Submitted to **Journal of the European Optical Society-Rapid Publications**
Template provided by EDP Sciences

Review papers

A Review of Binocular Vision Stereo Matching Algorithms

Yuandong Niu[1], Li Gao[2], Wenwen Yu[1], Zhexuan Huang[1,*], Huaqing Liu[1]

[1] Shijiazhuang Campus, Army Engineering University of PLA, Shijiazhuang 050003, PR China
[2] China Electronics Standardization Institute, Beijing 100007, PR China

Abstract. Binocular stereo vision features low cost and non-contact sensing, making it a mainstream approach to acquire 3D scene information in fields such as 3D reconstruction and autonomous driving. As its core module, stereo matching is the key determinant of system sensing accuracy and efficiency. This paper presents a comprehensive review of binocular stereo matching. It systematically elaborates the imaging principle of binocular stereo vision and clarifies the pivotal role of stereo matching in 3D reconstruction. According to technical development trends, existing stereo matching algorithms are classified into traditional methods and deep learning-based methods. This paper also summarizes two mainstream categories of datasets: real-scene datasets and synthetic datasets. Combined with mainstream evaluation metrics including end-point error and bad pixel rate, a standardized evaluation system for stereo matching algorithms is established. Finally, the current challenges are analyzed, such as poor scene generalization and insufficient datasets for complex working conditions. Future research directions are further prospected, including multi-modal dataset construction and cross-domain model optimization, aiming to provide a solid reference for related theoretical research and engineering applications.
Keywords: binocular stereo vision; stereo matching; 3D reconstruction; deep learning; autonomous driving

1. Introduction

With the rapid development of computer vision technology, industrial applications in fields such as 3D reconstruction, autonomous driving, digital twins, smart cities, and virtual reality have been continuously expanded, and the demand for accurate acquisition of 3D scene information has become increasingly urgent [1-3]. As a low-cost and non-contact 3D perception method, binocular vision technology features a simple structure, favorable real-time performance, and high measurement accuracy, making it one of the mainstream techniques for 3D information extraction [4-9]. The core of a binocular vision system is to simulate the visual mechanism of human eyes. Two cameras capture images of the same scene from different viewpoints, and stereo matching is employed to establish the correspondence mapping between the left and right images, so as to infer the 3D spatial information of the scene. As a core module of 3D reconstruction, stereo matching directly determines the overall performance of a binocular vision system in terms of matching accuracy and efficiency, and serves as a key bottleneck restricting the quality of 3D information acquisition.

In recent years, the continuous integration of deep learning and traditional computer vision methods has accelerated the development of stereo matching. Research has evolved from traditional global matching [10-31], local algorithms [32-43], and semi-global matching [44-49] to end-to-end matching models based on convolutional neural networks [50-90]. Relevant methods have been continuously optimized in terms of matching accuracy and robustness, and their application scenarios have been gradually expanded. However, complex real-world scenarios still present prominent challenges, such as illumination variations, weak textures, foreground occlusion, and multi-scale interference. Current algorithms involve inherent trade-offs among accuracy, computational efficiency, and real-time performance, while systematic summaries and horizontal comparative analyses remain insufficient.

Accordingly, this paper presents a review of binocular stereo matching technology. It summarizes the basic principles of binocular stereo perception and the technical development context, analyzes the technical characteristics of traditional and deep learning-based algorithms, and compares their respective merits, drawbacks, and application limitations. Combined with mainstream datasets and quantitative evaluation metrics [91-100], this paper summarizes the current technical difficulties and

* Corresponding author: lgdhuang@163.com

practical challenges, and prospects future development trends in this field. This study aims to provide theoretical references for related research and engineering applications, and promote the practical application of binocular stereo matching technology.

2. Binocular Stereo Vision

The core task of traditional stereo matching algorithms is to realize pixel-wise disparity estimation through classical mathematical models in accordance with binocular imaging geometric rules and image pixel features. The overall mathematical system mainly consists of three components: the binocular triangulation geometric principle, the grayscale similarity matching metrics, and the global energy minimization model. These three parts correspond respectively to the geometric basis, matching criteria and optimization solution mechanism of stereo matching, laying a theoretical foundation for readers to understand conventional disparity estimation algorithms.

2.1. Binocular Triangulation Geometric Principle

The binocular stereo vision system is an important means to acquire three-dimensional scene information, and stereo matching serves as the core step of 3D reconstruction [101, 102]. These two cited classic references systematically elaborate the fundamental theories of multi-view geometry, laying a solid theoretical foundation for the epipolar constraint, stereo rectification and triangulation introduced in this section. First, epipolar geometry is used to establish the geometric relationship between dual-view images. The left and right views are rectified into parallel views, which limits the search range of corresponding points to epipolar lines. Next, image features are utilized to complete corresponding point matching and establish the correspondence of 3D spatial points between the two views. Finally, disparity is solved based on the triangulation principle to obtain depth information. On this basis, we further deduce the mathematical relationship between disparity and scene depth, so as to clearly explain the imaging principle of binocular vision and facilitate readers to understand the subsequent various advanced stereo matching algorithms.

The binocular stereo vision system is shown in Figure 1, where L and R denote the left and right cameras with equal focal lengths. B is the camera baseline distance, and f is the camera focal length. The imaging planes and optical axes of the left and right cameras are completely parallel, and their X-axes are strictly aligned.

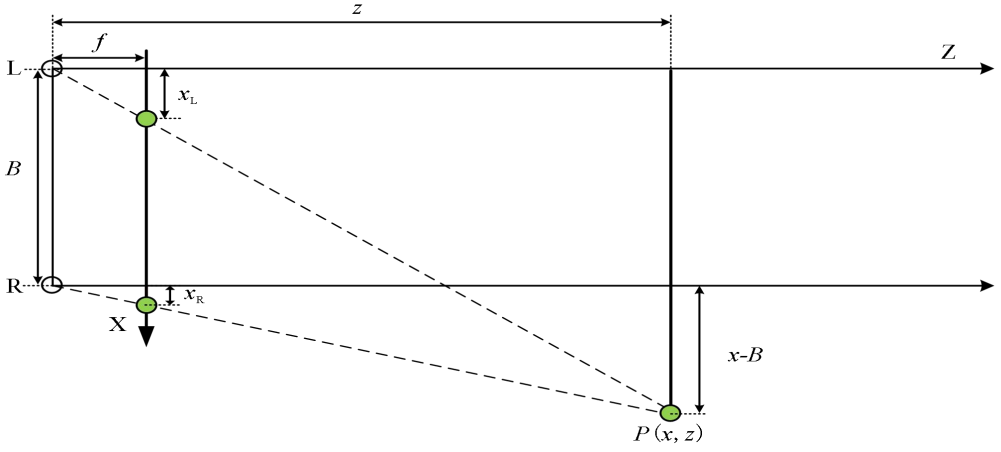

Fig. 1. Schematic of binocular vision system

The binocular system in Fig. 1 adopts the ideal pinhole camera model, a simplified camera ray projection model that ignores lens distortion and optical aberration. Any 3D spatial point emits light rays passing through the camera optical center and projects onto the 2D imaging plane. To constrain the matching search range and reduce computational overhead, the raw binocular images are geometrically rectified into a parallel forward optical configuration, where the optical axes of two cameras are strictly parallel and their X-axes are collinear. Typical binocular optical configurations mainly include convergent setup, vertical setup and rectified parallel forward setup. We select the rectified parallel configuration in this review, since it restricts corresponding points to horizontal epipolar lines, converting the 2D cross-image matching task into a 1D horizontal correlation search, which is friendly for signal processing researchers to analyze the pixel-wise cross-view correlation process.

Assume that the coordinates of the spatial point P in the X−Z plane are (x,z), where z denotes the depth value of the point. According to the principle of similar triangles:

$$\frac{z}{f}=\frac{x}{x_{\mathrm{L}}} \tag{1}$$

$$\frac{z}{f}=\frac{x-B}{x_{\mathrm{R}}} \tag{2}$$

Defining the disparity d=x_{L}-x_{R} and combining Equations (1) and (2), the depth value z can be derived as:

$$z=\frac{f\cdot B}{x_{\mathrm{L}}-x_{\mathrm{R}}}=\frac{f\cdot B}{d} \tag{3}$$

As shown in Equation (3), when the camera focal length f and baseline B are fixed, the depth of a spatial point can be calculated solely by obtaining its corresponding disparity. Furthermore, the closer the target is to the camera, the larger the disparity value. From the perspective of signal processing, rectified left and right images can be regarded as two groups of spatially correlated observation signals. Stereo matching essentially implements pixel-level cross-correlation matching between two signal maps. The horizontal offset of the maximum correlation position is defined as disparity d. According to the triangulation formula, depth is a monotonic inverse mapping function of disparity, which realizes the decoding from two-dimensional image signals to three-dimensional spatial depth information.

Based on the above binocular stereo vision framework, stereo matching aims to search for pixel-wise corresponding points on rectified binocular images to generate a dense disparity map. Essentially, it is a cross-view pixel correlation task. By locating the matching pixel of each pixel from the left image in the right image and calculating the horizontal coordinate difference (i.e., disparity), we can finally reconstruct the 3D depth information of the entire scene.

2.2. Grayscale Similarity Matching Metrics

Traditional local stereo matching relies on grayscale intensity to quantify the dissimilarity between corresponding pixels of rectified left and right images, where matching cost functions are defined to measure pixel-wise matching errors. Directly computing the matching cost of a single pixel is vulnerable to image noise. Accordingly, two mainstream optimization strategies are proposed: designing robust per-pixel cost functions and aggregating matching costs within a local support window.

2.2.1. Per-pixel Grayscale Cost Functions

In this subsection, $I_R(x, y)$ stands for the grayscale intensity of the rectified reference image at pixel coordinate (x, y), and $I_T(x+d, y)$ corresponds to the grayscale intensity of the rectified target image. Given the reference image I_R and the target image I_T, the matching cost quantifies the grayscale dissimilarity between the pixel (x, y) in the reference image and the pixel ($x+d$, y) in the target image under the candidate disparity d. Three typical per-pixel cost functions are elaborated as follows: Absolute Differences (AD), Squared Differences (SD), and Truncated Absolute Differences (TAD).

AD represents the absolute grayscale difference between two matched pixels, which can be formulated as:

$$AD(x,y,d)=\left|I_R(x,y)-I_T(x+d,y)\right| \tag{4}$$

This cost features low computational complexity yet suffers from high sensitivity to image noise.

SD denotes the squared value of pixel grayscale deviation, formulated as:

$$SD(x,y,d)=\left(I_R(x,y)-I_T(x+d,y)\right)^2 \tag{5}$$

It enlarges intensity discrepancies via squaring operation, but noise outliers will also be amplified and result in abnormal matching costs.

TAD refers to the absolute grayscale difference limited by a manually set threshold, which can be expressed as:

$$TAD(x,y,d)=\min\left\{\left|I_R(x,y)-I_T(x+d,y)\right|,T\right\} \tag{6}$$

where T is a predefined threshold. This function restricts the maximum matching cost to eliminate abnormal errors triggered by noise, occlusion or illumination variations, greatly enhancing the matching robustness of per-pixel strategies.

2.2.2. Aggregated Cost Based on Local Support Window

Per-pixel matching only utilizes individual pixel information and tends to produce ambiguous results in textureless regions. To address this issue, a local support window S centered on the target pixel is adopted to aggregate the matching cost of all

pixels inside the window, which leverages regional spatial consistency prior to improve matching reliability. Three typical aggregated cost functions are listed below: Sum of Absolute Differences (SAD), Sum of Squared Differences (SSD), and Sum of Truncated Absolute Differences (STAD).

SAD represents the cumulative sum of absolute grayscale differences of all pixels within the local support window, which can be formulated as:

$$SAD(x,y,d)=\sum_{x\in S}\left|I_R(x,y)-I_T(x+d,y)\right| \quad (7)$$

This cost makes full use of regional grayscale distribution information, which effectively suppresses random noise interference compared with per-pixel matching.

SSD denotes the sum of squared pixel grayscale deviations inside the support window, formulated as:

$$SSD(x,y,d)=\sum_{x\in S}\left(I_R(x,y)-I_T(x+d,y)\right)^2 \quad (8)$$

It amplifies local intensity mismatches through the square operation, yet extreme noise outliers will also be magnified and lead to unreasonable matching costs.

STAD refers to the accumulated truncated absolute grayscale difference constrained by a predefined threshold, which can be expressed as:

$$STAD(x,y,d)=\sum_{x\in S}\min\left\{\left|I_R(x,y)-I_T(x+d,y)\right|,T\right\} \quad (9)$$

where T is a predefined threshold. This function limits the maximum per-pixel matching error within the window to filter abnormal values caused by noise, occlusion or illumination variations, significantly improving the matching robustness of local window-based strategies.

