, Q., Kim, Y.-T., Lee, H.-S. (2021). Uasnet: Uncertainty adaptive sampling network for deep stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6311–6319
- Shen, Z., Dai, Y., Song, X., Rao, Z., Zhou, D., Zhang, L. (2022). Pcw-net: Pyramid combination and warping cost volume for stereo matching. In *European Conference on Computer Vision*, 280–297. Springer
- Chen, L., Wang, W., Mordohai, P. (2023). Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17235–17244
- Cheng, X., Zhong, Y., Harandi, M., Dai, Y., Chang, X., Li, H., Drummond, T., Ge, Z. (2020). Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, 33
- Wang, Q., Shi, S., Zhao, K., Chu, X. (2022). Easnet: searching elastic and accurate network architecture for stereo matching. In *European Conference on Computer Vision*, 437–453. Springer
- Hu, Y., Wang, W., Yu, H., Zhen, W., Scherer, S. (2021). Orstereo: Occlusion-aware recurrent stereo matching for 4k-resolution images. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5671–5678. IEEE
- Gong, R., Liu, W., Gu, Z., Yang, X., Cheng, J. (2020). Learning intra-view and cross-view geometric knowledge for stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Zhao, H., Zhou, H., Zhang, Y., Chen, J., Yang, Y., Zhao, Y. (2023). High-frequency stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1327–1336
- Zhao, H., Zhou, H., Zhang, Y., Zhao, Y., Yang, Y., Ouyang, T. (2022). Eai-stereo: Error aware iterative network for stereo matching. In *Proceedings of the Asian Conference on Computer Vision*, 315–332
- Xu, G., Wang, X., Ding, X., Yang, X. (2023). Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21919–21928
- Li, J., Wang, P., Xiong, P., Cai, T., Yan, Z., Yang, L., Liu, J., Fan, H., Liu, S. (2022). Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF*

- Conference on Computer Vision and Pattern Recognition*, 16263–16272
- Jing, J., Li, J., Xiong, P., Liu, J., Liu, S., Guo, Y., Deng, X., Xu, M., Jiang, L., Sigal, L. (2023). Uncertainty guided adaptive warping for robust and efficient stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3318–3327
- Wang, X., Xu, G., Jia, H., Yang, X. (2024). Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*
- Feng, M., Cheng, J., Jia, H., Liu, L., Xu, G., Yang, X. (2024). Mc-stereo: Multi-peak lookup and cascade search range for stereo matching
- Cheng, Z., Yang, J., Li, H. (2024). Stereo matching in time: 100+ fps video stereo matching for extended reality. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 8719–8728
- Chen, Z., Long, W., Yao, H., Zhang, Y., Wang, B., Qin, Y., Wu, J. (2024). Mocha-stereo: Motif channel attention network for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*
- Guo, W., Li, Z., Yang, Y., Wang, Z., Taylor, R.H., Unberath, M., Yuille, A., Li, Y. (2022). Context-enhanced stereo transformer. In *European Conference on Computer Vision*, 263–279. Springer
- Su, Q., Ji, S. (2022). Chitransformer: Towards reliable stereo from cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1939–1949
- Xu, H., Zhang, J., Cai, J., Rezatofghi, H., Yu, F., Tao, D., Geiger, A. (2023). Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
- Weinzaepfel, P., Lucas, T., Leroy, V., Cabon, Y., Arora, V., Brégier, R., Csurka, G., Antsfeld, L., Chidlovskii, B., Revaud, J. (2023). CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow. In: ICCV
- Lou, J., Liu, W., Chen, Z., Liu, F., Cheng, J. (2023). Elfnet: Evidential local-global fusion for stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17784–17793
- Liu, Z., Li, Y., Okutomi, M. (2024). Global occlusion-aware transformer for robust stereo matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3535–3544
- Knobelreiter, P., Sormann, C., Shekhovtsov, A., Fraundorfer, F., Pock, T. (2020). Belief propagation reloaded: Learning bp-layers for labeling problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7900–7909
- Guan, T., Wang, C., Liu, Y.-H. (2024). Neural Markov Random Field for Stereo Matching
- Yee, K., Chakrabarti, A. (2020). Fast deep stereo with 2d convolutional processing of cost signatures. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*
