remote sensing

MDPI

Review

Binocular Stereo Vision in Remote Sensing: A Review

Xing Li [1], Hongwei Zhou [1], Mingyu Sun [1], Bangshu Xiong [1], Yuchao Dai [2], Renjie He [2], Zhihua Chen [1] and Zhibo Rao [1,*]

[1] School of Information and engineering, Nanchang Hangkong University, Nanchang 330063, China; lixing36@foxmail.com (X.L.); hongwei_zhou26@163.com (H.Z.); mingyu61@foxmail.com (M.S.); xiongbs@126.com (B.X.); zhihua.chen@nchu.edu.cn (Z.C.)

[2] School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China; daiyuchao@nwpu.edu.cn (Y.D.); davidhrj@nwpu.edu.cn (R.H.)

* Correspondence: zhibo.rao@nchu.edu.cn or raoxi36@foxmail.com

Highlights

What are the main findings?

- This study synthesizes the technical progress and inherent limitations of representative models in the field of remote sensing stereo matching.
- The research provides a critical analysis of the characteristics and domain-specific constraints of major remote sensing stereo datasets.

What are the implications of the main findings?

- This study serves as a practical guide for researchers to select the most suitable models and datasets for specific remote sensing stereo matching tasks.
- This survey identifies the unresolved bottlenecks—large disparity ranges, ill-posed regions, cross-sensor domain shift, and label scarcity—that should guide the design of next-generation remote sensing stereo matching algorithms.

Abstract

Stereo vision leverages binocular imagery to emulate the human visual system in perceiving three-dimensional (3D) structures by estimating disparity from rectified image pairs and converting it to depth via geometric triangulation. In recent years, deep learning-based stereo matching has significantly advanced in accuracy, efficiency, and generalization, surpassing traditional methods and demonstrating great potential in remote sensing applications. However, stereo matching in remote sensing faces unique challenges not commonly seen in terrestrial datasets. These include limited access to satellite imagery, seasonal differences between image pairs, difficulty in identifying small objects, and widespread regions with repetitive textures, such as lakes and forests. Unlike prior surveys that primarily address ground-level scenes, this paper presents a comprehensive review of stereo matching techniques tailored for remote sensing. It synthesizes the progress and limitations of representative models, analyzes the characteristics and domain-specific constraints of remote sensing stereo datasets, and outlines future research directions and application prospects in this field.

Keywords: stereo matching; remote sensing; deep learning

Check for updates

Academic Editor: Riccardo Roncella

Received: 8 March 2026
Revised: 30 April 2026
Accepted: 6 May 2026
Published: 9 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Stereo matching using high-resolution satellite imagery is a fundamental technique in remote sensing and photogrammetry. It supports a variety of applications, including

3D reconstruction, terrain modeling, change detection, disaster assessment, image registration, and environmental monitoring [1–4]. Given a pair of rectified stereo images, the objective is to identify corresponding pixels—(x_l, y) in the left image and (x_r, y) in the right image—and compute the disparity as $d = x_l - x_r$. This disparity is triangulated to infer depth for subsequent remote sensing analysis.

Traditional stereo matching methods typically follow the classic four-step pipeline established by Scharstein and Szeliski [5]: matching cost computation for pixel-wise similarity measurement, cost aggregation to ensure spatial smoothness, disparity computation for initial estimation, and disparity refinement for final optimization. Representative algorithms within this framework include Graph Cuts [6], Semi-Global Matching [7], and PatchMatch [8]. Although these conventional approaches have achieved significant progress, several critical limitations continue to hinder their effectiveness in complex real-world scenarios:

(1) Empirically designed cost metrics: Traditional matching costs often rely on heuristic metrics, such as luminance difference and correlation coefficients. These hand-crafted measures are often insufficient for capturing complex radiometric variations in high-resolution images.
(2) Localized cost aggregation: Cost aggregation is typically performed within a finite neighborhood with a fixed window size. This localized approach lacks the flexibility to adapt to varying terrain scales and often fails in regions with repetitive textures or depth discontinuities.
(3) Dependency on post-processing: Disparity refinement heavily depends on a sequence of manual techniques, such as median filtering for smoothing and subpixel enhancement. These multi-stage pipelines increase computational complexity and require extensive parameter tuning.

Consequently, these hand-crafted schemes often falter in challenging environments and lack the scalability required to process the massive volumes of intricate data characteristic of modern satellite sensors. To transcend these inherent limitations, deep learning-based stereo matching has emerged as a transformative paradigm. By substituting manual feature engineering with end-to-end learnable architectures, these methods have demonstrated superior robustness and accuracy across diverse and demanding applications.

Since 2016, stereo matching methods based on deep learning [9–11] have made substantial advances. These methods overcome many of the limitations inherent in traditional algorithms, which often depend on hand-crafted features and rule-based optimization. Current deep models are typically categorized into three major types: cost volume-based methods, iterative optimization networks, and Transformer-based architectures. GC-Net [10] introduced cost volume construction as a central component and inspired numerous variants [11,12]. RAFT [13], originally proposed for optical flow, was later adapted for stereo with strong performance [14]. More recently, Transformer-based models [15,16] have emerged with promising capabilities in modeling global context and long-range dependencies.

Despite these advances, applying deep stereo models to remote sensing imagery remains highly challenging. Several domain-specific issues must be addressed:

(1) Computational inefficiency: High-resolution inputs from satellite sensors significantly increase computational demands. For instance, PSMNet [17] requires 576 ms to process a 1024×1024 remote sensing image on an NVIDIA GTX 1080Ti. BGA-Net [12] takes 1572 ms on the same GPU, and MaskCRNet [18] requires 724 ms on an NVIDIA RTX 3090. These requirements hinder real-time use and complicate deployment on edge devices.

(2) Reduced accuracy in remote sensing scenes: Remote sensing imagery is subject to various challenges, including occlusions in urban environments, repetitive textures such as forests and water surfaces, seasonal appearance changes, illumination variations, human activities, and the presence of small-scale targets. As illustrated in Figure 1, remote sensing stereo pairs often exhibit substantial appearance variations across views. These factors introduce significant ambiguities that degrade the performance of stereo matching models trained primarily on conventional terrestrial datasets. For example, PSMNet [17] achieves a three-pixel error rate of 24.81% on the remote sensing WHU-Stereo dataset [19], but yields a significantly lower error rate of 1.89% on the terrestrial KITTI 2012 dataset [20].

(3) Dependence on predefined priors: Many stereo matching methods rely on a fixed disparity range (e.g., 0–192), typically determined by the camera baseline and image resolution. This assumption limits model adaptability to the varying imaging geometries present in satellite-based systems. Moreover, the disparity distributions in remote sensing datasets, such as US3D [21] and WHU-Stereo [19], differ markedly from those in terrestrial datasets like KITTI [20] and ETH3D [22]. As illustrated in Figure 2a, remote sensing disparities often span $(-64, 64)$ or $(-112, 64)$, whereas terrestrial datasets commonly cover ranges such as $(0, 192)$ or $(0, 256)$. This disparity mismatch not only impedes model generalization but also complicates joint training across domains, thereby limiting the transferability of terrestrial datasets and pretrained models to remote sensing scenarios.

(4) Limited data: The scarcity of labeled stereo image pairs in remote sensing significantly constrains the generalization ability of deep learning models, particularly in supervised settings. As shown in Figure 2b, the WHU-Stereo dataset provides only 1757 annotated samples, whereas the widely used terrestrial SceneFlow dataset includes 39,049 samples. This severe data imbalance poses a major obstacle to training robust and transferable stereo matching models in remote sensing domains.

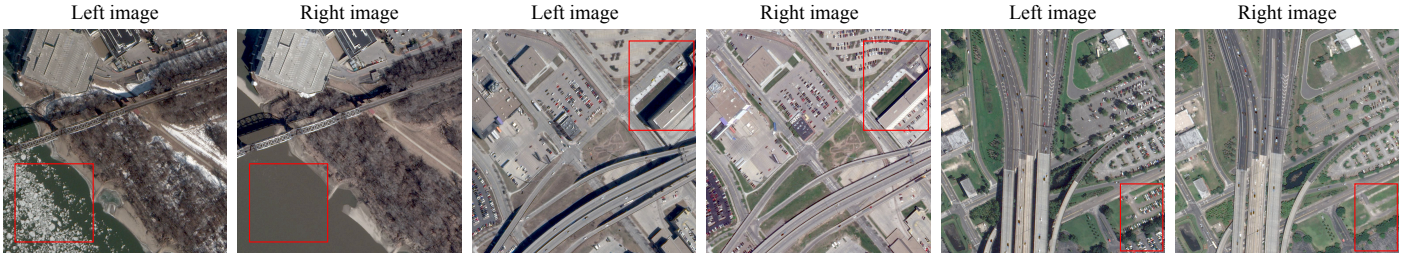
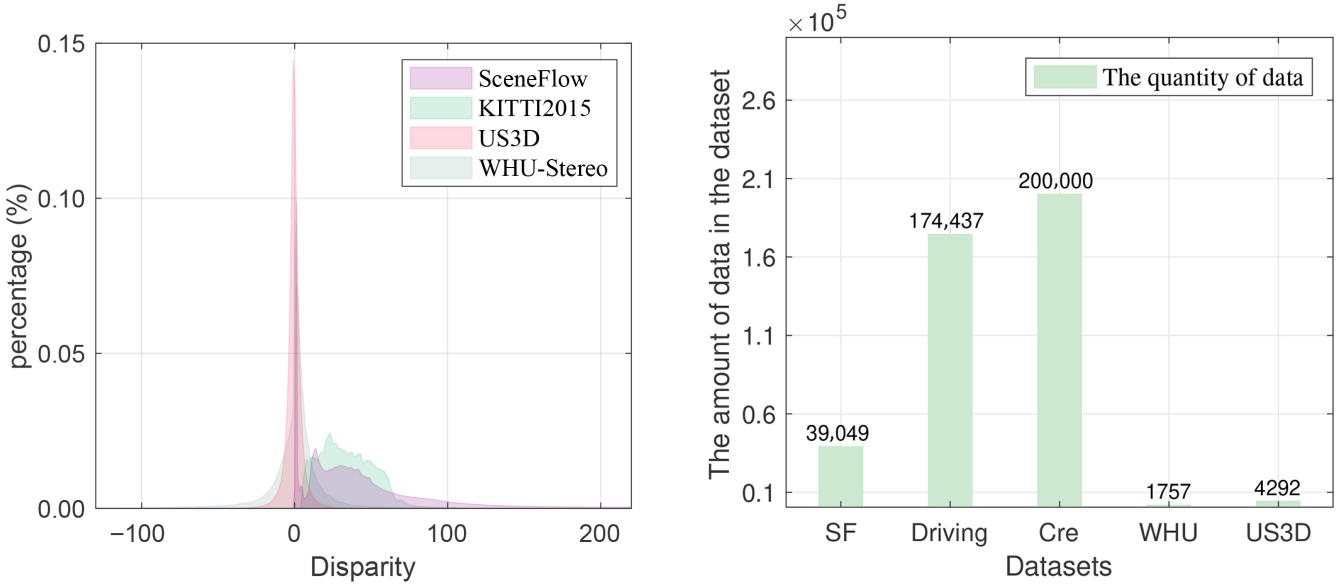


Figure 1. Appearance variations in remote sensing stereo pairs. (**a**–**c**) Examples of differences caused by seasonal changes, sun angle variations, and human activities. The red boxes highlight corresponding local regions in the left and right images where the appearance discrepancies are most pronounced, drawing attention to the matching ambiguities induced by these factors. These variations pose significant challenges for stereo matching and often lead to reduced accuracy in remote sensing applications.

Stereo matching in remote sensing (RS) has evolved largely by inheriting and adapting techniques from the computer vision (CV) community. Many representative RS algorithms are extensions of classical CV models, such as SGM, GC-Net, PSMNet, and RAFT-Stereo, because these frameworks provide strong geometric priors and scalable architectures. However, despite their successful transferability, CV models usually experience a noticeable performance drop when applied directly to RS imagery, mainly due to unique characteristics such as larger imaging baselines, radiometric inconsistencies, seasonal variations, and the presence of small repetitive patterns. As a result, most RS-oriented methods take CV

architectures as baselines and further introduce domain-specific adaptations, including radiometric normalization, adaptive disparity range modeling, and integration of auxiliary cues like DSM (Digital Surface Model) or LiDAR priors. Therefore, the development of RS stereo matching can be viewed as a progressive evolution from CV methods toward domain-adaptive designs that address the challenges of remote sensing data.



(**a**) Different domain attributes (**b**) Different dataset size

Figure 2. Key differences between remote sensing and terrestrial stereo datasets. (**a**) Disparity range: remote sensing datasets typically span $(-64, 64)$ or $(-112, 64)$, while terrestrial datasets cover $(0, 192)$ or $(0, 256)$. (**b**) Dataset size: remote sensing datasets such as WHU-Stereo [19] and US3D [21] contain significantly fewer images than terrestrial datasets like SceneFlow [9].

This survey presents a comprehensive review of binocular stereo vision techniques in remote sensing, including representative stereo matching methods, as shown in Table 1. We begin with traditional methods in Section 2, followed by deep learning approaches in Section 3. Acceleration strategies are discussed in Section 4, and publicly available datasets are summarized in Section 5. Finally, Section 6 outlines key findings and future research directions.

Table 1. Summary of representative stereo matching methods.

Type	Category	Characteristic	Representative Works (Common Data)	Representative Works (Remote Sensing)
Traditional	Classical	Interpretable, efficient, strong generalization	Scharstein and Szeliski [5], SGM [7], Hirschmuller et al. [23], Kordelas et al. [24], MGM [25], tMGM [26]	Lee et al. [27], Ghuffar [28], Qin et al. [29], SGM-ForestM [30], Wang et al. [31], Tatar et al. [32], DPSM [33], LGSM [34], Zhao et al. [35], L2GSM [36], Yue et al. [37]
Hybrid	Combine deep learning with traditional	Combines geometric priors with learning flexibility	MC-CNN [38], GA-Net [39], Deeppruner [40]	S2P-GANet [41], Gómez et al. [42], Albanwan and Qin [43], Sat-MVSF [44], Zheng et al. [45]

Table 1. *Cont.*

Type	Category	Characteristic	Representative Works (Common Data)	Representative Works (Remote Sensing)
Supervised	2D Convolution-based	Fast inference	DispNet [9], AANet [46], Bi3D [47], HITNet [48], SMD-Nets [49]	Ji et al. [50], Wang et al. [51]
	3D Convolution-based	High accuracy, high computing cost	Bosch et al. [21], GC-Net [10], CFNet [52], PCW-Net [53], Abc-Net [54], SDA [55], ADL [56], UPFNet [57]	Tao et al. [11], HMSM-Net [58], Jiang et al. [59], Tao et al. [60], PSMNet-FusionX3 [61], SRCV-Net [62], DBMSMNet [63]
	Iterative Optimization	High accuracy, iteratively updating	RAFT-Stereo [14], IGEV [64], IGEV++ [65], MoCha-Stereo [66], Selective-Stereo [67], TC-Stereo [68], MC-Stereo [69]	MaskCRNet [18], Patil and Guo [70]
	Transformer-based	Global context modeling	STTR [15], SSTTStereo [71], ELFNet [72], S^2M^2 [73]	Wei et al. [74], MaskCRNet [18], A-SATMVSNet [16]
	Multi-task Learning	Cross-task supervision	RTS^2Net [75], SGNet [76], NNNet [77], DWARF [78], Rao et al. [79]	BGA-Net [12], S^2Net [80], Chen et al. [81], Qin et al. [82], SDBF-Net [83], SemStereo [84]
	Vision Foundation Model Integration	Using priors from VFM, strong generalization	FoundationStereo [85], ViTAStereo [86], Monster [87], DEFOM-Stereo [88]	HDSM-Net [89]
	Other DL Models	Innovative, flexible	LEAStereo [90], NMRF-Stereo [91], CroCo v2 [92], DiffuVolume [93], DMIO [94]	DeepSim-Nets [95], MEMF-Net [96]
Alternative Supervised	Self-supervised	No GT, geometry-driven	SsSMnet [97], ActiveStereoNet [98], Flow2Stereo [99], ES^3Net [100], Yuan et al. [101], DualNet [102]	Knöbelreiter et al. [103], Igeta and Iwasaki [104], Chen et al. [105]
	Semi-supervised	Limited GT	Smolyanskiy et al. [106], SemiDepth [107], SSMF [108], Semi-Stereo [109]	HDADE [110]
	Weakly-supervised	Indirect cues	Tulyakov et al. [111], PA-Net [112], SUW-Stereo [113]	Albanwan and Qin [114]

2. Traditional Stereo Matching Methods

Stereo matching has long relied on a classical framework formalized by **Scharstein and Szeliski** [5] which decomposes the process into four key stages: matching cost computation, cost aggregation, disparity computation, and disparity refinement. This formulation has guided the development of traditional stereo matching algorithms for over two decades. In the first stage, similarity between pixel candidates in the left and right images is estimated, typically using photometric or gradient-based metrics. Lower costs signify higher correspondence likelihood. Cost aggregation then reinforces spatial consistency by integrating matching evidence from neighboring pixels. Disparity computation selects the disparity with minimum aggregated cost per pixel, and a final optimization stage refines results by correcting mismatches and enforcing global constraints.