After calculating the matching cost for all pixels within the predefined disparity range [$d_{\min}$, $d_{\max}$], all cost values are stored in a three-dimensional data structure named cost volume with the dimension of $H\times W\times D$, which serves as the fundamental input for subsequent cost aggregation and global disparity optimization. Besides the grayscale-based cost metrics, other alternatives such as Normalized Cross-Correlation (NCC), image gradient and Census transform are also widely applied to handle complex scenarios with dramatic illumination variations and repeated textures.

2.3. Global Energy Minimization Model

2.3.1. Global Energy Function Definition

To eliminate massive mismatches caused by local window-based matching in textureless and occluded regions, traditional global stereo matching converts disparity estimation into a constrained optimization problem. The global energy function is defined as:

$$E(d)=E_{\text{data}}(d)-E_{\text{smooth}}(d) \quad (10)$$

The data term $E_{\text{data}}(d)$ is derived from the precomputed cost volume, which restricts the matching cost of each candidate disparity to ensure the estimated result conforms to the grayscale feature of binocular images. Since the cost volume always contains noise and outliers, the smoothness term $E_{\text{smooth}}(d)$ is introduced as a global constraint. Based on the prior that scene depth changes continuously, this term imposes penalty on abrupt disparity jumps between adjacent pixels, allowing dramatic disparity variations only at object boundaries.

2.3.2. Classical Global Optimization Solvers

Minimizing the global energy function is an NP-hard problem, so a variety of approximate optimization algorithms have been extensively applied to seek feasible optimal solutions. Graph Cuts builds a graph model to find the minimum cut for globally optimal disparity configuration and performs remarkably well in processing occluded areas; Belief Propagation achieves probabilistic global optimization through the iterative transmission of contextual pixel information, and can be integrated with image segmentation to boost matching accuracy in ambiguous regions; Cooperative Optimization leverages region-based iterative constraints to address matching ambiguity arising from repeated textures; Dynamic Programming reduces computational overhead by limiting optimization to each independent scanline, yet it cannot make use of spatial constraints across different scanlines, which limits its overall matching performance.

2.3.3. Disparity Refinement and Traditional Stereo Matching Pipeline

Even after global optimization, the initial disparity map still contains invalid values caused by occlusion and noise. Therefore, disparity refinement is required to correct outliers, fill disparity holes and eliminate mismatches via left-right consistency check, median filtering and occlusion filling.

The complete pipeline of traditional stereo matching is summarized as follows: cost volume construction → cost computation → local cost aggregation or global energy optimization → disparity calculation via Winner-Takes-All (WTA) or

global optimization → disparity refinement. This entire mathematical framework lays the theoretical foundation for analyzing the inherent limitations of conventional stereo matching algorithms.

3. Stereo Matching Algorithms

As shown in Figure 2, stereo matching algorithms are divided into two categories based on their technical routes: traditional methods and deep learning-based methods. From the perspective of technical evolution, they represent two entirely distinct developmental paradigms for stereo matching, accompanied by clear technical inheritance between the two research lines. This chapter first analyzes handcrafted stereo matching algorithms with handcrafted cost functions in Section 3.1, which are classified into global, local, and semi-global branches. Subsequently, Section 3.2 systematically reviews mainstream deep learning stereo architectures evolved from early deep stereo approaches, including cost-volume-based CNNs and Transformer-based models.

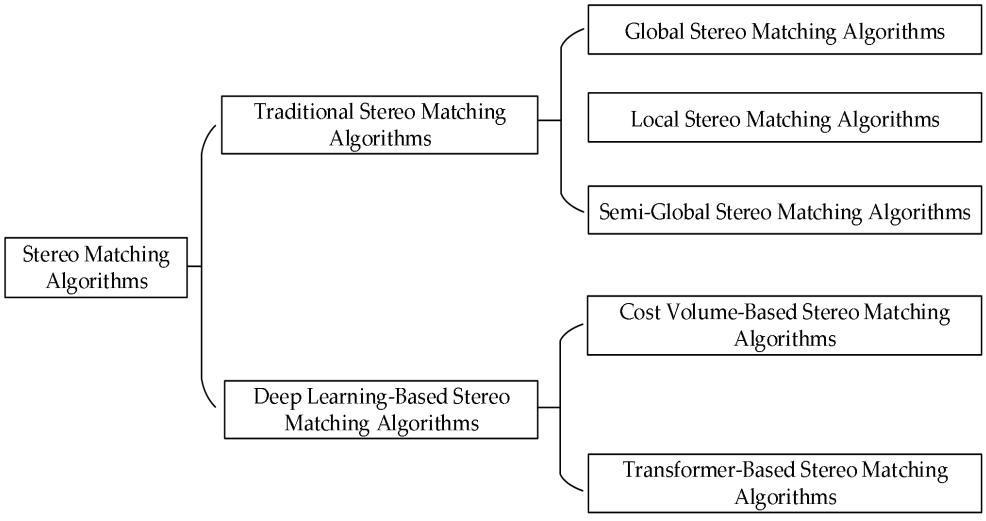

Fig. 2. Classification of stereo matching algorithms, which are categorized into traditional and deep learning-based methods with detailed subcategories illustrated in the framework.

3.1. Traditional Stereo Matching Algorithms

In 2002, Scharstein, the developer of the Middlebury dataset, systematically reviewed the framework and classification of stereo matching algorithms, and divided traditional methods into global and local stereo matching algorithms [10]. In 2005, to balance matching accuracy and computational efficiency, Hirschmüller integrated global and local optimization strategies and proposed the Semi-Global Matching (SGM) algorithm [11]. Accordingly, this section classifies traditional stereo matching algorithms into three categories: global methods, local methods, and semi-global methods, and elaborates on each in detail.

3.1.1. Global Stereo Matching Algorithms

Global stereo matching algorithms are based on global optimization theory. By constructing a global energy function composed of a data term and a smoothness term, they transform disparity estimation into an energy function minimization problem. With global smoothness constraints, these algorithms well cope with challenging scenarios including occlusions and textureless regions, and thus attain high matching accuracy. However, the global optimization process typically suffers from large memory usage, high computational complexity, and slow execution speed, making it difficult to meet real-time requirements. Due to their high precision, global stereo matching algorithms are widely used in applications requiring high accuracy but low real-time performance, such as cultural heritage preservation, industrial vision and high-precision mapping. Representative global stereo matching algorithms include Graph Cuts, Belief Propagation, and Dynamic Programming.

Among various global algorithms, Graph Cuts is one of the most widely adopted. Boykov et al. first proposed applying Graph Cuts to solve the energy function minimization problem and verified its effectiveness in stereo vision tasks [12]. Kolmogorov and Zabih applied Graph Cuts to stereo matching, effectively addressing the occlusion problem [13]. Kang et al. proposed a method combining local techniques with movable windows and dynamic selection, as well as Graph Cuts-based

global energy minimization, which also handles occlusions effectively [14]. Kolmogorov used Graph Cuts to solve multi-view 3D reconstruction problems [15]. Li et al. proposed a Graph Cuts-based stereo matching algorithm for homogeneous color regions, which achieves superior performance in textureless regions, disparity discontinuities, and occluded areas [16]. Taniai et al. presented a continuous Graph Cuts-based stereo matching algorithm, which propagates spatial information via locally shared labels for energy minimization, and verified its effectiveness on the Middlebury dataset [17].

Belief Propagation is another classic global optimization scheme. Sun et al. proposed a stereo matching algorithm based on a Bayesian framework, using Belief Propagation to solve the energy minimization problem [18]. Felzenszwalb and Huttenlocher proposed an energy minimization method based on max-product Belief Propagation to accelerate computation [19]. Sun et al. proposed a symmetric stereo model and adopted an iterative optimization scheme to approximate the minimum energy via Belief Propagation, with effectiveness validated on the Middlebury dataset [20]. Klaus et al. assigned image segments to appropriate disparity planes and solved the energy minimization problem using Belief Propagation [21]. Yang et al. developed a stereo matching algorithm integrating color-weighted correlation, hierarchical Belief Propagation, and occlusion handling to improve local robustness and achieve global optimization [22]. Yang et al. further proposed a constant-space Belief Propagation algorithm for stereo matching. It reduces memory overhead and computational complexity by narrowing the disparity search range and dynamically calculating data terms [23].

Dynamic Programming algorithms use a recursive optimization strategy to solve for disparity. Veksler proposed a global stereo matching algorithm based on dynamic programming. It minimizes the energy function by recursively solving the optimal disparity assignment for each node, and performs optimization on tree structures instead of a single scanline [24]. Aboali et al. compared Dynamic Programming-based stereo matching with basic block matching and found that the former yields superior smoothness and accuracy but lower computational efficiency [25].

Besides the above three typical schemes, many improved global stereo matching algorithms have also been developed. Luo et al. proposed an intensity-based cooperative bidirectional stereo matching algorithm. It adopts a hierarchical multi-scale structure and a two-stage relaxation algorithm to minimize the non-convex energy function, enabling 3D depth estimation for binocular and trinocular systems [26]. Scharstein and Szeliski proposed a stereo matching algorithm combining nonlinear diffusion and a Bayesian model. It adopts adaptive aggregation to optimize support region selection, achieving favorable performance in noise suppression and edge preservation [27]. Pritchett and Zisserman proposed a wide-baseline stereo matching algorithm that automatically estimates epipolar geometry and matches wide-baseline views, solving the problem that traditional small-baseline algorithms fail to handle large disparities [28]. Zitnick and Kanade proposed a cooperative stereo matching algorithm based on uniqueness and continuity assumptions, which resolves the mismatch problem of traditional window-based methods in low-texture, edge, and occluded regions [29]. Barron et al. proposed a global stereo matching algorithm based on bilateral space. It efficiently generates disparity maps suitable for synthetic defocus via low-dimensional mapping and convex optimization, ensuring edge consistency while running 10−100 times faster than traditional methods [30]. Li et al. proposed a PatchMatch-based superpixel segmentation algorithm that achieves leading performance on Middlebury datasets by combining CNN features, two-layer matching costs, and slanted support windows [31].

3.1.2. Local Stereo Matching Algorithms

Local stereo matching algorithms adopt a local optimization strategy for disparity estimation. Although they also adopt energy minimization for solution similar to global methods, their energy functions only contain data terms without smoothness terms. Since the disparity of each pixel is calculated independently without mutual interference, such algorithms feature simple computation, easy deployment and real-time capability. However, their core assumption - that all pixels within a local window share the same disparity - often fails at edges and depth discontinuities, which easily causes matching errors and yields inferior performance compared with global algorithms. Thanks to their high computational speed, local stereo matching algorithms are widely deployed in scenarios with strict real-time requirements, such as autonomous driving and UAV navigation.

Kanade et al. proposed an adaptive window stereo matching algorithm, which dynamically adjusts window size and shape by quantifying the effects of disparity and intensity variations via statistical modeling [32]. Koschan et al. presented a local algorithm combining image pyramid-based optimized block matching with active color illumination. It effectively improves matching quality and efficiency, especially performing well in homogeneous regions [33]. Veksler proposed a fast variable-window method based on integral images. It addresses the drawbacks of traditional fixed-window schemes, such as vulnerability to noise in low-texture areas and edge blurring at disparity discontinuities [34]. Yoon et al. proposed an adaptive support weight algorithm. By dynamically adjusting pixel support weights within the window according to color similarity and geometric proximity, it improves disparity accuracy in both uniform and depth-discontinuous regions without global reasoning [35]. Goesele et al. developed a method built on local matching, and enhanced its robustness and consistency via global constraints. It enables high-quality 3D reconstruction from community photo collections with drastic variations in illumination and scale [36]. Min et al. presented a stereo matching method based on weighted least squares. It achieves disparity estimation through weighted least-squares cost aggregation, multi-scale acceleration and asymmetric occlusion handling, and its effectiveness is validated on the Middlebury dataset without global optimization [37]. Geiger et al. proposed a fast stereo

matching algorithm for high-resolution images. It constructs disparity priors by triangulating support point sets, reduces matching ambiguity of remaining pixels, and achieves high-precision dense reconstruction without global optimization [38]. Mei et al. established a GPU-based stereo matching system, which realizes high-precision matching via AD-Census cost initialization, improved cross-region cost aggregation, scanline optimization and multi-step disparity refinement [39]. Bleyer et al. proposed a PatchMatch-based stereo matching algorithm. By estimating a 3D plane for each pixel, projecting support regions onto the plane to handle slanted surfaces, and optimizing via spatial, view, and temporal propagation, the method achieves sub-pixel accuracy [40]. Ma and He proposed a constant-time weighted median filtering algorithm for disparity refinement in stereo matching, and verified the essential role of disparity refinement by constructing a local stereo matching framework [41]. Hosni et al. presented a fast cost volume filtering framework based on guided filtering for local stereo matching [42]. Zhang et al. proposed a cross-scale cost aggregation framework, which unifies existing cost aggregation methods into a weighted least squares optimization problem. By introducing cross-scale regularization to fuse multi-scale cost volume information, it substantially boosts stereo matching performance in low-texture regions with only a slight increase in computational complexity [43].