- Yao, C., Jia, Y., Di, H., Li, P., Wu, Y. (2021). A decomposition model for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6091–6100
- Xu, G., Cheng, J., Guo, P., Yang, X. (2022). Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12981–12990
- Zeng, J., Yao, C., Yu, L., Wu, Y., Jia, Y. (2023) Parameterized cost volume for stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18347–18357
- Badki, A., Troccoli, A., Kim, K., Kautz, J., Sen, P., Gallo, O. (2020). Bi3d: Stereo depth estimation via binary classifications. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Li, X., Zhang, C., Su, W., & Tao, W. (2024). Iinet: Implicit intra-inter information fusion for real-time stereo matching. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4), 3225–3233. <https://doi.org/10.1609/aaai.v38i4.28107>
- Pang, J., Sun, W., Ren, J.S., Yang, C., Yan, Q. (2017). Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 887–895
- Xu, B., Xu, Y., Yang, X., Jia, W., Guo, Y. (2021). Bilateral grid learning for stereo matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12497–12506
- Xing, J., Qi, Z., Dong, J., Cai, J., Liu, H. (2020). Mabnet: a lightweight stereo network based on multibranch adjustable bottleneck module. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, 340–356. Springer
- Zhang, Y., Poggi, M., Mattoccia, S. (2023). Temporalstereo: Efficient spatial-temporal stereo matching network. In: IROS
- Chang, Q., Li, X., Xu, X., Liu, X., Li, Y., Miyazaki, J. (2023). Stereovae: A lightweight stereo-matching system using embedded gpus. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 1982–1988. IEEE
- Poggi, M., Tosi, F. (2024). Federated online adaptation for deep stereo. In: CVPR
- Bangunharcana, A., Cho, J.W., Lee, S., Kweon, I.S., Kim, K.-S., Kim, S. (2021). Correlate-and-excite: Real-time stereo matching via guided cost volume excitation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*
- Wang, Q., Shi, S., Zheng, S., Zhao, K., Chu, X. (2020). Fadnet: A fast and accurate network for disparity estimation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 101–107. IEEE
- Tankovich, V., Hane, C., Zhang, Y., Kowdle, A., Fanello, S., Bouaziz, S. (2021). Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14362–14372
- Cai, J., Qi, Z., Fu, K., Shi, X., Li, Z., Liu, X., Liu, H. (2022). Pbcstereo: A compressed stereo network with pure binary convolutional operations. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 4378–4394
- Chang, J.-R., Chang, P.-C., Chen, Y.-S. (2020). Attention-aware feature aggregation for real-time stereo matching on edge devices. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*
- Shamsafar, F., Woerz, S., Rahim, R., Zell, A. (2022). Mobilestereonet: Towards lightweight deep networks for stereo matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2417–2426
- Dovesi, P.L., Poggi, M., Andraghetti, L., Martí, M., Kjellström, H., Pieropan, A., Mattoccia, S. (2020). Real-time semantic stereo matching. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 10780–10787. IEEE
- Chen, S., Xiang, Z., Qiao, C., Chen, Y., Bai, T. (2020). Sgnet: Semantics guided deep stereo matching. In *Proceedings of the Asian Conference on Computer Vision*
- Kusupati, U., Cheng, S., Chen, R., Su, H. (2020). Normal assisted stereo depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Aleotti, F., Poggi, M., Tosi, F., Mattoccia, S. (2020). Learning end-to-end scene flow by distilling single tasks knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 10435–10442
- Jiao, Y., Tran, T.D., Shi, G. (2021). Effiscene: Efficient per-pixel rigidity inference for unsupervised joint learning of optical flow, depth, camera pose and motion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5538–5547

- Chi, C., Wang, Q., Hao, T., Guo, P., Yang, X. (2021). Feature-level collaboration: Joint unsupervised learning of optical flow, stereo depth and camera motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2463–2473