Traditional methods are broadly categorized into local and global approaches. Local methods restrict computation to finite support windows around each pixel, making implicit smoothness assumptions via cost aggregation. These methods are computationally efficient and well-suited to real-time applications, but their performance often degrades in low-texture or occluded regions due to their limited contextual modeling. Global methods, by contrast, formulate stereo matching as an energy minimization problem, typically using Markov Random Fields (MRFs), where the cost function incorporates both data fidelity

and smoothness terms. These models achieve higher accuracy in challenging regions but incur higher computational costs.

To balance these trade-offs, Hirschmüller proposed **Semi-Global Matching (SGM)** [7], which performs multiple 1D global optimizations along multiple paths to approximate global regularization while maintaining efficiency. Due to its effectiveness, SGM quickly became the method of choice for many real-world applications. In parallel, **Hirschmüller et al.** [23] conducted a comprehensive study of matching cost functions, offering valuable guidance for algorithm design. Since its introduction, SGM has been extensively extended and adapted for diverse datasets and domains [24–26].

Building on the SGM framework, a series of methods were specifically developed for remote sensing stereo applications. **Lee et al.** [27] pioneered the use of stereo matching in satellite 3D reconstruction, incorporating a conjugate search strategy and correlation-based patch design to improve runtime efficiency and matching quality. **Ghuffar et al.** [28] applied SGM to WorldView-3 imagery and qualitatively evaluated its reconstruction performance.

Machine learning has also been introduced to augment traditional pipelines. **Qin et al.** [29] trained a Support Vector Machine (SVM) [115] to predict the potential stereo matching quality. **SGM-ForestM** [30] leveraged random forests to estimate optimal disparities per scanline, significantly enhancing SGM performance. **Wang et al.** [31] developed a feature-based SGM acceleration method that reduced noise and improved robustness in textureless or discontinuous regions. Similarly, **Tatar et al.** [32] improved aggregation by introducing homogeneity weights and edge-guided filtering, which were particularly effective for urban scenes.

Recent research has increasingly focused on integrating geometric and semantic priors. **LGSM** [34] (LiDAR-Guided Semi-Global Matching) constrained disparity search ranges using LiDAR data, thereby reducing mismatches in difficult matching areas and refining the boundaries of objects. **DPSM** [33] (Dual Propagation Stereo Matching) introduced a novel bidirectional propagation strategy that exploited building regularities and shape constraints, optimizing disparity via an energy minimization framework. **Zhao et al.** [35] incorporated building edge cues into cost aggregation using a multi-order Census transform to mitigate the effects of noise. **L2GSM** [36] fused LiDAR data and depth discontinuity lines with SGM, yielding improved results in low-textured and depth discontinuity regions. Most recently, **Yue et al.** [37] proposed a hierarchical aggregation strategy guided by semantic and geometric edge features. Their method adaptively constrained local disparity candidates, improving accuracy in complex urban environments.

In summary, traditional stereo matching methods, especially those based on the SGM framework, have laid a solid foundation for stereo vision in remote sensing. By integrating geometric priors, semantic cues, and learning-based enhancements, these methods have continued to evolve. However, their reliance on hand-crafted features and limited accuracy in complex scenes underscores the need for more intelligent, data-driven approaches.

3. Deep Learning-Based Stereo Matching Methods

Deep stereo matching has progressed rapidly with the advancement of deep learning. In this section, we present a comprehensive taxonomy of deep stereo methods, grouped into three major categories: (1) hybrid approaches that combine deep learning with traditional stereo algorithms, (2) fully supervised end-to-end methods, and (3) alternative supervision strategies, including self-supervised, semi-supervised, and weakly supervised learning.

The fully supervised category is further subdivided based on architectural design, encompassing 2D convolutional models, 3D convolutional networks, iterative refinement frameworks, Transformer-based architectures, multi-task learning strategies, and methods integrating vision foundation models.

This survey is organized following the evolution from general computer-vision stereo algorithms toward domain-adaptive designs for remote sensing. Each methodological category is discussed first in the CV context and then extended to RS scenarios. A detailed review of each category, along with representative methods, is provided in the following subsections.

3.1. Combining Deep Learning and Traditional Algorithms

Deep learning techniques have demonstrated remarkable capabilities in tasks such as object recognition, classification, and semantic segmentation since 2012. The first application of deep learning to stereo matching can be traced back to 2015. From 2015 to 2020, researchers began integrating deep learning modules into traditional stereo pipelines, leading to what are commonly called hybrid methods [38–40]. These approaches aim to retain the robustness and interpretability of classical algorithms while leveraging the learning power of deep networks.

In the field of remote sensing, several notable works have adopted this hybrid strategy. To address the integration gap between deep learning models and satellite stereo pipelines, as well as the lack of systematic performance evaluation in complex multi-temporal scenarios, **S2P-GANet** [41] introduced a stereo processing pipeline tailored to satellite imagery, in which Rational Polynomial Coefficients (RPCs) were utilized to model satellite viewing geometry. Within this framework, GA-Net [39] was employed to regress bidirectional disparity maps. Furthermore, by integrating this pipeline with multi-view stereo techniques such as COLMAP [116], the method achieved enhanced 3D reconstruction performance. To address the sub-optimal performance caused by heuristic pair selection and noisy data integration in traditional satellite multi-view stereo pipelines, **Gómez et al.** [42] proposed an iterative refinement approach incorporating bilateral filtering and multi-view fusion, further enhancing the geometric accuracy of remote sensing reconstructions. In a broader evaluation, **Albanwan and Qin** [43] systematically compared deep stereo networks with classical algorithms such as SGM. Their analysis revealed two key observations: (1) deep learning models typically outperform classical methods under in-distribution scenarios, but often suffer from poor generalization to unseen domains; and (2) traditional approaches exhibit more consistent performance across domains. These findings underscore a fundamental trade-off between the adaptability of learned models and the robustness of handcrafted priors.

Recent studies have investigated integrating deep learning with traditional geometric constraints to address the challenges of geometric adaptation in satellite multi-view stereo matching, the scarcity of training samples, and the limited generalization across multi-temporal imagery. Specifically, **Sat-MVSF** [44] introduced a self-optimization framework that leverages pseudo-labels guided by structural consistency, demonstrating strong robustness across diverse and seasonal datasets. Additionally, **Zheng et al.** [45] proposed a two-stage stereo pipeline, wherein a deep network produces an initial disparity estimate, which is subsequently refined via handcrafted post-processing.

Overall, these approaches integrate the robustness of handcrafted geometric priors with the representational flexibility of deep neural networks, thereby improving generalization across domains. Nevertheless, their multi-stage pipeline is relatively complex, and the overall performance gain remains limited when compared with fully learning frameworks.

3.2. End-to-End Supervised Algorithms

End-to-end supervised stereo matching methods take a pair of stereo images as input and directly predict the corresponding disparity map. We categorize existing end-to-end stereo matching methods into seven architectural groups: (1) 2D convolution-

based networks, (2) 3D convolution-based networks, (3) iterative refinement frameworks, (4) Transformer-based architectures, (5) multi-task learning models, (6) methods integrating vision foundation models, and (7) other emerging designs.

3.2.1. 2D Convolution-Based Models

One of the earliest milestones in end-to-end stereo matching was **DispNet** [9], which introduced a fully convolutional architecture integrating a 3D cost volume within a 2D convolutional framework. The model comprises four key components: feature extraction using 2D convolutions, construction of a 3D cost volume from stereo image features, cost volume regularization via an encoder-decoder structure, and disparity regression. Subsequent methods, including AANet [46], Bi3D [47], HITNet [48], and SMD-Nets [49], followed this design paradigm by leveraging 2D convolutions for efficient representation learning while maintaining the expressiveness of volumetric cost aggregation.

In the remote sensing domain, **Ji et al.** [50] extended two-view deep stereo networks (GC-Net [10] , MC-CNN [38]) into a dense multi-view matching framework with multi-view geometry constraints, tailored for strip-overlap acquisition of aerial photogrammetry. Their evaluation demonstrated a consistent superiority of deep models over traditional methods. **Wang et al.** [51] surveyed dense matching cost computation for satellite stereo and, through comparative analysis including DispNet-C and SGM, reinforced the benefits of learning-based approaches in remote sensing scenarios.

2D convolution-based stereo matching models were the first to enable fully end-to-end disparity estimation, outperforming traditional approaches in terms of accuracy. Owing to their exclusive use of 2D convolution operations, these models offer notable advantages in inference efficiency, making them suitable for real-time and edge computing applications. However, they still face limitations in accuracy and robustness, particularly in geometrically complex or textureless regions. As high-precision 3D convolutional architectures emerged, research interest in 2D convolution-based methods has gradually declined.

3.2.2. 3D Convolution-Based Models

3D convolution-based stereo models closely follow the classical four-step stereo matching pipeline and aim to enhance spatial regularization through 4D cost volume construction, as shown in Figure 3. A seminal contribution in this category is **GC-Net** [10], which first extracts features from left and right images using 2D convolutional layers. These features are concatenated across the disparity dimension to construct a 4D cost volume, which is then regularized using 3D convolutions, enabling the aggregation of both spatial and disparity cues. The final disparity map is obtained via a differentiable Soft ArgMin operation. GC-Net established a foundational design that inspired numerous successors [21,52–57], improving accuracy across a variety of scenes.

In remote sensing domain, **Ji et al.** [117] demonstrated that GC-Net [10] and MC-CNN [38] significantly outperform traditional algorithms when applied to satellite imagery. To address the specific challenge of occlusions in satellite scenes, **Tao et al.** [11] integrated an unsupervised bidirectional loss into PSMNet [17] via bidirectional pyramid network, achieving notable improvements in heavily occluded regions. Further extending this direction, **HMSM-Net** [58] introduced a hierarchical multi-scale stereo matching framework designed to handle intractable regions caused by repetitive patterns, texture-less areas, and disparity discontinuities. By constructing multi-scale cost volumes, the model effectively enhances disparity estimation in these challenging scenarios. Moreover, recognizing that the scarcity of training data is a major bottleneck hindering the deployment of CNN-based techniques, the authors released the Gaofen-7 dataset, establishing a new benchmark for evaluating remote sensing stereo algorithms.

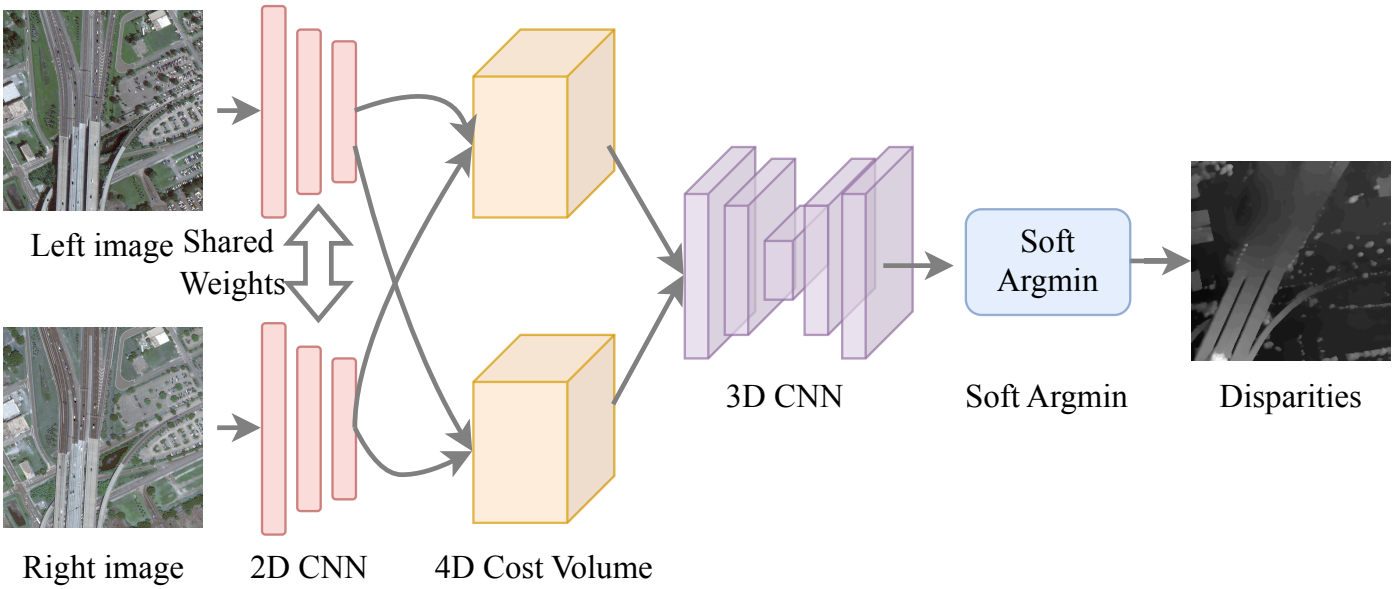


Figure 3. Pipeline of 3D Convolution-based Stereo Matching Methods. The process includes four main stages: feature extraction, 4D cost volume construction, cost volume regularization using 3D convolutions, and disparity regression.

Subsequently, to address the high cost of acquiring LiDAR ground truth and the limited generalization of stereo networks across different sensors and scenarios, **Jiang et al.** [59] introduced a systematic training strategy aimed at enhancing the generalization ability of remote sensing stereo matching networks. Their approach was validated on several state-of-the-art architectures, including CFNet [52], HMSM-Net [58], and PASM-Net [118], and demonstrated consistent performance improvements across diverse satellite scenes. To address the challenges of matching multi-scale objects in large scenes and the ambiguity of multi-modal probability distributions in occluded or textureless regions, **Tao et al.** [60] proposed a confidence-aware cascade refinement strategy, which progressively refines disparity maps from coarse to fine resolution. To address the limitations of noisy photogrammetric surfaces and the excessive sparsity of LiDAR point clouds in aerial 3D reconstruction, **PSMNet-FusionX3** [61] introduced a fusion framework combining stereo image features and LiDAR point clouds via triangular irregular interpolation, improving performance in sparse regions. To address the problem of disparity shifts caused by repetitive textures and textureless regions in satellite imagery, **SRCV-Net** [62] introduced a Cost Volume Refinement Strategy (CVRS) that uses a "left-left" cost volume as a reference to suppress false matches in the standard "left-right" cost volume. Integrated into their SRCV-Net, this method significantly improves disparity consistency and accuracy for satellite stereo matching, especially in repetitive and textureless areas. To address the challenges of disparity estimation in complex intractable regions, such as textureless, repeated texture, and occlusion areas, **DBMSMNet** [63] proposed a lightweight stereo matching network featuring a dual-branch module for multiscale feature extraction. It uses a coarse-to-fine cost aggregation strategy with disparity-channel attention for enhanced fusion, and a final refinement step guided by image intensity and gradients to produce accurate disparity maps for satellite images.

3D convolution-based stereo matching networks closely adhere to the classical pipeline, offering strong interpretability. Compared to 2D convolution-based network, these models achieve higher disparity estimation accuracy by constructing and regularizing a 4D cost volume. However, the computational overhead associated with 3D convolutions imposes

significant constraints on inference speed, limiting their applicability in real-time or large-scale remote sensing scenarios.

3.2.3. Iterative Optimization-Based Models

Iterative optimization-based models draw inspiration from classical optimization-based methods by refining disparity estimates through multiple recurrent updates. A seminal contribution in this direction is **RAFT** [13], originally developed for optical flow and awarded Best Paper at ECCV 2020. RAFT extracts 2D features at one-eighth the input resolution, computes pairwise similarity to build a correlation volume, and iteratively refines the flow or disparity using a Gated Recurrent Unit (GRU). The final disparity map is generated through upsampling to the full resolution.