3.1.3. Semi-Global Stereo Matching Algorithms

Local methods tend to blur object boundaries and are sensitive to illumination changes, while global methods achieve high accuracy but suffer from high computational complexity and low speed. To address the limitations of both local and global stereo matching algorithms, researchers have proposed semi-global stereo matching algorithms and their improved schemes. Semi-global stereo matching algorithms integrate the advantages of local and global algorithms. On the one hand, they retain the global energy function framework to ensure matching accuracy. On the other hand, they replace the two-dimensional optimization of global algorithms with multi-directional one-dimensional path aggregation. By using locally optimal results to approximate the global optimum, they achieve a good balance between accuracy and efficiency. Due to their ability to balance accuracy and efficiency, semi-global stereo matching methods are widely applied in fields such as digital surface model generation and high-precision mapping for autonomous driving.

In 2005, Hirschmüller proposed the SGM algorithm, which pioneered semi-global stereo matching [11]. Based on mutual information for pixel-wise matching, it employs multiple one-dimensional constraints to approximate global smoothness, preserving both local detail and global consistency. It achieves better performance than local methods, with accuracy close to that of global methods and much higher computational efficiency. Mattoccia et al. achieved accurate dense disparity estimation and boundary localization via segmentation-based disparity refinement [44]. This method exhibited superior performance on the Middlebury dataset. Gehrig et al. adopted fractional disparity sampling, disparity smoothing, and gravity constraints to enhance the sub-pixel accuracy of semi-global stereo matching in low-texture areas [45]. On the basis of the SGM algorithm, Hirschmüller introduced a post-processing module to improve its robustness, and verified the method via engineering experiments on aerial and airborne scanning images [46]. Mattoccia post-processed disparity maps from semi-global stereo matching algorithms by relaxing local consistency constraints and shrinking active support windows to significantly improve efficiency [47]. Mattoccia also proposed a superpixel-constrained local consistency method to improve the accuracy of semi-global stereo matching algorithms [48]. Michael et al. presented an improved SGM method to overcome the limited adaptability of conventional SGM with fixed penalty parameters, thus enhancing matching accuracy in textureless and edge regions [49].

3.2. Deep Learning-Based Stereo Matching Algorithms

Traditional stereo matching algorithms have achieved promising performance, but suffer from limited feature extraction capability. They perform poorly in occluded and weak-texture regions, and fail to fully exploit prior information from datasets, creating obvious performance bottlenecks. Since the breakthrough of convolutional neural networks in 2012, deep learning has developed rapidly in computer vision, providing effective solutions to the inherent shortcomings of conventional stereo matching methods [50].

In 2015, Žbontar et al. first applied CNNs to matching cost computation, and completed stereo matching by adopting cost aggregation, semi-global matching and other post-processing strategies [51]. They further designed diverse network architectures and enhanced dataset adaptability to optimize matching accuracy and efficiency [52]. Chen et al. constructed a deep visual correspondence embedding model for data-driven cost estimation to achieve efficient disparity prediction [53]. Luo et al. optimized network architectures and loss functions to significantly improve stereo matching efficiency [54]. Shaked et al. introduced residual modules to address the convergence problem of deep networks in stereo matching tasks [55]. Park adopted pyramid pooling modules to enlarge the receptive field and aggregate multi-scale features, thereby improving overall matching performance [56]. To quantitatively evaluate the performance of the above representative traditional and early deep-learning stereo matching algorithms, we summarize their experimental results on mainstream public datasets in Table 1. End-Point

Error (EPE) and Bad-Pixel Rate are adopted as the two core evaluation metrics. This table reports the quantitative results of these classical stereo matching methods, where the pioneering end-to-end DispNet proposed by Mayer et al. [57] is introduced as a benchmark for fair horizontal comparison.

Table 1. Performance comparison of representative traditional and early deep-learning stereo matching algorithms on public datasets.

References	Dataset	Operating Environment	EPE（px）	Bad-Pixel Rate(%)
Hirschmuller et al, 2005 [11]	SceneFlow	Xeon	8.7	—
	KITTI 2012		10.06	—
	KITTI2015		7.21	10.86
Žbontar et al, 2015 [51]	KITTI2012	Nvidia GeForce GTX Titan GPU	—	2.61
Žbontar et al, 2016 [52]	KITTI2012	Nvidia GeForce GTX Titan X GPU	—	2.43
	KITTI2015			3.89
	Middlebury			8.29
Chen et al, 2015 [53]	KITTI2012	Nvidia GeForce GTX Titan GPU	0.9	3.10
Luo et al, 2016 [54]	KITTI2012	Nvidia GeForce GTX Titan X GPU	0.8	3.07
	KITTI2015		1.84	7.23
Shaked et al, 2017 [55]	KITTI2012	Nvidia GeForce GTX Titan X GPU	—	2.27
	KITTI2015			2.91
Park et al, 2017 [56]	Middlebury	Nvidia GeForce GTX Titan GPU	—	8.45
Mayer et al, 2016 [57]	SceneFlow KITTI2012 KITTI2015	Nvidia GeForce GTX Titan X GPU	1.0 (KITTI2012 test)	4.34 (KITTI2015 test)

As summarized in Table 1, the experimental results reveal that CNN-based stereo matching algorithms achieve higher matching accuracy than traditional methods. Nevertheless, most early CNN schemes still adopt the traditional multi-stage pipeline and rely on handcrafted modules for feature extraction and cost calculation; these handcrafted optimization steps impose an upper bound on matching performance, resulting in severe performance degradation and poor cross-dataset generalization. To break this bottleneck, researchers converted discrete processing steps into differentiable computations and proposed end-to-end stereo learning frameworks. In 2016, Mayer et al. put forward the encoder–decoder-based DispNet, the first end-to-end stereo matching network [57]. Shortly afterwards, Flynn et al. developed DeepStereo to implement end-to-end depth estimation via plane sweep volume reprojection [58]. These architectures greatly boost disparity prediction accuracy in repetitive-texture, occluded and textureless regions by leveraging dataset prior knowledge. At present, modern deep stereo matching algorithms fall into two mainstream categories: cost volume-based frameworks and Transformer-based models.

3.2.1. Cost Volume-Based Stereo Matching Algorithms

Kendall et al. proposed GC-Net, which innovatively employs 3D convolution to build a cost volume and regresses disparity maps from matching cost volumes with contextual information [59]. Its network architecture is shown in Figure 3. Pang et al. established a cascaded residual learning framework and improved disparity estimation accuracy through multi-scale residual learning, achieving competitive performance on the KITTI 2015 dataset [60]. Ye et al. adopted a multi-scale pooling module to expand the receptive field while preserving image resolution, thereby enhancing stereo matching performance [61]. Khamis et al. presented StereoNet. Leveraging high-precision sub-pixel matching, this method adopts low-resolution cost volumes and hierarchical edge-aware refinement to achieve real-time stereo matching [62]. Liang et al. proposed iResNet, which integrates four key steps of stereo matching: cost computation, cost aggregation, disparity calculation, and disparity optimization [63]. This method obtains basic disparity information through multi-scale shared feature extraction and initial disparity estimation, and further optimizes disparity based on feature consistency. Chang et al. proposed PSM-Net, which aggregates multi-scale contextual information using the spatial pyramid pooling (SPP) module and regularizes cost volumes via stacked 3D hourglass CNNs to achieve end-to-end stereo matching [64]. Guo et al. proposed SegStereo, which constructs cost volumes by computing feature correlation between left and right images [65]. Song et al. designed EdgeStereo and further optimized it in subsequent work, replacing traditional cascaded architectures and 3D convolutional regularization modules with a residual pyramid structure [66,67]. Guo et al. presented GwcNet, which divides features along channels to construct group-wise correlation cost volumes and adopts an improved stacked 3D hourglass network for cost aggregation to achieve end-to-end stereo matching [68]. Based on PSM-Net, Nie et al. proposed a multi-level contextual hyper-aggregation strategy to optimize feature aggregation in the cost computation stage [69]. Zhang et al. constructed GA-Net by combining GC-Net with the

traditional SGM algorithm, in which semi-global and local guided aggregation layers replace 3D convolutions for cost aggregation, with its effectiveness validated on SceneFlow, KITTI and other datasets [70]. Zhang et al. proposed DSMNet. With domain normalization and structure-preserving graph filters, this method achieves excellent performance on real datasets like KITTI by training solely on synthetic data without target-domain fine-tuning [71]. Duggal et al. developed DeepPruner, which uses a differentiable PatchMatch module and a confidence range predictor to narrow the disparity search space, thus realizing efficient stereo matching via cost aggregation and image-guided refinement [72]. Xu et al. proposed AANet, which adopts adaptive single-scale and cross-scale cost aggregation to replace 3D convolution operations [73]. Lipson et al. presented RAFT-Stereo, which constructs a 3D correlation volume via the dot product of left and right feature vectors at the same height and builds a correlation pyramid to fully exploit feature information [74]. Li et al. proposed CREStereo, which constructs cost volumes via adaptive group-wise correlation layers for local correlation calculation and completes disparity estimation with cascaded recurrent networks [75]. Shen et al. designed PCW-Net, which extracts domain-invariant features and optimizes disparity maps via pyramid and deformable 3D cost volumes, achieving excellent cross-domain generalization and matching accuracy [76]. Xu et al. proposed IGEV-Stereo, which constructs a combined geometric encoding volume integrating global geometric clues and local matching details, and iteratively updates disparity through ConvGRU [77]. Jing et al. proposed Match-Stereo-Videos, which formulates stereo matching as local matching and global aggregation to achieve temporally consistent dynamic stereo matching [78]. Chen et al. presented MoCha-Stereo based on cost volume learning to solve edge matching degradation caused by the loss of geometric structural information in feature channels [79]. Bartolomei et al. constructed a two-branch framework combining stereo matching geometric constraints and monocular depth priors from foundation models, with cost volume fusion and enhancement strategies enabling robust depth estimation under complex zero-shot conditions [80]. Cheng et al. proposed MonSter, which integrates monocular depth estimation and stereo matching into a two-branch architecture for mutual iterative optimization. Its stereo matching module is developed based on the IGEV framework [81]. To quantitatively evaluate the performance of various stereo matching algorithms with different cost volume construction strategies, we summarize their experimental results on mainstream public benchmark datasets in Table 2.

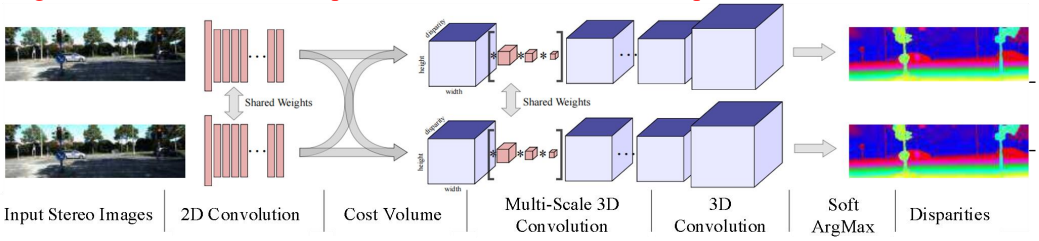

Fig. 3. The network architecture of GC-Net [59]. This network extracts binocular image features via shared-weight 2D convolution, constructs a 4D cost volume, and adopts multi-scale 3D convolution and Soft ArgMax regression to predict high-precision disparity maps, which serves as a classic baseline for cost-volume-based stereo matching methods.

Table 2. Performance comparison of cost volume-based stereo matching algorithms on public datasets.