- You, Y., Wang, Y., Chao, W.-L., Garg, D., Pleiss, G., Hariharan, B., Campbell, M., Weinberger, K.Q. (2020). Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. In: ICLR
- Zhang, J., Ramanagopal, M.S., Vasudevan, R., Johnson-Roberson, M. (2020). Listero: Generate dense depth maps from lidar and stereo imagery. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 7829–7836. IEEE
- Huang, Y.-K., Liu, Y.-C., Wu, T.-H., Su, H.-T., Chang, Y.-C., Tsou, T.-L., Wang, Y.-A., Hsu, W.H. (2021). S3: Learnable sparse signal superdensity for guided depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16706–16716
- Yin, H., Deng, L., Chen, Z., Chen, B., Sun, T., Yusen, X., Xiao, J., Fu, Y., Deng, S., Li, X. (2022). Lsm-d-net: Lidar-stereo fusion with mixture density network for depth sensing. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 552–568
- Bartolomei, L., Poggi, M., Tosi, F., Conti, A., Mattoccia, S. (2023). Active stereo without pattern projector. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18470–18482
- Tulyakov, S., Fleuret, F., Kiefel, M., Gehler, P., Hirsch, M. (2019). Learning an event sequence embedding for event-based deep stereo. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. To appear. <https://fleuret.org/papers/tulyakov-et-al-iccv2019.pdf>
- Nam, Y., Mostafavi, M., Yoon, K.-J., Choi, J. (2022). Stereo depth from events cameras: Concentrate and focus on the future. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6114–6123
- Cho, H., Yoon, K.-J. (2022). Selection and cross similarity for event-image deep stereo. In *European Conference on Computer Vision*, 470–486. Springer
- Zhang, K., Che, K., Zhang, J., Cheng, J., Zhang, Z., Guo, Q., Leng, L. (2022). Discrete time convolution for fast event-based stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8676–8686
- Cho, H., Cho, J., Yoon, K.-J. (2023). Learning adaptive dense event stereo from the image domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17797–17807
- Mostafavi, M., Yoon, K.-J., Choi, J. (2021). Eventintensity stereo: Estimating depth by the best of both worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4258–4267
- Cho, H., Yoon, K.-J. (2022). Event-image fusion stereo using cross-modality feature propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, 454–462
- Chen, X., Weng, W., Zhang, Y., Xiong, Z. (2024). Depth from asymmetric frame-event stereo: A divide-and-conquer approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3045–3054
- Walz, S., Bijelic, M., Ramazzina, A., Walia, A., Mannan, F., Heide, F. (2023). Gated stereo: Joint depth estimation from gated and wide-baseline active stereo cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13252–13262
- Zhang, Y., Khamis, S., Rhemann, C., Valentin, J., Kowdle, A., Tankovich, V., Schoenberg, M., Izadi, S., Funkhouser, T., Fanello, S. (2018). Activestereonet: End-to-end self-supervised learning for active stereo systems. In *Proceedings of the European Conference on Computer Vision (ECCV)*
- Baek, S.-H., Heide, F. (2021). Polka lines: Learning structured illumination and reconstruction for active stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5757–5767
- Liu, I., Yang, E., Tao, J., Chen, R., Zhang, X., Ran, Q., Liu, Z., Su, H. (2022). Activezero: Mixed domain learning for active stereovision with zero annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13033–13042
- Xu, Y., Yang, X., Yu, Y., Jia, W., Chu, Z., Guo, Y. (2022). Depth estimation by combining binocular stereo and monocular structured-light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1746–1755
- Chen, R., Liu, I., Yang, E., Tao, J., Zhang, X., Ran, Q., Liu, Z., Xu, J., Su, H. (2023). Activezero++: Mixed domain learning stereo and confidence-based depth completion with zero annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
- Zhi, T., Pires, B.R., Hebert, M., Narasimhan, S.G. (2018). Deep material-aware cross-spectral stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1916–1925
- Liang, M., Guo, X., Li, H., Wang, X., Song, Y. (2019). Unsupervised cross-spectral stereo matching by learning to synthesize. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 8706–8713