RAFT-Stereo [14] adapted this framework for stereo matching by tailoring the similarity computation to disparity-specific cues, achieving competitive results on multiple benchmarks. Its success has inspired a series of follow-up models, such as IGEV [64], IGEV++ [65], MoCha-Stereo [66], Selective-Stereo [67], TC-Stereo [68], and MC-Stereo [69], which further improve matching precision, runtime, or domain robustness through architectural refinements.

In the remote sensing domain, Patil and Guo [70] introduced the Stellar dataset to address the challenges of significant visual appearance variations caused by non-simultaneous imaging and the scarcity of large-scale annotated data. By evaluating RAFT-Stereo and 3D convolutional networks on this benchmark—which provides rectified stereo pairs with true disparity and semantic labels—they demonstrated that both iterative refinement frameworks and 3D architectures are capable of producing dense and accurate disparity maps even under complex radiometric configurations. Building upon this direction, to address the challenges of imperfect epipolar rectification, missing data, and significant domain differences in high-resolution satellite imagery, **MaskCRNet** [18] proposed a cascaded recurrent stereo architecture specifically designed for satellite images, as shown in Figure 4. The model integrates a Transformer-based encoder, a CNN-based encoder, multi-scale cascaded recurrent refinement modules, and a self-supervised image reconstruction branch, achieving SOTA accuracy on high-resolution remote sensing benchmarks.

Iterative optimization-based models provide a compelling balance between computational efficiency and refinement capacity. Their recurrent architecture enables accurate and memory-efficient updates, supporting robust performance across diverse domains. However, due to their inherently sequential inference process, these models may encounter latency bottlenecks in high-resolution or large-scale deployments, posing a critical challenge for real-time remote sensing applications.

3.2.4. Transformer-Based Models

The Transformer architecture, originally introduced by Vaswani et al. [119] for sequence modeling in natural language processing, was later adapted to vision tasks via the Vision Transformer (ViT) [120]. ViT segments an image into fixed-size patches, encodes each patch into a feature token, and leverages multi-head attention to capture long-range dependencies across the spatial domain. Motivated by its success in image classification, the Transformer paradigm has since been extended to stereo matching.

As illustrated in Figure 5, we present four types of Transformer-based stereo matching models: (a) direct attention-based disparity estimation, exemplified by STTR [15]; (b) Transformer-based feature extraction, as in A-SATMVSNet [16]; (c) hybrid models combining Transformer and CNN backbones, as in MaskCRNet [18]; and (d) Transformer-based cost volume regularization, such as SSTTStereo [71].

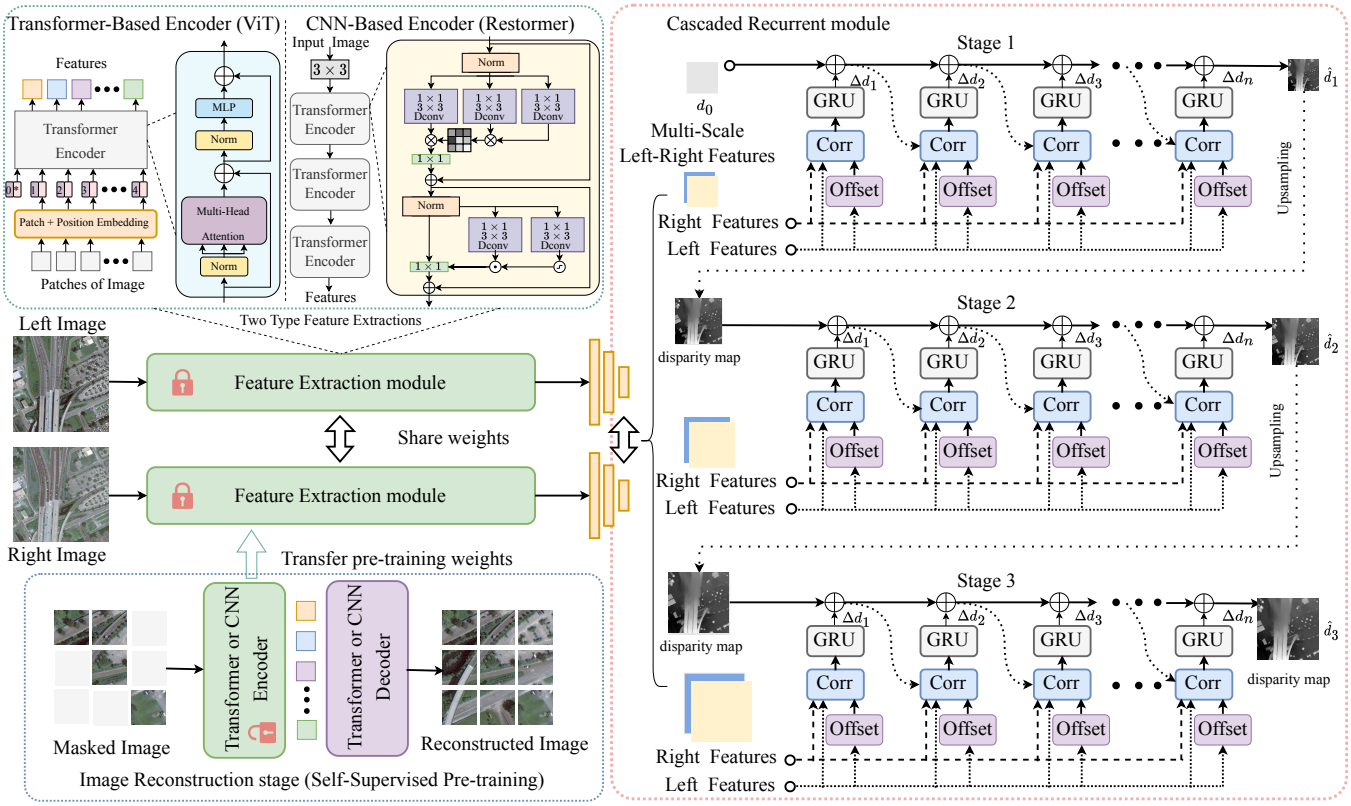


Figure 4. Architecture of MaskCRNet [18]. The model comprises a Transformer-based encoder, a CNN-based encoder, cascaded recurrent networks operating at multiple scales, and self-supervised image reconstruction stage. Solid arrows indicate the forward feature/data flow through the network, dashed arrows denote the recurrent skip connections inside each cascaded stage, and dotted arrows represent the disparity that is propagated from a coarser stage to the next finer stage. The asterisk (∗) marks the [CLS] token of the ViT-based encoder, and the lock icon indicates parameters that are frozen from the pre-trained backbone.

STTR [15] represents one of the earliest stereo matching frameworks based on attention mechanisms. It employs a convolutional stem to extract features from input images, followed by a Transformer module that performs both self-attention and cross-attention to estimate initial disparity values, which are subsequently refined through a context adjustment stage. **SSTTStereo** [71] proposes a Sliding Space-Disparity Transformer that regularizes the 4D cost volume using localized attention, balancing accuracy and computational efficiency. **ELFNet** [72] introduces a deep evidence learning framework that fuses Transformer-based context encoding with cost-volume representations, enabling two-stage fusion and generating both aleatoric and epistemic uncertainty maps alongside the predicted disparity. **S^{2}M^2** [73] is a scalable global stereo matching model that leverages a Multi-Resolution Transformer (MRT) to efficiently handle high-resolution inputs. It combines optimal transport for robust correspondence with a novel Probabilistic Mode Concentration (PMC) loss, achieving state-of-the-art accuracy and reliable depth estimation.

In the field of remote sensing, **A-SATMVSNet** [16] incorporated attention mechanisms into the feature extraction stage to alleviate the problem of incomplete surface detail recovery in satellite images. Similarly, **MaskCRNet** [18] combined a ViT-based encoder with masked representation learning in a cascaded recurrent framework to handle missing data, domain shift, and rectification errors in high-resolution satellite stereo. **Wei et al.** [74] further improved robustness by constructing multi-scale cost volumes and applying feature-level attention to suppress image noise and enhance disparity map quality for satellite stereo.

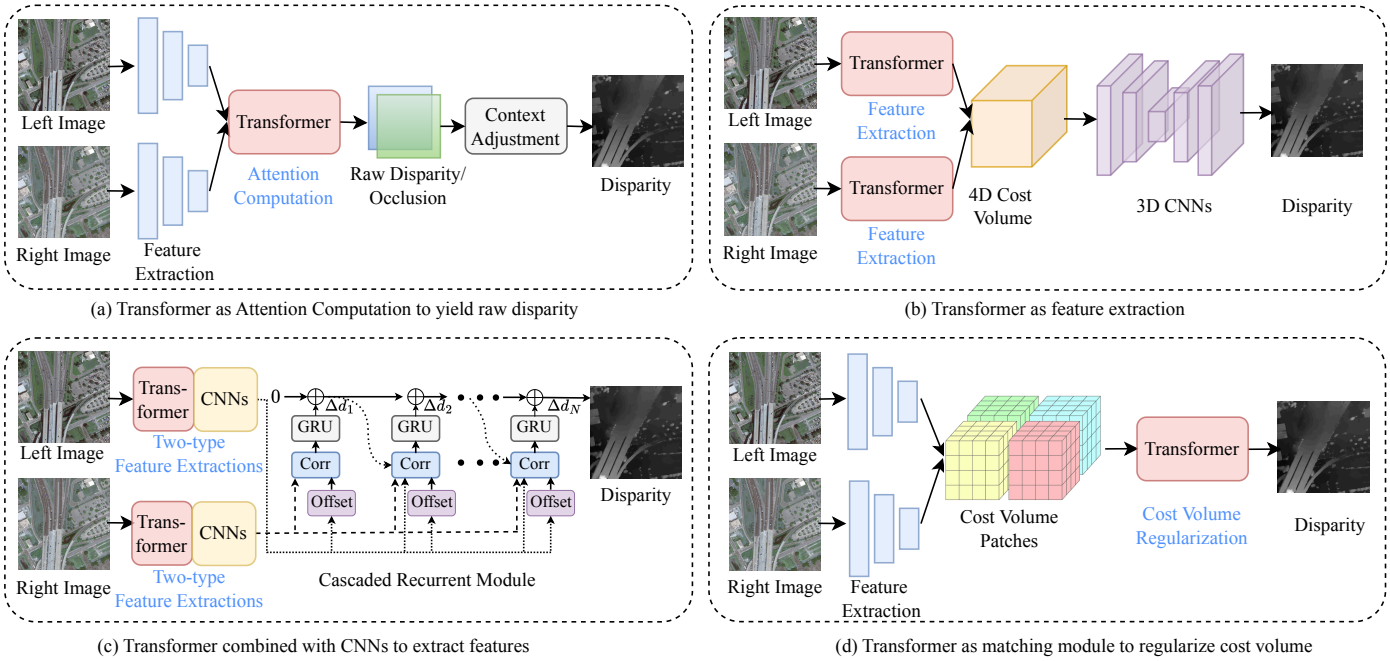


Figure 5. Four types of Transformer-based stereo matching methods. (**a**) Using Transformer for direct attention-based disparity estimation, as in STTR [15]; (**b**) employing Transformer for feature extraction, as in A-SATMVSNet [16]; (**c**) integrating Transformer with CNNs for hybrid feature extraction, as in MaskCRNet [18]; and (**d**) applying Transformer to cost volume regularization, as in SSTTStereo [71]. In all four sub-architectures, solid arrows indicate the forward propagation direction of features and intermediate tensors between successive modules; the dashed arrows in (**c**) additionally denote the iterative refinement loop of the cascaded recurrent module.

Transformers offer strong capabilities in hierarchical feature extraction and global context modeling. On terrestrial datasets, Transformer-based stereo matching models have progressively demonstrated SOTA performance. However, their application in remote sensing stereo tasks remains limited. Given the complex textures and the presence of small-scale targets in satellite imagery, the ability of Transformers to capture fine-grained and long-range dependencies makes them particularly well-suited for this domain. As such, developing Transformer-based stereo models represents a promising direction for advancing stereo matching in remote sensing.

3.2.5. Multi-Task Learning-Based Models

Multi-task learning (MTL) seeks to jointly optimize multiple related tasks, enabling shared feature representations and cross-task supervision. In the context of stereo matching, MTL is commonly combined with auxiliary tasks such as semantic segmentation, surface normal estimation, or optical flow prediction. This integration facilitates a more comprehensive understanding of scene geometry and semantics, thereby improving disparity estimation performance.

Several notable methods have demonstrated the efficacy of MTL. **RTS^{2}Net** [75] proposed a lightweight, real-time architecture that performs joint semantic segmentation and disparity estimation using a shared encoder to balance accuracy and computational efficiency. **SGNet** [76] introduced a multi-task framework incorporating three modules: a confidence module assessing consistency between semantic and disparity features, a residual initial disparity refinement module guided by semantic categories, and a fusion module for final disparity adjustment. **NNNet** [77] added a surface normal estimation branch to improve disparity accuracy via a normal consistency loss and refinement scheme. **DWARF** [78] presented a coarse-to-fine architecture that simultaneously estimates disparity, optical flow,

and disparity change using compact correlation volumes and hierarchical feature warping. More recently, **Rao et al.** [79] integrated masked self-supervised learning into a pseudo-multi-task framework to enhance model robustness and cross-domain generalization.

In the domain of remote sensing, a key milestone was the 2019 IEEE GRSS Data Fusion Contest (DFC) [121], which, for the first time, provided paired disparity and semantic segmentation annotations for satellite imagery via the US3D dataset. The competition aimed to assess the potential of joint learning strategies for stereo matching in satellite scenarios, where models must contend with drastic object-scale variation (from vehicles to large buildings), weak-texture regions (flat rooftops, farmland), heavy occlusions and shadows, and the imperfect rectification typical of incidental satellite imagery. Chen et al. [81], the winning team, proposed a multi-receptive-field contextual fusion module to aggregate context across heterogeneous object scales, solving the scale-variance problem inherent to high-resolution satellite scenes. Qin et al. [82] introduced a dynamic loss-weighting scheme and an ensemble voting strategy to jointly train U-Net and PSMNet [17], balancing the convergence rates of the semantic and disparity branches and stabilizing predictions in low-confidence regions caused by satellite-specific noise, securing second place. SDBF-Net [83] combined DeepLab-v3 and GC-Net [10] as backbones and introduced a semantic–disparity bidirectional fusion mechanism, leveraging semantic priors to disambiguate disparity in weak-texture and occluded regions on incidental satellite images.

Subsequent research further advanced these ideas. **BGA-Net** [12] introduced top-down and bottom-up bidirectional guided attention to share information between semantic segmentation and disparity estimation, alleviating the seasonal-appearance variation that hampers cross-temporal 3D detection in remote sensing imagery. **S^{2}Net** [80] built a dual-task framework in which semantic features supervise cost-volume construction to stabilize disparity under data disturbance, while disparity-derived RGB-D features in turn disambiguate foreground–background confusion in satellite RGB imagery. Figure 6 illustrates the architectures of BGA-Net and S^{2}Net, along with an example of downstream 3D reconstruction enabled by their outputs. **SemStereo** [84] argued that earlier multi-task designs—being either loosely parallel or only implicitly interacting—fail to capture the inherent coupling between the two heterogeneous tasks in remote sensing. It therefore tightly couples them through a Semantic-Guided Cascade (SGC) that propagates deep semantic features into disparity estimation, a Semantic Selective Refinement (SSR) branch for explicit disparity refinement, and a Left–Right Semantic Consistency (LRSC) loss enforcing cross-view consistency, demonstrating the effectiveness of tightly coupling semantic and stereo tasks.

MTL offers a promising strategy for enhancing stereo matching through the integration of complementary tasks. Its application in remote sensing has demonstrated notable improvements in both accuracy and generalization. However, future work must address challenges in task balancing, annotation cost, and computational overhead to fully realize its potential in operational settings.

3.2.6. Integrating Vision Foundation Models

Vision Foundation Models (VFMs), such as CLIP [122], DINO [123,124], and SAM [125], have recently garnered significant attention due to their impressive generalization capabilities across diverse tasks and domains. These models are pretrained on large-scale vision-language datasets or on unlabeled image data using self-supervised learning, serving as powerful task-agnostic feature extractors and promptable learners. While VFMs have been extensively applied to classification, segmentation, object detection, and text-to-image, their integration into stereo matching remains a nascent yet promising research direction.