References	Dataset	Operating Environment	EPE（px）	Bad-Pixel Rate(%)
Kendall et al, 2017 [59]	KITTI2012	Nvidia GeForce GTX Titan X GPU	0.6	1.77
	KITTI2015		—	2.61
Pang et al, 2017 [60]	SceneFlow	Nvidia GeForce GTX Titan X GPU	1.32	6.20
	Middlebury		1.40	9.13
	KITTI2015		0.68	2.10
Ye et al, 2017 [61]	Middlebury	Nvidia GeForce GTX Titan X GPU	—	3.92
Khamis et al, 2018 [62]	SceneFlow	Nvidia GeForce GTX Titan X GPU	0.768	—
	KITTI2012		0.8	4.91
	KITTI2015			4.83
Liang et al, 2018 [63]	SceneFlow	Nvidia GeForce GTX Titan X GPU	1.40	4.57
	KITTI2012		—	1.71
	KITTI2015		—	2.19
Chang et al, 2018 [64]	SceneFlow	Nvidia GeForce GTX Titan GPU	1.09	—
	KITTI2012		0.5	1.49
	KITTI2015		—	1.83
Yang et al, 2018 [65]	SceneFlow	Nvidia GeForce GTX Titan	1.45	3.50

	KITTI2012	X GPU	0.5	1.68
	KITTI2015		1.96	2.08
	SceneFlow		1.11	4.97
Song et al, 2018 [66]	KITTI2012	Nvidia GeForce GTX Titan X GPU	—	1.73
	KITTI2015		—	2.59
	SceneFlow		0.74	3.96
Song et al, 2020 [67]	KITTI2012	Nvidia GTX 1080Ti GPU	0.4	1.46
	KITTI2015		—	2.08
	Middlebury		1.0	5.84
Guo et al, 2019 [68]	KITTI2012	Nvidia Titan Xp GPU	0.5	1.32
	KITTI2015		—	2.11
	SceneFlow		0.56	—
Nie et al, 2019 [69]	KITTI2012	Nvidia Titan Xp GPU	0.4	1.26
	KITTI2015		—	2.09
	SceneFlow		0.84	9.9
Zhang et al, 2019 [70]	KITTI2012	TESLA P40 GPU	0.5	1.36
	KITTI2015		—	2.03
	KITTI2012			3.9
Zhang et al, 2020 [71]	KITTI2015	GPU memory: >=5G (for testing); >=11G (for training)	—	4.1
	Middlebury			20.1
	ETH3D			6.0
Duggal et al, 2019 [72]	SceneFlow	Nvidia Titan Xp GPU	0.85	—
	KITTI2015		—	1.80
	SceneFlow		0.83	—
Xu et al, 2020 [73]	KITTI2012	NVIDIA V100 GPU	—	1.71
	KITTI2015		—	2.24
	SceneFlow			7.92
Lipson et al, 2021 [74]	ETH3D	RTX 6000 GPU	—	2.44
	Middlebury			4.74
	KITTI2015			1.96
Xu et al, 2020 [75]	ETH3D	NVIDIA GTX 2080Ti GPU	—	1.71
	Middlebury			0.98
Shen et al, 2022 [76]	KITTI2012	NVIDIA V100 GPU	—	1.37
	KITTI2015			1.67
	SceneFlow		0.47	—
	KITTI2012		0.4	1.44
Xu et al, 2023 [77]	KITTI2015	NVIDIA RTX 3090 GPU	—	1.59
	ETH3D		—	3.6
	Middlebury		—	6.2
Jing et al, 2024 [78]	SceneFlow	NVIDIA A100 GPU	0.062	0.22
	SceneFlow		0.41	—
Chen et al, 2024 [79]	KITTI2012	NVIDIA A6000 GPU	0.4	1.36
	KITTI2015		—	1.53
	KITTI2012		0.83	3.90
Bartolomei, 2025 [80]	KITTI2015		0.97	3.93
	ETH3D	NVIDIA A100 GPU	0.24	1.66
	Middlebury2012		0.94	6.96
	Middlebury2014		1.08	7.97
	SceneFlow		0.37	—
	KITTI2012		—	1.09
Cheng, 2025 [81]	KITTI2015	NVIDIA RTX 3090 GPU	—	1.41
	ETH3D		—	0.72
	Middlebury		—	6.14

As listed in Table 2, we summarize the benchmark performance of mainstream cost-volume-based stereo matching algorithms on widely adopted public datasets. With continuous optimization of network architectures and cost-volume feature aggregation strategies, most advanced methods achieve gradually lower EPE and Bad-Pixel Rate on standard test sets, which

greatly boosts disparity estimation accuracy. However, the performance of these algorithms fluctuates significantly across different datasets. Many models achieve outstanding results on their training datasets yet suffer from obvious accuracy degradation when generalized to unseen scenarios. In addition, the pursuit of higher matching precision often requires high-end GPUs and substantial memory overhead, restricting the deployment of such algorithms on embedded edge devices for practical industrial applications. These bottlenecks further motivate researchers to explore Transformer-based stereo matching solutions for better cross-domain generalization and scene robustness.

3.2.2. Transformer-Based Stereo Matching Algorithms

In 2017, Vaswani et al. proposed Transformer, a novel network architecture that replaces traditional recurrent and convolutional operations with the attention mechanism. It outperforms conventional ones in machine translation tasks and features higher training efficiency [82]. In 2021, Dosovitskiy et al. presented the Vision Transformer (ViT), which directly applies the Transformer to image recognition tasks [83]. Li et al. developed STTR, the first Transformer-based stereo matching network, which formulates stereo matching as a sequence-to-sequence problem and alleviates limitations of traditional methods such as fixed disparity ranges, poor occlusion handling, and lack of uniqueness constraints [84]. Guo et al. proposed CEST, which captures long-range global information to improve the generalization and robustness of stereo depth estimation in challenging regions [85], and its network architecture is shown in Figure 4. Ding et al. designed a multi-view stereo matching network integrated with Transformer. Through a feature-matching Transformer to aggregate global contextual information, it achieves state-of-the-art performance in multi-view stereo matching tasks on datasets such as DTU [6]. Xu et al. combined CNN and Transformer as feature extractors, using Transformer to learn high-quality discriminative features and realize cross-task transfer [86]. Jia et al. constructed a dense feature extraction network based on the Transformer architecture, which fully utilizes the powerful global modeling capability of Transformer to achieve high-quality stereo matching in weak-texture and occluded regions [87]. Jiang et al. proposed DEFOM-Stereo, a recursive stereo matching framework integrated with depth foundation models. By integrating monocular depth cues into the stereo matching architecture, it significantly enhances the robustness and zero-shot generalization ability of stereo matching algorithms in complex scenarios [88]. Wang et al. presented RobuSTereo, which adopts a robust feature encoder integrating denoising ViT and VGG19. This encoder can extract stable and reliable features from degraded images, and the network further improves stereo matching accuracy under harsh weather conditions with a disparity refinement network [89]. Min et al. proposed S2M2, a network built on a multi-resolution Transformer. With high scalability, it can stably process challenging data with high resolution and a wide disparity range [90]. To quantitatively evaluate the performance of various Transformer-based stereo matching algorithms, we summarize their experimental results on mainstream public benchmark datasets in Table 3.

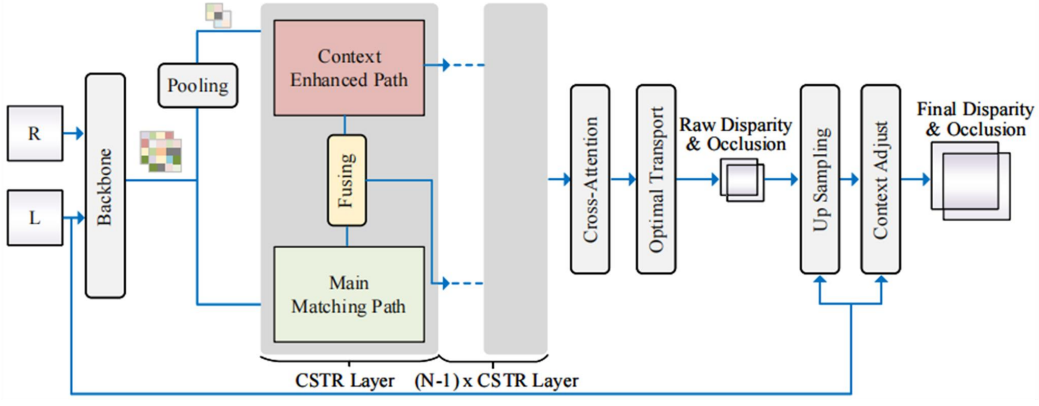

Fig. 4. The network architecture of CEST [85]. The model first extracts binocular features using a shared backbone network, and iteratively enhances matching information via stacked dual-path CSTR layers. Relying on cross-attention and optimal transport modules, it obtains initial disparity and occlusion predictions, which are further refined through upsampling and context adjustment to generate the final prediction results.

Table 3. Performance comparison of Transformer-based stereo matching algorithms on public datasets.

References	Dataset	Operating Environment	EPE（px）	Bad-Pixel Rate(%)
Li, et al, 2021 [84]	MPI Sintel	Nvidia GeForce GTX Titan GPU	3.01	5.75
	KITTI2015		1.50	6.74
	Middlebury		2.33	6.19

	SCARED		1.57	3.69
	SceneFlow		0.42	1.18
Guo, et al, 2022 [85]	MPI Sintel	Tesla A100 GPU	2.58	5.51
	KITTI2015		1.43	5.78
	Middlebury		1.16	5.16
	ETH3D	Tesla A100 GPU Tesla V100 GPU	—	1.83
Xu, et al, 2023 [86]	KITTI2015		—	1.77
	Middlebury		1.31	7.14
	SceneFlow		0.32	0.89
Jia, et al, 2024 [87]	MPI Sintel	NVIDIA RTX3090 GPU	2.63	3.70
	KITTI2015		—	1.80
	Middlebury		1.58	7.27
	KITTI2012		0.3	0.94
Jiang, 2025 [88]	KITTI2015	NVIDIA RTX4090 GPU	—	1.41
	ETH3D		0.11	0.70
	Middlebury		0.79	2.39
Wang, 2025 [89]	DrivingStereo	—	0.836	1.598
	SeeingThroughFog		3.359	20.881
	KITTI2015		—	3.23
Min, 2025 [90]	ETH3D	Nvidia H100 GPU	0.10	0.22
	Middlebury		0.69	3.57

As listed in Table 3, we summarize the benchmark performance of mainstream Transformer-based stereo matching algorithms across widely adopted public datasets. With the powerful global feature modeling capability of self-attention mechanisms, most Transformer-based methods achieve gradually lower EPE and Bad-Pixel Rate on standard test sets, which further improves disparity estimation accuracy compared with cost-volume-based approaches. However, it can also be observed that the performance of these algorithms varies greatly across different datasets. Many models obtain state-of-the-art results on conventional benchmark datasets but suffer from severe accuracy degradation when applied to extreme scenarios such as foggy driving environments. In addition, the global self-attention mechanism requires high-end GPUs and enormous memory resources for model inference, which restricts the deployment of such algorithms on embedded edge devices for real-world industrial applications. These inherent drawbacks of Transformer-based schemes, together with the limitations of cost-volume-based methods analyzed above, jointly reveal the core bottlenecks restricting the practical application of deep stereo matching algorithms.

Despite the remarkable progress achieved by deep learning-based stereo matching algorithms, the two dominant technical routes still have an unavoidable accuracy-efficiency trade-off determined by their intrinsic network design logic. In terms of computational overhead and feature perception scope, the two frameworks show completely different application boundaries. Cost-volume frameworks built upon 3D convolutions deliver low memory consumption and short inference latency, making them highly suitable for real-time autonomous driving and other edge-side tasks, yet they can hardly capture long-range contextual information, leading to unsatisfactory matching results in weak-texture and heavily occluded regions. In contrast, the self-attention mechanism embedded in Transformer-based architectures excels at modeling global dependencies and yields competitive performance in complex challenging scenes, but its quadratic computational complexity inevitably brings excessive hardware costs and inference delays, which greatly limits its popularization in resource-constrained industrial scenarios. Such contradictory characteristics of the two mainstream architectures leave considerable room for further algorithm optimization in future research.

4. Dataset

In 3D reconstruction tasks, datasets provide essential data support for the quantitative evaluation of stereo matching

algorithms. According to data sources, datasets can be divided into two categories: real-captured datasets and synthetic datasets.

4.1. Real-Captured Datasets

Real-captured datasets are collected by cameras in real-world environments. Depth information is annotated through RGB-D cameras, depth cameras or LiDAR point clouds to obtain disparity maps that characterize depth distribution. Typical real-captured stereo matching datasets include the Middlebury dataset, the KITTI dataset and the ETH3D dataset.

