remote sensing

MDPI

Article

Stereo Matching in Satellite Imagery: A Depth Estimation Foundation Model-Assisted Iterative Approach

Kunpeng Hu and Wei Zhao *

School of Electronic and Information Engineering, Beihang University, Beijing 100191, China; hkp2000@buaa.edu.cn
* Correspondence: zhaowei203@buaa.edu.cn

Highlights

What are the main findings?

- We propose IFMA-Stereo, a novel iterative-based foundation model-assisted stereo matching method for remote sensing. This method incorporates scene structure priors from a depth estimation foundation model into the iterative disparity optimization process at both the feature and pixel levels. It achieves mutual guidance and collaborative optimization between depth information and disparity estimation, thereby overcoming the limitations of existing remote sensing disparity estimation approaches.
- The model achieves superior accuracy in texture-less, repetitive-pattern and occluded regions, and effectively mitigates errors caused by spatio-temporal heterogeneity.

What are the implications of the main findings?

- By leveraging geometric priors from a foundation model (feature-level) and iterative refinement (pixel-level), the approach provides a new pathway to overcome dataset scarcity and generalizes robustly across unseen urban areas.
- The successful coupling of monocular depth and stereo matching offers a framework to address binocular matching failures (e.g., from occlusions or scene changes), enhancing the reliability of large-scale 3D reconstruction from satellite imagery.

Abstract

In optical remote sensing 3D reconstruction, high-resolution satellite stereo matching is a critical task, yet it is challenged by extreme imaging geometries, texture-less and repetitive patterns, occlusions, and scene variations caused by spatio-temporal heterogeneity. To address these issues, we propose IFMA-Stereo, an innovative binocular disparity estimation method that leverages a monocular depth foundation model. Our approach constructs a multi-scale spatial information pyramid to jointly integrate the foundation model with a disparity extraction network. At the feature level, an attention interaction mechanism captures multi-dimensional contextual dependencies and transforms general scene understanding priors into long-range associative features suitable for stereo cost volume construction. At the pixel level, a cyclic iterative refinement module embeds depth information from the foundation model throughout the iteration process and performs joint optimization, enhancing the model's adaptability in geometrically complex regions. Experiments on the US3D and GaoFen-7 datasets demonstrate that IFMA-Stereo achieves superior performance in challenging areas (texture-less regions, disparity discontinuities, repetitive patterns) and effectively mitigates prediction errors caused by spatio-temporal heterogeneity, albeit at the cost of increased inference time compared to baseline methods. Quantitatively, the method achieves an end-point error (EPE) of 1.347 and a D1 error of 7.26% on the US3D dataset, and an EPE of 1.585 and a D1 error of 13.41% on the GaoFen-7 dataset. Notably, the method

Check for updates

Academic Editor: Rongjun Qin

Received: 4 May 2026
Revised: 11 June 2026
Accepted: 30 June 2026
Published: 7 July 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.

also yields precise predictions for unseen urban areas, indicating strong generalization. These results confirm that IFMA-Stereo achieves state-of-the-art accuracy in remote sensing disparity estimation.

Keywords: stereo matching; disparity estimation; foundation model; iterative optimization

1. Introduction

Stereo matching of satellite images constitutes a prominent research focus in remote sensing photogrammetry. Over the past three decades, commercial software packages such as ENVI, ERDAS, and Catalyst (formerly PCI Geomatica) have been widely adopted for operational matching and topographic mapping from satellite imagery, forming the backbone of many production workflows. The core task of stereo matching is to infer pixel-wise correspondences across multi-view satellite images, thereby generating dense disparity maps. Here, disparity is formally defined as the horizontal coordinate shift in a homologous point between a reference image (e.g., left image) and a target image (e.g., right image) after epipolar rectification. These disparity maps are directly linked to the geometric reconstruction pipeline: based on the principles of photogrammetric triangulation and given accurate knowledge of satellite orbital parameters, disparity values can be converted into elevation information relative to a reference surface, thereby enabling the generation of dense 3D point clouds, digital surface models (DSMs), and digital elevation models (DEMs). Consequently, accurate disparity estimation is a prerequisite for deriving reliable height measurements from stereo satellite imagery. It finds wide application in large-scale topographic mapping, 3D city modeling, disaster damage assessment, and infrastructure monitoring [1–5].

In traditional stereo matching, Semi-Global Matching (SGM) [6] is a highly representative technique that has significantly influenced many subsequent studies. Its pipeline mainly consists of four steps: matching cost computation, cost aggregation, disparity computation, and disparity refinement. However, SGM and similar traditional methods exhibit notable limitations. First, their computational demands are often substantial, leading to low efficiency and high time costs. Second, these methods are heavily constrained by the handcrafted cost computation strategies employed in the initial step, making them less adaptable to complex and variable remote sensing scenarios (e.g., large terrain slopes, urban canyons, and multi-source satellite image pairs).

In recent years, learning-based stereo matching algorithms have become dominant, particularly in close-range scenarios, driven by the availability of large-scale, high-quality datasets. Models such as RAFT-Stereo [7], IGEV-Stereo [8], and Selective-IGEV [9] employ iterative optimization, combining cost volumes with all-pairs correlations to mitigate ambiguity in ill-posed regions and over-smoothing at boundaries and fine details, establishing state-of-the-art performance in close-range stereo matching.

However, due to fundamental differences in spatial geometry, environmental conditions, and system design constraints, remote sensing imaging scenarios differ substantially from close-range imaging scenarios [10]. Consequently, the design requirements for close-range models and remote sensing models exhibit significant distinctions. Firstly, remote sensing images cover vast ground areas featuring complex and highly unpredictable scenes. Secondly, texture-less regions (e.g., roads, rivers), repetitive patterns (e.g., forests, farmland), and occlusion areas (e.g., tall buildings) are more prevalent and widely distributed in satellite imagery. Thirdly, satellite image pairs often possess inherent spatio-temporal heterogeneity, as the two images are typically not captured simultaneously, leading to po-

tential physical scene changes between acquisitions [11]. Finally, compared to close-range scenarios, which have benefited from the early release of ample, high-quality standard datasets [12–15], stereo matching in remote sensing started later. Although several excellent datasets [16–18] have been made public, the overall volume and variety of data remain relatively limited, which constrains the development of stereo matching techniques in this domain to some extent [19,20].

Our architecture is inspired by monocular depth estimation foundation models. These models are trained via a semi-supervised approach on large-scale unlabeled real-world images alongside high-quality synthetic data, resulting in depth foundation models with strong generalization capabilities in monocular scenarios [21]. After applying a depth estimation foundation model to remote sensing data, we observed that it already possesses a certain capability for relative depth extraction. This indicates that the model exhibits a degree of scene generalization ability that is not constrained by stereo image pairs. However, this relative depth extraction capability primarily reflects macroscopic depth variations and generally underperforms in fine-grained textural details. Based on these observations, we propose IFMA-Stereo, an iterative stereo matching network that integrates a depth foundation model to enhance generalization at both feature and pixel levels.

Our architecture incorporates a multi-scale spatial information pyramid for synergistic feature fusion between the foundation model and disparity extraction network. An alternating self- and cross-attention mechanism encodes stereo structural information into a bidirectional cost cube adapted to satellite disparity ranges. A cascaded recurrent network aligns foundation model depth maps with disparity via learnable transformation. Finally, a monocular–stereo joint iterative module leverages pixel-level depth priors to guide stereo matching in ill-posed regions, while stereo matching refines fine details in depth predictions, achieving mutual co-optimization.

In summary, our main contributions are as follows:

- We propose a novel iterative-based foundation model-assisted stereo matching method for remote sensing, termed IFMA-Stereo. It introduces a depth estimation foundation model into remote sensing stereo matching and integrates it through a joint multi-scale spatial information pyramid. This enables a coarse-to-fine incorporation of the foundation model's scene-generalized features into the disparity extraction model By leveraging additional prior knowledge provided by the foundation model, the proposed approach partially mitigates the challenge of dataset scarcity in remote sensing and offers a novel strategy to enhance generalization in this field.
- We introduce a feature-level attention interaction mechanism designed to capture multi-dimensional contextual dependencies. This extends the contextual feature representation and integrates specialized remote sensing scene structural priors—derived from previously incorporated generalized features—into the cost volume. This use of generalized model representations significantly improves prediction accuracy in spatio-temporal heterogeneity regions commonly encountered in remote sensing while enhancing overall robustness.
- We propose a pixel-level monocular–stereo joint iterative network to refine the disparity map. Within the optimized cost volume, non-local geometric, contextual, and stereo-structural information are jointly encoded into recurrent features, enabling mutual guidance and co-optimization between relative depth information and disparity estimates, which further advances the accuracy of the stereo matching technique. The proposed method effectively improves disparity estimation performance across various complex regions in remote sensing imagery.

- Extensive experiments on the US3D and GaoFen-7 datasets demonstrate that our IFMA-Stereo achieves competitive performance compared to state-of-the-art methods in benchmark evaluations.

2. Related Works

2.1. Disparity Estimation Method for Close-Range Images

In the field of disparity estimation, Semi-Global Matching (SGM) [6] stands as the most classic traditional method. This method primarily constructed matching costs based on manually designed matching criteria, followed by cost aggregation along multiple one-dimensional paths around matching pixels, thereby approximating a two-dimensional solution through the summation of multiple one-dimensional paths. This approach achieved relatively high-precision disparity maps within a limited computational budget. Subsequently, traditional methods such as FBS [22] further incorporated edge information from the original images, operating under the assumption that structural edges in images typically correspond to significant disparity variations. This led to the creation of bilateral space filtering for global optimization, which emphasized overall accuracy and smoothness relative to the reference image while ensuring precise structural separation. However, traditional methods often relied heavily on manually designed prior knowledge, resulting in poor adaptability to complex environments. Additionally, although SGM offered optimizations in computational time compared to global matching, the overall time cost remained substantial.

Given these limitations, recent research has increasingly shifted toward deep learning-based disparity estimation methods [23]. One foundational work in this area is GC-Net [24], which was the first to construct a 4D cost volume from the features of left and right images and employed 3D convolution to process this cost volume. It was discovered that using 3D convolutions helps extract richer contextual information from the cost volume. Following this, StereoNet [25] adopted a coarse-to-fine refinement strategy, initially generating a relatively coarse disparity map using ideas from GC-Net and subsequently designing a hierarchical, edge-sensitive refinement network to progressively optimize the disparity map. Furthermore, PSMNet [26] utilized a Spatial Pyramid Pooling (SPP) module during the feature extraction stage to capture multi-scale contextual features for constructing the cost volume. It then introduced a stacked hourglass architecture for cost volume regularization, achieving cost optimization with reduced computational complexity and enhancing the robustness of the final output.

Beyond integer-pixel disparity estimation, subpixel matching techniques have been extensively studied to overcome the discretization limitation of pixel-level matching. Traditional approaches achieve subpixel accuracy through disparity-space parabolic fitting [6], phase correlation methods [27], or gradient-based optimization [28,29]. Sabater demonstrated that block matching can theoretically achieve up to 1/20 pixel accuracy under ideal conditions [30]. In the deep learning era, differentiable soft-argmin operations enable end-to-end subpixel disparity regression [24,26]. More recent advances explicitly learn subpixel disparity distributions via uncertainty-aware focal loss [31], achieving state-of-the-art performance on benchmarks.

Inspired by Gated Recurrent Unit (GRU) modules [32], the latest research has increasingly adopted an iterative refinement approach for disparity estimation. Notable examples include RAFT-Stereo [7], IGEV-Stereo [8], and Selective-Stereo [9], which introduced specialized cyclic iterative modules. By encoding non-local geometric and contextual information into these iterative modules, these methods gradually guide the model toward deriving the optimal disparity results.

However, unlike close-range images, satellite images exhibit significant differences at multiple levels: instance-level, structural-level, image-level, and data-level. Firstly, at the instance level, satellite images cover vast and highly unpredictable ground scenes with complex variations. Secondly, at the structural level, satellite images contain extensive texture-less regions, repetitive patterns, and occlusion areas. Thirdly, at the image level, satellite image pairs are often subject to spatio-temporal imaging constraints, leading to potential scene variations between pairs. Finally, at the data level, close-range scenarios benefit from an abundance of high-quality datasets such as SceneFlow [12], Middlebury [13], KITTI [14], and ETH3D [33]. In contrast, datasets in the remote sensing domain remain relatively scarce. Although recent releases such as US3D [16], WHU [18], and GaoFen-7 [17] datasets have significantly advanced algorithm development in this field, it must be acknowledged that current research in remote sensing stereo matching remains constrained to some extent by limited data volume and dataset diversity. Therefore, developing specialized stereo matching modules and networks tailored for satellite imagery is of paramount importance.