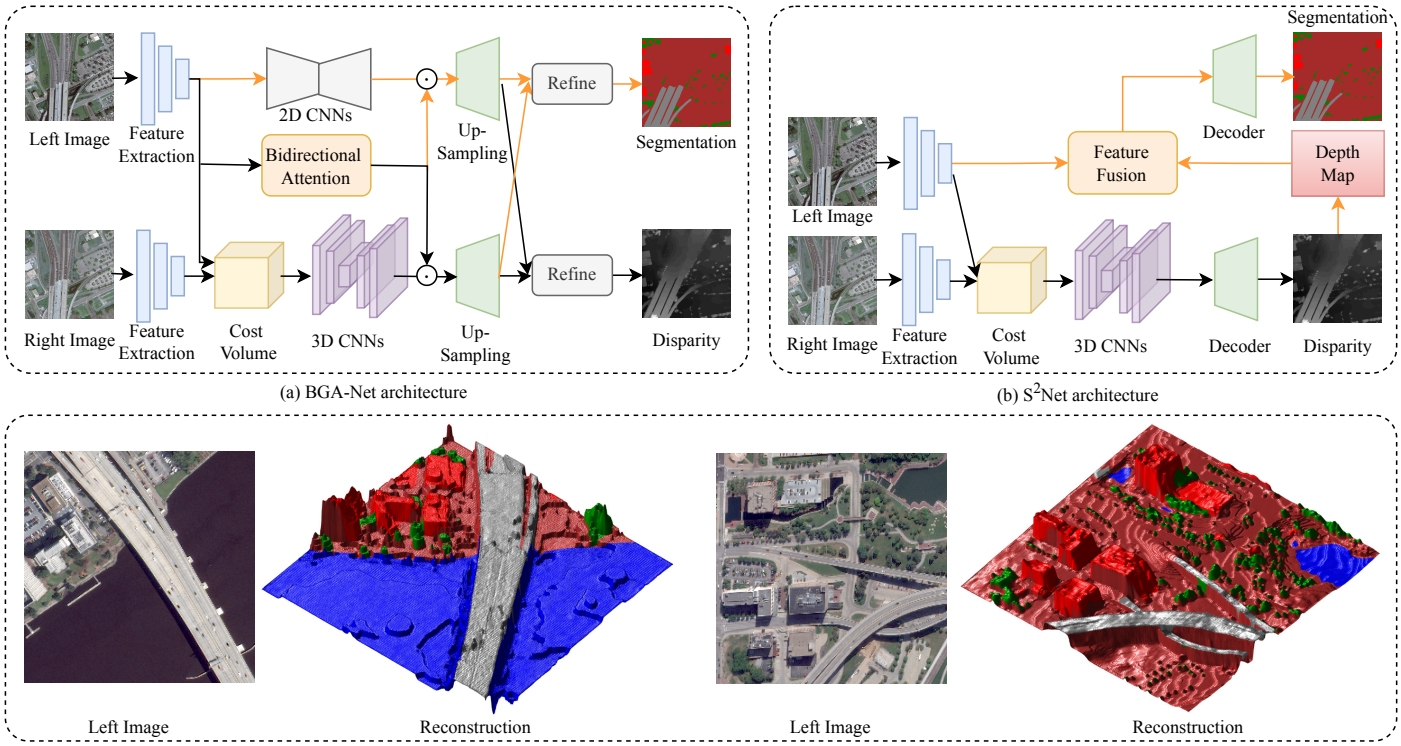


Figure 6. Examples of multi-task learning architectures for remote sensing. (**a**) BGA-Net [12] employs a shared backbone with bidirectional attention to extract unified features for both stereo matching and segmentation, followed by a refinement module that integrates initial disparity and semantic maps for further improvement. (**b**) S^2Net [80] adopts a 3D convolutional-based pipeline that first estimates disparity from binocular images and then fuses disparity and RGB features via weighted accumulation to enhance semantic segmentation. (**c**) Illustration of 3D reconstruction using combined disparity and segmentation outputs, showcasing a downstream application enabled by multi-task learning. In (**a**,**b**), black arrows indicate the forward data flow of the stereo matching branch (feature extraction → cost volume → 3D CNNs → disparity), while orange arrows indicate the data flow of the semantic segmentation branch and the bidirectional information exchange between the two branches.

Current approaches typically adapt VFMs in two main ways: (1) as high-capacity feature extractors for stereo matching pipelines; and (2) as providers of monocular depth priors that guide stereo estimation. **ViTAStereo** [86] represents one of the earliest explorations in this domain, leveraging features from the vision foundation model DINOv2 [124] to enhance the generalization of stereo models. **MonSter** [87] introduced a dual-branch architecture consisting of a monocular depth estimation module and a stereo matching module, coupled via a mutual refinement mechanism. The monocular branch utilizes DepthAnythingV2 [126], while the stereo branch adopts the IGEV framework [64]; both outputs are iteratively refined through cross-task feedback. **DEFOM-Stereo** [88] integrated monocular depth predictions into traditional iterative optimization-based stereo networks, using pretrained depth priors to initialize disparity estimation in a recurrent matching process. Later, **FoundationStereo** [85] proposed a comprehensive framework comprising a large-scale synthetic dataset (one million stereo pairs), monocular priors from DepthAnythingV2, and an attentive hybrid filter that adaptively fuses spatial and disparity cues. The architectures of these four approaches are depicted in Figure 7.

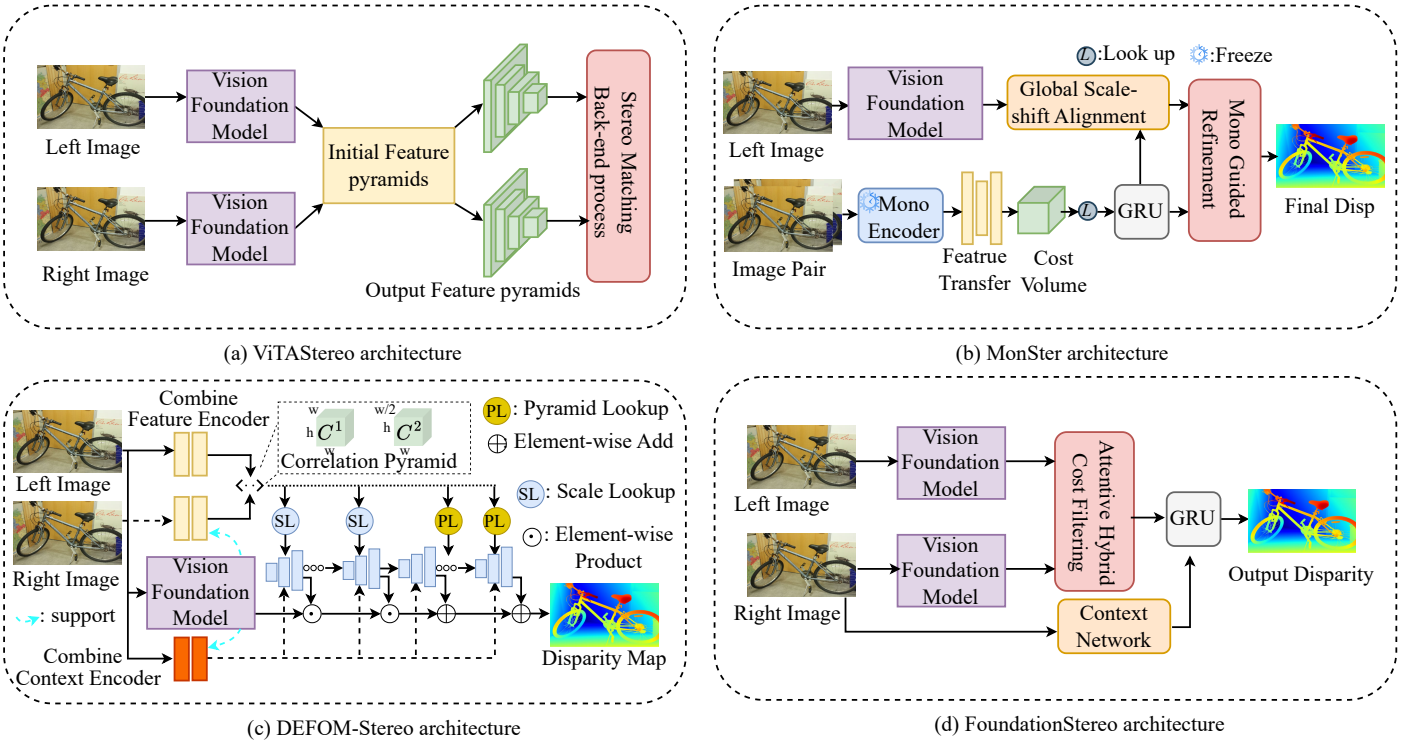


Figure 7. Four types of approaches adapting VFMs for stereo matching. (**a**) Leveraging VFM features directly for enhanced generalization, as in ViTAStereo [86]; (**b**) employing a dual-branch mutual refinement mechanism between monocular VFM and stereo modules, as in MonSter [87]; (**c**) integrating monocular depth priors from VFMs to initialize iterative disparity estimation, as in DEFOM-Stereo [88]; and (**d**) adaptively fusing monocular priors from VFMs via an attentive hybrid filter, as in FoundationStereo [85]. In all four sub-architectures, solid black arrows denote the forward propagation direction of features and intermediate tensors between successive modules; in (**c**), dashed cyan arrows denote auxiliary support connections through which the frozen VFM provides guidance features to the encoders, and dashed black arrows denote the iterative refinement loop of the recurrent matching module.

In the field of remote sensing, **HDSM-Net** [89] addressed the constraints of fixed disparity ranges and the scarcity of high-quality ground-truth labels by proposing a hierarchical vision transformer (ViT) framework for satellite stereo matching. This approach integrates self-supervised DINO for robust feature learning and a Context-Enhanced Path (CEP) to fuse global contextual information. Furthermore, by employing a pixel-by-pixel matching strategy and improved position encoding, the model enhances spatial accuracy while effectively mitigating the geometric limitations inherent in traditional satellite imagery.

Collectively, these pioneering studies underscore the potential of VFMs in advancing stereo matching, particularly in enhancing cross-domain generalization and fusing multi-modal representations. As VFMs continue to evolve with richer geometric understanding and multi-scale semantic encoding, their application in remote sensing stereo tasks represents a fertile ground for future research, especially for achieving robust and semantically-aware 3D reconstruction in open-world environments.

3.2.7. Other Emerging Models

In addition to convolutional, recurrent, and Transformer-based models, recent research has explored a variety of alternative strategies for stereo matching, drawing inspiration from architecture search, probabilistic modeling, universal visual pretraining, generative learning paradigms, and contrastive learning.

LEAStereo [90] leveraged a hierarchical Neural Architecture Search (NAS) framework [127], which incorporated stereo-specific priors into the search process to discover architectures optimized for disparity estimation. This approach yielded improved task-adaptive performance and enhanced model accuracy. **NMRF-Stereo** [91] introduced a Neural Markov Random Field (NMRF) formulation that learns pixel-level dependencies in a fully data-driven manner, mitigating the inefficiency of manual design and enabling more flexible cost aggregation. **CroCo v2** [92] extends the original CroCo framework [128] by pretraining a unified architecture that consists of a monocular encoder and a binocular decoder. The model is trained using a cross-view completion task, enabling it to learn dense correspondence. Notably, this approach eliminates the need for conventional components such as cost volumes, feature warping, or multi-scale refinement.

Generative learning has also made inroads into stereo vision. **DiffuVolume** [93] was the first to apply diffusion models to stereo matching, modeling the disparity estimation process as an iterative denoising procedure over a latent cost volume. Building on this, **DMIO** [94] combined diffusion-based reasoning with an iterative optimization framework, introducing a time-aware gated recurrent unit (T-GRU) to simultaneously capture disparity and temporal dynamics.

In the remote sensing domain, to address the over-reliance on specific scene geometries and the limited generalization of end-to-end models in unseen scenarios, **DeepSimNets** [95] proposed three multi-scale architectures that learn pixel-level similarity using contrastive loss. By remaining explicitly geometry-agnostic, these models decouple similarity learning from the underlying scene geometry, ensuring robust and flexible matching performance across diverse aerial and satellite imagery datasets.

Furthermore, to address the loss of global contextual information caused by patch-wise processing in high-complexity Transformers and the difficulty of matching multiscale objects in ill-posed regions, **MEMF-Net** [96] introduced a Mamba-based stereo matching framework for high-resolution satellite imagery. By leveraging a Mamba-based feature extractor for efficient global context modeling and a channel-spatial attention-enhanced multifrequency fusion module, the network enhances disparity estimation in challenging areas. Moreover, it incorporates a gradient-based convex upsampling module to refine output details.

Collectively, these emerging approaches reflect a growing interest in rethinking stereo matching beyond traditional pipelines. By incorporating principles from architecture search, probabilistic reasoning, foundational visual pretraining, generative modeling, and contrastive learning, they provide new pathways for building more interpretable, generalizable, and scalable stereo vision systems.

3.3. Weakly-, Semi-, and Self-Supervised Algorithms

Most deep stereo matching models rely heavily on large-scale, densely annotated disparity maps for supervised training. However, acquiring such annotations is particularly challenging in real-world scenarios, especially in remote sensing, due to the significant labor, time, and cost involved in generating high-quality ground truth data. To mitigate this dependency, recent research has focused on weakly-, semi-, and self-supervised approaches that aim to reduce the reliance on labeled data while maintaining competitive model performance.

3.3.1. Self-Supervised Models

Self-supervised learning refers to a class of approaches that generate supervisory signals directly from the input data, without requiring explicit ground truth. In stereo matching, most self-supervised methods are grounded in the principle of geometric con-

sistency: given a left view and a predicted disparity map, the right view is reconstructed via image warping. The photometric difference between the reconstructed and original right views is then used as a loss function to guide network optimization [97–101]. Recently, **DualNet** [102] proposed a self-supervised stereo framework that combines robust self-supervised teacher learning and pseudo-label supervised student training.

In the remote sensing domain, **Knöbelreiter et al.** [103] generated training data directly by applying their own 3D reconstruction method to the target dataset, and introduced an outlier filtering step to ensure data quality. To circumvent the high cost and difficulty of acquiring ground-truth labels for high-resolution satellite imagery, **Igeta et al.** [104] proposed an unsupervised stereo matching network. Their approach addresses the inherent model selection challenge in unsupervised settings by introducing a novel criterion for identifying the optimal training epoch without the need for validation labels, ensuring robust performance even in the absence of reference data. To enhance robustness in texture-less and discontinuous regions, **Chen et al.** [105] developed a self-supervised method that incorporates a superpixel random walk pre-matching (SRWP) strategy alongside a parallax-channel attention mechanism (PCAM), as shown in Figure 8.

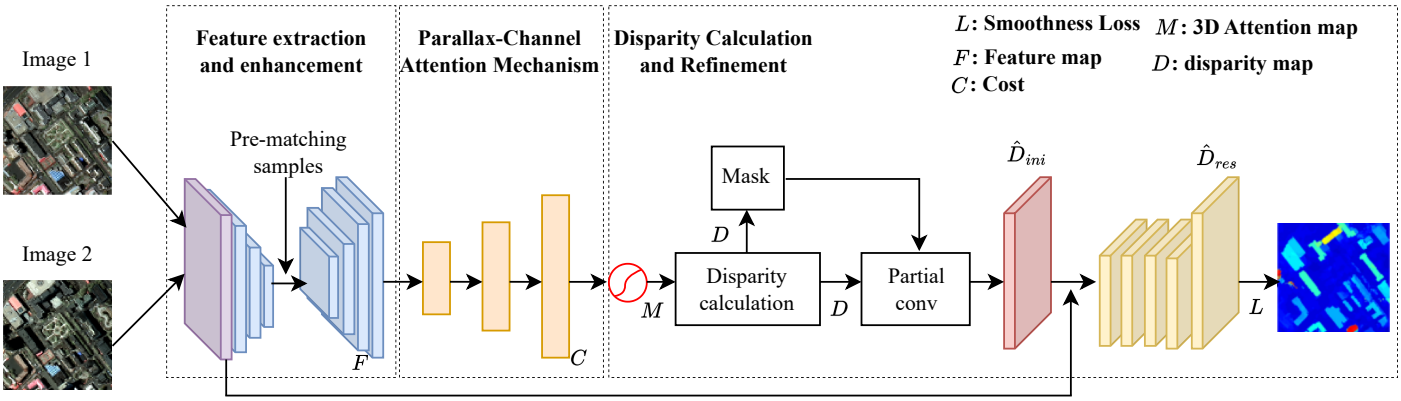


Figure 8. Architecture of the self-supervised framework [105]. This method integrates a superpixel random walk pre-matching (SRWP) strategy and a parallax-channel attention mechanism (PCAM) to enhance robustness in texture-less and discontinuous regions.