- Walters, C., Mendez, O., Johnson, M., Bowden, R. (2021). There and back again: Self-supervised multispectral correspondence estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 5147–5154. IEEE
- Tosi, F., Ramirez, P.Z., Poggi, M., Salti, S., Mattoccia, S., Di Stefano, L. (2022). Rgb-multispectral matching: Dataset, learning methodology, evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15958–15968
- Tian, C., Pan, W., Wang, Z., Mao, M., Zhang, G., Bao, H., Tan, P., Cui, Z. (2023). Dps-net: Deep polarimetric stereo depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3569–3579
- Brucker, S., Walz, S., Bijelic, M., Heide, F. (2024). Cross-spectral gated-rgb stereo depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Guo, X., Yang, K., Yang, W., Wang, X., Li, H. (2019). Group-wise correlation stereo network. In: CVPR
- Teed, Z., Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 402–419. Springer
- Liang, Z., & Li, C. (2024). Any-stereo: Arbitrary scale disparity estimation for iterative stereo matching. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4), 3333–3341. <https://doi.org/10.1609/aaai.v38i4.28119>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30
- Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C. (2023). Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13229–13239
- Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P. (2020). Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Pro-*

- ceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520
- Tonioni, A., Tosi, F., Poggi, M., Mattoccia, S., Stefano, L.D. (2019). Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Zhan, W., Ou, X., Yang, Y., Chen, L. (2019). Dsnet: Joint learning for scene segmentation and disparity estimation. In *2019 International Conference on Robotics and Automation (ICRA)*, 2946–2952. IEEE
- Yang, G., Zhao, H., Shi, J., Deng, Z., Jia, J. (2018). Segstereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 636–651
- Jiang, H., Sun, D., Jampani, V., Lv, Z., Learned-Miller, E., Kautz, J. (2019). Sense: A shared encoder network for scene-flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3195–3204
- Wu, Z., Wu, X., Zhang, X., Wang, S., Ju, L. (2019). Semantic stereo matching with pyramid cost volumes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7484–7493
- Chang, J.-R., Chen, Y.-S. (2018). Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5410–5418
- Vedula, S., Baker, S., Rander, P., Collins, R., Kanade, T. (1999). Three-dimensional scene flow. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 2, 722–729. IEEE
- Teed, Z., Deng, J. (2021). Raft-3d: Scene flow using rigid-motion embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Liu, H., Lu, T., Xu, Y., Liu, J., Li, W., Chen, L. (2022). Camliflow: bidirectional camera-lidar fusion for joint optical flow and scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5791–5801
- Li, A., Hu, A., Xi, W., Yu, W., Zou, D. (2024). Stereo-lidar depth estimation with deformable propagation and learned disparity-depth conversion. arXiv preprint [arXiv:2404.07545](https://arxiv.org/abs/2404.07545)
- Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Tabá, B., Censi, A., Leutenegger, S., Davison, A. J., Conradt, J., Daniilidis, K., & Scaramuzza, D. (2022). Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1), 154–180. <https://doi.org/10.1109/TPAMI.2020.3008413>
- Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A. (2017). End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, 66–75
- Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2337–2346
- Zhu, A., Wang, Z., Khant, K., Daniilidis, K. (2021). Eventgan: Leveraging large scale image datasets for event cameras. In *2021 IEEE International Conference on Computational Photography (ICCP)*, 1–11
- Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D. (2019). High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell. (T-PMI)*
- Andersson, P. (2006). Long-range three-dimensional imaging using range-gated laser radar images. *Optical Engineering*, 45(3), 034301–034301.