2.2. Disparity Estimation Method for Satellite Images

Disparity estimation methods for satellite imagery emerged relatively late, primarily due to the initial lack of standardized datasets for this task. In recent years, researchers have released several high-quality satellite disparity estimation datasets, which have spurred the rapid advancement of subsequent methods in this field. Bidir-EPNet [34], building upon PSMNet, introduced a network architecture that employs a bidirectional feature propagation mechanism to fuse multi-scale contextual information for satellite disparity estimation. DSM-Net [35] adopted a dual-branch structure to process features at different scales: a coarse-scale branch captures global context, while a fine-scale branch preserves local details, thereby enhancing the accuracy of disparity estimation in satellite images. BGA-Net [36] considered the mutual reinforcement between disparity estimation and semantic segmentation tasks, jointly addressing them to achieve complementary optimization of both results. The HMSM-Net network [17] constructed cost volumes at three different scales, performed feature aggregation separately, and progressively predicted disparity maps in a coarse-to-fine manner. It also incorporated a disparity refinement module utilizing auxiliary information, significantly improving performance in challenging areas. PCWLAD [37] combined phase congruency with weighted least absolute deviation to achieve subpixel accuracy of approximately 0.4 pixels on multimodal remote sensing datasets, addressing the specific challenges of satellite image matching such as multimodal radiometric differences and complex geometric distortions. UGC-Net [38] modeled the variance uncertainty within the cost volume, linking it to disparity estimation confidence to distinguish between regions where disparity estimation is difficult and those where it is reliable. By combining this with a pseudo-volume called the 4D context volume—created through context feature expansion during the aggregation stage—it selectively leveraged contextual information to assist disparity estimation in difficult regions.

In summary, the release of high-quality satellite disparity estimation datasets has given rise to numerous effective methods, consistently advancing the precision of disparity estimation. However, challenges remain in two critical aspects. First, the accuracy of disparity estimation in difficult regions—such as texture-less areas, repetitive patterns, and occlusions—still faces significant hurdles. Second, variations in geographic locations or camera parameters can lead to substantial differences between datasets, making models trained on known urban datasets challenging to generalize to new urban environments. More importantly, most existing methods assume spatio-temporal consistency between satellite image pairs, rendering them less robust to real-world factors such as seasonal

changes and ground object variations. To enable the practical application of satellite disparity estimation, these issues must be thoroughly considered and addressed.

2.3. Vision Foundation Model in Disparity Estimation

In recent years, with the rise and remarkable success of large language models, increasing attention has been paid to visual foundation models in the field of computer vision. This has further accelerated research efforts toward developing visual foundation models that are more capable and exhibit stronger generalization. DINO [39] was a representative work in self-supervised learning based on the Transformer [40] architecture. It leveraged self-supervised learning to utilize vast amounts of unlabeled data, learning a set of transferable visual feature representations through contrastive learning on large-scale datasets. This approach facilitated the realization of various downstream image tasks, including disparity estimation. Marigold [41] was a Latent Diffusion Model (LDM) based on Stable Diffusion. It achieved superior and more generalized depth estimation by leveraging the broad priors captured in generative diffusion models. Building upon DINO, the Depth Anything series of models [21,42] further investigated the application of large-scale unlabeled data. They first trained a stable teacher model on high-quality data and then used it to generate pseudo-labels from large-scale real-world data to guide the learning of a student model. This process ultimately yielded a highly generalized depth estimation foundation model with strong generalization capabilities.

3. Materials and Methods

3.1. Overview

The proposed IFMA-Stereo introduces a visual foundation model into remote sensing disparity estimation, leveraging its generalized representations at both the feature level and the outcome level to enhance the disparity estimation process. As illustrated in Figure 1, the model consists of four main components: feature extraction based on a depth estimation foundation model, generation of a cost volume via alternating self-attention and cross-attention mechanisms, monocular–stereo joint iterative module, and disparity optimization based on multiple factors.

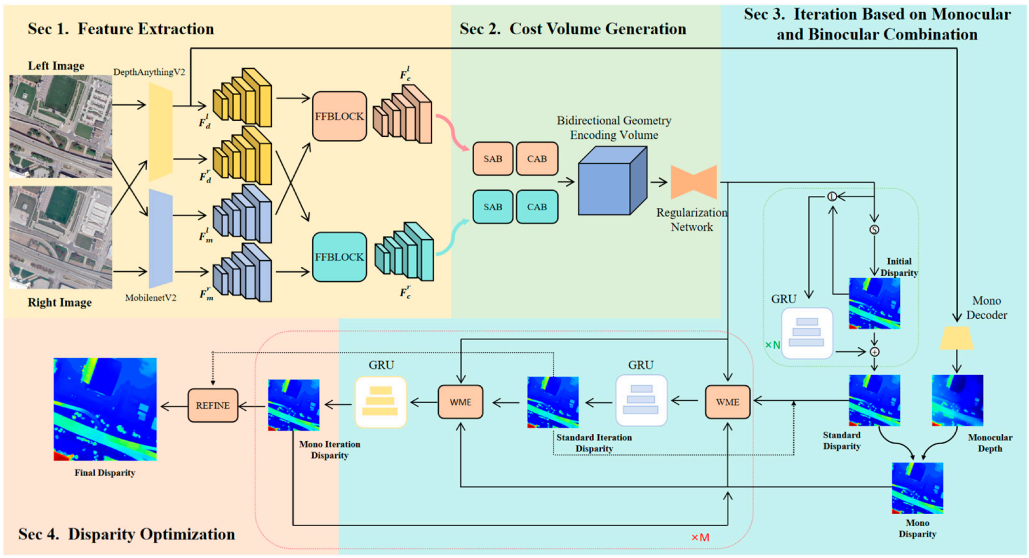


Figure 1. The overall architecture of the proposed IFMA-Stereo.

Compared to previous remote sensing stereo matching methods, our approach incorporates specialized designs based on visual foundation models. Specifically, in Section 1,

we employ a MobileNetV2 [43] network for conventional stereo matching cost volume construction and enhance it with feature-level interaction using DepthAnythingV2—a model built upon the DINO visual foundation model. Through a specially designed feature fusion block (FFBLOCK) together with the attention modules (SAB and CAB) in Section 2, monocular structural information is incorporated into the stereo cost volume, complementing the purely matching-based cues. Furthermore, in Section 3, pixel-level depth predictions are utilized to assist the binocular iterative process, guiding the model to focus more effectively on relative scene modeling. The dashed box represents the iterative process, WME represents warped motion encoder, and $\times N$ or $\times M$ denote N or M iterations. Finally, Section 4 includes the disparity optimization step that produces the final output.

3.2. Feature Extraction Based on Vision Foundation Model

For feature extraction, we construct two branches. One branch adopts a trainable lightweight CNN (MobileNetV2), following the feature extraction choice of outstanding disparity prediction models such as IGEV-Stereo [8]. The other branch employs a frozen, pre-trained DepthAnythingV2 foundation model as an auxiliary branch to introduce geometric structure priors. According to the illustration in Figure 2, a multi-scale integration is performed on the CNN features and the prior-informed features from the foundation model. The multi-scale features $\{\mathbf{F}_{d,i}^{l(r)} | i = 1,2,3,4\}$ extracted by DepthAnythingV2 are integrated with the corresponding multi-scale features $\{\mathbf{F}_{m,i}^{l(r)} | i = 1,2,3,4\}$ extracted by MobileNetV2 through the feature fusion block, where $l(r)$ denotes the possible left-view (right-view). The design enables a coarse-to-fine incorporation of the generalization information from the depth estimation foundation model, progressively yielding refined generalized features $\{\mathbf{F}_{c,i}^{l(r)} | i = 1,2,3,4\}$. Mathematically,

$$\mathbf{F}_{c,i}^{l(r)} = \text{FFBlock}_i(\mathbf{F}_{d,i}^{l(r)}, \mathbf{F}_{m,i}^{l(r)}, \mathbf{F}_u^{l(r)}(i)) \tag{1}$$

$$\mathbf{F}_u^{l(r)}(i) = \begin{cases} 0\,, & i = 1 \\ \text{upconv}_i(\mathbf{F}_{c,i-1}^{l(r)}), & i = 2,3,4 \end{cases} \tag{2}$$

where FFBlock_i denotes the i-th feature fusion block, which performs dimensional concatenation followed by convolutional fusion of features from different levels, and upconv_i represents the i-th feature alignment layer, designed to spatially align the generalized features with the multi-scale features of the adjacent level in terms of resolution.

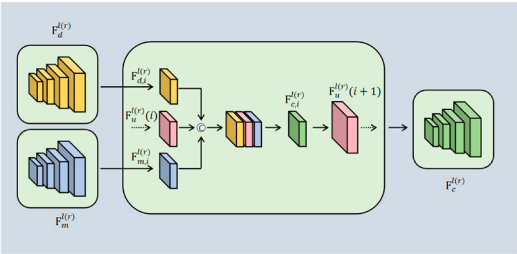


Figure 2. Flowchart of the multi-scale feature extraction process.

3.3. Cost Volume Generation Based on Self- and Cross-Attention

During feature extraction, the original foundation model DepthAnythingV2 is primarily designed for monocular depth estimation. Consequently, the multi-scale features it extracts encapsulate an understanding of the scene's inherent geometry but are not explicitly guided for disparity reasoning between stereo pairs. Therefore, this paper introduces an attention mechanism to capture long-range contextual dependencies between feature positions across the left and right views for more effective cost volume construction. By

computing attention weights between query and key vectors, the set of vectors originally dedicated to monocular depth extraction are weighted and transformed into features more suitable for stereo cost matching.

In detail, the generalized features $\{\mathbf{F}_{c,i}^{l(r)} | i = 1, 2, 3, 4\}$ obtained from feature extraction first undergo positional encoding to incorporate relative positional relationships into the spatial information. This enhancement aids the model in better understanding edge structures and supports reasoning over larger spatial contexts, which is particularly beneficial for constructing and interpreting geometric features in extensive texture-less regions. This paper employs sinusoidal relative positional encoding. Given that spatial information in images exhibits translation invariance and disparity depends solely on image content [44], relative positional encoding is applied only to the Query and Key vectors. Mathematically,

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{(\mathbf{Q} + \mathbf{R}_Q)(\mathbf{K} + \mathbf{R}_K)^{\mathrm{T}}}{\sqrt{C}})\mathbf{V} \tag{3}$$

Since $\mathbf{R}_Q$ and $\mathbf{R}_K$ represent relative positional encodings and carry no image content information—and considering that disparity is determined exclusively by image content—the attention mechanism can be correspondingly expanded and reformulated as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}} + \mathbf{R}_Q\mathbf{K}^{\mathrm{T}} + \mathbf{Q}\mathbf{R}_K^{\mathrm{T}}}{\sqrt{C}})\mathbf{V} \tag{4}$$

After obtaining the generalized features for the left and right images through the feature extraction step, we perform weighted aggregation along the epipolar line within the disparity search range to construct the attention mechanism. Following the approach described in prior attention-related literature [45], we first apply self-attention to the generalized features of the left and right images separately, resulting in refined features. Subsequently, we employ a cross-attention mechanism for further processing, yielding feature representations that are more suitable for stereo cost matching. The specific procedures of self-attention block SAB and cross-attention block CAB are illustrated in Figure 3.

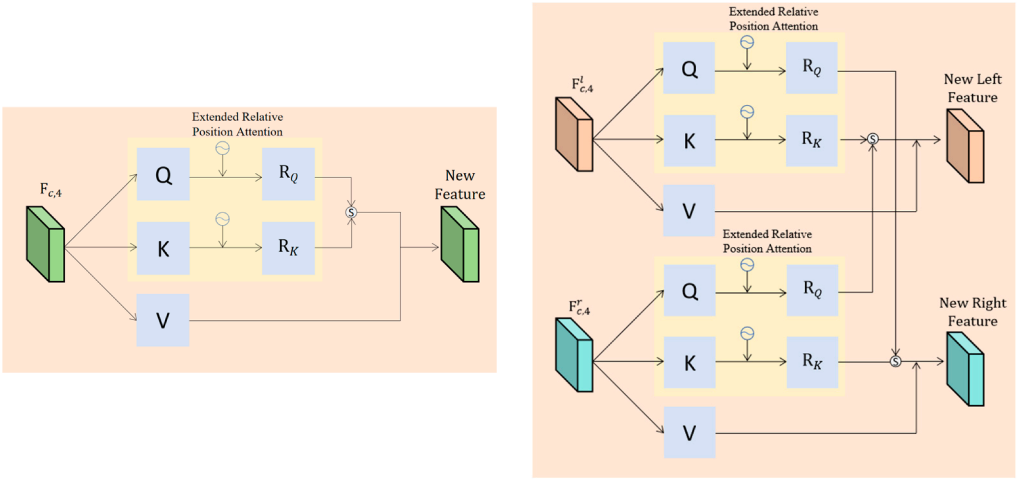


Figure 3. Flowchart of applying a single self-attention operation and a cross-attention operation.