Despite recent advancements, self-supervised stereo matching methods still face significant challenges in achieving consistent and accurate results, especially in remote sensing. Unlike terrestrial datasets, remote sensing imagery often exhibits: (1) significant seasonal or temporal differences between left and right views; and (2) large homogeneous or repetitive textures (e.g., forests, water bodies, and building facades). These conditions undermine the reliability of photometric and geometric consistency assumptions, thereby limiting the effectiveness of traditional self-supervised loss formulations. Developing self-supervised algorithms that can robustly handle such domain-specific challenges remains an important and open research problem in remote sensing stereo vision.

3.3.2. Semi-Supervised Models

Semi-supervised learning serves as a promising compromise between fully supervised and self-supervised paradigms, aiming to leverage both labeled and unlabeled data to improve model performance. **Smolyanskiy et al.** [106] employed a semi-supervised loss function that combined LiDAR supervision with photometric consistency to train a deep stereo neural network. **SemiDepth** [107] leveraged left-right consistency in stereo reconstruction and integrated ground truth depth derived from LiDAR to guide network training. **Xu et al.** [108] proposed a unified framework that incorporates consistency regularization and entropy minimization to effectively utilize large-scale unlabeled data.

Semi-Stereo [109] introduced a semi-supervised stereo matching framework based on the teacher–student paradigm, wherein both networks are co-trained in a mutually beneficial manner. They further proposed a consistency-based pseudo-labeling regularization strategy with weak-strong data augmentation to better exploit information from noisy and incomplete data.

In the remote sensing domain, **HDADE** [110] tackles the two key challenges of domain shifts and the high cost of annotations by implementing a semi-supervised hierarchical domain adaptation framework. This multi-stage approach systematically facilitates feature alignment across disparate satellite sensors, proving that robust disparity estimation is achievable in cross-domain environments with minimal supervision, as shown in Figure 9.

Semi-supervised approaches show great potential in mitigating data scarcity issues and improving cross-domain generalization. However, further research is needed to address challenges such as pseudo-label noise, domain shifts, and the stability of consistency training, especially in complex and heterogeneous remote sensing environments.

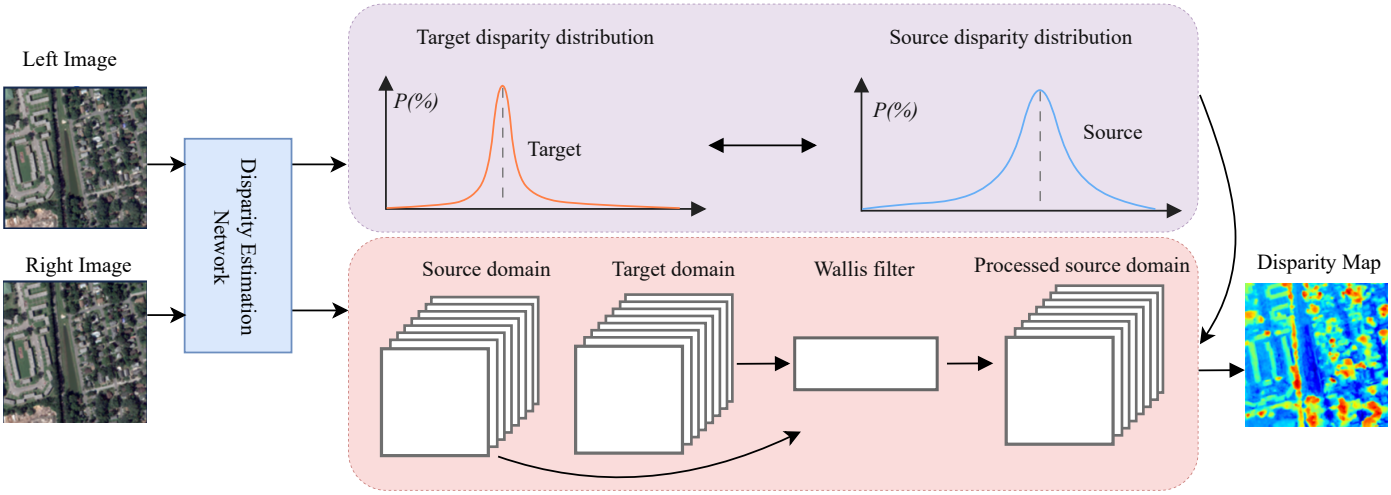


Figure 9. Architecture of HDADE [110]. This multi-stage hierarchical framework aligns source and target domains through disparity distribution matching and domain-level feature processing (e.g., Wallis filtering) to achieve robust, semi-supervised disparity estimation across disparate satellite sensors.

3.3.3. Weakly-Supervised Models

Weakly-supervised stereo matching seeks to estimate disparities using indirect or imprecise supervisory signals, such as semantic segmentation, geometric layout, or camera configuration priors, rather than dense ground-truth disparity labels. This paradigm is especially appealing for domains like remote sensing, where acquiring accurate pixel-wise annotations at scale is impractical.

Tulyakov et al. [111] pioneered this direction by integrating coarse scene and optical flow priors to guide disparity estimation from stereo image pairs, effectively mitigating label noise during training. Following this work, additional weakly-supervised methods have been proposed for general-purpose stereo vision [112], demonstrating the feasibility of learning disparity under limited annotation regimes. **SUW-Stereo** [113] jointly combined supervised learning, unsupervised learning, and weakly-supervised learning for disparity estimation.

In the remote sensing domain, **Albanwan and Qin** [114] addressed the high cost of acquiring ground truth for diverse global locations and the limited transferability of deep learning models by introducing a weakly-supervised strategy. To circumvent the dependence on precise labels, this approach utilizes filtered disparity maps derived from the traditional Semi-Global Matching (SGM) algorithm as weak supervision signals. Specifically,

the framework employs a confidence-based selection scheme that integrates SGM energy maps with image texture analysis to extract reliable disparity measurements, effectively mitigating the noise inherent in traditional matching results. By fine-tuning well-established models—such as GC-Net, PSMNet, and LEAStereo—on these high-confidence pseudo-labels, the method significantly enhances model robustness across 20 diverse geographical sites without requiring any ground-truth information, as shown in Figure 10.

Despite its potential, weakly-supervised stereo matching remains underexplored in remote sensing. This is largely due to the absence of suitable indirect labels and standardized benchmarks tailored to satellite imagery. Consequently, progress in this area has stalled, highlighting an urgent need to construct remote sensing-specific datasets and evaluation protocols that facilitate weakly-supervised learning.

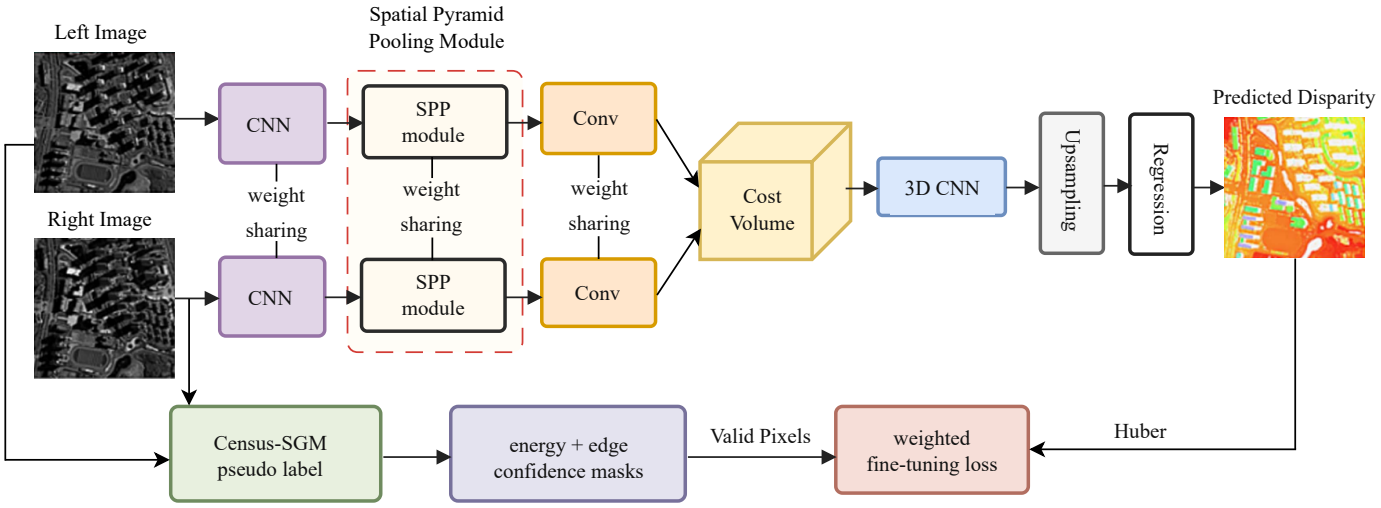


Figure 10. Architecture of the weakly-supervised framework [114]. This pipeline utilizes filtered Census-SGM pseudo-labels and confidence-based masks (integrating energy and edge information) to provide weak supervision for a Siamese network, enabling robust model fine-tuning across diverse locations without requiring ground-truth disparities.

3.4. Advantages and Limitations

Different categories of stereo matching algorithms present distinct strengths and weaknesses. For clarity, we summarize the main advantages and limitations of each representative approach in Table 2, which provides a concise comparison to highlight their applicability in remote sensing scenarios. Beyond category-specific designs, several symmetry/consistency-based refinements (e.g., left–right semantic/photometric consistency) have shown tangible gains and are directly applicable to RS scenarios.

Table 2. Advantages and Disadvantages (A&D) of different categories of stereo matching algorithms.

Category	Advantages	Disadvantages
Hybrid methods	Combine robustness of priors with flexibility of deep learning; improved cross-domain generalization	Complex pipeline; limited improvement compared to fully learning networks
2D convolution-based models	Efficient and lightweight; suitable for real-time and edge devices	Limited accuracy and robustness in complex or textureless regions
3D convolution-based models	Strong spatial regularization; higher disparity accuracy	High computation and memory cost; difficult for large-scale or real-time use

Table 2. *Cont.*

Category	Advantages	Disadvantages
Iterative optimization-based models	Accurate and memory-efficient; support progressive refinement	Sequential inference introduces latency, especially on high-resolution RS data
Transformer-based models	Capture long-range dependencies; strong semantic awareness	High computational cost; limited applications in RS domain
Multi-task learning models	Leverage semantic/geometry cues; benefit downstream tasks	Task balancing is challenging; increased annotation and computation cost
Vision Foundation Model Integration	Strong generalization; effective cross-domain adaptation via large-scale pretraining	Integration into stereo is still preliminary; training/fine-tuning is resource-demanding
Self/semi/weakly supervised methods	Reduce dependence on dense GT disparity; more practical in RS	Sensitive to seasonal/illumination changes; accuracy gap with fully supervised remains

4. Acceleration of Stereo Matching Models

End-to-end deep stereo matching models have substantially improved disparity estimation accuracy. However, their reliance on computationally intensive operations, particularly large 3D convolutions, often demands high-performance GPUs, thereby limiting their deployment in real-time or resource-constrained environments. To overcome these challenges, recent research has increasingly emphasized model acceleration, aiming to reduce runtime and memory consumption while maintaining competitive accuracy.

4.1. Lightweight Design

Although 3D convolution-based architectures offer superior accuracy, they incur significant computational overhead due to the high complexity of cost volume regularization. This results in parameter redundancy and constrains their practical application in real-world scenarios. Consequently, a significant number of studies have focused on designing lightweight network architectures to reduce computational overhead while preserving accuracy.

SCV-Net [129] introduced sparse cost volume construction to reduce computational demands, enabling stereo models to retain high accuracy with lower memory usage and faster inference. To mitigate the high GPU memory footprint of the Guided Aggregation Network (GA-Net), **Xia et al.** [130] proposed a pyramid optimization module that progressively refines disparity predictions. Their method significantly reduced memory overhead and improved runtime efficiency, while effectively accommodating large disparity ranges commonly encountered in remote sensing imagery.

Current lightweight approaches mainly center around architectural refinements and computational optimizations. While some models have demonstrated real-time performance on GPUs, achieving similar efficiency on embedded platforms or mobile devices remains an open challenge for future exploration.

4.2. Compression

4.2.1. Model Pruning

Model pruning aims to eliminate redundant parameters by identifying and removing less informative weights, thereby reducing network size and computational complexity while maintaining inference accuracy. Importance criteria for pruning can be based on weight magnitudes, gradient sensitivity, or activation statistics.

Xiang et al. [131] introduced a fast iterative shrinkage-thresholding algorithm to learn soft pruning masks for 2D convolutional layers, enabling more efficient stereo reconstruc-

tion. Despite the success of pruning in other vision domains, its application to stereo matching remains limited. Nonetheless, these initial efforts suggest that tailored pruning strategies could significantly enhance the computational efficiency of stereo models. In the context of remote sensing, where high-resolution data and large disparity ranges are common, designing pruning techniques that preserve spatial detail and domain-specific robustness remains a valuable direction for future research.

4.2.2. Model Quantization

Model quantization aims to compress deep neural networks by reducing numerical precision, thereby decreasing model size and accelerating inference. The primary challenge lies in minimizing performance degradation resulting from reduced bit width and dynamic range. Although quantization has been extensively applied to tasks such as image classification and object detection, its use in stereo matching remains relatively underexplored.

Bi3D [47] proposed a binary disparity estimation network, demonstrating a linear trade-off between precision and computational complexity. **Chen et al.** [132] developed a stereo matching framework using binary descriptors, which achieved real-time performance on FPGA hardware. More recently, **Wang et al.** [133] introduced an 8-bit quantized stereo model optimized for deployment on wearable edge devices such as smart glasses.

While quantization offers significant advantages in terms of hardware deployment and runtime efficiency, it often entails accuracy degradation. Furthermore, quantized stereo models tailored for remote sensing imagery remain largely unexplored. Given the unique challenges posed by large disparity ranges, repetitive textures, and environmental noise in satellite imagery, developing robust quantized stereo methods is both an urgent and impactful research direction.

5. Remote Sensing Stereo Matching Datasets

5.1. Development of Stereo Datasets

Public datasets have played a pivotal role in advancing stereo matching models, especially in the deep learning era. Numerous benchmark datasets have been developed for terrestrial scenarios. For instance, in 2012 and 2015, Geiger et al. [20,134] introduced the KITTI 2012 and KITTI 2015 datasets, which leverage LiDAR data to provide sparse ground truth for autonomous driving applications. Scharstein et al. [135] presented the Middlebury dataset in 2014, using structured light to obtain dense disparity maps in indoor environments. In 2016, Mayer et al. [9,136] proposed SceneFlow, a large-scale synthetic dataset created using Blender 4.2, which became widely used for pre-training stereo networks. This contributed to the now-standard training paradigm: pre-training on synthetic datasets followed by fine-tuning on smaller, real-world datasets. That same year, Huang et al. introduced Apolloscape [137], a high-resolution street-view dataset. More recently, DrivingStereo [138] and CRE-Stereo [139] further enriched large-scale stereo datasets with diverse weather conditions and synthetic scenarios.

In contrast to the abundance of stereo datasets available for natural image scenarios, benchmark datasets for remote sensing stereo matching remain scarce. This limitation primarily stems from the substantial cost and technical challenges associated with generating accurate disparity ground truth at large scales. To address this gap, the IEEE GRSS Data Fusion Contest (DFC) in 2019, organized by Le et al. [21], introduced the Urban Semantic 3D (US3D) dataset, providing annotated data for both stereo matching and semantic segmentation. As summarized in Table 3, a number of representative methods have been evaluated on this benchmark. For reference, Figure 11 presents the qualitative results of [18] on the US3D dataset, as reported in the original paper. In addition, Patil et al. [140]

proposed the SatStereo dataset, constructed from WorldView-3 satellite imagery, further facilitating research in satellite-based stereo reconstruction.