- Zhang, F., Qi, X., Yang, R., Prisacariu, V., Wah, B., Torr, P. (2020). Domain-invariant stereo matching networks. In *Europe Conference on Computer Vision (ECCV)*
- Zhang, J., Wang, X., Bai, X., Wang, C., Huang, L., Chen, Y., Gu, L., Zhou, J., Harada, T., Hancock, E.R. (2022). Revisiting domain generalized stereo matching networks from a feature consistency perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13001–13011
- Liu, B., Yu, H., Qi, G. (2022). Graftnet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13012–13021
- Chuah, W., Tennakoon, R., Hoseinnezhad, R., Bab-Hadiashar, A., Suter, D. (2022). Itsa: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13022–13032
- Chang, T., Yang, X., Zhang, T., Wang, M. (2023). Domain generalized stereo matching via hierarchical visual transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9559–9568
- Rao, Z., Xiong, B., He, M., Dai, Y., He, R., Shen, Z., Li, X. (2023). Masked representation learning for domain generalized stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5435–5444
- Cai, C., Poggi, M., Mattoccia, S., Mordohai, P. (2020). Matching-space stereo networks for cross-domain generalization. In *2020 International Conference on 3D Vision (3DV)*, 364–373. <https://doi.org/10.1109/3DV50981.2020.00046>
- Cheng, K., Wu, T., & Healey, C. (2022). Revisiting non-parametric matching cost volumes for robust and generalizable stereo matching. *Advances in Neural Information Processing Systems*, 35, 16305–16318.
- Aleotti, F., Tosi, F., Ramirez, P.Z., Poggi, M., Salti, S., Mattoccia, S., Di Stefano, L. (2021). Neural disparity refinement for arbitrary resolution stereo. In *2021 International Conference on 3D Vision (3DV)*, 207–217. IEEE
- Tosi, F., Aleotti, F., Ramirez, P.Z., Poggi, M., Salti, S., Mattoccia, S., Di Stefano, L. (2024). Neural disparity refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
- Pilzer, A., Hou, Y., Loppi, N., Solin, A., Kannala, J. (2023). Expansion of visual hints for improved generalization in stereo matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5840–5849
- Watson, J., Aodha, O.M., Turmukhambetov, D., Brostow, G.J., Firman, M. (2020). Learning stereo from single images. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 722–740. Springer
- Tosi, F., Tonioni, A., De Gregorio, D., Poggi, M. (2023). Nerf-supervised deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 855–866
- Zhang, J., Li, J., Huang, L., Yu, X., Gu, L., Zheng, J., Bai, X. (2024). Robust synthetic-to-real transfer for stereo matching. arXiv preprint [arXiv:2403.07705](https://arxiv.org/abs/2403.07705)
- Liu, P., King, I., Lyu, M.R., Xu, J. (2020). Flow2stereo: Effective self-supervised learning of optical flow and stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Aleotti, F., Tosi, F., Zhang, L., Poggi, M., Mattoccia, S. (2020). Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 614–632. Springer
- Chen, Z., Ye, X., Yang, W., Xu, Z., Tan, X., Zou, Z., Ding, E., Zhang, X., Huang, L. (2021). Revealing the reciprocal relations between self-supervised stereo and monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15529–15538
- Yuan, W., Zhang, Y., Wu, B., Zhu, S., Tan, P., Wang, M.Y., Chen, Q. (2021). Stereo matching by self-supervision of multiscope

- vision. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5702–5709. IEEE
- Shen, Z., Song, X., Dai, Y., Zhou, D., Rao, Z., Zhang, L. (2023). Digging into uncertainty-based pseudo-label for robust stereo matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
- Liu, R., Yang, C., Sun, W., Wang, X., Li, H. (2020). Stereogan: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Song, X., Yang, G., Zhu, X., Zhou, H., Wang, Z., Shi, J. (2021). Adastereo: A simple and efficient approach for adaptive stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10328–10337
- Zhang, C., Tian, K., Fan, B., Meng, G., Zhang, Z., Pan, C. (2022). Continuous stereo matching of continuous driving scenes with growing architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18901–18910
- Wang, H., Wang, X., Song, J., Lei, J., Song, M. (2020). Faster self-adaptive deep stereo. In *Proceedings of the Asian Conference on Computer Vision*
- Poggi, M., Tonioni, A., Tosi, F., Mattoccia, S., & Di Stefano, L. (2021). Continual adaptation for deep stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9), 4713–4729.