Subsequently, we need to leverage the generalized features obtained from the foundation model to construct a cost volume that can be integrated into the main stereo matching pipeline. A potential assumption underlying previous approaches that incorporate general features from foundation models into stereo matching is the strict correspondence between views and viewing angles—namely, the left view is observed by the left device and the right view by the right device. This assumption holds in almost all conventional scenarios. How-

ever, due to the specific characteristics of remote sensing image pairs, including variations in acquisition devices and viewing angles, the disparity range of such image pairs can vary drastically from negative to positive values, thereby breaking the strict correspondence between views and viewing angles. To address this issue, we introduce negative disparities into the cost volume construction and extend the view-restricted geometry encoding volume to a more generalized bidirectional geometry encoding volume that no longer relies on strict view correspondence. Figure 4 illustrates a comparison of pixel matching between general scenes and remote sensing scenes, and the horizontal and vertical axes represent the potential matching horizontal coordinates of the epipolarly rectified left and right images, respectively. Within a specified disparity range ΔD, features from the left and right views are horizontally shifted and subsequently processed through dot-product operations between feature vectors, resulting in the construction of a bidirectional geometry encoding volume. Furthermore, we adopt the all-pairs correlation approach proposed by RAFT-Stereo to enhance image detail preservation and improve discriminative capability in challenging edge regions.

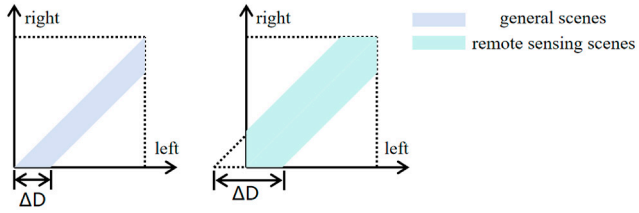


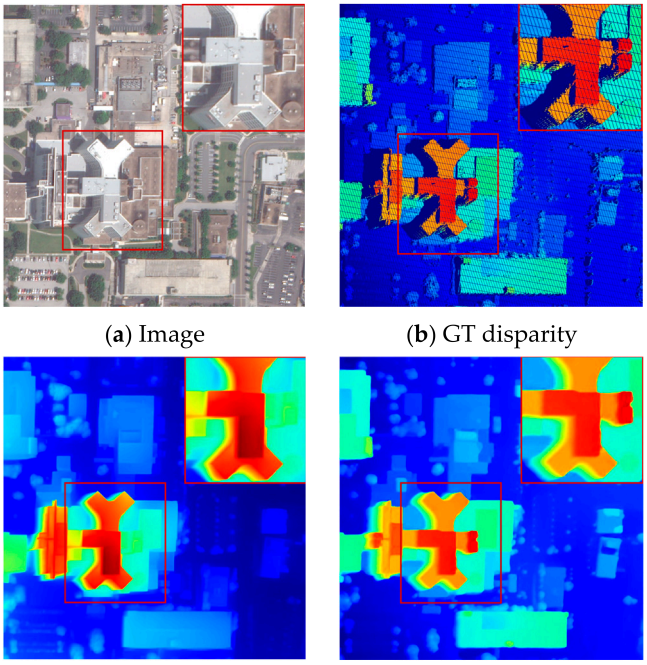
Figure 4. Comparison of the matching range using dot products of feature vectors between general scenes and remote sensing scenes.

3.4. Iteration Based on Monocular and Binocular Combination

Monocular depth estimation foundation models possess a certain capability to regress depth information. This pixel-level depth information can, from a different perspective, help overcome challenges such as significant spatio-temporal variations and difficulties in generalizing across images from different cities, thereby providing a relatively reliable initial depth estimate. On the other hand, iterative-based methods do not directly regress the final disparity but aim to iteratively predict disparity residuals, gradually approaching the optimal estimate. While iterative methods primarily operate at the disparity level, the depth information estimated from monocular images only contains relative relationships and cannot be directly converted. Therefore, we need to extract the relative relationships from this depth information and align them with disparity.

Our experiments reveal that the disparity results obtained from standard iterative methods are more conducive to recovering local fine details, yet they tend to produce blurry predictions in certain non-local geometric structures. In contrast, foundation models, which are primarily adapted to natural images, exhibit strong recognition capability for global geometric contours. However, due to the lack of specialized training on remote sensing data, the domain gap leads to low reliability of local details when transferring these models to remote sensing scenarios. As illustrated in Figure 5, the unaligned depth extracted from the foundation model is visualized in color and referred to as Relative depth, while the color-visualized disparity results obtained from the standard iterative method are denoted as Predicted disparity. The ground-truth disparity maps are also provided for reference (GT disparity). Comparing Relative depth with Predicted disparity reveals that Relative depth is more sensitive to structural edges and distinguishes well between macro-scale structures due to its understanding of scene geometry. In the figure, this manifests as a tendency to produce relatively straight edges that reflect the relative depth relationships, which aligns well with the structural information of the buildings in this scene. In contrast, Predicted

disparity tends to yield relatively blurrier edges, appearing as less straight boundaries. However, Relative depth exhibits deficiencies in capturing local details. For instance, as shown in the red box in Figure 5, the ground-truth disparity indicates that the right side of the X-shaped structure should be at a similar depth level to the two ends of the X-shape. The Predicted disparity from the standard iterative method accurately reflects this. In contrast, the Relative depth map incorrectly assigns a significantly lower depth value to the right side of the X-shape, in stark discrepancy with the depth levels at the two ends of the X-shape. This undoubtedly represents a severe distortion in the local depth prediction.



(**a**) Image (**b**) GT disparity

(**c**) Relative depth (**d**) Predicted disparity

Figure 5. Color visualization comparison of results from the depth estimation foundation model and the conventional iterative method.

Therefore, we decide to combine the advantages of both approaches by leveraging the macroscopic accuracy of the depth estimation foundation model to guide the iterative method in further refining the structural geometric information. This requires extracting the relative relationships from the depth information to establish an appropriate mapping. We propose to construct a simple linear relationship for mapping depth to disparity. The analytical solution to this linear regression problem can be obtained via the least squares method, ultimately reducing to solving a system of linear equations. We need to compute the optimal transformation parameters $\theta \in \mathbb{R}^{2\times 1}$, including a weight term and a bias term, to establish the conversion relationship between the flattened monocular depth $\mathbf{D} \in \mathbb{R}^{L\times 1}$ and the target flattened disparity $\mathbf{d} \in \mathbb{R}^{L\times 1}$. Mathematically, this is derived as follows:

$$\mathbf{d} = [\mathbf{D}, \mathbf{I}_{L\times 1}]\theta \tag{5}$$

where $\mathbf{I}_{L\times 1}$ represents an all-1 matrix with the size of $L \times 1$. By applying the least squares method, the final parameters θ can be obtained:

$$\min_{\theta} \|\mathbf{d} - [\mathbf{D}, \mathbf{I}]\theta\|_2^2 \tag{6}$$

$$\frac{\partial}{\partial \theta}\|\mathbf{d} - [\mathbf{D}, \mathbf{I}]\theta\|_2^2 = 0 \tag{7}$$

$$[\mathbf{D},\mathbf{I}]^{\mathrm{T}}[\mathbf{D},\mathbf{I}]\theta = [\mathbf{D},\mathbf{I}]^{\mathrm{T}}\mathbf{d} \tag{8}$$

$$\theta = \left([\mathbf{D},\mathbf{I}]^{\mathrm{T}}[\mathbf{D},\mathbf{I}]\right)^{-1}[\mathbf{D},\mathbf{I}]^{\mathrm{T}}\mathbf{d} \tag{9}$$

To establish the conversion relationship between depth and disparity, the standard disparity $\mathbf{d}$ must be known. However, the standard disparity is the target output and cannot be directly used in the conversion formula. Therefore, an intermediate result is employed as an approximation here, adopting a segmented iterative approach. Initially, a single-path iteration is constructed to predict a disparity, followed by a monocular–stereo joint iteration for higher-precision disparity estimation.

Specifically, as illustrated in Figure 6, the initial disparity is first regressed from the geometry encoding volume obtained by the regularization network as a relatively coarse disparity. Subsequently, following the idea of IGEV, in each iteration k, the current disparity value is used to index the features in the geometry encoding volume, and the corresponding correlations are computed via linear interpolation. These indexed features are then fed into the subsequent GRU-based update operator (ConvGRU), which works in conjunction with the context network to jointly update the hidden state of the ConvGRU and the current disparity value. The resulting disparity, which is relatively accurate, is referred to as the standard disparity.

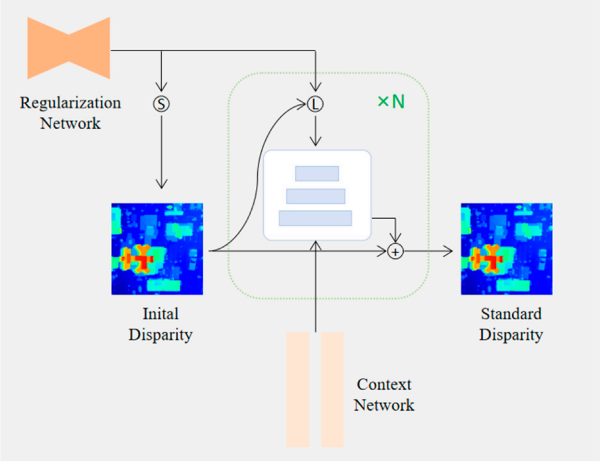


Figure 6. Architecture of the single-path iterative disparity prediction process.

The architecture of the monocular–stereo joint iteration is illustrated in Figure 7. The depth information is processed via linear regression, where the standard disparity obtained from the initial single-path iteration is substituted as $\mathbf{d}$ into Equation (9) to compute the optimal transformation parameters θ. This ultimately yields the mono disparity corresponding to the monocular depth. Meanwhile, to mitigate the impact of erroneous initial values on subsequent iterative updates, we employ the final-level generalized features of the left and right images constructed via the attention mechanism. These features are independently aligned using a warped motion encoder (WME) to compute the current iterative disparity residual, which is then fed into the ConvGRU update operator, thereby progressively refining the current iterative disparity. Specifically, the current iterative disparity residual is first constructed by computing the difference between the left image features $\mathbf{F}_{\mathrm{A}}^{\mathrm{l}}$ and the right image features $\mathbf{F}_{\mathrm{A}}^{r}$—which have been processed through the attention mechanism in the previous step—after warping the right features using the current disparity. Mathematically,

$$\mathbf{flaw}^{s(m)} = \mathrm{warp}\left(\mathbf{F}_{\mathrm{A}}^{r}, \mathbf{d}_{k}^{s(m)}\right) - \mathbf{F}_{\mathrm{A}}^{\mathrm{l}} \tag{10}$$

where warp($\mathbf{F}_{\mathrm{A}}^{r}, \mathbf{d}_{k}^{s(m)}$) denotes the result of resampling the right image features along the epipolar direction according to the current disparity.

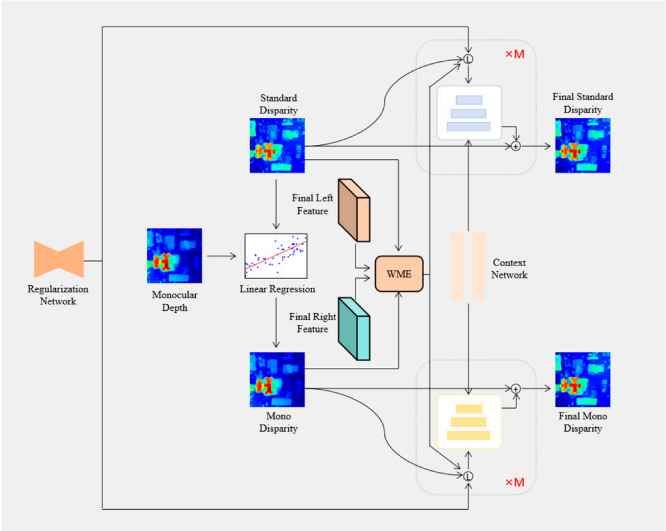


Figure 7. Architecture of the monocular–stereo joint iteration.

Subsequently, the iterative disparity residual is utilized for warped motion encoding and then incorporated into a GRU-based update operator. Finally, a new x_k' is computed by combining the standard disparity $\mathbf{d}_k^s$ and the mono disparity $\mathbf{d}_k^m$. Mathematically,

$$x_k' = \left[\mathrm{Encoder}_g^s\left(\left[\mathbf{G}_f^s, \mathbf{flaw}^s\right]\right), \mathrm{Encoder}_g^m\left(\left[\mathbf{G}_f^m, \mathbf{flaw}^m\right]\right), \mathrm{Encoder}_d^s(\mathbf{d}_k^s), \mathrm{Encoder}_d^m(\mathbf{d}_k^m), \mathbf{d}_k^s, \mathbf{d}_k^m\right] \tag{11}$$

where s denotes the standard disparity branch, m denotes the mono disparity branch. By substituting the new x_k' into the GRU update formula, a new dual-branch iterative disparity update can be achieved.