Table 3. Evaluation of the stereo matching methods on the US3D datasets. EPE: End-Point Error. D1: percentage of erroneous pixels. The downward arrow (↓) next to a metric indicates that lower values correspond to better performance. * denotes self-supervision, † represents baseline results reported in [21]. Note: MRFCNet is named after the multi-receptive-field contextual network proposed in [81]; the original paper does not provide an explicit acronym.

Method	Year	Publication	EPE (px) ↓	D_1 (%) ↓	Times (ms) ↓	Equipment
SGM † [141]	2005	CVPR	10.34	43.00	-	Xeon
iResNet-i2 † [142]	2018	CVPR	3.05	33.00	-	NVIDIA Titan X
DenseMapNet † [143]	2018	ICRA	3.51	35.00	-	NVIDIA GTX 1080Ti
MRFCNet [81]	2019	IGARSS	1.34	9.01	670	-
BGA-Net [12]	2020	TGRS	1.20	7.20	1572	NVIDIA 1080Ti
RAFT-Stereo [14]	2021	3DV	1.29	8.01	575	NVIDIA RTX 6000
HMSM-Net [58]	2022	ISPRS	1.19	7.16	511	NVIDIA RTX 3090
S^2Net [80]	2023	TGRS	1.439	10.051	-	NVIDIA Tesla V100
MaskCRNet [18]	2024	ISPRS	1.12	7.01	724	NVIDIA RTX 3090
S^3Net [144]	2024	IGARSS	1.403	9.579	-	NVIDIA Tesla V100
MEMF-Net [96]	2025	JSTARS	1.18	6.64	-	NVIDIA RTX 4090
HDSM-Net * [89]	2025	ISPRS	2.09	18.6	-	NVIDIA RTX 4090
SRCV-Net [62]	2025	JSTARS	1.387	8.615	-	NVIDIA RTX A6000
DBMSMNet [63]	2025	JSTARS	1.48	7.41	490	NVIDIA RTX 4090

Subsequent efforts have progressively expanded the availability of remote sensing stereo benchmarks through diverse annotation strategies. In 2021, Wu et al. [145] constructed the ISPRS 2021 dataset by integrating 3D reconstructions with LiDAR point clouds to generate sparse disparity annotations. In 2022, He et al. [58] leveraged GaoFen-7 stereo image pairs and airborne LiDAR measurements to obtain dense disparity ground truth, further improving annotation completeness. To alleviate the difficulty of real-world annotation, Reyes et al. [146] proposed SyntCities, a synthetic remote sensing dataset generated using CityEngine and specifically designed for stereo vision research. More recently, Li et al. [19] introduced WHU-Stereo, a benchmark constructed from airborne optical imagery and LiDAR point clouds. As summarized in Table 4, this dataset has become an important benchmark for evaluating representative remote sensing stereo methods. For reference, Figure 12 presents the qualitative results of [18] on the WHU-Stereo dataset, as reported in the original paper. Furthermore, Zhang et al. [147] proposed UAVStereo, a multi-resolution and multi-scene dataset collected from low-altitude UAV platforms, extending stereo benchmarks to aerial imaging scenarios.

5.2. Dataset Characteristics

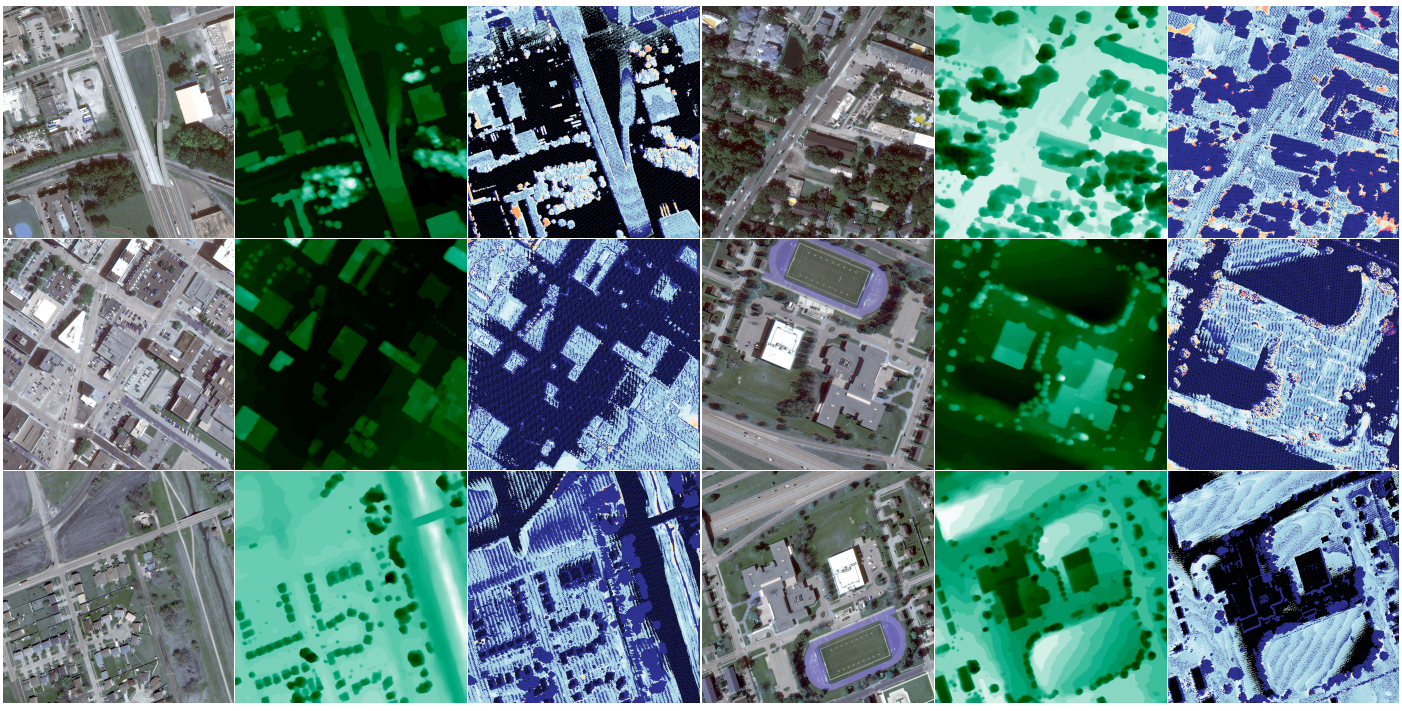
Notably, remote sensing datasets exhibit several unique characteristics that substantially influence stereo matching performance. The main characteristics are listed in follows:

(1) Ground Sampling Distance (GSD). Remote sensing datasets exhibit diverse spatial resolutions—for instance, US3D at approximately 0.3 m, WHU-Stereo at around 0.8 m, GaoFen-7 at about 0.8 m, and UAVStereo ranging from 0.05 m to 0.2 m. The GSD defines the spatial resolution of a remote sensing image, representing the real-world distance between two adjacent pixels on the ground. In satellite stereo imagery, GSD is

determined by the imaging altitude H, focal length f, and sensor pixel size p, following the relation GSD $= H \times p/f$. A smaller GSD provides finer spatial detail, allowing more accurate disparity estimation and 3D reconstruction, but also increases the data volume and computational burden. Conversely, larger GSD values reduce spatial precision and may obscure small objects such as vehicles or trees, thereby complicating stereo correspondence.

(2) Acquisition mode and processing methods. A further source of variability arises from acquisition strategies. Certain datasets (e.g., US3D, SatStereo) are constructed from incidental multi-temporal views of the same geographic location. These inevitably introduce seasonal and illumination changes, thereby creating radiometric inconsistencies between the left and right images. By contrast, other datasets (e.g., WHU-Stereo, GaoFen-7, ISPRS 2021) are based on near-simultaneous acquisitions, which mitigate such inconsistencies but still retain geometric and atmospheric challenges.

(3) Disparity range. Remote sensing datasets also exhibit distinctive disparity distributions. Unlike computer vision benchmarks such as KITTI (0–192) or SceneFlow (0–256), the disparity range in RS data is often narrower and may even include negative values (e.g., −64 to 64 in US3D). This property has direct implications for algorithm design: methods based on fixed disparity assumptions become less suitable, and instead adaptive cost volume construction, hierarchical search, or multi-scale disparity estimation strategies are required.



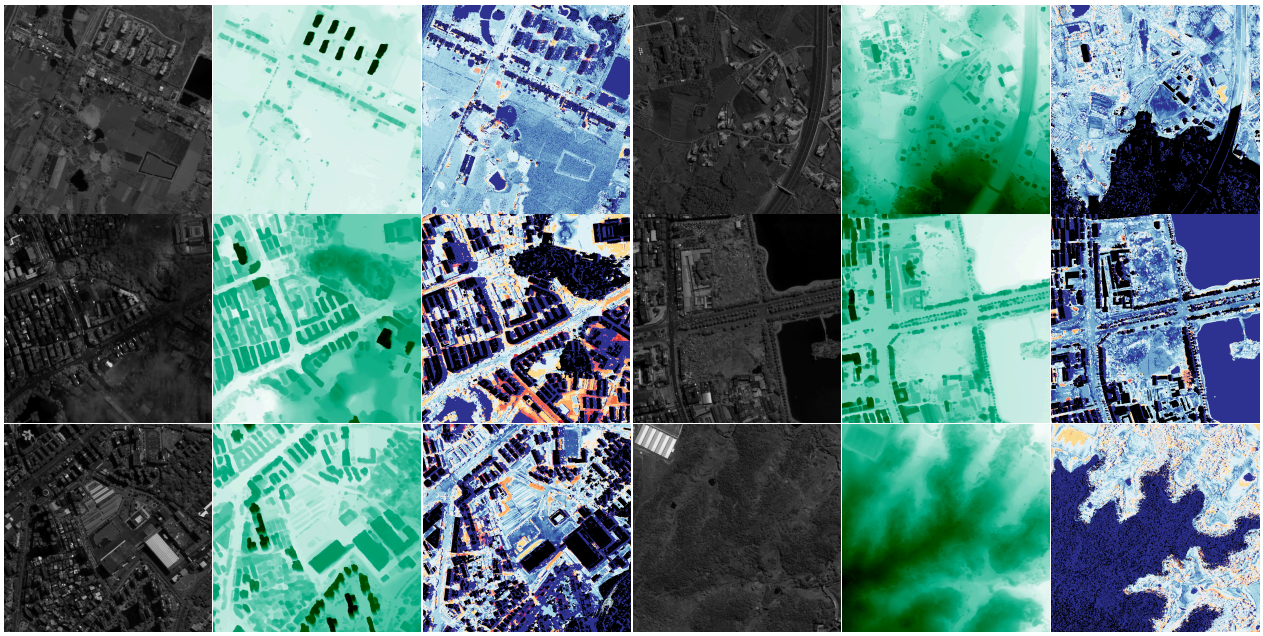
(a) left image (b) disparity map (c) error map (d) left image (e) disparity map (f) error map

Figure 11. Visualization examples on US3D dataset. Here, depicting correct estimates (<3 px) in blue and wrong estimates in red color tones.

Table 5 provides a summary of commonly used RS stereo datasets, highlighting their resolutions, acquisition modes, and ground-truth types. For example, SatStereo and US3D rely on temporally displaced image pairs from the same region, thus introducing seasonal differences, whereas GaoFen-7, WHU-Stereo, and ISPRS 2021 leverage simultaneous image capture, which helps to avoid such inconsistencies.

Table 4. Evaluation of the stereo matching methods on the WHU-Stereo datasets. The downward arrow (↓) next to a metric indicates that lower values correspond to better performance. EPE: End-Point Error. D1: percentage of erroneous pixels. *: Denotes self-supervision.

Method	Year	Publication	EPE (px) ↓	D_1 (%) ↓	Times (ms) ↓	Equipment
SGM [141]	2005	CVPR	4.88	50.79	506	Xeon
StereoNet [148]	2018	ECCV	2.45	25.12	238	NVIDIA Titan X
PSMNet [17]	2018	CVPR	2.48	24.81	614	NVIDIA Titan-Xp
RAFT-Stereo [14]	2021	3DV	1.77	14.35	575	NVIDIA RTX 6000
HMSM-Net [58]	2022	ISPRS	1.67	12.94	511	-
MaskCRNet [18]	2024	ISPRS	1.66	12.87	724	NVIDIA RTX 3090
MEMF-Net [96]	2025	JSTARS	1.524	11.6	-	NVIDIA RTX 4090
HDSM-Net * [89]	2025	ISPRS	3.310	29.0	-	NVIDIA RTX 4090
SRCV-Net [62]	2025	JSTARS	1.7523	14.15	-	NVIDIA RTX A6000
DBMSMNet [63]	2025	JSTARS	2.13	13.46	-	NVIDIA RTX 4090



(a) left image (b) disparity map (c) error map (d) left image (e) disparity map (f) error map

Figure 12. Visualization examples on WHU-Stereo dataset. Here, depicting correct estimates (<3 px) in blue and wrong estimates in red color tones.

Table 5. Stereo matching dataset for remote sensing images.

Datasets	Year	Resolution	Mode	Total Data	Training Size	Testing Size	Label Type
US3D [21]	2019	1024 × 1024	RGB	4292	4242	50	Dense
SatStereo [140]	2019	500 × 500	Panchromatic	72	–	–	Dense
ISPRS 2021 [145]	2021	1024 × 1024	IRRG	1092	585	507	Sparse
GaoFen-7 [58]	2022	1024 × 1024	Panchromatic	490	400	90	Dense
SyntCities [146]	2022	1024 × 1024	RGB	8100	6480	1620	Dense
WHU-Stereo [19]	2023	1024 × 1024	Panchromatic	1757	1220	537	Dense
UAVStereo [147]	2023	960 × 540∼8192 × 5460	RGB	38,781	30,924	7757	Dense

5.3. Dataset Challenges

Based on these dataset-level properties, remote sensing stereo faces additional domain-specific challenges. We have summarized the following main challenges:

(1) Limited data volume. Most RS stereo datasets are relatively small compared to terrestrial benchmarks, which restricts the training of deep stereo networks and limits their ability to learn complex spatial and semantic patterns.
(2) Seasonal and environmental variations. In datasets with multi-temporal acquisition, differences in season, weather, and illumination between left and right images substantially complicate disparity estimation and require models with stronger generalization capability.
(3) Distinct disparity statistics. The disparity distributions in RS imagery differ markedly from those of everyday scenes, diminishing the effectiveness of pre-training on large-scale synthetic datasets designed for CV applications.
(4) High frequency of small targets. RS scenes often contain numerous fine-scale objects, such as buildings, trees, and vehicles. Accurate disparity estimation for such targets requires stereo models to maintain high spatial resolution and fine-grained matching precision.
(5) Irregular and repetitive textures. Surfaces such as vegetation, water bodies, and urban facades frequently exhibit repetitive or ambiguous patterns. These characteristics create challenges for traditional correspondence search and necessitate algorithms that integrate semantic and geometric context.

Taken together, these characteristics make stereo matching in remote sensing a uniquely challenging problem, distinct from terrestrial scenarios. Addressing them calls for continued development of large-scale and well-annotated benchmark datasets, as well as the design of novel model architectures and learning strategies tailored specifically for remote sensing imagery.

6. Future Work and Suggestions

Despite the remarkable progress achieved in recent years, stereo matching for remote sensing imagery still faces critical challenges in accuracy, generalization, and computational efficiency. To address these limitations, future research should focus on two complementary aspects: (1) model design, to enhance robustness and adaptability across diverse remote sensing scenarios; and (2) dataset construction, to provide more representative and standardized benchmarks for algorithm development and evaluation.

6.1. Model Design

Future stereo matching models for remote sensing should be both accurate and efficient, capable of adapting to diverse imaging geometries and environmental conditions. Several directions are particularly promising:

(1) Development of high-accuracy and remote sensing oriented algorithms. Remote sensing imagery exhibits distinctive characteristics such as seasonal variation, small object size, and repetitive textures across large areas. Current models, often adapted from terrestrial benchmarks, struggle to capture these patterns effectively. Future work should focus on designing specialized architectures that explicitly model the geometric, spectral, and temporal properties of remote sensing imagery, improving both robustness and precision in complex scenes. Meanwhile, the integration of large-scale vision foundation models and multi-modal satellite data (e.g., SAR–optical fusion) will likely define the next phase of RS stereo research.
(2) Lightweight and efficient stereo matching networks. Although deep learning-based methods have achieved remarkable accuracy, their high computational cost limits

practical deployment. With increasing demands for onboard processing on UAVs, satellites, and embedded systems, future models must achieve an optimal balance between accuracy, complexity, and runtime. Lightweight multi-scale architectures, efficient cost aggregation, and adaptive attention mechanisms could facilitate real-time inference in resource-constrained environments. Although 3D reconstruction tasks emphasize accuracy, edge-based inference is essential for onboard satellite or UAV applications requiring immediate geometric feedback.