- Kim, K., Park, J., Lee, J., Min, D., Sohn, K. (2022). Pointfix: Learning to fix domain bias for robust online stereo adaptation. In *European Conference on Computer Vision*, 568–585. Springer
- Chen, C., Chen, X., Cheng, H. (2019). On the over-smoothing problem of cnn based disparity estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8997–9005
- Zhang, Y., Chen, Y., Bai, X., Yu, S., Yu, K., Li, Z., Yang, K. (2020). Adaptive unimodal cost volume filtering for deep stereo matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 12926–12934
- Garg, D., Wang, Y., Hariharan, B., Campbell, M., Weinberger, K. Q., & Chao, W.-L. (2020). Wasserstein distances for stereo disparity estimation. *Advances in Neural Information Processing Systems*, 33, 22517–22529.
- Liu, B., Yu, H., Long, Y. (2022). Local similarity pattern and cost self-reassembling for deep stereo matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, 1647–1655
- Tosi, F., Liao, Y., Schmitt, C., Geiger, A. (2021). Smd-nets: Stereo mixture density networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*
- Xu, P., Xiang, Z., Qiao, C., Fu, J., Zhao, X. (2024). Adaptive multimodal cross-entropy loss for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*
- Chai, C.-Y., Wu, Y.-P., Tsao, S.-L. (2020). Deep depth fusion for black, transparent, reflective and texture-less objects. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 6766–6772. IEEE
- Wu, Z., Su, S., Chen, Q., Fan, R. (2023). Transparent objects: A corner case in stereo matching. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 12353–12359. IEEE
- Costanzino, A., Ramirez, P.Z., Poggi, M., Tosi, F., Mattoccia, S., Di Stefano, L. (2023). Learning depth estimation for transparent and mirror surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9244–9255
- Liu, Y., Ren, J., Zhang, J., Liu, J., Lin, M. (2020). Visually imbalanced stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2029–2038
- Chen, X., Xiong, Z., Cheng, Z., Peng, J., Zhang, Y., Zha, Z.-J. (2022). Degradation-agnostic correspondence from resolution-asymmetric stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12962–12971
- Song, T., Kim, S., Sohn, K. (2023). Unsupervised deep asymmetric stereo matching with spatially-adaptive self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13672–13680
- Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J. (2012). A naturalistic open source movie for optical flow evaluation. In *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part VI 12*, 611–625. Springer
- Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P. (2014). High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2–5, 2014, Proceedings 36*, 31–42. Springer
- Duggal, S., Wang, S., Ma, W.-C., Hu, R., Urtasun, R. (2019). Deep-pruner: Learning efficient stereo matching via differentiable patch-match. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4384–4393
- Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V. (2022). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3)
- Müller, T., Evans, A., Schied, C., & Keller, A. (2022). Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4), 1–15.
- Godard, C., Mac Aodha, O., Brostow, G.J. (2017). Unsupervised monocular depth estimation with left-right consistency. In: *CVPR*
- Zhou, Z., Dong, Q. (2023). Two-in-one depth: Bridging the gap between monocular and binocular self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9411–9421
- Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T. (2016). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4040–4048
- Cheng, X., Wang, P., & Yang, R. (2019). Learning depth with convolutional spatial propagation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10), 2361–2379.
- Zhang, F., Prisacariu, V., Yang, R., Torr, P.H. (2019). Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 185–194
- Yang, G., Manela, J., Happold, M., Ramanan, D. (2019). Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5515–5524
- Žbontar, J., & LeCun, Y. (2016). Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(65), 1–32.