3.5. Disparity Optimization Based on Multiple Factors

Previous works [17,46] have demonstrated the effectiveness of disparity optimization. Some utilize image gradients to assist disparity refinement, others selectively combine effective information by separating high- and low-frequency components, while still others integrate multi-scale information. In general, these approaches incorporate additional auxiliary techniques after the disparity estimation step, selectively combining specific information to guide the model toward focusing on more critical details. In this paper, we extend this line of thought but tailor it to the characteristics of our model by introducing a remote sensing image-specific monocular–stereo joint optimization strategy. Specifically, the final binocular branch disparity $\mathbf{D}$ and monocular branch disparity are horizontally resampled onto the original right image $\mathbf{I}_r$, and their errors relative $\mathbf{flaw}_{\mathrm{I}}$ to the left image $\mathbf{I}_l$ are computed. Mathematically,

$$\mathbf{flaw}_{\mathrm{I}} = \mathrm{warp}(\mathbf{I}_r, \mathbf{D}) - \mathbf{I}_l \tag{12}$$

Furthermore, inspired by the approach of [46], we construct high-frequency and low-frequency separation channels to decouple the high- and low-frequency features within the integrated information. The decoupled information is then adaptively recombined to predict the residual of the initial disparity, thereby further refining the disparity. Building upon this, we additionally introduce the disparity deviation map between the original left and right images as a confidence measure for disparity reliability. Moreover, the previously

derived monocular and binocular branch disparities are integrated and jointly fed into the disparity refinement module. Through high-low frequency separation and adaptive recombination, the final optimized disparity is obtained.

3.6. Training Loss

The output of this study comprises three components: the initial disparity map $\mathbf{d}_0$ predicted by the initial cost volume, the initial iterative disparity map $\mathbf{d}_i$ output by the initial iterative model, and the iterative disparity maps $\mathbf{D}_i$ simultaneously generated by both branches during each iteration cycle. All output disparity maps are optimized using the Smooth_{L_1} loss function.

$$L_1 = \text{Smooth}_{L_1}(\mathbf{d}_0 - \mathbf{d}_{gt}) \tag{13}$$

$$L_2 = \sum_{i=1}^{N} \lambda_i * \text{Smooth}_{L_1}(\mathbf{d}_i - \mathbf{d}_{gt}) \tag{14}$$

$$L_3 = \sum_{i=1}^{2M} \mu_i * \text{Smooth}_{L_1}(\mathbf{D}_i - \mathbf{d}_{gt}) \tag{15}$$

where N denotes the number of iterations for the initial iterative model, and M denotes the number of iterations for the dual-branch iterative model. The Smooth_{L_1} loss function is defined as:

$$\text{Smooth}_{L_1}(x) = \begin{cases} 0.5x^2, \text{if } |x| < 1 \\ |x| - 0.5, \text{otherwise} \end{cases} \tag{16}$$

Finally, the total loss function is

$$L_{\text{total}} = L_1 + L_2 + L_3 \tag{17}$$

4. Results

4.1. Experiment Settings

4.1.1. Datasets and Evaluation Metrics

US3D [16] is a large-scale remote sensing dataset captured by the American WorldView-3 satellite. It comprises 2139 image pairs from the Jacksonville area and 2153 image pairs from the Omaha area, all of which have been epipolarly rectified, with each image having a resolution of 1024 × 1024 pixels. The ground truth disparity data are provided by airborne LiDAR, and the disparity range of the dataset is $[-96, 96]$. We randomly selected 1600 image pairs from the Jacksonville region as the training set, with the remaining 539 pairs serving as the validation set. All 2153 image pairs from the Omaha area are used as a separate test set to evaluate the model's cross-city generalization capability.

GaoFen-7 (GF-7) [17] is a remote sensing dataset captured by the Chinese GaoFen-7 satellite, containing 490 epipolarly rectified stereo image pairs from Yingde City, Guangdong Province, with the same resolution of 1024 × 1024 pixels. The ground truth disparity data are also provided by airborne LiDAR, and the disparity range of the dataset is $[-112, 16]$. In addition to urban scenes, this dataset includes a wide variety of non-urban terrains such as farmlands, lakes, forests, rivers, and mountains, enabling performance evaluation in more diverse and open scenarios. Following previous work, we allocate 400 image pairs for training and the remaining 90 pairs for testing.

In this study, end-point error (EPE) and 3-pixel error D1 are used as evaluation indexes for disparity. Mathematically,

$$EPE = \frac{1}{N} \sum_{(x,y)} \left| \hat{d}(x,y) - \widetilde{d}(x,y) \right| \tag{18}$$

$$D1 = \frac{1}{N}\sum_{(x,y)}\left[\left|\hat{d}(x,y) - \tilde{d}(x,y)\right| > 3\right] \tag{19}$$

where x and y represent the horizontal and vertical coordinates of the picture, $\hat{d}$ represents the estimated disparity at a specific pixel, and $\tilde{d}$ represents the ground truth at a specific pixel. D1 is a pixel error with an evaluation difference threshold of 3.

In addition to the standard disparity metrics (EPE and D1), we introduce the height-based accuracy metrics Root-Mean-Square Error (RMSE) and Mean Absolute Error (MAE) to directly evaluate the practical utility of the proposed method for DSM generation and height measurement. They are defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(h_i^{\mathrm{pred}} - h_i^{\mathrm{ref}}\right)^2} \tag{20}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|h_i^{\mathrm{pred}} - h_i^{\mathrm{ref}}\right| \tag{21}$$

In the above definitions, h_i^{pred} denotes the height estimate obtained by applying a calibration to the predicted disparity map, while h_i^{ref} is the corresponding ground-truth AGL (Above Ground Level) value, which represents the relative height of objects above the local terrain surface. It should be noted that the primary focus of this study remains on disparity estimation, and the height-based evaluation presented here serves as a supplementary photogrammetric assessment rather than the main contribution. The core quantitative comparisons and conclusions of this work are still based on the standard disparity metrics (EPE and D1), which directly reflect the stereo matching accuracy. The elevation analysis is intended to provide additional insight into the practical applicability of our method for DSM generation, but it is not intended to replace the primary disparity-based evaluation.

4.1.2. Implementation Details

This study employs the PyTorch framework (version 2.0.0) and conducts experiments using NVIDIA RTX 4090 GPUs (NVIDIA Corporation, Santa Clara, CA, USA). During training, simple data augmentation strategies including random scaling and saturation adjustment are applied to both the US3D and GaoFen-7 datasets. The training image pairs are subsequently randomly cropped to a size of 320 × 736 pixels for preprocessing. Given the inherent differences in disparity distributions between the US3D and GaoFen-7 datasets, the primary disparity search ranges are set to $[-96, 96]$ and $[-112, 16]$, respectively. Additionally, both datasets undergo uniform normalization prior to training.

Regarding training parameters, a batch size of 4 is used across our methods. The AdamW optimizer is adopted along with a one-cycle learning rate scheduling policy, with the initial learning rate set to 0.0002. The model is trained from scratch on the US3D dataset for 50,000 iterations. For the GaoFen-7 dataset, the model is initialized with weights pre-trained on US3D and similarly trained for 50,000 iterations. After training, performance is evaluated on the respective test sets.

4.2. Experiment Result

4.2.1. Results on US3D

To evaluate the cross-city generalization capability of the model, we compare the proposed method with various classical and state-of-the-art approaches. All methods are trained exclusively on a subset of images from the Jacksonville area and tested on image pairs from the Omaha area. Among the classical methods, NCC (Normalized Cross-Correlation) suffered from prohibitively long matching times due to the high resolution

of the original images and the large disparity range—typically requiring approximately 20 min per image pair, while the US3D test set contains over 2000 pairs. Consequently, NCC was evaluated only on the US3D dataset after compressing the images to a resolution of 128 × 128 pixels. Additionally, all deep learning-based methods used for comparison were trained following the parameter settings described in [17].

Table 1 presents the quantitative comparison of different methods on the US3D test set. As a classic and widely used traditional method, SGM shows a noticeable performance gap in both EPE and D1 metrics compared to various deep learning-based approaches. Among the deep learning-based methods, our proposed IFMA-Stereo achieves improved accuracy over the others, reducing EPE and D1 by 0.067 and 0.56%, respectively, compared to the leading HMSM-Net. This gain comes with an increase in inference time from 497 ms to 843 ms.

Table 1. Quantitative comparison of different methods on US3D.

Method	EPE (Pixel)	D1 (%)	Time (ms)
NCC	4.041	29.13	13,688
SGM	3.201	35.14	-
DenseMapNet	2.030	17.37	162
BGA-Net	1.663	10.55	588
StereoNet	1.627	11.04	229
PSMNet	1.489	9.08	368
HMSM-Net	1.414	7.82	497
IFMA-Stereo (Ours)	1.347	7.26	843

4.2.2. Results on GaoFen-7

The GaoFen-7 dataset is a domestic remote sensing dataset that covers part of Yingde City, Guangdong Province, China. Similarly to the experiments conducted on the US3D dataset, we evaluate the performance of various classical and state-of-the-art methods and quantitatively compare them with our proposed approach. Notably, all methods were initialized using their pre-trained weights on the US3D dataset. Finally, the test results for different methods on the GaoFen-7 dataset are presented in Table 2. Compared to existing methods, our approach demonstrates improvements in both EPE and D1 metrics. Specifically, relative to HMSM-Net, our method achieves reductions of 0.177 in EPE and 1.33% in D1. These results indicate that the proposed method attains the best performance among the compared models.

Table 2. Quantitative comparison of different methods on GaoFen-7.

Method	EPE (Pixel)	D1 (%)
SGM	5.189	51.20
DenseMapNet	3.066	35.01
StereoNet	2.091	21.07
PSMNet	1.971	18.62
HMSM-Net	1.762	14.74
IFMA-Stereo (Ours)	1.585	13.41

4.2.3. Results with Lower Simulation Resolution

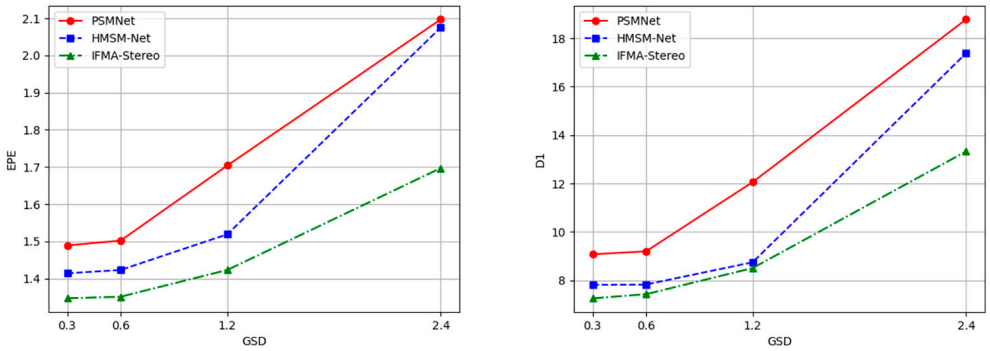
To further investigate the sensitivity of our proposed IFMA-Stereo to variations in ground sampling distance (GSD) and to provide a preliminary assessment of its robustness under reduced spatial resolution, we conducted a set of controlled simulation experiments. It is worth noting that publicly available satellite stereo matching datasets specifically de-

signed for low-resolution scenarios are extremely scarce—to the best of our knowledge, no publicly dedicated low-resolution satellite stereo benchmark currently exists. Consequently, directly evaluating models on authentic low-resolution satellite imagery is infeasible at this stage. To circumvent this limitation while still gaining insight into the model's behavior under resolution degradation, we adopted an engineering-oriented simplification.

Specifically, we artificially degraded the high-resolution US3D images by downsampling them to several lower resolutions—namely, we sequentially compressed the original image dimensions to 1/2, 1/4, and 1/8 of their original size. After downsampling, the images were upsampled back to the original resolution using nearest interpolation to ensure compatibility with the existing network architecture. Since the GSD of the US3D data is approximately 0.3 m, the simulated GSDs of our four test sets at different resolutions are 0.3 m, 0.6 m, 1.2 m, and 2.4 m, respectively.

It is important to acknowledge that this simple downsampling–upsampling operation does not fully replicate the physical characteristics of genuine low-resolution satellite imagery (e.g., the specific Modulation Transfer Function (MTF) of sensors, atmospheric scattering effects, or quantization noise). This experiment serves primarily as an engineering-oriented simplification to simulate a reduction in GSD and to evaluate the relative performance degradation of different models under a controlled scale-variation scenario, rather than to mimic authentic low-resolution imaging conditions. The goal is to analyze the trend of accuracy loss rather than absolute performance on native low-resolution data.