(3) Intelligent and adaptive learning under label scarcity. The limited availability of annotated stereo pairs and large domain gaps between terrestrial and aerial imagery restrict supervised learning. Future research should explore adaptive and self-evolving learning paradigms, such as self-supervised, weakly supervised, or meta-learning frameworks, integrated with foundation models and domain adaptation to enhance generalization across sensors and conditions.

(4) Unified multi-task stereo matching frameworks. Stereo matching in remote sensing is often coupled with tasks like semantic segmentation, object detection, and 3D reconstruction. Developing unified architectures that enable mutual task reinforcement while maintaining interpretability and computational efficiency could lead to comprehensive and scene-aware 3D understanding in remote sensing.

6.2. Dataset Construction

Progress in remote sensing stereo matching also depends on the availability of diverse, high-quality benchmark datasets. Future datasets should aim to capture the complexity of real-world scenarios and support fair, scalable evaluation. Two key directions are outlined below:

(1) Large-scale and diverse benchmarks. Future datasets should encompass a broad range of Ground Sampling Distances (GSDs)—from centimeter-level UAV imagery to meter-level satellite data—and include both simultaneous and multi-temporal acquisitions to assess robustness under illumination, seasonal, and geometric variations. Integrating negative disparity ranges and metadata such as viewing geometry or land-cover categories will further enhance their utility.

(2) Task-driven and standardized dataset design. Beyond disparity estimation alone, future benchmarks should be constructed with downstream applications in mind, such as 3D reconstruction, digital surface modeling, and change detection. Building datasets with consistent annotation standards, unified evaluation metrics, and open-source protocols will foster fair comparison, reproducibility, and progress across the community.

Taken together, these future directions emphasize that advancing stereo matching in remote sensing requires synergistic efforts in both algorithmic innovation and data foundation—balancing accuracy, efficiency, and generalization toward practical and scalable deployment.

7. Conclusions

This survey has reviewed binocular stereo vision in remote sensing, encompassing traditional algorithms, deep learning-based models, acceleration strategies, and datasets. While deep stereo networks developed in the computer vision domain have achieved remarkable success, they face clear limitations in remote sensing, including poor cross-domain generalization, high computational costs for very high-resolution images, sensitivity to seasonal and illumination variations, difficulty in capturing small objects, and reliance on scarce labeled data. Future progress requires models tailored to remote sensing, with lightweight and multi-scale architectures, integration of geometric and semantic priors, self-

and weakly supervised training to mitigate annotation scarcity, and improved robustness through attention mechanisms and domain adaptation. Addressing these challenges will enable more accurate and efficient stereo matching for applications such as 3D reconstruction, terrain modeling, and environmental monitoring.

Author Contributions: Conceptualization, X.L. and Z.R.; methodology, X.L.; software, Z.R.; investigation, X.L. and H.Z.; resources, M.S. and H.Z.; data curation, M.S. and H.Z.; writing—original draft preparation, X.L.; writing—review and editing, X.L., B.X., Y.D., R.H. and Z.R.; visualization, X.L., H.Z., Z.C. and Z.R.; supervision, X.L. and Z.R.; project administration, X.L.; funding acquisition, X.L. and Z.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Natural Science Foundation of China (62401244, 62473187, 52405022), the Natural Science Foundation of Jiangxi Province of China (20252BAC200195 and 20252BAC240225), and the Early-Career Young Scientists and Technologists Project of Jiangxi Province (20244BCE52116 and 20244BCE52117).

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pourrahmati, M.R.; Baghdadi, N.; Scolforo, H.F.; Alvares, C.A.; Stape, J.L.; Fayad, I.; Le Maire, G. Integration of very high-resolution stereo satellite images and airborne or satellite LiDAR for eucalyptus canopy height estimation. *Sci. Remote Sens.* **2024**, *10*, 100170. [CrossRef]
2. Venkatesan, V.; Panangian, D.; Reyes, M.F.; Bittner, K. SyntStereo2Real: Edge-aware GAN for remote sensing image-to-image translation while maintaining stereo constraint. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 17–21 June 2024; pp. 512–521.
3. Wang, X.; Jiang, L.; Wang, F.; You, H.; Xiang, Y. Disparity refinement for stereo matching of high-resolution remote sensing images based on GIS data. *Remote Sens.* **2024**, *16*, 487. [CrossRef]
4. Zhang, J.; Huang, L.; Bai, X.; Zheng, J.; Gu, L.; Hancock, E. Exploring the usage of pre-trained features for stereo matching. *Int. J. Comput. Vis.* **2024**, *132*, 4305–4326. [CrossRef]
5. Scharstein, D.; Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42. [CrossRef]
6. Boykov, Y.Y.; Jolly, M.P. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Vancouver, BC, Canada, 7–14 July 2001; Volume 1, pp. 105–112.
7. Hirschmuller, H. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 328–341. [CrossRef]
8. 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, 29 August–2 September 2011; Volume 11, pp. 1–11.
9. Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; Brox, T. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4040–4048.
10. Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; Bry, A. End-to-end learning of geometry and context for deep stereo regression. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 66–75.
11. Tao, R.; Xiang, Y.; You, H. Stereo matching of VHR remote sensing images via bidirectional pyramid network. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Waikoloa, HI, USA, 26 September–2 October 2020; pp. 6742–6745.
12. Rao, Z.; He, M.; Zhu, Z.; Dai, Y.; He, R. Bidirectional guided attention network for 3-D semantic detection of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6138–6153. [CrossRef]
13. Teed, Z.; Deng, J. RAFT: Recurrent all-pairs field transforms for optical flow. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 402–419.
14. Lipson, L.; Teed, Z.; Deng, J. RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. In Proceedings of the International Conference on 3D Vision (3DV), London, UK, 1–3 December 2021; pp. 218–227.

15. Li, Z.; Liu, X.; Drenkow, N.; Ding, A.; Creighton, F.X.; Taylor, R.H.; Unberath, M. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 6197–6206.
16. Lin, L.; Zhang, Y.; Wang, Z.; Zhang, L.; Liu, X.; Wang, Q. A-SATMVSNet: An attention-aware multi-view stereo matching network based on satellite imagery. *Front. Earth Sci.* **2023**, *11*, 1108403. [CrossRef]
17. Chang, J.R.; Chen, Y.S. Pyramid stereo matching network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 5410–5418.
18. Rao, Z.; Li, X.; Xiong, B.; Dai, Y.; Shen, Z.; Li, H.; Lou, Y. Cascaded recurrent networks with masked representation learning for stereo matching of high-resolution satellite images. *ISPRS J. Photogramm. Remote Sens.* **2024**, *218*, 151–165. [CrossRef]
19. Li, S.; He, S.; Jiang, S.; Jiang, W.; Zhang, L. WHU-Stereo: A challenging benchmark for stereo matching of high-resolution satellite images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5603914. [CrossRef]
20. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the KITTI vision benchmark suite. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
21. Bosch, M.; Foster, K.; Christie, G.; Wang, S.; Hager, G.D.; Brown, M. Semantic stereo for incidental satellite images. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1524–1532.
22. Schops, T.; Schonberger, J.L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; Geiger, A. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3260–3269.
23. Hirschmuller, H.; Scharstein, D. Evaluation of cost functions for stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 18–23 June 2007; pp. 1–8.
24. Kordelas, G.A.; Alexiadis, D.S.; Daras, P.; Izquierdo, E. Content-based guided image filtering, weighted semi-global optimization, and efficient disparity refinement for fast and accurate disparity estimation. *IEEE Trans. Multimed.* **2016**, *18*, 155–170. [CrossRef]
25. Facciolo, G.; De Franchis, C.; Meinhardt, E. MGM: A significantly more global matching for stereovision. In Proceedings of the British Machine Vision Conference (BMVC), Swansea, UK, 7–10 September 2015; pp. 1–13.
26. Patil, S.; Prakash, T.; Comandur, B.; Kak, A. A comparative evaluation of SGM variants (including a new variant, tMGM) for dense stereo matching. *arXiv* **2019**, arXiv:1911.09800. [CrossRef]
27. Lee, H.Y.; Kim, T.; Park, W.; Lee, H.K. Extraction of digital elevation models from satellite stereo images through stereo matching based on epipolarity and scene geometry. *Image Vis. Comput.* **2003**, *21*, 789–796. [CrossRef]
28. Ghuffar, S. Satellite stereo based digital surface model generation using semi global matching in object and image space. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 63–68.
29. Qin, R. A critical analysis of satellite stereo pairs for digital surface model generation and a matching quality prediction model. *ISPRS J. Photogramm. Remote Sens.* **2019**, *154*, 139–150. [CrossRef]
30. Xia, Y.; d'Angelo, P.; Tian, J.; Fraundorfer, F.; Reinartz, P. Multi-label learning based semi-global matching forest. *Remote Sens.* **2020**, *12*, 1069.
31. Wang, Y.; Qin, A.; Hao, Q.; Dang, J. Semi-global stereo matching of remote sensing images combined with speeded up robust features. *Acta Opt. Sin.* **2020**, *40*, 1628003–1628012. [CrossRef]
32. Tatar, N.; Arefi, H.; Hahn, M. High-resolution satellite stereo matching by object-based semiglobal matching and iterative guided edge-preserving filter. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1841–1845. [CrossRef]
33. Zhao, L.; Liu, Y.; Men, C.; Men, Y. Double propagation stereo matching for urban 3-D reconstruction from satellite imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5601717.
34. Zhang, Y.; Zou, S.; Liu, X.; Huang, X.; Wan, Y.; Yao, Y. LiDAR-guided stereo matching with a spatial consistency constraint. *ISPRS J. Photogramm. Remote Sens.* **2022**, *183*, 164–177.
35. Zhao, J.; Chen, X.; Hou, W.; Han, J. Stereo matching based on urban satellite remote sensing image pair. *Opt. Precis. Eng.* **2022**, *30*, 830–839. [CrossRef]
36. Zou, S.; Liu, X.; Huang, X.; Zhang, Y.; Wang, S.; Wu, S.; Zheng, Z.; Liu, B. Edge-preserving stereo matching using LiDAR points and image line features. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
37. Yue, Y.; Fang, T.; Li, W.; Chen, M.; Xu, B.; Ge, X.; Hu, H.; Zhang, Z. Hierarchical edge-preserving dense matching by exploiting reliably matched line segments. *Remote Sens.* **2023**, *15*, 4311. [CrossRef]
38. Žbontar, J.; LeCun, Y. Computing the stereo matching cost with a convolutional neural network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1592–1599.
39. Zhang, F.; Prisacariu, V.; Yang, R.; Torr, P.H. GA-Net: Guided aggregation net for end-to-end stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 185–194.

40. Duggal, S.; Wang, S.; Ma, W.C.; Hu, R.; Urtasun, R. DeepPruner: Learning efficient stereo matching via differentiable PatchMatch. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4384–4393.
41. Gómez, A.; Randall, G.; Facciolo, G.; von Gioi, R.G. An experimental comparison of multi-view stereo approaches on satellite images. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2022; pp. 844–853.
42. Gómez, A.; Randall, G.; Facciolo, G.; von Gioi, R.G. Improving the pair selection and the model fusion steps of satellite multi-view stereo pipelines. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–7 January 2023; pp. 6344–6353.
43. Albanwan, H.; Qin, R. A comparative study on deep-learning methods for dense image matching of multi-angle and multi-date remote sensing stereo-images. *Photogramm. Rec.* **2022**, *37*, 385–409. [CrossRef]
44. Gao, J.; Liu, J.; Ji, S. A general deep learning based framework for 3D reconstruction from multi-view stereo satellite images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 446–461. [CrossRef]
45. Zheng, Z.; Wan, Y.; Zhang, Y.; Hu, Z.; Wei, D.; Yao, Y.; Zhu, C.; Yang, K.; Xiao, R. Digital surface model generation from high-resolution satellite stereos based on hybrid feature fusion network. *Photogramm. Rec.* **2024**, *39*, 36–66. [CrossRef]
46. Xu, H.; Zhang, J. AANet: Adaptive aggregation network for efficient stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 1959–1968.
47. Badki, A.; Troccoli, A.; Kim, K.; Kautz, J.; Sen, P.; Gallo, O. Bi3D: Stereo depth estimation via binary classifications. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 1600–1608.
48. Tankovich, V.; Hane, C.; Zhang, Y.; Kowdle, A.; Fanello, S.; Bouaziz, S. HITNet: Hierarchical iterative tile refinement network for real-time stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 14362–14372.
49. Tosi, F.; Liao, Y.; Schmitt, C.; Geiger, A. SMD-Nets: Stereo mixture density networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 8942–8952.
50. Ji, S.; Liu, J.; Lu, M. CNN-based dense image matching for aerial remote sensing images. *Photogramm. Eng. Remote Sens.* **2019**, *85*, 415–424. [CrossRef]
51. Wang, Y.; Gong, D.; Hu, H.; Wang, S.; Han, Y.; Wang, Y.; Ma, X. State of the art in dense image matching cost computation for high-resolution satellite stereo. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2021**, *43*, 109–114. [CrossRef]
52. Shen, Z.; Dai, Y.; Rao, Z. CFNet: Cascade and fused cost volume for robust stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 13906–13915.
53. Shen, Z.; Dai, Y.; Song, X.; Rao, Z.; Zhou, D.; Zhang, L. PCW-Net: Pyramid combination and warping cost volume for stereo matching. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 280–297.
54. Li, X.; Fan, Y.; Lv, G.; Ma, H. Area-based correlation and non-local attention network for stereo matching. *Vis. Comput.* **2022**, *38*, 3881–3895. [CrossRef]
55. Li, X.; Fan, Y.; Rao, Z.; Lv, G.; Liu, S. Synthetic-to-real domain adaptation joint spatial feature transform for stereo matching. *IEEE Signal Process. Lett.* **2022**, *29*, 60–64. [CrossRef]
56. Xu, P.; Xiang, Z.; Qiao, C.; Fu, J.; Pu, T. Adaptive multi-modal cross-entropy loss for stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 5135–5144.
57. Chen, Q.; Ge, B.; Quan, J. Unambiguous pyramid cost volumes fusion for stereo matching. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 9223–9236. [CrossRef]
58. He, S.; Li, S.; Jiang, S.; Jiang, W. HMSM-Net: Hierarchical multi-scale matching network for disparity estimation of high-resolution satellite stereo images. *ISPRS J. Photogramm. Remote Sens.* **2022**, *188*, 314–330. [CrossRef]
59. Jiang, L.; Wang, F.; Zhang, W.; Li, P.; You, H.; Xiang, Y. Rethinking the Key Factors for the Generalization of Remote Sensing Stereo Matching Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 4936–4948. [CrossRef]
60. Tao, R.; Xiang, Y.; You, H. A confidence-aware cascade network for multi-scale stereo matching of very-high-resolution remote sensing images. *Remote Sens.* **2022**, *14*, 1667. [CrossRef]
61. Wu, T.; Vallet, B.; Pierrot-Deseilligny, M. PSMNet-FusionX3: LiDAR-guided deep learning stereo dense matching on aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 18–22 June 2023; pp. 6527–6536.
62. Kim, J.; Cho, S.; Chung, M.; Kim, Y. Improving Disparity Consistency With Self-Refined Cost Volumes for Deep Learning-Based Satellite Stereo Matching. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 9262–9278. [CrossRef]
63. Xu, Z.; Jiang, Y.; Wang, J.; Wang, Y. A Dual Branch Multiscale Stereo Matching Network for High-Resolution Satellite Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 949–964. [CrossRef]