4.1.1. Middlebury Dataset

In 2001, Scharstein et al. from Middlebury College first released the Middlebury stereo matching dataset [91]. This dataset was constructed using a high-resolution Canon G1 camera mounted on a horizontally movable platform. High-quality images were captured periodically from a view perpendicular to the moving direction. The dataset consists of six planar scenes, with each set containing nine images and two disparity maps. Since all images are captured in planar scenes, disparity estimation can achieve high precision. It provides favorable conditions for pixel-level matching verification and pioneers the quantitative evaluation of stereo matching algorithms based on standard datasets. In 2003, to address the overfitting issue existing in the 2001 dataset during algorithm validation, Scharstein and Szeliski adopted structured light to acquire ground truth (GT) and established a new dataset with two scenes, namely Cones and Teddy [92]. Each sub-dataset contains nine color images and two disparity maps. Different from the 2001 version, the 2003 dataset is no longer limited to planar scenes and covers complex surface textures, shadows and depth-discontinuous regions. Benefiting from the high-precision acquisition strategy of the 2003 version, Scharstein and Pal further constructed a high-accuracy dataset with nine groups of samples in 2005 [93]. In 2006, Hirschmüller et al. followed the same pipeline and released a dataset containing 21 groups of high-precision data [94]. Compared with the 2003 version, the datasets of 2005 and 2006 feature richer scenes and larger disparity ranges, making them more suitable for quantitative evaluation of stereo matching algorithms. In 2014, Scharstein et al. improved the previous acquisition scheme and built an updated dataset covering 33 new indoor scenes [95]. Relevant examples are presented in Figure 5.

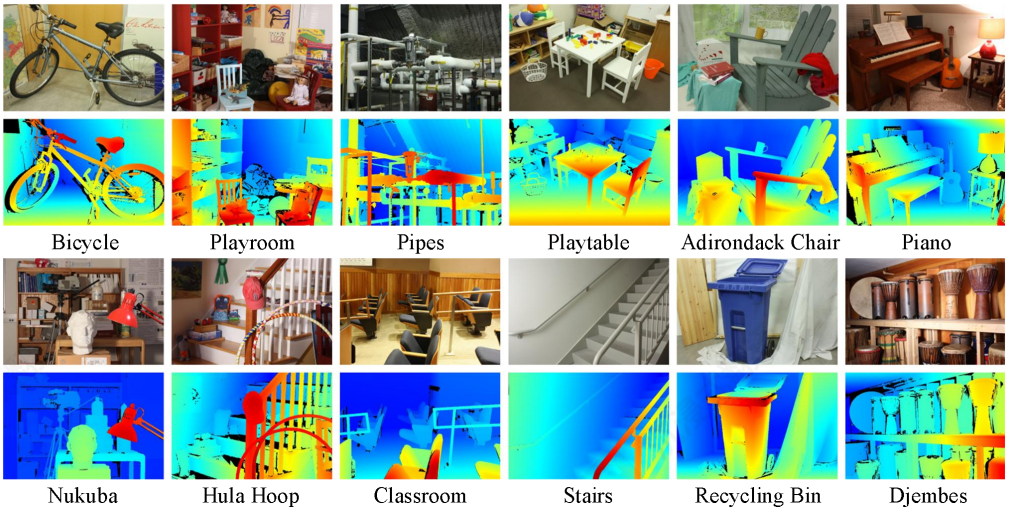

Fig. 5. The examples of Middlebury 2014 dataset [95]. Each group consists of an original binocular scene image and its corresponding ground-truth disparity heatmap, covering indoor scenarios with complex textures, occlusions and different depth distributions, which is widely adopted to evaluate the performance of stereo matching algorithms.

In terms of data acquisition for the Middlebury 2014 dataset, a portable stereo photography system was utilized. The system consists of two Canon DSLR cameras (EOS 450 equipped with 18-55 mm lenses), two digital cameras, and displacement guide rails, and captures laboratory scenes at a medium resolution of 6 megapixels. By fusing disparity estimates obtained from multiple projector positions, this dataset achieves a disparity accuracy of 0.2 pixels in most regions, including semi-occluded areas. From 2018 to 2019, Dai et al. jointly developed a stereo photography acquisition system based on mobile devices (iPod Touch 6G, 8 megapixels). The mobile device was mounted on a URS robotic arm for data collection. Following the structured light scheme adopted in the 2014 version, disparity maps were generated. Using this system, a stereo matching dataset containing 11 scenes and 21 sets of data was constructed between 2019 and 2021. The Middlebury dataset is mainly

collected in indoor static scenes. As a key benchmark for the validation of stereo matching algorithms, it remains widely adopted for algorithm evaluation amid the rapid iteration of existing technologies.

4.1.2. KITTI Dataset

With the rapid development of autonomous driving technology, research on sensors including LiDAR, stereo vision systems and millimeter-wave radars for autonomous driving perception has attracted extensive attention. This subsection introduces the KITTI dataset, the most representative benchmark in the autonomous driving domain, with relevant examples illustrated in Figure 6. Unlike the Middlebury stereo matching datasets that focus primarily on indoor scenes, the KITTI dataset offers stereo matching benchmarks collected in medium-scale urban, rural and highway environments.

The data acquisition platform is mounted on a Volkswagen Passat B6 vehicle and equipped with an eight-core i7 computer running Ubuntu, an inertial navigation system, a 64-line LiDAR, two grayscale cameras, two color cameras and four zoom lenses. Designed for outdoor data collection, the platform acquires precise 3D information through LiDAR point clouds and establishes the KITTI 2012 stereo matching dataset with 194 training image pairs and 195 test image pairs [96,97]. The static scene data adopted in the KITTI 2012 dataset only accounts for a small part of the raw data. The authors mentioned their intention to conduct deeper exploration and processing on raw data to improve dataset diversity, which laid the groundwork for the proposal of the KITTI 2015 dataset. In 2015, Menze and Geiger annotated 400 dynamic scenes with moving objects based on the original data from 2012, providing optical flow and disparity ground truth for consecutive frames [7]. This dataset contains 200 training scenes and 200 test scenes, serving as a high-quality benchmark for stereo matching and scene flow estimation tasks.

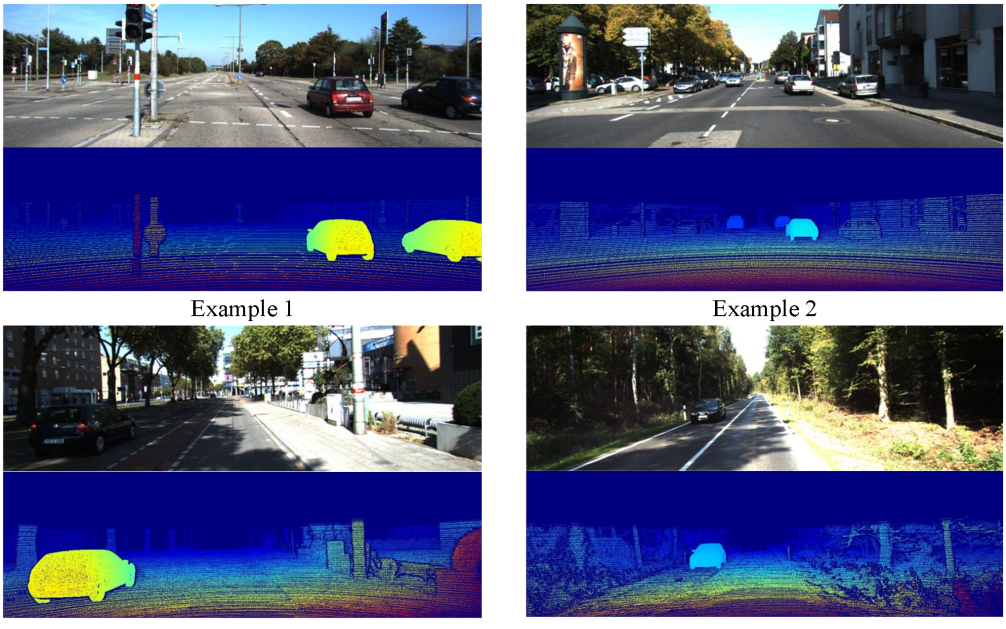

Fig. 6. The examples of KITTI dataset [96]. Each sample contains a real-world outdoor street binocular image and its corresponding ground-truth disparity heatmap. This dataset covers diverse traffic scenes and is widely used to verify the practical generalization performance of stereo matching algorithms in autonomous driving scenarios.

4.1.3. ETH3D Dataset

In 2017, Schöps et al. from ETH Zurich constructed the ETH3D dataset [98], as illustrated in Figure 7. The dataset acquires depth ground truth using a high-precision laser scanner and captures binocular images with four synchronized cameras to form two stereo image pairs, including 27 training sets and 20 test sets. Compared with the Middlebury dataset, ETH3D breaks the constraints of elaborately arranged laboratory scenes and covers office environments, man-made outdoor scenarios, and natural vegetation scenes.

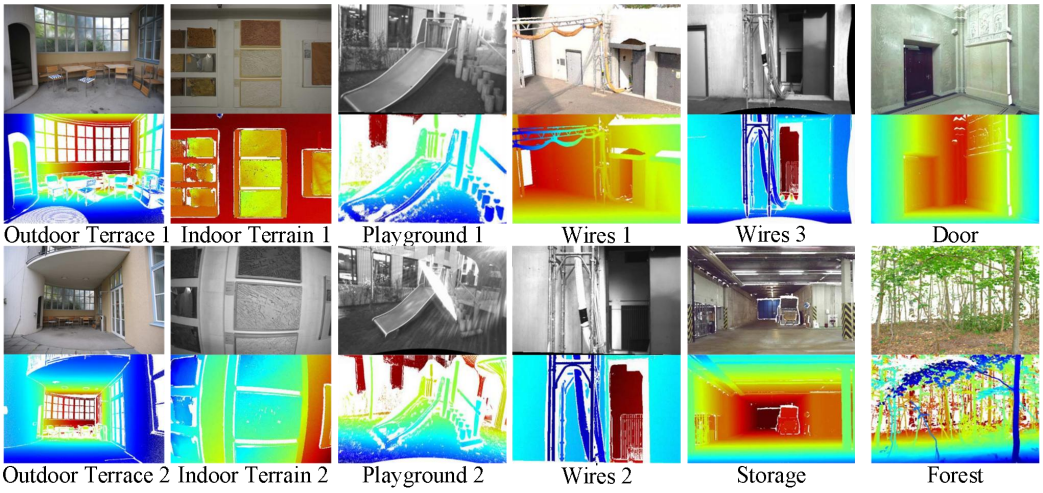

Fig. 7. The examples of ETH3D dataset [98]. Each group includes the original binocular image and the corresponding ground-truth disparity heatmap. Captured via high-precision laser scanning and multi-camera synchronization, this dataset contains diverse indoor and realistic outdoor scenarios, which is widely used to evaluate the robustness of stereo matching algorithms under complex natural scenes.

4.2. Synthetic Datasets

In the field of stereo matching, the KITTI and Middlebury datasets have long been vital resources for researchers. However, with the rapid development of deep learning in stereo matching, the limitations of real-captured datasets have gradually emerged. Limited training data cannot meet the demand of deep learning models for large-scale samples, which restricts the further improvement of algorithm performance to a certain extent. Moreover, the collection and production of real-scene stereo matching datasets are complicated and costly, making it impractical to build large-scale datasets based on real scenarios. With the development of computer graphics, especially the continuous optimization of Blender software, it has become feasible to produce large-scale stereo matching datasets through synthetic methods. Common synthetic stereo matching datasets include the MPI Sintel dataset and the SceneFlow dataset.

4.2.1. MPI Sintel Dataset

In 2012, Butler et al. built the MPI Sintel optical flow estimation dataset based on the open-source animated film Sintel created with Blender [99,100]. Based on this optical flow dataset, Butler et al. further expanded and constructed a stereo matching dataset, whose scale is more than two orders of magnitude larger than that of the Middlebury. Representative examples are shown in Figure 8.

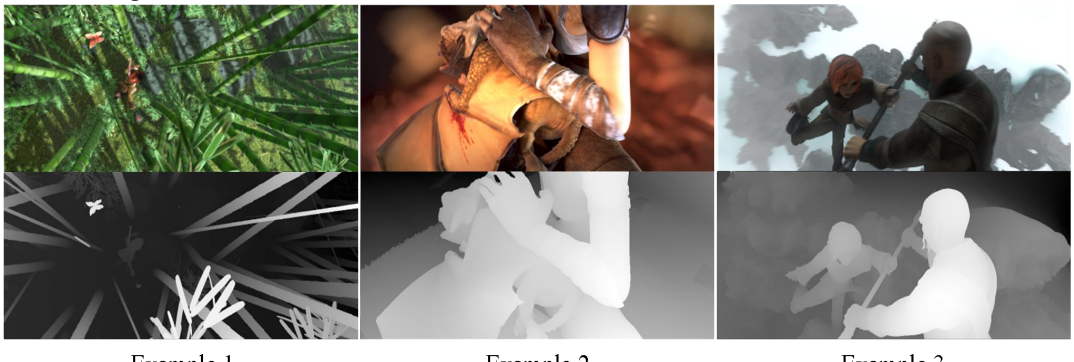

Fig. 8. The examples of MPI Sintel dataset [100]. Each sample consists of a rendered animation scene image and its corresponding grayscale ground-truth depth map. This synthetic dataset contains abundant motion blur, large displacement, complex occlusion and textureless regions, which is widely adopted to evaluate the generalization capability of stereo matching algorithms under challenging virtual scenarios.