Figure 8 presents the quantitative results of different methods under simulated low-resolution conditions. As expected, a clear trend can be observed: the performance of all models consistently deteriorates as the resolution decreases. This degradation is primarily attributed to the irreversible loss of fine-grained textural and edge information during downsampling, which directly affects the reliability of pixel-wise matching in the cost volume. Furthermore, it can be observed that when the GSD is within the sub-meter range, the test results are relatively less affected by the reduction in GSD, whereas a sharp performance drop may occur once the GSD exceeds the meter level. Nevertheless, across all tested resolution levels, our IFMA-Stereo model consistently outperforms all baseline methods and exhibits the smallest performance decline beyond the meter-level GSD, demonstrating superior resilience to resolution degradation. These results confirm that, although stereo matching models are generally sensitive to extreme reductions in GSD, our method remains more competitive and robust than existing approaches under simulated low-resolution conditions.



(**a**) EPE (**b**) D1

Figure 8. Comparison of matching results under different simulated GSDs.

4.2.4. Height-Based Evaluation

To further bridge the gap between disparity estimation and practical photogrammetric applications, we conducted a simple height-based evaluation. Specifically, for each stereo

pair in the US3D validation set, we established a mapping between the ground-truth disparity values and the AGL (Above Ground Level) reference values to convert the predicted disparity maps generated by each method into estimated height values. The resulting height estimates were compared against the AGL ground truth to compute two complementary accuracy metrics: Root-Mean-Square Error (RMSE) and Mean Absolute Error (MAE). RMSE is sensitive to large outliers and reflects the overall error magnitude, while MAE provides an intuitive measure of the average height deviation. The elevation analysis is intended to provide additional insight into the practical applicability of our method for DSM generation.

Table 3 reports the height estimation accuracy of different methods on the US3D validation set. Among all compared methods, IFMA-Stereo achieves the lowest RMSE (2.838 m) and MAE (1.719 m), outperforming both PSMNet (RMSE: 3.355 m, MAE: 1.979 m) and HMSM-Net (RMSE: 2.995 m, MAE: 1.761 m). The proposed method reduces RMSE by 5.2% compared to HMSM-Net and by 15.4% compared to PSMNet, while MAE is reduced by 2.4% and 13.1%, respectively. The reduction in RMSE is particularly noteworthy, indicating that IFMA-Stereo is more robust in handling large height estimation errors that typically occur in challenging regions such as occlusion boundaries and sudden terrain changes.

Table 3. Height-based accuracy comparison on the US3D validation set.

Model	RMSE (m)	MAE (m)
PSMNet	3.355	1.979
HMSM-Net	2.995	1.761
IFMA-Stereo (Ours)	2.838	1.719

Overall, the above height-based preliminary evaluation corroborates the effectiveness of IFMA-Stereo from a photogrammetric perspective, confirming that the improvements in disparity estimation translate directly into more reliable DSM/height estimates. These results strengthen the relevance of our contribution to the remote sensing and photogrammetry community.

4.3. Complex Regions

Satellite images contain extensive texture-less regions, disparity discontinuity regions, and repetitive-pattern regions. Additionally, remote sensing image pairs often suffer from scene variation issues due to inherent spatio-temporal heterogeneity during acquisition. These challenges collectively constitute the primary difficulties in disparity prediction for remote sensing, frequently leading to ambiguous disparity estimates in such areas and significantly impacting model performance. To thoroughly investigate the impact of various complex conditions, this section compares the experimental results of previous methods with those of our proposed method across different scenarios. For each pair of original images, as direct disparity estimation results make it difficult to visually assess the disparity estimation error at each pixel, we primarily present microscopic error visualization maps to analyze local details. On this basis, we provide macroscopic percentage of error distribution to analyze the distribution. Additionally, we present the number of resulting matches (with an error within 3 pixels) for local challenging regions. Finally, we present evaluation metrics in tabular form for global analysis. In the microscopic error visualization maps, pixels in blue indicate an error of generally less than 3 (errors within 3 are typically considered accurate prediction regions), with darker blue representing more accurate predictions. Green and yellow pixels generally correspond to errors exceeding 3, indicating a certain degree of prediction deviation. Red pixels, with errors generally above 6, indicate regions where predictions contain substantial errors. In the percentage of error distribution, the

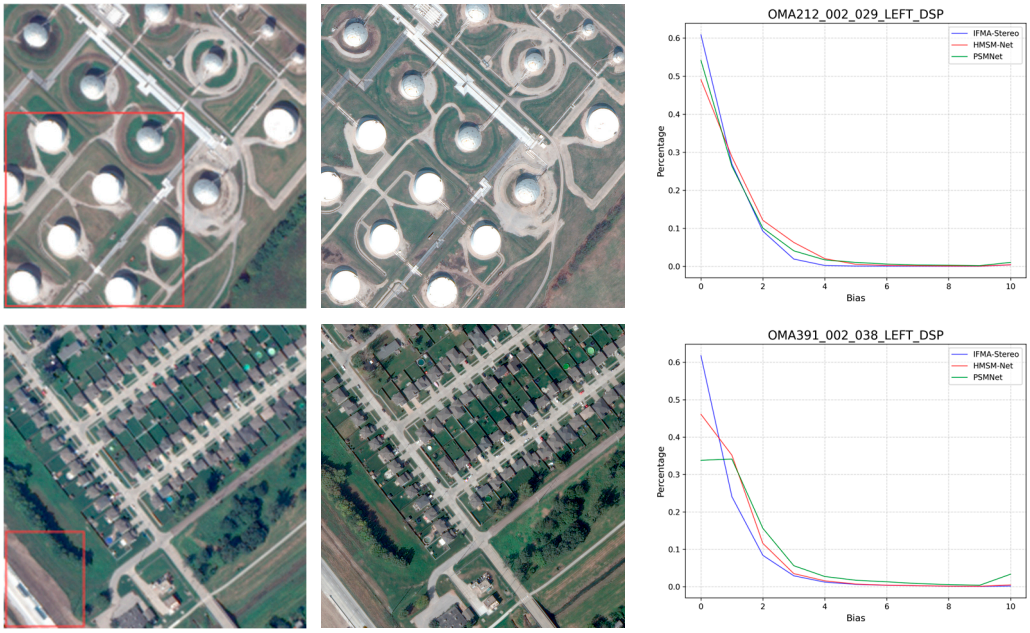
horizontal coordinates 0–9 correspond to ten intervals of length 1, ranging from [0, 1) to [9, 10), while coordinate 10 denotes the interval [10, ∞). The vertical axis represents the percentage of errors falling into each interval.

4.3.1. Texture-Less Regions

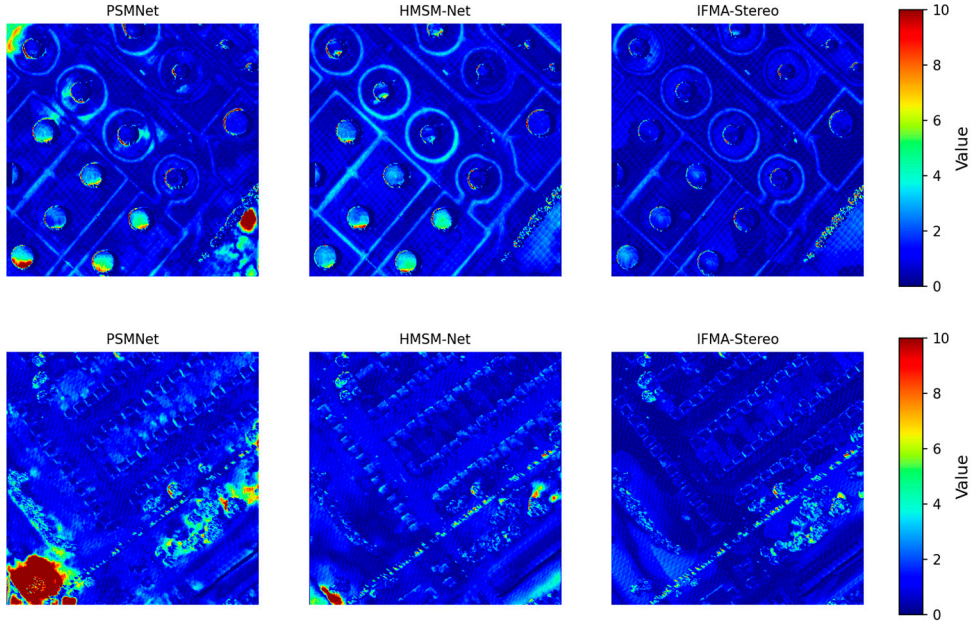
Texture-less regions, or weakly textured regions, refer to areas in an image with minimal brightness variation and a lack of distinctive feature points, primarily including surfaces such as roads, roofs, and water bodies. Due to the high similarity of pixel-level information within adjacent areas in these regions, establishing precise matching correspondences for pixels becomes challenging, often resulting in blurred or ambiguous disparity predictions.

Figure 9 presents the test results of different models in such regions, featuring two rows of images corresponding to two different scenes, which sequentially correspond to the two rows in Table 4. Regions of interest are highlighted with red boxes, and the error visualization maps of these locations are examined. In the first scene, the area where disparity prediction is notably more difficult is clearly identifiable as the white building rooftop within the red box, representing a typical smooth building rooftop commonly encountered in remote sensing imagery. Comparing the error visualization maps, the prediction error map of PSMNet contains a large number of green and yellow pixels, along with a small number of red pixels, indicating that the predictions over large areas of the building rooftop are inaccurate, with even a few regions exhibiting substantial errors. In contrast, the error map of HMSM-Net primarily contains some green and yellow pixels, performing slightly better than PSMNet. For IFMA-Stereo, the error map over the building rooftop is almost entirely composed of dark blue pixels (indicating highly accurate predictions) and light blue pixels (indicating relatively accurate predictions), demonstrating that the proposed method achieves higher accuracy in these difficult-to-predict rooftop areas.

In the second scene, the red box in the lower-left corner contains a white road and brown land, representing the monochromatic road and land surfaces commonly seen in remote sensing imagery. The error map of PSMNet contains a large number of red pixels, indicating very poor predictions in this region. The error map of HMSM-Net contains a small number of red pixels along with some yellow-green pixels, suggesting that predictions remain relatively poor in some areas. For IFMA-Stereo, the error map over the monochromatic road and land surfaces is almost entirely composed of blue pixels, representing accurate predictions in these regions. Furthermore, comparing the percentage of error distribution in Figure 9, IFMA-Stereo achieves the highest percentage in the [0, 1) interval and the lowest percentage at higher coordinate values in both scenes, further validating the accuracy of IFMA-Stereo. Additionally, the quantitative analysis in Table 4 demonstrates that IFMA-Stereo achieves the best EPE and D1 metrics in both scenes compared to PSMNet and HMSM-Net. Notably, the D1 metric shows a substantial improvement over both PSMNet and HMSM-Net in the two scenarios, reflecting the strong matching capability of our model in texture-less regions. Furthermore, by observing the number of resulting matches in the local complex regions in Figure 10, the improvement in matching performance achieved by our method can be observed very directly. The above comparisons indicate that our approach exhibits strong adaptability in texture-less regions where precise matching is difficult, primarily due to its enhanced capability in modeling stereo scenes. By mitigating severe errors caused by incorrect matches and leveraging contextual scene relationships, our method constructs more accurate geometric information for texture-less areas.



(**a**) Left image (**b**) Right image (**c**) Percentage of error distribution

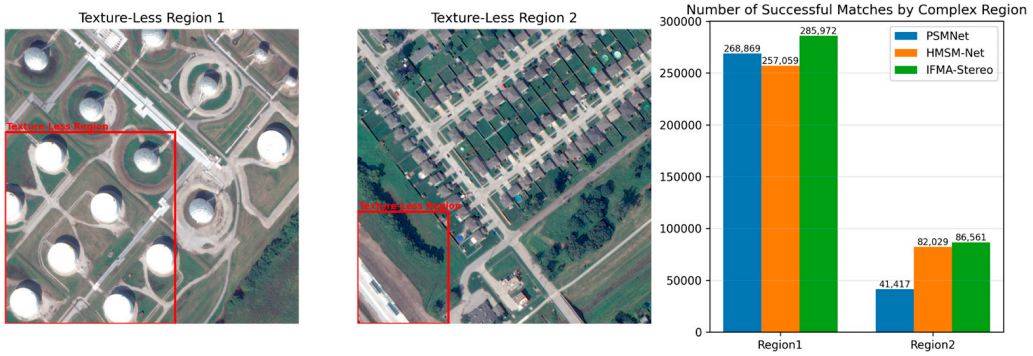


(**d**) Error visualization maps

Figure 9. Visualization of prediction results in texture-less regions.

Table 4. Evaluation metrics for prediction in texture-less regions.