64. Xu, G.; Wang, X.; Ding, X.; Yang, X. Iterative geometry encoding volume for stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 21919–21928.
65. Xu, G.; Wang, X.; Zhang, Z.; Cheng, J.; Liao, C.; Yang, X. IGEV++: Iterative multi-range geometry encoding volumes for stereo matching. *arXiv* **2024**, arXiv:2409.00638. [CrossRef]
66. Chen, Z.; Long, W.; Yao, H.; Zhang, Y.; Wang, B.; Qin, Y.; Wu, J. MoCha-Stereo: Motif channel attention network for stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 27768–27777.
67. Wang, X.; Xu, G.; Jia, H.; Yang, X. Selective-Stereo: Adaptive frequency information selection for stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 19701–19710.
68. Zeng, J.; Yao, C.; Wu, Y.; Jia, Y. Temporally consistent stereo matching. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 341–359.
69. Feng, M.; Cheng, J.; Jia, H.; Liu, L.; Xu, G.; Yang, X. MC-Stereo: Multi-peak lookup and cascade search range for stereo matching. In Proceedings of the International Conference on 3D Vision (3DV), Davos, Switzerland, 18–21 March 2024; pp. 344–353.
70. Patil, S.; Guo, Q. Stellar: A large satellite stereo dataset for digital surface model generation. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2023**, *48*, 433–440. [CrossRef]
71. Rao, Z.; He, M.; Dai, Y.; Shen, Z. Sliding space-disparity transformer for stereo matching. *Neural Comput. Appl.* **2022**, *34*, 21863–21876. [CrossRef]
72. Lou, J.; Liu, W.; Chen, Z.; Liu, F.; Cheng, J. ELFNet: Evidential local-global fusion for stereo matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 17784–17793.
73. Min, J.; Jeon, Y.; Kim, J.; Choi, M. S^2M^2: Scalable Stereo Matching Model for Reliable Depth Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025; pp. 26729–26739.
74. Wei, K.; Huang, X.; Li, H. Stereo matching method for remote sensing images based on attention and scale fusion. *Remote Sens.* **2024**, *16*, 387. [CrossRef]
75. Dovesi, P.L.; Poggi, M.; Andraghetti, L.; Martí, M.; Kjellström, H.; Pieropan, A.; Mattoccia, S. Real-time semantic stereo matching. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 10780–10787.
76. Chen, S.; Xiang, Z.; Qiao, C.; Chen, Y.; Bai, T. SGNet: Semantics guided deep stereo matching. In Proceedings of the Asian Conference on Computer Vision (ACCV), Kyoto, Japan, 30 November–4 December 2020; pp. 106–122.
77. Kusupati, U.; Cheng, S.; Chen, R.; Su, H. Normal assisted stereo depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 2189–2199.
78. Aleotti, F.; Poggi, M.; Tosi, F.; Mattoccia, S. Learning end-to-end scene flow by distilling single tasks knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10435–10442.
79. Rao, Z.; Xiong, B.; He, M.; Dai, Y.; He, R.; Shen, Z.; Li, X. Masked representation learning for domain generalized stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 5435–5444.
80. Liao, P.; Zhang, X.; Chen, G.; Wang, T.; Li, X.; Yang, H.; Zhou, W.; He, C.; Wang, Q. S^2Net: A multi-task learning network for semantic stereo of satellite image pairs. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–13. [CrossRef]
81. Chen, H.; Lin, M.; Zhang, H.; Yang, G.; Xia, G.S.; Zheng, X.; Zhang, L. Multi-level fusion of the multi-receptive fields contextual networks and disparity network for pairwise semantic stereo. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Yokohama, Japan, 28 July–2 August 2019; pp. 4967–4970.
82. Qin, R.; Huang, X.; Liu, W.; Xiao, C. Pairwise stereo image disparity and semantics estimation with the combination of U-Net and pyramid stereo matching network. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Yokohama, Japan, 28 July–2 August 2019; pp. 4971–4974.
83. Rao, Z.; He, M.; Zhu, Z.; Dai, Y.; He, R. SDBF-Net: Semantic and disparity bidirectional fusion network for 3D semantic detection on incidental satellite images. In Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 438–444.
84. Chen, C.; Zhao, L.; He, Y.; Long, Y.; Chen, K.; Wang, Z.; Hu, Y.; Sun, X. SemStereo: Semantic-Constrained Stereo Matching Network for Remote Sensing. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 15758–15766.
85. Wen, B.; Trepte, M.; Aribido, J.; Kautz, J.; Gallo, O.; Birchfield, S. FoundationStereo: Zero-Shot Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 5249–5260.
86. Liu, C.W.; Chen, Q.; Fan, R. Playing to Vision Foundation Model's Strengths in Stereo Matching. *IEEE Trans. Intell. Veh.* **2024**, 1–12. [CrossRef]

87. Cheng, J.; Liu, L.; Xu, G.; Wang, X.; Zhang, Z.; Deng, Y.; Zang, J.; Chen, Y.; Cai, Z.; Yang, X. MonSter: Marry Monodepth to Stereo Unleashes Power. *arXiv* **2025**, arXiv:2501.08643.
88. Jiang, H.; Lou, Z.; Ding, L.; Xu, R.; Tan, M.; Jiang, W.; Huang, R. DEFOM-Stereo: Depth Foundation Model Based Stereo Matching. *arXiv* **2025**, arXiv:2501.09466. [CrossRef]
89. He, X.; Yang, M.; Jiang, S.; Jiang, W.; Li, Q. Stereo Matching of High-Resolution Satellite Images via Hierarchical ViT and Self-Supervised DINO. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2025**, *X-G-2025*, 357–364. [CrossRef]
90. Cheng, X.; Zhong, Y.; Harandi, M.; Dai, Y.; Chang, X.; Li, H.; Drummond, T.; Ge, Z. Hierarchical neural architecture search for deep stereo matching. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22158–22169.
91. Guan, T.; Wang, C.; Liu, Y.H. Neural markov random field for stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 5459–5469.
92. Weinzaepfel, P.; Lucas, T.; Leroy, V.; Cabon, Y.; Arora, V.; Brégier, R.; Csurka, G.; Antsfeld, L.; Chidlovskii, B.; Revaud, J. CroCo v2: Improved cross-view completion pre-training for stereo matching and optical flow. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 17969–17980.
93. Zheng, D.; Wu, X.M.; Liu, Z.; Meng, J.; Zheng, W.s. DiffuVolume: Diffusion model for volume based stereo matching. *Int. J. Comput. Vis.* **2025**, *133*, 3807–3821. [CrossRef]
94. Shi, Y. Rethinking iterative stereo matching from diffusion bridge model perspective. *arXiv* **2024**, arXiv:2404.09051. [CrossRef]
95. Chebbi, M.A.; Rupnik, E.; Pierrot-Deseilligny, M.; Lopes, P. DeepSim-Nets: Deep similarity networks for stereo image matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 18–22 June 2023; pp. 2097–2105.
96. Yang, M.; Jiang, S.; Jiang, W.; Li, Q. Mamba-Based Feature Extraction and Multifrequency Information Fusion for Stereo Matching of High-Resolution Satellite Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 23273–23288. [CrossRef]
97. Zhong, Y.; Dai, Y.; Li, H. Self-supervised learning for stereo matching with self-improving ability. *arXiv* **2017**, arXiv:1709.00930.
98. Zhang, Y.; Khamis, S.; Rhemann, C.; Valentin, J.; Kowdle, A.; Tankovich, V.; Schoenberg, M.; Izadi, S.; Funkhouser, T.; Fanello, S. ActiveStereoNet: End-to-end self-supervised learning for active stereo systems. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 784–801.
99. Liu, P.; King, I.; Lyu, M.R.; Xu, J. Flow2Stereo: Effective self-supervised learning of optical flow and stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 6648–6657.
100. Fang, I.S.; Wen, H.C.; Hsu, C.L.; Jen, P.C.; Chen, P.Y.; Chen, Y.S. ES3Net: Accurate and efficient edge-based self-supervised stereo matching network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 18–22 June 2023; pp. 4472–4481.
101. Yuan, W.; Zhang, Y.; Wu, B.; Zhu, S.; Tan, P.; Wang, M.Y.; Chen, Q. Stereo matching by self-supervision of multiscopic vision. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 5702–5709.
102. Wang, Y.; Zheng, J.; Zhang, C.; Zhang, Z.; Li, K.; Zhang, Y.; Hu, J. DualNet: Robust Self-Supervised Stereo Matching with Pseudo-Label Supervision. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 8178–8186.
103. Knöbelreiter, P.; Vogel, C.; Pock, T. Self-supervised learning for stereo reconstruction on aerial images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 22–27 July 2018; pp. 4379–4382.
104. Igeta, T.; Iwasaki, A. An unsupervised network for stereo matching of very high resolution satellite imagery. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 971–974.
105. Chen, W.; Chen, H.; Yang, S. Self-supervised stereo matching method based on SRWP and PCAM for urban satellite images. *Remote Sens.* **2022**, *14*, 1636. [CrossRef]
106. Smolyanskiy, N.; Kamenev, A.; Birchfield, S. On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1007–1015.
107. Amiri, A.J.; Loo, S.Y.; Zhang, H. Semi-supervised monocular depth estimation with left-right consistency using deep neural network. In Proceedings of the IEEE International Conference on Robotics and Biomimetics (ROBIO), Dali, China, 6–8 December 2019; pp. 602–607.
108. Xu, F.; Wang, L.; Li, H. A unified and efficient semi-supervised learning framework for stereo matching. *Pattern Recognit.* **2024**, *147*, 110129. [CrossRef]
109. Yue, X.; Lu, Z.; Lin, X.; Ren, W.; Shao, Z.; Hu, H.; Zhang, Y.; Liao, Q. Semi-Stereo: A universal stereo matching framework for imperfect data via semi-supervised learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 17–21 June 2024; pp. 646–655.

110. Zhang, G.; Jiang, Y.; Wei, S.; Wang, Y.; Chu, J.; Tan, M.; Li, Z. Hierarchical domain adaptation framework for disparity estimation in optical satellite stereo imagery: Bridging spatiotemporal-sensor heterogeneity. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4704516. [CrossRef]
111. Tulyakov, S.; Ivanov, A.; Fleuret, F. Weakly supervised learning of deep metrics for stereo reconstruction. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1339–1348.
112. Rao, Z.; He, M.; Dai, Y.; Shen, Z. Patch attention network with generative adversarial model for semi-supervised binocular disparity prediction. *Vis. Comput.* **2022**, *38*, 77–93. [CrossRef]
113. Ren, H.; El-Khamy, M.; Lee, J. Stereo disparity estimation via joint supervised, unsupervised, and weakly supervised learning. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 2760–2764.
114. Albanwan, H.; Qin, R. Fine-tuning deep learning models for stereo matching using results from semi-global matching. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *V-2-2022*, 39–46. [CrossRef]
115. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Appl.* **1998**, *13*, 18–28. [CrossRef]
116. Schönberger, J.L.; Zheng, E.; Frahm, J.M.; Pollefeys, M. Pixelwise view selection for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 501–518.
117. Ji, S.; Luo, C.; Liu, J. A Review of Dense Stereo Image Matching Methods Based on Deep Learning. *Geomat. Inf. Sci. Wuhan Univ.* **2021**, *46*, 193–202.
118. Wang, L.; Guo, Y.; Wang, Y.; Liang, Z.; Lin, Z.; Yang, J.; An, W. Parallax attention for unsupervised stereo correspondence learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 2108–2125. [CrossRef]
119. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
120. Dosovitskiy, A. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
121. Kunwar, S.; Chen, H.; Lin, M.; Zhang, H.; D'Angelo, P.; Cerra, D.; Azimi, S.M.; Brown, M.; Hager, G.; Yokoya, N.; et al. Large-scale semantic 3-D reconstruction: Outcome of the 2019 IEEE GRSS Data Fusion Contest—Part A. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 922–935. [CrossRef]
122. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; pp. 8748–8763.
123. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 9650–9660.
124. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193.
125. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 4015–4026.
126. Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; Zhao, H. Depth Anything v2. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 21875–21911.
127. Liu, H.; Simonyan, K.; Yang, Y. DARTS: Differentiable architecture search. *arXiv* **2018**, arXiv:1806.09055.
128. Weinzaepfel, P.; Leroy, V.; Lucas, T.; Brégier, R.; Cabon, Y.; Arora, V.; Antsfeld, L.; Chidlovskii, B.; Csurka, G.; Revaud, J. CroCo: Self-supervised pre-training for 3D vision tasks by cross-view completion. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 3502–3516.
129. Lu, C.; Uchiyama, H.; Thomas, D.; Shimada, A.; Taniguchi, R.i. Sparse cost volume for efficient stereo matching. *Remote Sens.* **2018**, *10*, 1844. [CrossRef]
130. Xia, Y.; d'Angelo, P.; Fraundorfer, F.; Tian, J.; Fuentes Reyes, M.; Reinartz, P. GA-Net-Pyramid: An efficient end-to-end network for dense matching. *Remote Sens.* **2022**, *14*, 1942. [CrossRef]
131. Xiang, X.; Wang, Z.; Lao, S.; Zhang, B. Pruning multi-view stereo net for efficient 3D reconstruction. *ISPRS J. Photogramm. Remote Sens.* **2020**, *168*, 17–27. [CrossRef]
132. Chen, G.; Ling, Y.; He, T.; Meng, H.; He, S.; Zhang, Y.; Huang, K. StereoEngine: An FPGA-based accelerator for real-time high-quality stereo estimation with binary neural network. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2020**, *39*, 4179–4190. [CrossRef]
133. Wang, J.; Scharstein, D.; Bapat, A.; Blackburn-Matzen, K.; Yu, M.; Lehman, J.; Alsisan, S.; Wang, Y.; Tsai, S.; Frahm, J.M.; et al. A practical stereo depth system for smart glasses. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 21498–21507.

134. 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, 7–12 June 2015; pp. 3061–3070.
135. Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; Westling, P. High-resolution stereo datasets with subpixel-accurate ground truth. In Proceedings of the Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, 2–5 September 2014; Proceedings 36 (GCPR); pp. 31–42.
136. Mayer, N.; Ilg, E.; Fischer, P.; Hazirbas, C.; Cremers, D.; Dosovitskiy, A.; Brox, T. What makes good synthetic training data for learning disparity and optical flow estimation? *Int. J. Comput. Vis.* **2018**, *126*, 942–960. [CrossRef]
137. Huang, X.; Cheng, X.; Geng, Q.; Cao, B.; Zhou, D.; Wang, P.; Lin, Y.; Yang, R. The ApolloScape dataset for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 954–960.
138. Yang, G.; Song, X.; Huang, C.; Deng, Z.; Shi, J.; Zhou, B. DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 899–908.
139. Li, J.; Wang, P.; Xiong, P.; Cai, T.; Yan, Z.; Yang, L.; Liu, J.; Fan, H.; Liu, S. Practical stereo matching via cascaded recurrent network with adaptive correlation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 16263–16272.
140. Patil, S.; Comandur, B.; Prakash, T.; Kak, A.C. A new stereo benchmarking dataset for satellite images. *arXiv* **2019**, arXiv:1907.04404. [CrossRef]
141. Hirschmuller, H. Accurate and efficient stereo processing by semi-global matching and mutual information. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; Volume 2, pp. 807–814.
142. Liang, Z.; Feng, Y.; Guo, Y.; Liu, H.; Chen, W.; Qiao, L.; Zhou, L.; Zhang, J. Learning for disparity estimation through feature constancy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2811–2820.
143. Atienza, R. Fast disparity estimation using dense networks. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 3207–3212.
144. Yang, Q.; Chen, G.; Tan, X.; Wang, T.; Wang, J.; Zhang, X. S^3Net: Innovating stereo matching and semantic segmentation with a single-branch semantic stereo network in satellite epipolar imagery. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Athens, Greece, 7–12 July 2024; pp. 8737–8740.
145. Wu, T.; Vallet, B.; Pierrot-Deseilligny, M.; Rupnik, E. A new stereo dense matching benchmark dataset for deep learning. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2021**, *43*, 405–412. [CrossRef]
146. Reyes, M.F.; d'Angelo, P.; Fraundorfer, F. SyntCities: A large synthetic remote sensing dataset for disparity estimation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 10087–10098. [CrossRef]
147. Zhang, X.; Cao, X.; Yu, A.; Yu, W.; Li, Z.; Quan, Y. UAVStereo: A multiple resolution dataset for stereo matching in UAV scenarios. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 2942–2953. [CrossRef]
148. Khamis, S.; Fanello, S.; Rhemann, C.; Kowdle, A.; Valentin, J.; Izadi, S. StereoNet: Guided hierarchical refinement for real-time edge-aware depth prediction. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 573–590.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.