4.2.2. SceneFlow Dataset

In 2016, Mayer et al. from the University of Freiburg and the Technical University of Munich created the SceneFlow stereo matching dataset using Blender [57], with representative examples shown in Figure 9.

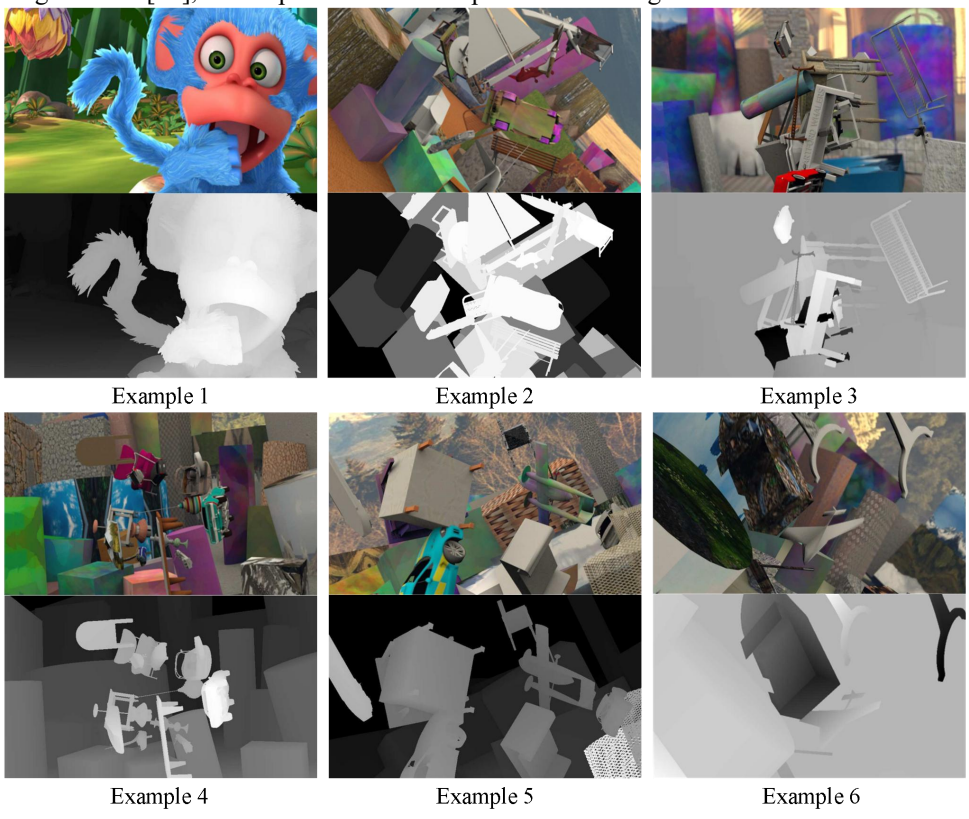

Fig. 9. The examples of SceneFlow Dataset [57]. Each sample includes a rendered RGB stereo image and its corresponding grayscale ground-truth depth map. Covering a large number of virtual indoor scenes with diverse objects, this dataset provides dense and accurate depth annotations, which is commonly used for the training and quantitative benchmark evaluation of stereo matching algorithms.

For data generation, the dataset predefines camera intrinsic parameters (focal length, principal point) and rendering parameters (image resolution, format, and sensor specifications) in Blender. It projects 3D motion vectors into 2D pixel motion vectors coplanar with the imaging plane. Meanwhile, depth information is derived from 3D pixel coordinates. Combined with the intrinsic parameters of virtual stereo cameras, depth values are converted into disparities to generate the final disparity maps. Compared with real-captured datasets, SceneFlow presents clear advantages in data scale and scene diversity. Consisting of over 35,000 stereo pairs, it can effectively support large-scale model training.

As of May 2026, statistics from online stereo matching benchmark platforms show that 279 algorithms have been submitted for evaluation on the KITTI 2012 dataset, 373 on KITTI 2015, and 830 on ETH3D, while 237 online results have been reported for the Middlebury 2014 dataset. These statistics fully demonstrate the essential role of benchmark datasets in advancing stereo matching research. They not only provide unified training and evaluation criteria for algorithm testing, but also drive continuous innovation and technical breakthroughs in this field. In addition, a prominent domain gap exists between synthetic datasets and real-captured datasets due to their different data distributions. When models trained on one type of dataset are deployed to another, their performance will drop sharply. Conventional data augmentation methods can only alleviate this problem slightly. In contrast, feature alignment and domain normalization strategies are more effective for enhancing cross-domain generalization. At present, most stereo matching algorithms still struggle to achieve stable performance on unfamiliar scenes and datasets. The obvious performance fluctuation of all algorithms between SceneFlow, KITTI and Middlebury in Table 1-3 further verifies the severe domain gap problem.

5. Evaluation Metrics

To accurately evaluate the performance of stereo matching algorithms, two metrics are commonly adopted: End-Point Error and Bad Pixel Rate [84,91].

5.1. EPE

End-Point Error reflects the overall deviation between the predicted disparity and the ground-truth disparity of a stereo matching algorithm. It is defined as the average error between predicted and ground-truth disparity values over all valid pixels and is formulated as:

$$\mathrm{EPE} = \frac{1}{N}\sum_{(x,y)\in\Omega}\left|d_{\mathrm{pred}}(x,y) - d_{\mathrm{gt}}(x,y)\right| \quad (4)$$

where N is the total number of valid pixels for calculation, Ω denotes the set of all valid pixels, $d_{\mathrm{pred}}(x,y)$ is the predicted disparity at pixel (x,y) and $d_{\mathrm{gt}}(x,y)$ is the corresponding ground-truth disparity. As a key metric for overall matching accuracy, a smaller EPE indicates a lower deviation between predicted and ground-truth disparity. In model training, EPE loss is widely adopted as a direct supervision signal to guide the network in learning accurate disparity mapping.

5.2. Bad Pixel Rate

The bad pixel rate quantifies the proportion of pixels with disparity error exceeding a predefined threshold and serves as a core metric for evaluating algorithm robustness. Its mathematical definition is given as:

$$B = \frac{1}{N}\sum_{(x,y)}\left(\left|d_{pred}(x,y) - d_{gt}(x,y)\right| > \delta_d\right) \quad (5)$$

where δ_d denotes the error threshold (typically 1 pixel or 3 pixels). The term inside the parentheses represents a logical judgment function, which equals 1 when the disparity error exceeds the threshold and 0 otherwise. A lower value of this metric indicates a lower mismatch rate, as well as better robustness against interference factors such as noise, weak textures, and occlusion.

6. Challenges and Future Directions

6.1. Current Key Challenges

Despite tremendous advances in stereo matching from traditional handcrafted methods to deep learning-based frameworks, several critical barriers still limit the practical deployment and generalization capability of existing algorithms.

The first prominent barrier lies in poor cross-domain generalization. Restricted by the inherent data distribution gap between synthetic datasets (SceneFlow) and real-world benchmarks (KITTI, ETH3D, Middlebury), models trained on simulated data degrade sharply when deployed in unseen real scenarios. Most algorithms require target-domain fine-tuning to guarantee accuracy, which greatly reduces environmental adaptability.

A core limitation is the unavoidable trade-off between matching accuracy and computational efficiency. Traditional global methods yield precise disparity results in occluded and textureless regions but suffer from high memory cost and slow inference. Local algorithms run in real time yet blur object edges. 3D convolution-based cost volume networks balance speed and accuracy but fail to capture long-range global context, while Transformer models deliver global modeling ability at the cost of heavy computation, limiting edge-side deployment.

The third limitation originates from insufficient environmental robustness and scarce diverse annotated datasets. Most mainstream datasets are collected under ideal conditions, lacking samples for low-light, foggy, motion-blurred and high-reflection scenes. Repeated textures, severe occlusion and abrupt depth changes easily cause mismatches, and downsampling-induced feature loss damages boundary prediction accuracy. Besides, high pixel-level labeling costs hinder the construction of multi-scenario datasets and restrict model iterative optimization.

6.2. Future Research Directions

To tackle the above practical barriers, this section summarizes several targeted research directions to advance theoretical innovation and large-scale engineering applications of stereo matching technologies.

First, universal foundation models combined with domain adaptation techniques effectively address weak cross-domain generalization. Methods like domain normalization, feature alignment and generative data augmentation narrow the data distribution gap between synthetic and real scenes and lower reliance on target-domain fine-tuning. Building large multi-modal datasets covering extreme weather and industrial scenarios also helps train models with strong zero-shot generalization capacity.

Second, lightweight CNN-Transformer hybrid architectures balance matching accuracy and real-time inference. Convolutional layers extract local fine-grained features, and simplified Transformer blocks capture global context by cutting redundant computation. Equipped with pruning, quantization and knowledge distillation, high-precision networks can be deployed on vehicle-mounted and embedded devices to meet real-time 3D perception demands.

Third, multi-source prior constraints greatly enhance robustness in challenging regions. Monocular depth priors, semantic information and binocular geometric constraints are embedded into stereo matching pipelines to realize region-adaptive cost aggregation, which reduces mismatches and retains structural details in textureless, occluded and boundary regions.

Beyond single-model optimization, multi-task joint learning extends the application scope of stereo matching. Joint optimization with downstream tasks including 3D reconstruction, visual SLAM and object detection achieves mutually boosted performance, promoting the practical use of stereo matching in autonomous driving and industrial inspection.

7. Conclusion

This paper systematically reviews the evolution of binocular stereo matching and establishes a unified classification framework for traditional stereo matching algorithms and two mainstream deep learning architectures: cost-volume and Transformer-based models. By summarizing mainstream datasets and quantitative metrics such as EPE and Bad-Pixel Rate, this work constructs a standardized benchmark system for algorithm performance comparison. Different from previous incomplete surveys, this review contributes by comparing state-of-the-art Transformer-based methods, standardizing the evaluation system, and comprehensively analyzing practical deployment bottlenecks.

From the perspective of practical industrial deployment, three high-priority research gaps are summarized in this review. Poor cross-domain generalization becomes the primary barrier to real-scene deployment, followed by the inherent conflict between matching precision and inference efficiency that restricts edge computing. Insufficient robustness under extreme conditions and the shortage of diverse annotated datasets also hinder continuous model optimization.

Accordingly, this paper offers phased research suggestions. Short-term efforts should focus on cross-domain optimization and lightweight network design to balance generalization and real-time performance. In the medium term, expanding multi-scenario datasets and applying multi-source prior constraints will enhance robustness in complex environments. Long-term research can explore multi-task collaborative learning to broaden the industrial application scope of binocular stereo matching.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that helped improve the quality of manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest

The authors declare that they have no competing interests.

Data availability statement

All datasets adopted in this review are publicly accessible, and relevant details are elaborated in the text.

Author contribution statement

Methodology, Yuandong Niu; validation, Huaqing Liu; investigation, Li Gao; data curation, Wenwen Yu; writing—original draft preparation, Yuandong Niu; writing—review and editing, Zhexuan Huang. All authors have read and agreed to the published version of the manuscript.

Supplementary material

No supplementary material is associated with this article.

References

[1] Haris M, Watanabe T, Fan L, et al., Superresolution for UAV Images via Adaptive Multiple Sparse Representation and Its Application to 3-D Reconstruction, *IEEE Trans. Geosci. Remote Sens*. 55, 4047-4058 (2017). https://doi.org/10.1109/TGRS.2017.2687419.

[2] Cui Y, Li Q, Yang B, et al., Automatic 3-D Reconstruction of Indoor Environment With Mobile Laser Scanning Point Clouds, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 12, 3117-3130 (2019). https://doi.org/10.1109/JSTARS.2019.2918937.

[3] Li W, Hu Z, Meng L, et al., Weakly Supervised 3-D Building Reconstruction From Monocular Remote Sensing Images, *IEEE Trans. Geosci. Remote Sens.* 62, 1-15 (2024). https://doi.org/10.1109/TGRS.2024.3377694.