Image		PSMNet	HMSM-Net	IFMA-Stereo (Ours)
OMA212_002_029	EPE (pixel)	1.4335	1.4066	1.0677
	D1 (%)	9.434	9.988	4.401
OMA391_002_038	EPE (pixel)	2.2668	1.3863	1.0043
	D1 (%)	16.47	7.164	4.892



(**a**) Region 1 (**b**) Region 2 (**c**) Number of matches

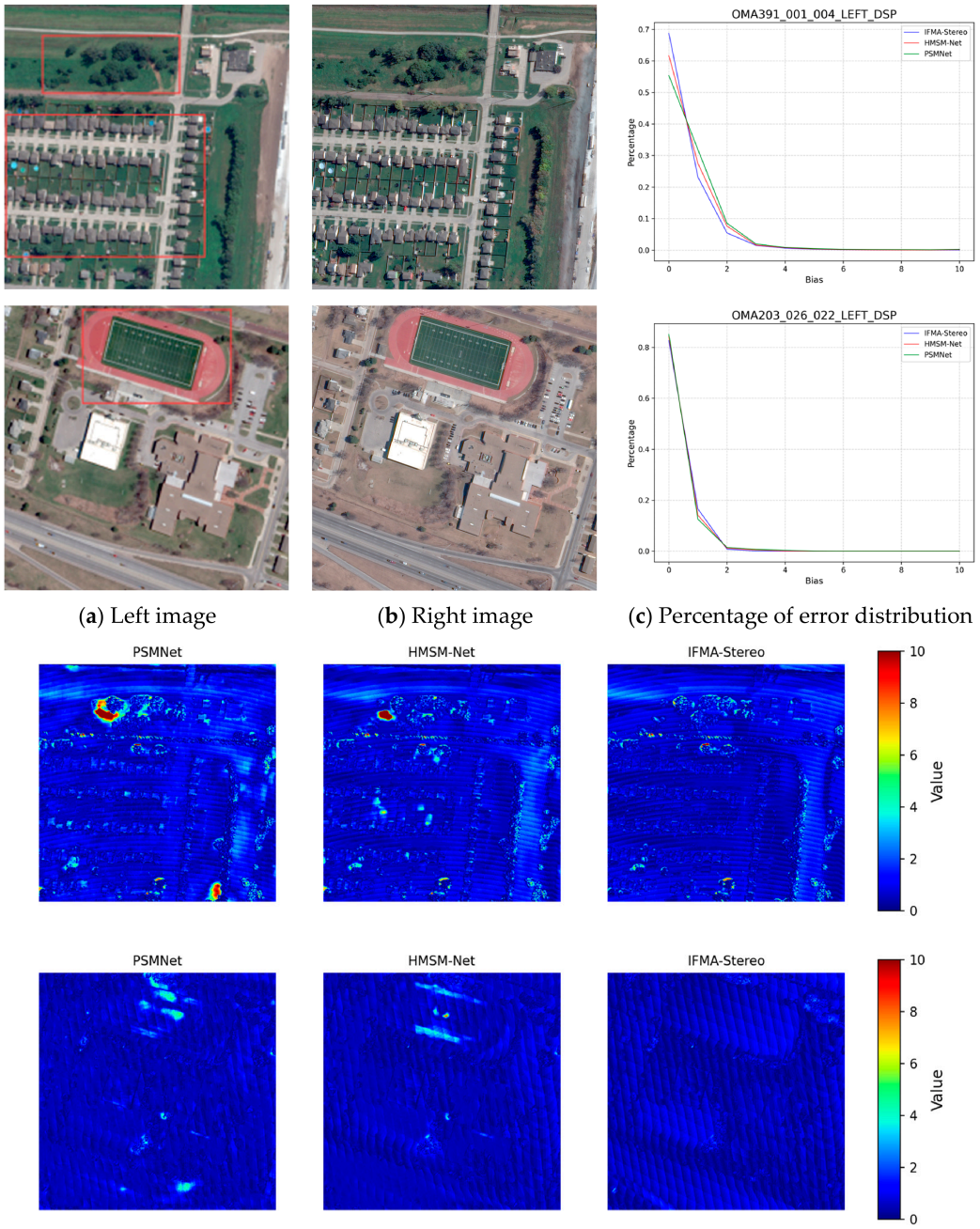
Figure 10. Comparison of number of resulting matches for local texture-less regions.

4.3.2. Repetitive-Pattern Regions

These refer to areas in an image containing large-scale repeated structural information, such as extensive tree canopies, dense building complexes, and repetitive patterns. While pixel-level distinctions may exist to some extent in these regions, the extensive structural repetition makes it difficult to establish accurate correspondences of structural information between stereo image pairs, leading to large-scale matching errors.

Figure 11 presents the test results of different models, using two distinct representative scenes for illustrative analysis. In the first scene, two regions of interest are highlighted with red boxes: the dense tree canopies in the upper part of the image and the dense building complexes in the mid-lower part. These represent typical forest and building complex scenes commonly encountered in remote sensing imagery. Comparing the error visualization maps, both PSMNet and HMSM-Net exhibit patchy clusters of dark red pixels in the tree canopy areas, indicating substantial prediction errors in these regions. In contrast, the error map of IFMA-Stereo is predominantly composed of dark blue and light blue pixels, demonstrating accurate predictions in the tree canopy areas. Furthermore, examining the building complexes below, although the two baseline methods achieve relatively accurate predictions, they still contain some noticeable light blue pixels. IFMA-Stereo, however, exhibits almost entirely dark blue pixels in the building complex, indicating its ability to achieve high-precision predictions in this region.

In the second scene, the red box encloses a stadium area, which contains numerous white ground marking lines on the central green field and multiple parallel white lane lines on the surrounding red track. This represents a repetitive-pattern scenario commonly found in remote sensing imagery. Comparing the error visualization maps, both PSMNet and HMSM-Net exhibit some light blue and greenish pixels on the central field and surrounding track. In contrast, the error map of IFMA-Stereo consists almost entirely of dark blue pixels, indicating prediction errors close to zero and demonstrating the remarkable predictive accuracy of IFMA-Stereo in repetitive-pattern regions. Furthermore, comparing the percentage of error distribution in Figure 11, IFMA-Stereo achieves optimal performance across all intervals in the first scene. In the second scene, IFMA-Stereo exhibits a higher proportion of errors within the $\leq$2-pixel range (i.e., the highest proportion of high-precision predictions represented by dark blue pixels) and a lower proportion of errors exceeding 2 pixels. Additionally, the quantitative analysis in Table 5 shows that, compared with PSMNet and HMSM-Net, IFMA-Stereo achieves a modest performance improvement in the first scene. For the second scene, our method achieves an extremely low D1 metric, demonstrating its high-precision prediction capability in repetitive-pattern regions under that scenario. Furthermore, Figure 12 compares the matching points in local repetitive-pattern regions, and our method yields the highest number of matches.



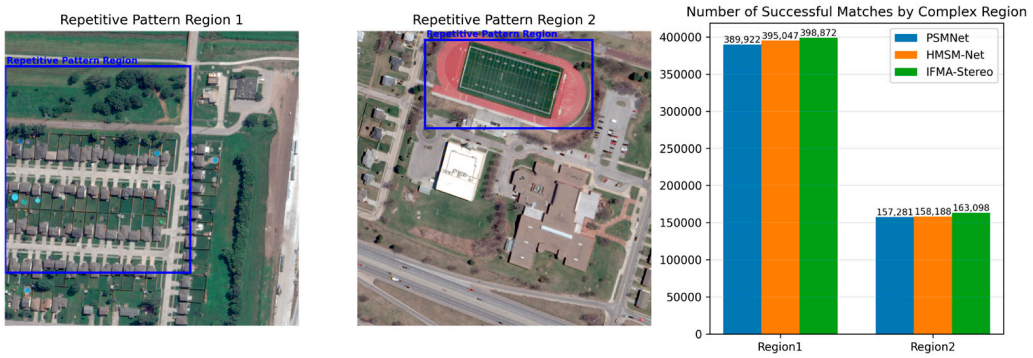
(**a**) Left image (**b**) Right image (**c**) Percentage of error distribution

(**d**) Error visualization maps

Figure 11. Visualization of prediction results in repetitive-pattern regions.

Table 5. Evaluation metrics for prediction in repetitive-pattern regions.

Image		PSMNet	HMSM-Net	IFMA-Stereo (Ours)
OMA391_001_004	EPE (pixel)	1.1613	1.0354	0.8842
	D1 (%)	4.283	3.241	2.559
OMA203_026_022	EPE (pixel)	0.6042	0.5728	0.5246
	D1 (%)	1.012	0.583	0.027



(**a**) Region 1 (**b**) Region 2 (**c**) Number of matches

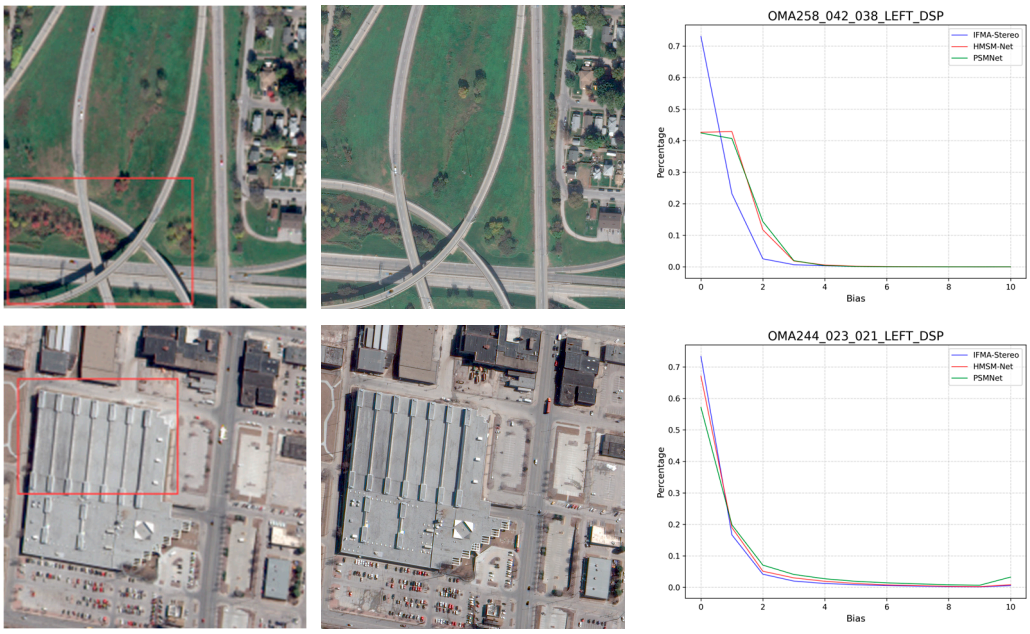
Figure 12. Comparison of number of resulting matches for local repetitive-pattern regions.

4.3.3. Disparity Discontinuity Regions

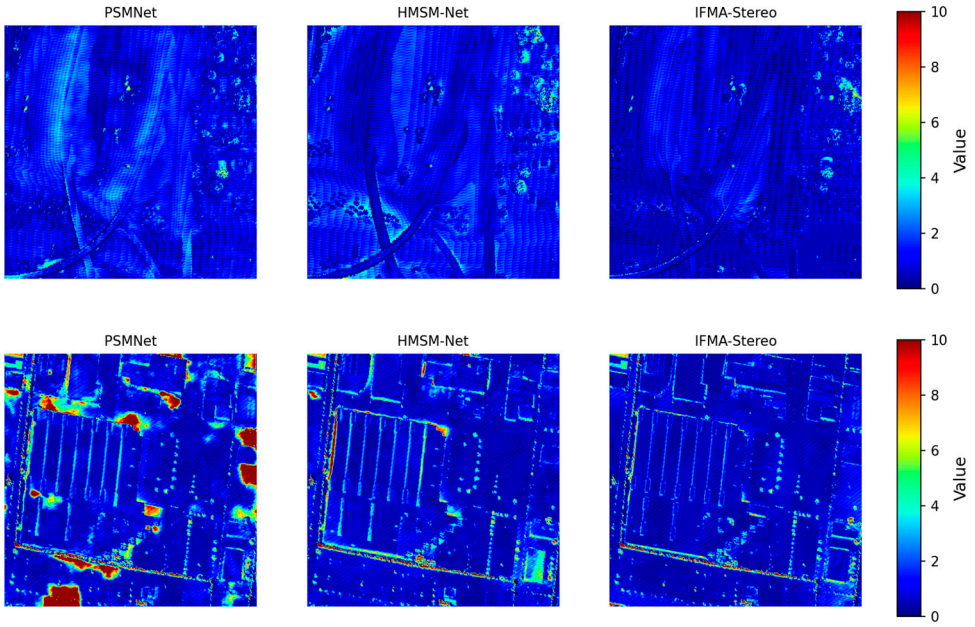
These refer to areas where the local disparity undergoes abrupt changes. In such regions, significant height variations lead to differences in the visible areas between the left and right stereo images, often manifesting as occlusion in one of the images. The discrepancy in image content between the two views can easily induce local matching errors, thereby affecting the prediction accuracy of the model.