- Liang, Z., Feng, Y., Guo, Y., Liu, H., Chen, W., Qiao, L., Zhou, L., Zhang, J. (2018). Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2811–2820
- Hirschmuller, H. (2007). Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2), 328–341.
- Jiang, H., Xu, R., Jiang, W. (2022). An improved raftstereo trained with a mixed dataset for the robust vision challenge 2022. arXiv preprint [arXiv:2210.12785](https://arxiv.org/abs/2210.12785)
- Rao, Z., Dai, Y., Shen, Z., He, R. (2022). Rethinking training strategy in stereo matching. *IEEE Transactions on Neural Networks and Learning Systems*
- Mehlretter, M., Heipke, C. (2019). Cnn-based cost volume analysis as confidence measure for dense matching. In *Proceedings of the*

- IEEE/CVF International Conference on Computer Vision Workshops*, 0–0
- Ilg, E., Saikia, T., Keuper, M., Brox, T. (2018). Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 614–630
- Batsos, K., Cai, C., Mordohai, P. (2018). CbmV: A coalesced bidirectional matching volume for disparity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2060–2069
- Geiger, A., Roser, M., Urtasun, R. (2010). Efficient large-scale stereo matching. In *Asian Conference on Computer Vision*, 25–38. Springer
- Yin, Z., Darrell, T., Yu, F. (2019). Hierarchical discrete distribution decomposition for match density estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6044–6053
- Schulze, M. (2011). A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Social Choice and Welfare*, 36(2), 267–303.
- Ramirez, P.Z., Tosi, F., Poggi, M., Salti, S., Mattoccia, S., Di Stefano, L. (2022). Open challenges in deep stereo: The booster dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21168–21178
- Ramirez, P.Z., Costanzino, A., Tosi, F., Poggi, M., Salti, S., Mattoccia, S., Di Stefano, L. (2023). Booster: a benchmark for depth from images of specular and transparent surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
- Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H. (2024). Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR
- Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H. (2024). Depth anything v2. [arXiv:2406.09414](https://arxiv.org/abs/2406.09414)
- Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J. (2024). Dust3r: Geometric 3d vision made easy. In: CVPR
- Leroy, V., Cabon, Y., Revaud, J. (2024). Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, 71–91. Springer
- Zhang, J., Herrmann, C., Hur, J., Jampani, V., Darrell, T., Cole, F., Sun, D., Yang, M.-H. (2024). MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion. <https://arxiv.org/abs/2410.03825>
- Shin, U., Park, J., Kweon, I.S. (2023). Deep depth estimation from thermal image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1043–1053
- Bao, W., Wang, W., Xu, Y., Guo, Y., Hong, S., & Zhang, X. (2020). Instereo2k: A large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63, 1–11.
- Chaney, K., Cladera, F., Wang, Z., Bisulco, A., Hsieh, M.A., Korpela, C., Kumar, V., Taylor, C.J., Daniilidis, K. (2023). M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 4015–4022
- Gehrig, M., Aarents, W., Gehrig, D., & Scaramuzza, D. (2021). Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3), 4947–4954.
- Zabih, R., Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. In *Computer Vision-ECCV'94: Third European Conference on Computer Vision Stockholm, Sweden, May 2–6 1994 Proceedings*, Volume II 3, 151–158. Springer
- Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C. (2015). Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3279–3286
- Žbontar, J., & LeCun, Y. (2016). Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(65), 1–32.