[4] Peng Y, Yang M, Zhao G, et al., Binocular-Vision-Based Structure From Motion for 3-D Reconstruction of Plants, *IEEE Geosci. Remote Sens. Lett.* 19, 1-15 (2022). https://doi.org/10.1109/LGRS.2021.3105106.

[5] Jiang L, Wang F, Zhang W, et al., Rethinking the Key Factors for the Generalization of Remote Sensing Stereo Matching Networks, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 18, 4936-4948 (2025). https://doi.org/10.1109/JSTARS.2024.3524078.

[6] Ding Y, Yuan W, Zhu Q, et al., TransMVSNet: Global Context-aware Multi-view Stereo Network with Transformers, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA (IEEE, New York, 2022), pp. 8575-8584. https://doi.org/10.1109/CVPR52688.2022.00839.

[7] Menze M, Geiger A, Object scene flow for autonomous vehicles, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA (IEEE, New York, 2015), pp. 3061-3070. https://doi.org/10.1109/CVPR.2015.7298925.

[8] Huang C, Guan H, Jiang A, et al., Registration Based Few-Shot Anomaly Detection, In Proceedings of the European Conference on Computer Vision (ECCV), Israel (Springer, Cham, 2022), pp. 303-319. https://doi.org/10.1007/978-3-031-20053-3_18.

[9] Hesch JA, Roumeliotis SI, A Direct Least-Squares (DLS) method for PnP, In Proceedings of the International Conference on Computer Vision (ICCV), Barcelona, Spain (IEEE, New York, 2011), pp. 383-390. https://doi.org/10.1109/ICCV.2011.6126266.

[10] Scharstein D, Szeliski R, A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms, *Int. J. Comput. Vis.* 47, 7-42 (2002). https://doi.org/10.1023/A:1014573219977.

[11] Hirschmuller H, Accurate and Efficient Stereo Processing by Semi-global Matching and Mutual Information, In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA (IEEE, New York, 2005), pp. 807-814. https://doi.org/10.1109/CVPR.2005.56.

[12] Boykov Y, Veksler O, Zabih R, Fast Approximate Energy Minimization Via Graph Cuts, *IEEE Trans. Pattern Anal. Mach. Intell*. 23, 1222-1239 (2001). https://doi.org/10.1109/34.969114.

[13] Kolmogorov V, Zabih R, Computing Visual Correspondence with Occlusions Using Graph Cuts, In Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV), Vancouver, BC, Canada (IEEE, New York, 2001), pp. 508-515. https://doi.org/10.1109/ICCV.2001.937668.

[14] Kang SB, Szeliski R, Chai JX, Handling Occlusions in Dense Multi-view Stereo, In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Kauai, HI, USA (IEEE, New York, 2001), pp. 1-35. https://doi.org/10.1109/CVPR.2001.990462.

[15] Kolmogorov V, Graph Based Algorithms for Scene Reconstruction from Two or More Views, Ph.D. Thesis, Cornell University, Ithaca, NY, USA, 2004.

[16] Li H, Chen G, Segment-based Stereo Matching Using Graph Cuts, In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA (IEEE, New York, 2004), pp. 1-8. https://doi.org/10.1109/CVPR.2004.1315016.

[17] Taniai T, Matsushita Y, Naemura T, Graph Cut Based Continuous Stereo Matching Using Locally Shared Labels, In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA (IEEE, New York, 2014), pp. 1613-1620. https://doi.org/10.1109/CVPR.2014.209.

[18] Sun J, Zheng NN, Shum HY, Stereo Matching Using Belief Propagation, *IEEE Trans. Pattern Anal. Mach. Intell*. 25, 787-800 (2003). https://doi.org/10.1109/TPAMI.2003.1206509.

[19] Felzenszwalb PF, Huttenlocher DR, Efficient Belief Propagation for Early Vision, In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA (IEEE, New York, 2004), pp. 1-8. https://doi.org/10.1109/CVPR.2004.1315041.

[20] Sun J, Li Y, Kang SB, et al., Symmetric Stereo Matching for Occlusion Handling, In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA (IEEE, New York, 2005), pp. 399-406. https://doi.org/10.1109/CVPR.2005.337.

[21] Klaus A, Sormann M, Karner K, Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure, In Proceedings of the 18th International Conference on Pattern Recognition (ICPR), Hong Kong, China (IEEE, New York, 2006), pp. 15-18. https://doi.org/10.1109/ICPR.2006.1033.

[22] Yang Q, Wang L, Yang R, et al., Stereo Matching with Color-Weighted Correlation, Hierarchical Belief Propagation, and Occlusion Handling, *IEEE Trans. Pattern Anal. Mach. Intell*. 31, 492-514 (2009). https://doi.org/10.1109/TPAMI.2008.99.

[23] Yang QX, Wang L, Ahuja N, A Constant-space Belief Propagation Algorithm for Stereo Matching, In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA (IEEE, New York, 2010), pp. 1457-1465. https://doi.org/10.1109/CVPR.2010.5539797.

[24] Veksler O, Stereo Correspondence by Dynamic Programming on a Tree, In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA (IEEE, New York, 2005), pp. 384-390. https://doi.org/10.1109/CVPR.2005.334.

[25] Aboali M, Manap NA, Darsono AM, et al., Performance Analysis between Basic Block Matching and Dynamic Programming of Stereo Matching Algorithm, *J. Telecommun. Electron. Comput. Eng*. 9, 7-16 (2017). https://jtec.utem.edu.my/jtec/article/view/2558.

[26] Luo A, Burkhardt H, An Intensity-based Cooperative Bidirectional Stereo Matching with Simultaneous Detection of Discontinuities and Occlusions, *Int. J. Comput. Vis*. 15, 177-188 (1995). https://doi.org/10.1007/BF01451740.

[27] Scharstein D, Szeliski R, Stereo Matching with Non-linear Diffusion, In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA (IEEE, New York, 1996), pp. 343-350. https://doi.org/10.1109/CVPR.1996.517095.

[28] Pritchett P, Zisserman A, Wide Baseline Stereo Matching, In Proceedings of the Sixth International Conference on Computer Vision (ICCV), Bombay, India (IEEE, New York, 1998), pp. 754-760. https://doi.org/10.1109/ICCV.1998.710802.

[29] Zitnick CL, Kanade T, A Cooperative Algorithm for Stereo Matching and Occlusion Detection, *IEEE Trans. Pattern Anal. Mach. Intell*. 22, 675-684 (2000). https://doi.org/10.1109/34.865184.

[30] Barron JT, Adams A, Shih Y, et al., Fast Bilateral-space Stereo for Synthetic Defocus, In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA (IEEE, New York, 2015), pp. 4466-4474. https://doi.org/10.1109/CVPR.2015.7299076.

[31] Li LC, Zhang SL, Yu X, et al., PMSC: PatchMatch-Based Superpixel Cut for Accurate Stereo Matching, *IEEE Trans. Circuits Syst. Video Technol*. 28, 679-692 (2018). https://doi.org/10.1109/TCSVT.2016.2628782.

[32] Kanade T, Okutomi M, A Stereo Matching Algorithm with An Adaptive Window: Theory and Experiment, In Proceedings of the 1991 IEEE International Conference on Robotics and Automation, Sacramento, CA, USA (IEEE, New York, 1991), pp. 1088-1095. https://doi.org/10.1109/ROBOT.1991.131738.

[33] Koschan A, Rodehorst V, Spiller K, Color Stereo Vision Using Hierarchical Block Matching and Active Color Illumination, In Proceedings of the 13th International Conference on Pattern Recognition (ICPR), Vienna, Austria (IEEE, New York, 1996), pp. 835-839. https://doi.org/10.1109/ICPR.1996.546141.

[34] Veksler O, Fast Variable Window for Stereo Correspondence Using Integral Images, In Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Madison, WI, USA (IEEE, New York, 2003), pp. 1-6. https://doi.org/10.1109/CVPR.2003.1211403.

[35] Yoon K, Kweon IS, Adaptive Support-weight Approach for Correspondence Search, *IEEE Trans. Pattern Anal. Mach. Intell.* 28, 650-656 (2006). https://doi.org/10.1109/TPAMI.2006.70.

[36] Goesele M, Snavely N, Curless BH, et al., Multi-View Stereo for Community Photo Collections, In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil (IEEE, New York, 2007), pp. 1-8. https://doi.org/10.1109/ICCV.2007.4408933.

[37] Min D, Sohn K, Cost Aggregation and Occlusion Handling With WLS in Stereo Matching, *IEEE Trans. Image Process*. 17, 1431-1442 (2008). https://doi.org/10.1109/TIP.2008.925372.

[38] Geiger A, Roser M, Urtasun R, Efficient Large-Scale Stereo Matching, In Proceedings of the 10th Asian Conference on Computer Vision (ACCV), Queenstown, New Zealand (Springer, Cham, 2010), pp. 25-38. https://doi.org/10.1007/978-3-642-19315-6_3.

[39] Mei X, Sun X, Zhou M, et al., On Building an Accurate Stereo Matching System on Graphics Hardware, In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). Barcelona, Spain (IEEE, New York, 2011), pp. 467-474. https://doi.org/10.1109/ICCVW.2011.6130280.

[40] Bleyer M, Rhemann C, Rother C, PatchMatch Stereo-Stereo Matching with Slanted Support Windows, In Proceedings of the British Machine Vision Conference (BMVC), Dundee, UK (BMVA, Durham, 2011), pp. 1-11. http://dx.doi.org/10.5244/C.25.14.

[41] Ma ZY, He KM, Wei YC, et al., Constant Time Weighted Median Filtering for Stereo Matching and Beyond, In Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV), Sydney, NSW, Australia (IEEE, New York, 2013), pp. 49-56. https://doi.org/10.1109/ICCV.2013.13.

[42] Hosni A, Rhemann C, Bleyer M, et al., Fast Cost-Volume Filtering for Visual Correspondence and Beyond, *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 504-511 (2013). https://doi.org/10.1109/TPAMI.2012.156.

[43] Zhang K, Fang Y, Min D, et al., Fast Cross-scale Cost Aggregation for Stereo Matching, *IEEE Trans. Circuits Syst. Video Technol.* 27, 965-976 (2017). https://doi.org/10.1109/TCSVT.2015.2513663.

[44] Mattoccia S, Tombari F, Stefano LD, Stereo Vision Enabling Precise Border Localization within a Scanline Optimization Framework, In Proceedings of the 8th Asian Conference on Computer Vision-Volume Part II, Tokyo, Japan (Springer, Cham, 2007), pp. 517-527. https://doi.org/10.1007/978-3-540-76390-1_51.

[45] Gehrig SK, Franke U, Improving Stereo Sub-Pixel Accuracy for Long Range Stereo, In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil (IEEE, New York, 2007), pp. 1-7. https://doi.org/10.1109/ICCV.2007.4409212.

[46] Hirschmuller H, Stereo Processing by Semiglobal Matching and Mutual Information, *IEEE Trans. Pattern Anal. Mach. Intell*. 30, 328-341 (2008). https://doi.org/10.1109/TPAMI.2007.1166.

[47] Mattoccia S, Fast Locally Consistent Dense Stereo on Multicore, In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA (IEEE, New York, 2010), pp. 69-76. https://doi.org/10.1109/CVPRW.2010.5543767.

[48] Mattoccia S, Accurate Dense Stereo by Constraining Local Consistency on Superpixels, In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey (IEEE, New York, 2010), pp. 1832-1835. https://doi.org/10.1109/ICPR.2010.452.

[49] Michael M, Salmen J, Stallkamp J, et al., Real-time Stereo Vision: Optimizing Semi-Global Matching, In Proceedings of the 2013 IEEE Intelligent Vehicles Symposium (IV), Gold Coast, QLD (IEEE, New York, 2013), pp. 1197-1202. https://doi.org/10.1109/IVS.2013.6629629.

[50] Krizhevsky A, Sutskever I, Hinton GE, ImageNet Classification with Deep Convolutional Neural Networks, In Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, Nevada, United States (ACM, New York, 2012), pp. 1097-1105. https://doi.org/10.1145/3065386.

[51] Žbontar J, LeCun Y, Computing the Stereo Matching Cost with a Convolutional Neural Network, In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA (IEEE, New York, 2015), pp. 1592-1599. https://doi.org/10.1109/CVPR.2015.7298767.

[52] Žbontar J, LeCun Y, Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches, *J. Mach. Learn. Res*. 17, 2287-2318 (2016). https://dl.acm.org/doi/abs/10.5555/2946645.2946710.

[53] Chen ZY, Sun X, Wang L, A Deep Visual Correspondence Embedding Model for Stereo Matching Costs, In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile (IEEE, New York, 2015), pp. 972-980. https://doi.org/10.1109/ICCV.2015.117.