In Figure 13, the red boxes in the two scenes highlight an overpass area and a high-rise building area, representing two common types of remote sensing scenes characterized by significant local height variations in urban environments. Comparing the error visualization maps, both PSMNet and HMSM-Net achieve accurate predictions on the overpass structure itself (as indicated by dark blue pixels on the bridge deck). However, once the positions of height variation at the edges of the overpass are involved, their prediction accuracy is notably affected, transitioning to large areas of light blue or even greenish pixels. In contrast, the prediction results of IFMA-Stereo over the overpass area consist almost entirely of dark blue pixels, demonstrating its robust adaptability to disparity jumps at overpass edges.

In the second scene, the high-rise building within the red box exhibits large height jumps at its edges, as well as smaller height variations at multiple raised strip structures on the building rooftop. The error maps of PSMNet and HMSM-Net contain conspicuous large patches of green and red pixels at the building edges, indicating significant prediction errors in these regions of drastic height variation. In comparison, IFMA-Stereo achieves substantially better results, exhibiting only a very small number of error pixels of other colors at the edges of the high-rise building, demonstrating its effectiveness in mitigating the impact of severe height changes on disparity prediction. Furthermore, examining the multiple raised strip structures on the building rooftop, which involve smaller height variations, the error maps of the two baseline methods exhibit distinct continuous cyan-green stripes, whereas IFMA-Stereo presents only faint light blue stripes. This indicates that IFMA-Stereo achieves higher-precision disparity prediction within a certain range of height variations. Additionally, comparing the percentage of error distribution in Figure 13 and the macroscopic quantitative analysis in Table 6, IFMA-Stereo achieves the best performance in both scenarios. Notably, in the first scenario, PSMNet achieves an EPE of 1.2476 and a D1 of 2.569%, while HMSM-Net achieves an EPE of 1.2399 and a D1 of 2.760%. In contrast, our method achieves an EPE of 0.7106 and a D1 of 1.032%, demonstrating substantial improvements in both metrics. Furthermore, Figure 14 clearly demonstrates that our method achieves the highest number of matching points. These results demonstrate that our method exhibits higher accuracy and robustness in predicting disparity in regions characterized by significant disparity variations.



(**a**) Left image (**b**) Right image (**c**) Percentage of error distribution

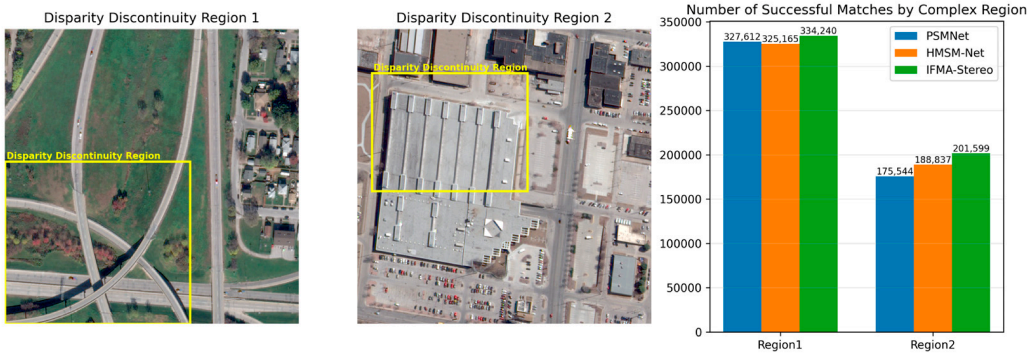


(**d**) Error visualization maps

Figure 13. Visualization of prediction results in disparity discontinuity regions.

Table 6. Evaluation metrics for prediction in disparity discontinuity regions.

Image		PSMNet	HMSM-Net	IFMA-Stereo (Ours)
OMA258_042_038	EPE (pixel)	1.2476	1.2399	0.7106
	D1 (%)	2.569	2.760	1.032
OMA244_023_021	EPE (pixel)	1.8637	1.1801	1.0011
	D1 (%)	15.92	8.807	5.504



(**a**) Region 1 (**b**) Region 2 (**c**) Number of matches

Figure 14. Comparison of number of resulting matches for local disparity discontinuity regions.

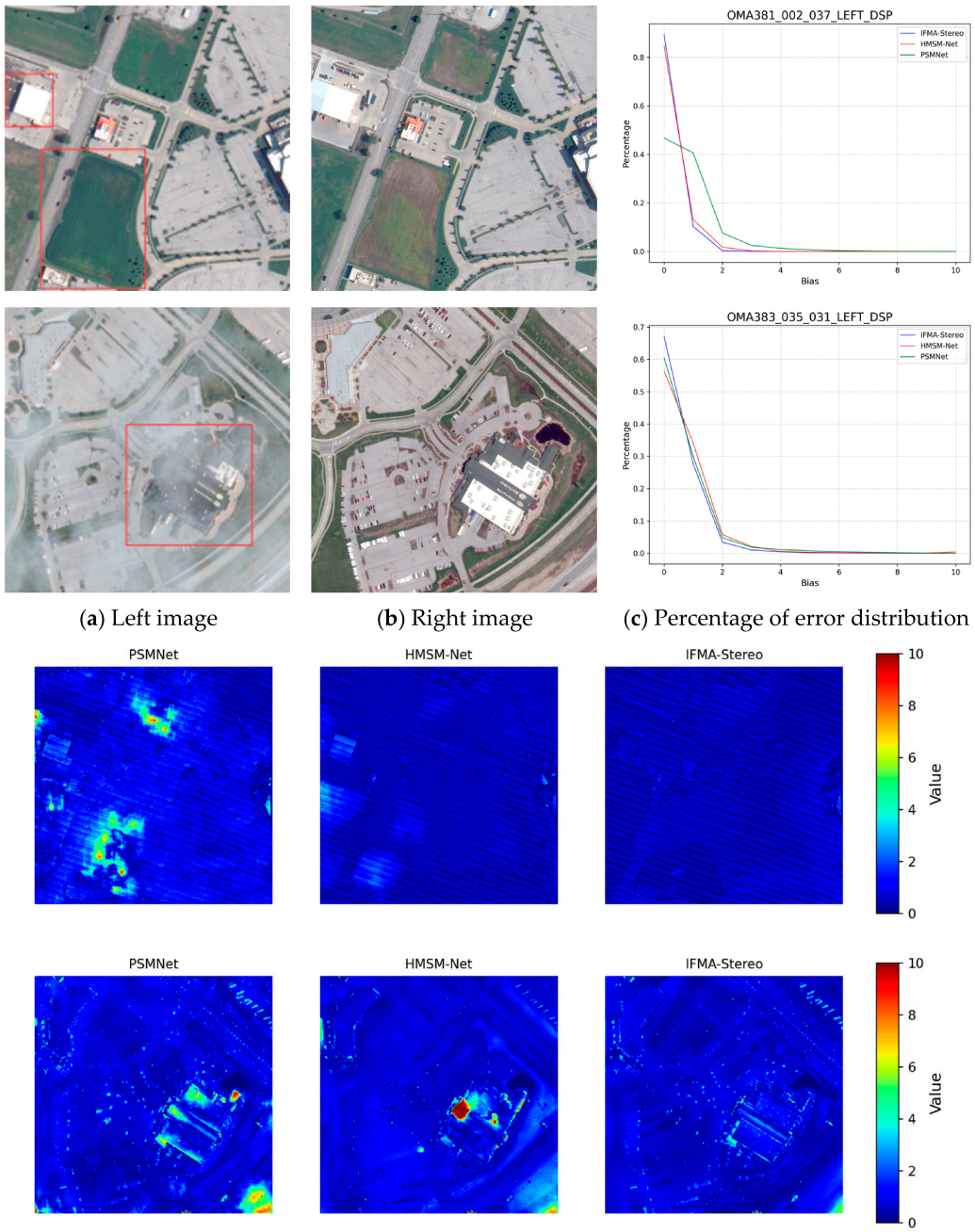
4.3.4. Spatio-Temporal Heterogeneity Regions

These refer to inconsistencies in morphology between remote sensing image pairs caused by actual differences in time and space. On one hand, the acquisition times of stereo satellite images are often not simultaneous but separated by a certain interval—ranging from momentary differences to seasonal variations—leading to changes in the physical scene. On the other hand, significant differences in viewing angles between images may cause variations in ground sample distance (GSD), while atmospheric conditions and other environmental factors can also influence imaging. These factors collectively cause physical discrepancies between the left and right images, posing greater challenges for disparity prediction in satellite imagery.

In Figure 15, the red boxes in the two scenes highlight areas where actual changes have occurred due to the passage of time and areas affected by weather conditions, respectively, representing two common types of remote sensing scenarios: vegetation or building changes and weather-induced factors such as fog or cloud occlusion. In the first scene, evident signs of temporal change can be observed between the image pair. The vegetation within the lower red box has transformed from lush to withered, partially exposing brown land surfaces, while the building in the red box at the left edge of the image has been expanded in the right image, resulting in substantial changes to the real physical scene. Comparing the error visualization maps in Figure 15, PSMNet and HMSM-Net exhibit clear degradation in the aforementioned vegetation change area and the expanded building area, whereas the results of IFMA-Stereo remain almost unaffected (consisting almost entirely of dark blue pixels).

In the second scene, weather conditions cause the entire left image to appear hazy, with the building within the red box being severely affected. The error maps of PSMNet and HMSM-Net contain large areas of yellow-green or even red pixels, indicating substantial prediction errors. In contrast, although weather conditions also have a certain impact on the results of IFMA-Stereo, the main body of the error map still consists of blue pixels representing extremely low errors, with only a small number of greenish pixels indicating minor errors, demonstrating its significant resilience to the effects of weather factors. Meanwhile, Figure 16 demonstrates that our method achieves the optimal number of matching points in regions exhibiting local spatio-temporal heterogeneity. Furthermore, comparing the percentage of error distribution in Figure 15 and the macroscopic quantitative analysis in Table 7, it can be observed that IFMA-Stereo achieves a certain degree of performance improvement over the other methods in both scenarios. Notably, the exceptionally low D1 metric in the first scenario indicates its strong adaptability to changes in conventional physical scenes (with almost no error pixels), further confirming the superiority of IFMA-

Stereo. Its ability to learn scene geometric information enables it to focus more on global consistency, thereby better adapting to the effects of spatio-temporal heterogeneity.



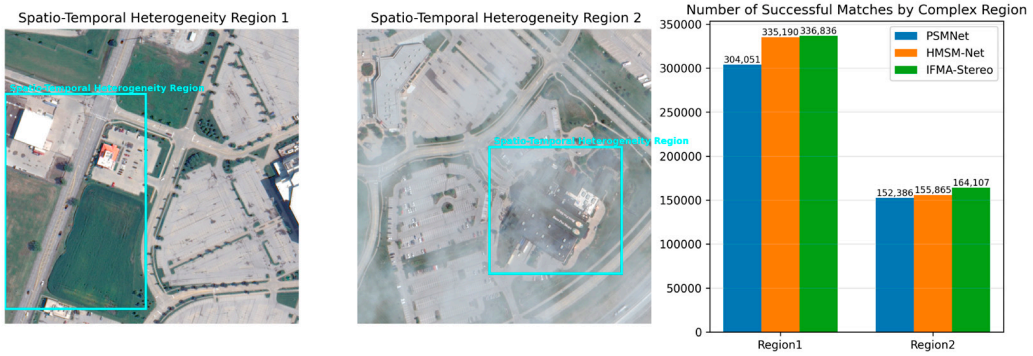
(**a**) Left image (**b**) Right image (**c**) Percentage of error distribution

(**d**) Error visualization maps

Figure 15. Visualization of prediction results in spatio-temporal heterogeneity regions.

Table 7. Evaluation metrics for prediction in spatio-temporal heterogeneity regions.

Image		PSMNet	HMSM-Net	IFMA-Stereo (Ours)
OMA381_002_037	EPE (pixel)	1.2169	0.6055	0.5450
	D1 (%)	4.997	0.206	0.027
OMA383_035_031	EPE (pixel)	1.0564	1.0786	0.8892
	D1 (%)	4.713	3.761	2.093



(**a**) Region 1 (**b**) Region 2 (**c**) Number of matches

Figure 16. Comparison of number of resulting matches for local spatio-temporal heterogeneity regions.

5. Discussion

5.1. Ablation Studies

In this section, to validate the performance of each component proposed in this paper, we conduct ablation studies on the US3D dataset. Since our method is built upon the iterative architecture of IGEV, we adopt the basic IGEV iterative architecture—with only the necessary adaptations for remote sensing image matching—as the baseline model for this section (i.e., Model 1 in Table 8). Corresponding to the constituent modules of our approach, we systematically ablate the following four components: foundation feature extraction based on the depth estimation foundation model, generation of the cost volume via alternating self-attention and cross-attention mechanisms, the monocular–stereo joint iterative module integrating monocular and binocular complementary information, and the multi-factor disparity refinement module. The results are summarized in Table 8.