- Chen, Z., Sun, X., Wang, L., Yu, Y., Huang, C. (2015). A deep visual correspondence embedding model for stereo matching costs. In *Proceedings of the IEEE International Conference on Computer Vision*, 972–980
- Luo, W., Schwing, A.G., Urtasun, R. (2016). Efficient deep learning for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5695–5703
- Park, M.-G., Yoon, K.-J. (2015). Leveraging stereo matching with learning-based confidence measures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 101–109
- Spyropoulos, A., Komodakis, N., Mordohai, P. (2014). Learning to detect ground control points for improving the accuracy of stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1621–1628
- Poggi, M., Mattoccia, S. (2016). Learning a general-purpose confidence measure based on o(1) features and a smarter aggregation strategy for semi global matching. In *2016 Fourth International Conference on 3D Vision (3DV)*, 509–518. IEEE
- Shaked, A., Wolf, L. (2017) Improved stereo matching with constant highway networks and reflective confidence learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4641–4650
- Gidaris, S., Komodakis, N. (2017). Detect, replace, refine: Deep structured prediction for pixel wise labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5248–5257
- Batsos, K., Mordohai, P. (2018). Recresnet: A recurrent residual cnn architecture for disparity map enhancement. In *2018 International Conference on 3D Vision (3DV)*, 238–247. IEEE
- Ronneberger, O., Fischer, P., Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-assisted intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, 234–241. Springer
- Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J., Izadi, S. (2018). Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 573–590
- Wang, Y., Lai, Z., Huang, G., Wang, B.H., Van Der Maaten, L., Campbell, M., Weinberger, K.Q. (2019). Anytime stereo image depth estimation on mobile devices. In *2019 International Conference on Robotics and Automation (ICRA)*, 5893–5900. IEEE
- Song, X., Zhao, X., Hu, H., Fang, L. (2019). Edgestereo: A context integrated residual pyramid network for stereo matching. In *Computer Vision—ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, 20–35. Springer
- Cabon, Y., Murray, N., Humenberger, M. (2020). Virtual kitti 2. arXiv preprint [arXiv:2001.10773](https://arxiv.org/abs/2001.10773)
- Wang, Q., Zheng, S., Yan, Q., Deng, F., Zhao, K., Chu, X. (2021). Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE
- Mehl, L., Schmalfluss, J., Jahedi, A., Nalivayko, Y., Bruhn, A. (2023). Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Hua, Y., Kohli, P., Uplavikar, P., Ravi, A., Gunaseelan, S., Orozco, J., Li, E. (2020). Holopix50k: A large-scale in-the-wild stereo image dataset. arXiv preprint [arXiv:2003.11172](https://arxiv.org/abs/2003.11172)
- Treible, W., Saponaro, P., Sorensen, S., Kolagunda, A., O’Neal, M., Phelan, B., Sherbondy, K., Kambhamettu, C. (2017). Cats: A color and

- thermal stereo benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*
- Zhu, A. Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., & Daniilidis, K. (2018). The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3), 2032–2039. <https://doi.org/10.1109/LRA.2018.2800793>
- Zhang, J., Singh, S. (2014). Loam: Lidar odometry and mapping in real-time. In: *Robotics: Science and Systems*, 2, 1–9 Berkeley, CA
- Bai, C., Xiao, T., Chen, Y., Wang, H., Zhang, F., & Gao, X. (2022). Faster-lio: Lightweight tightly coupled lidar-inertial odometry using parallel sparse incremental voxels. *IEEE Robotics and Automation Letters*, 7(2), 4861–4868. <https://doi.org/10.1109/LRA.2022.3152830>
- Yang, G., Manela, J., Happold, M., Ramanan, D. (2019). Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V. (2017). Carla: An open urban driving simulator. In *Conference on Robot Learning*, 1–16. PMLR
- Gaidon, A., Wang, Q., Cabon, Y., Vig, E. (2016). Virtual worlds as proxy for multi-object tracking analysis. In: *CVPR*
- Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S. (2020). Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4909–4916. IEEE
- Jospin, L., Antony, A., Xu, L., Laga, H., Boussaid, F., & Bennamoun, M. (2022). Active-passive simstereo-benchmarking the cross-generalization capabilities of deep learning-based stereo methods. *Advances in Neural Information Processing Systems*, 35, 29235–29247.
- Wu, C.-Y., Wang, J., Hall, M., Neumann, U., Su, S. (2022). Toward practical monocular indoor depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3814–3824

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.