[54] Luo W, Schwing AG, Urtasun R, Efficient Deep Learning for Stereo Matching, In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA (IEEE, New York, 2016), pp. 5695-5703. https://doi.org/10.1109/CVPR.2016.614.

[55] Shaked A, Wolf L, Improved Stereo Matching with Constant Highway Networks and Reflective Confidence Learning, In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA (IEEE, New York, 2017), pp. 6901-6910. https://doi.org/10.1109/CVPR.2017.730.

[56] Park H, Lee KM, Look Wider to Match Image Patches With Convolutional Neural Networks, *IEEE Trans. Pattern Anal. Mach. Intell*. 24, 1788-1792 (2017). https://doi.org/10.1109/LSP.2016.2637355.

[57] Mayer N, Ilg E, Häusser P, A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation, In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA (IEEE, New York, 2016), pp. 4040-4048. https://doi.org/10.1109/CVPR.2016.438.

[58] Flynn J, Neulander I, Philbin J, et al., Deep Stereo: Learning to Predict New Views from the World's Imagery, In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA (IEEE, New York, 2016), pp. 5515-5524. https://doi.org/10.1109/CVPR.2016.595.

[59] Kendall A, Martirosyan H, Dasgupta S, et al., End-to-End Learning of Geometry and Context for Deep Stereo Regression, In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy (IEEE, New York, 2017), pp. 66-75. https://doi.org/10.1109/ICCV.2017.17.

[60] Pang J, Sun W, Ren JS, et al., Cascade Residual Learning: A Two-Stage Convolutional Neural Network for Stereo Matching, In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy (IEEE, New York, 2017), pp. 878-886. https://doi.org/10.1109/ICCVW.2017.108.

[61] Ye XQ, Li JM, Wang H, et al., Efficient Stereo Matching Leveraging Deep Local and Context Information, *IEEE Access* 5, 18745-18755 (2017). https://doi.org/10.1109/ACCESS.2017.2754318.

[62] Khamis S, Fanello S, Rhemann C, et al., StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction, In Proceedings of the Computer Vision-ECCV: 15th European Conference (ECCV), Munich, Germany (Springer, Cham, 2018), pp. 596-613. https://doi.org/10.1007/978-3-030-01267-0_35.

[63] Liang ZF, Feng YL, Guo YL, et al., Learning for Disparity Estimation Through Feature Constancy, In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA (IEEE, New York, 2018), pp. 2811-2820. https://doi.org/10.1109/CVPR.2018.00297.

[64] Chang JR, Chen YS, Pyramid Stereo Matching Network, In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA (IEEE, New York, 2018), pp. 5410-5418. https://doi.org/10.1109/CVPR.2018.00567.

[65] Yang GR, Zhao HS, Shi JP, et al., SegStereo: Exploiting Semantic Information for Disparity Estimation, In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany (Springer, Cham, 2018), pp. 660-676. https://doi.org/10.1007/978-3-030-01234-2_39.

[66] Song X, Zhao X, Hu HW, et al., EdgeStereo: A Context Integrated Residual Pyramid Network for Stereo Matching, In Proceedings of the 14th Asian Conference on Computer Vision (ACCV), Perth, Australia (Springer, Cham, 2018), pp. 20-35. https://doi.org/10.1007/978-3-030-20873-8_2.

[67] Song X, Zhao X, Fang LJ, et al., EdgeStereo: An Effective Multi-task Learning Network for Stereo Matching and Edge Detection, *Int. J. Comput. Vis.* 128, 910-930 (2020). https://doi.org/10.1007/s11263-019-01287-w.

[68] Guo XY, Yang K, Yang WK, et al., Group-Wise Correlation Stereo Network, In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA (IEEE, New York, 2019), pp. 3268-3277. https://doi.org/10.1109/CVPR.2019.00339.

[69] Nie GY, Cheng MM, Liu Y, et al., Multi-Level Context Ultra-Aggregation for Stereo Matching, In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA (IEEE, New York, 2019), pp. 3278-3286. https://doi.org/10.1109/CVPR.2019.00340.

[70] Zhang FH, Prisacariu V, Yang R, et al., GA-Net: Guided Aggregation Net for End-To-End Stereo Matching, In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA (IEEE, New York, 2019), pp. 185-194. https://doi.org/10.1109/CVPR.2019.00027.

[71] Zhang FH, Qi XJ, Yang RG, et al., Domain-Invariant Stereo Matching Networks, In Proceedings of the 16th European Conference on Computer Vision (ECCV), Glasgow, UK (Springer, Cham, 2020), pp. 420-439. https://doi.org/10.1007/978-3-030-58536-5_25.

[72] Duggal S, Wang S, Ma WC, et al., DeepPruner: Learning Efficient Stereo Matching via Differentiable PatchMatch, In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea (IEEE, New York, 2019), pp. 4383-4392. https://doi.org/10.1109/ICCV.2019.00448.

[73] Xu HF, Zhang JY, AANet: Adaptive Aggregation Network for Efficient Stereo Matching, In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA (IEEE, New York, 2020), pp. 1956-1965. https://doi.org/10.1109/CVPR42600.2020.00203.

[74] Lipson L, Teed Z, Deng J, RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching, In Proceedings of the 2021 International Conference on 3D Vision (3DV), London, United Kingdom (IEEE, New York, 2021), pp. 218-227. https://doi.org/10.1109/3DV53792.2021.00032.

[75] Li JK, Wang PS, Xiong PF, et al., Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation, In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA (IEEE, New York, 2022), pp. 16242-16251. https://doi.org/10.1109/CVPR52688.2022.01578.

[76] Shen ZL, Dai YC, Song XB, et al., PCW-Net: Pyramid Combination and Warping Cost Volume for Stereo Matching, In Proceedings of the 17th European Conference on Computer Vision (ECCV), Tel Aviv, Israel (Springer, Cham, 2022), pp. 280-297. https://doi.org/10.1007/978-3-031-19824-3_17.

[77] Xu GW, Wang XQ, Ding XH, et al., Iterative Geometry Encoding Volume for Stereo Matching, In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada (IEEE, New York, 2023), pp. 21919-21928. https://doi.org/10.1109/CVPR52729.2023.02099.

[78] Jing JP, Mao Y, Mikolajczyk K, Match-Stereo-Videos: Bidirectional Alignment for Consistent Dynamic Stereo Matching, In Proceedings of the 18th European Conference on Computer Vision (ECCV), Milan, Italy (Springer, Cham, 2024), pp. 415-432. https://doi.org/10.1007/978-3-031-73027-6_24.

[79] Chen ZY, Long W, Yao H, et al., MoCha-Stereo: Motif Channel Attention Network for Stereo Matching, In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA (IEEE, New York, 2024), pp. 27768-27777. https://doi.org/10.1109/CVPR52733.2024.02623.

[80] Bartolomei L, Tosi F, Poggi M, et al., Stereo Anywhere: Robust Zero-Shot Deep Stereo Matching Even Where Either Stereo or Mono Fail, In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA (IEEE, New York, 2025), pp. 1013-1027. https://doi.org/10.1109/CVPR52734.2025.00103.

[81] Cheng JD, Liu LL, Xu GW, et al., MonSter: Marry Monodepth to Stereo Unleashes Power, In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA (IEEE, New York, 2025), pp. 6273-6282. https://doi.org/10.1109/CVPR52734.2025.00588.

[82] Vaswani A, Shazeer N, Parmar N, et al., Attention is All You Need, In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), Long Beach California USA (ACM, New York, 2017), pp. 6000-6010. https://dl.acm.org/doi/10.5555/3295222.3295349.

[83] Dosovitskiy A, Beyer L, Kolesnikov A, et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, In Proceedings of the International Conference on Learning Representations (LCLR), Virtual Event (ICLR, OpenReview, 2021), pp. 1-21. https://openreview.net/forum?id=YicbFdNTTy.

[84] Li ZS, Liu XT, Drenkow N, et al., Revisiting Stereo Depth Estimation From a Sequence-to-Sequence Perspective with Transformers, In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada (IEEE, New York, 2021), pp. 6177-6186. https://doi.org/10.1109/ICCV48922.2021.00614.

[85] Guo WY, Li ZS, Yang YK, et al., Context-Enhanced Stereo Transformer, In Proceedings of the 17th European Conference on Computer Vision (ECCV), Tel Aviv, Israel (Springer, Cham, 2022), pp. 263-279. https://doi.org/10.1007/978-3-031-19824-3_16.

[86] Xu HF, Zhang J, Cai JF, et al., Unifying Flow, Stereo and Depth Estimation, *IEEE Trans. Pattern Anal. Mach. Intell*. 45, 13941-13958 (2023). https://doi.org/10.1109/TPAMI.2023.3298645.

[87] Jia D, Cai P,Wang Q, et al., A Transformer-Based Architecture for High-Resolution Stereo Matching, *IEEE Trans. Comput. Imaging*. 10, 83-92 (2024). https://doi.org/10.1109/TCI.2024.3350884.

[88] Jiang HL, Lou ZQ, Ding LY, et al., DEFOM-Stereo: Depth Foundation Model Based Stereo Matching, In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA (IEEE, New York, 2025), pp. 21857-21867. https://doi.org/10.1109/CVPR52734.2025.02036.

[89] Wang YR, Liang YP, Hu YT, et al., RobuSTereo: Robust Zero-Shot Stereo Matching under Adverse Weather, In Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA (IEEE, New York, 2025), pp. 25134-25144. https://doi.org/10.1109/ICCV51701.2025.02331.

[90] Min J, Jeon Y, Kim J, et al., S^2M^2: Scalable Stereo Matching Model for Reliable Depth Estimation, In Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA (IEEE, New York, 2025), pp. 26729-26739. https://doi.org/10.1109/ICCV51701.2025.02481.

[91] Scharstein D, Szeliski R, Zabih R, A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms, In Proceedings of the IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV), Kauai, HI, USA (IEEE, New York, 2001), pp. 131-140. https://doi.org/10.1109/SMBV.2001.988771.

[92] Scharstein D, Szeliski R, High-accuracy Stereo Depth Maps Using Structured Light, In Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Madison, WI, USA (IEEE, New York, 2003), pp. 195-202. https://doi.org/10.1109/CVPR.2003.1211354.

[93] Scharstein D, Pal C, Learning Conditional Random Fields for Stereo, In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA (IEEE, New York, 2007), pp. 1-8. https://doi.org/10.1109/CVPR.2007.383191.

[94] Hirschmuller H, Scharstein D, Evaluation of Cost Functions for Stereo Matching, In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA (IEEE, New York, 2007), pp. 1-8. https://doi.org/10.1109/CVPR.2007.383248.

[95] Scharstein D, Hirschmuller H, Kitajima Y, et al., High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth, In Proceedings of the German Conference on Pattern Recognition (GCPR 2014), Münster, Germany (Springer, Cham, 2014), pp. 31-42. https://doi.org/10.1007/978-3-319-11752-2_3.

[96] Geiger A, Lenz P, Urtasun R, Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite, In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA (IEEE, New York, 2012), pp. 3354-3361. https://doi.org/10.1109/CVPR.2012.6248074.

[97] Geiger A, Lenz P, Stiller C, Vision Meets Robotics: The KITTI Dataset, *Int. J. Robot. Res*. 32, 1231-1237 (2013). https://doi.org/10.1177/0278364913491297.

[98] Schöps T, Schönberger JL, Galliani S, A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos, In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA (IEEE, New York, 2017), pp. 2538-2547. https://doi.org/10.1109/CVPR.2017.272.

[99] Butler DJ, Wulff J, Stanley GB, et al., A Naturalistic Open Source Movie for Optical Flow Evaluation, In Proceedings of the 12th European Conference on Computer Vision (ECCV), Florence, Italy (Springer, Cham, 2012), pp. 611-625. https://doi.org/10.1007/978-3-642-33783-3_44.

[100] Wulff J, Butler DJ, Stanley GB, et al., Lessons and Insights from Creating a Synthetic Optical Flow Benchmark, In Proceedings of the 12th European Conference on Computer Vision (ECCV), Florence, Italy (Springer, Cham, 2012), pp. 168-177. https://doi.org/10.1007/978-3-642-33868-7_17.

[101] Lazaros, N.; Sirakoulis, G.C.; Gasteratos, A. Review of Stereo Vision Algorithms: From Software to Hardware. Int. J. Optomechatron. 2008, 2, 435-462. [CrossRef]

[102] Niu, Y.D.; Liu, L.M.; Huang, F.Y.; Huang, S.Y.; Chen, S.Y. Overview of Image-Based 3D Reconstruction Technology. J. Eur. Opt. Soc.-Rapid Publ. 2024, 20, 18. [CrossRef]