Table 8. Comparison of ablation study results for different modules ("√" indicates the module is enabled).

Model	Components				EPE (Pixel)	D1 (%)	Time (ms)
	Foundation Extraction	Self- and Cross-Attention	Iteration Complementary	Disparity Optimization			
1					1.442	8.06	523
2	√				1.412	7.97	714
3	√		√		1.379	7.88	796
4	√	√	√		1.355	7.33	823
5	√	√	√	√	1.347	7.26	843

First, our baseline model (referred to as Model 1) follows an architecture similar to IGEV-Stereo, employing a fundamental iterative framework without the assistance of any depth estimation foundation model or enhancement modules. The only adaptation to the specific characteristics of remote sensing data is the adoption of a bidirectional geometry encoding volume. Subsequently, we extend the baseline by incorporating feature-level assistance from the depth estimation foundation model. Building upon this, we integrate pixel-level assistance from the depth estimation foundation model along with a complementary iterative module. We then introduce self-attention and cross-attention mechanisms to enhance cross-view feature interaction, and finally add a disparity refinement module. By comparing the results from Model 1 to Model 5, it can be observed that:

Compared to the baseline, Model 2 (with feature-level interaction) reduces EPE by 2.08% (from 1.442 to 1.412) and D1 by 1.12% (from 8.06 to 7.97). Model 3 (with additional pixel-level interaction) further reduces EPE by 4.37% (from 1.442 to 1.379) and D1 by 2.23% (from 8.06 to 7.88). Model 4, leveraging optimized matching features and incorporating

attention-guided cost volume generation, brings more substantial gains: EPE decreases by 6.03% (from 1.442 to 1.355) and D1 by 9.06% (from 8.06 to 7.33). Finally, Model 5, which integrates all modules and includes the disparity refinement module, achieves the best results, reducing EPE by 6.59% (from 1.442 to 1.347) and D1 by 9.93% (from 8.06 to 7.26) compared to the baseline.

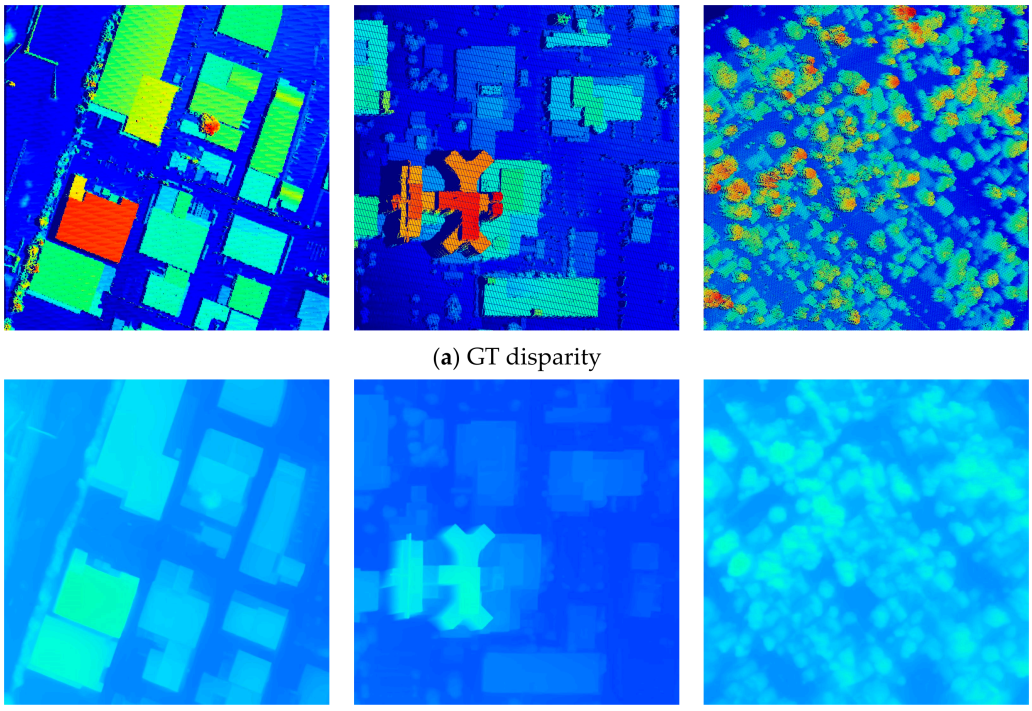
In terms of inference time, the average time cost of Model 1 (baseline) is 523 ms. Comparing the time costs after introducing the four modules, it can be clearly observed that the main increase in inference time is contributed by the foundation extraction module. This is because this module additionally introduces the encoder–decoder of the foundation model, which incurs substantial computational complexity. However, this does not imply that the foundation extraction module has the lowest cost-effectiveness, since the features extracted by the foundation model also serve subsequent modules, forming the cornerstone of the overall architecture. As for the disparity optimization module, although it improves the global EPE by only 0.008 and D1 by 0.07 percentage points, it does not introduce significant computational overhead; compared to Model 4, its time cost increases by only 2.4%. More importantly, the primary role of this module is not to merely reduce the global average error, but to perform a final-step refinement on local error regions through an adaptive error-correction mechanism—these error regions typically correspond exactly to the complex regions that are of central interest. Given its low computational cost and its targeted effectiveness on local error regions, we consider the inclusion of the disparity optimization module to be reasonably justified, unless extreme time constraints apply.

These results indicate that the proposed modules leveraging the depth estimation foundation model are crucial for achieving high precision. The relatively larger reductions in D1 compared to EPE suggest that our method focuses not only on refining regions with minor disparity deviations but, more importantly, on correcting severely erroneous predictions. This is attributed to the enhanced scene understanding capability introduced by our architecture. By effectively guiding the model to focus on the relative relationships of scene information from multiple dimensions, our approach strengthens the model's geometric scene modeling capacity rather than relying solely on cross-view matching. This effectively mitigates the negative impact of matching errors caused by complex regions and spatio-temporal heterogeneity.

5.2. The Influence of Bias in Foundation Model-Based Depth Prediction on Disparity Estimation

It is foreseeable that, on one hand, the disparity predicted by conventional iterative models may suffer from inaccuracies in challenging regions for stereo matching. On the other hand, monocular depth estimation may exhibit certain deviations in a small number of difficult local details within monocular scenes. These inaccuracies collectively affect Equation (5), resulting in the target flattened disparity obtained through linear alignment of monocular depth failing to strictly align with the true disparity. To investigate this issue, we conducted a comparative experiment on the US3D validation set. Specifically, we mapped the monocular depth maps from the foundation model to the disparity space using the linear transformation described in Equation (5), and compared the resulting target flattened disparity with the ground-truth disparity. As shown in Figure 17, consistent with the discussion in Section 3.4, the aligned disparity map achieves accurate discrimination only at the level of macro-scale structures, while its reliability for local details remains low—particularly for some locally extreme results (e.g., the disparities corresponding to dark red and yellow pixels), which are not well captured by the depth information. Furthermore, we performed a quantitative analysis of the initial aligned monocular disparity, as presented in Table 9. It can be observed that the aligned monocular disparity exhibits

considerable quantitative deviations, which are also attributable to the low reliability of numerous complex local details.



(**a**) GT disparity

(**b**) Aligned Mono disparity

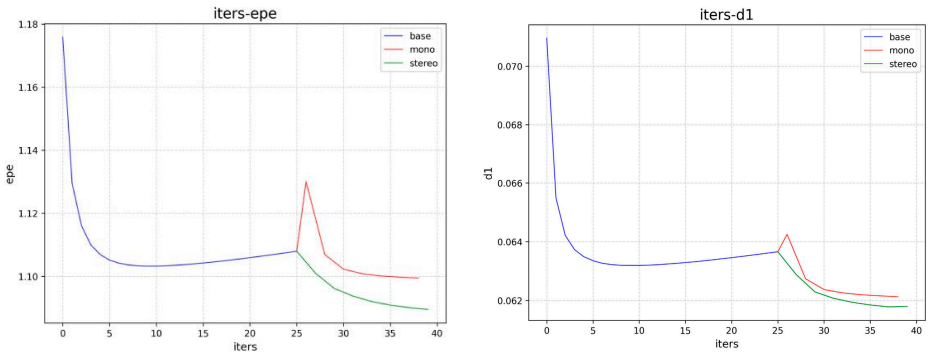
Figure 17. Visualization of the aligned monocular disparity versus the ground-truth disparity.

Table 9. Quantitative comparison between the initial aligned monocular disparity and the final results on the US3D validation set.

	EPE (Pixel)	D1 (%)
Aligned Mono Disparity	3.544	39.80
Final Disparity	1.060	5.86

However, the goal of aligning monocular depth with disparity is not an end in itself, but rather a means to assist in the iterative refinement of disparity estimates. The fundamental challenge faced by existing disparity iteration methods lies in the failure of binocular matching caused by factors such as occlusion or physical scene variations in binocular scenarios. In contrast, the core value of monocular depth lies in its ability to capture the relative relationships of scene geometry from a monocular perspective, independent of binocular constraints. The key is to incorporate this additional information—which is not limited by binocular correspondence—into the iterative process to assist in refining the disparity estimates. In other words, minor local deviations in monocular depth may cause subtle misalignments in the globally linear mapping, but this does not compromise the macro-level structural information that we intend to introduce—information that remains unaffected by binocular constraints. Such macro-structural information can help resolve the binocular matching failures caused by occlusion or physical scene changes. Furthermore, through the subsequent monocular–stereo joint iteration, the standard disparity from the iterative model is used to refine the overall alignment and difficult local details of the monocular disparity, which may suffer from slight misalignments. Conversely, the geometric structural information from the monocular disparity is used to refine regions where the standard disparity suffers from binocular matching failures.

We evaluated the standard iterative process and the monocular–stereo joint iteration including two disparity refinement strategies on the validation set of the US3D dataset [16], as shown in Figure 18. Base denotes the conventional iterative process, mono refers to the branch that leverages monocular disparity for optimization, and stereo represents the branch that utilizes standard disparity for optimization. The conventional iterative process was first performed for 25 iterations. It can be observed that the conventional iterative method rapidly approaches near-optimal performance after only a small number of iterations, while further iterations without any additional auxiliary strategies may even lead to a slight degradation in performance. At the 25-iteration mark, the curve representing monocular disparity (red) exhibits a substantial negative impact in the short term, which can be attributed to the previously mentioned misalignment introduced by linear mapping—a misalignment influenced by the amount of erroneous local details in the depth map. Furthermore, comparing the EPE and D1 metrics, we find that the EPE decreases much more significantly than the D1. EPE reflects the overall average error, whereas D1 reflects the percentage of pixels with a large error. The smaller reduction in D1 combined with the larger reduction in EPE indicates that a few pixels have a very large error. This observation is consistent with the low reliability of local detail recognition shown in Figure 17, confirming that the validation data contain a relatively large number of such erroneous local details.



(**a**) EPE (**b**) D1

Figure 18. Evolution of intermediate evaluation metrics throughout the overall iterative process on the validation set.

Furthermore, to further investigate whether the aforementioned metric drop phenomenon is generally present across different scenes or merely a reflection of average performance, we recorded the average metrics at the 25th iteration (the final result of the standard iterative process), the 26th iteration (the first iteration of the mono branch), and the 27th iteration (the first iteration of the stereo branch). In addition, we computed the percentage of image pairs in which the latter two exhibit a metric drop compared to the former. The results are presented in Table 10. It can be observed that, in terms of average performance, the mono branch is considerably affected by the misalignment in depth estimation, resulting in an increase in both EPE and D1. However, this does not have a strong negative impact on the stereo branch; on the contrary, it promotes an improvement in the average metrics of the stereo branch. Moreover, examining the drop percentage, we find that the metric drop in the mono branch does not occur for every image pair. Specifically, 69.57% of the image pairs exhibit an EPE drop, and 55.28% exhibit a D1 drop. Furthermore, the degradation is predominantly concentrated in regions with complex local details.

Table 10. Performance comparison at the initial stage of the monocular–stereo joint iteration.

	Standard (25)	**Mono (26)**	**Stereo (27)**
EPE (pixel)	1.108	1.130	1.101
D1 (%)	6.37	6.43	6.29
EPE Drop Percentage	\	69.57%	12.80%
D1 Drop Percentage	\	55.28%	16.69%

Nevertheless, even with this pronounced negative impact from the monocular disparity, the subsequent monocular–stereo joint iteration rapidly restored the metrics by the 28th iteration (after just two monocular–stereo joint iterations). Thereafter, both the monocular and standard disparities quickly surpassed the previously achieved optimal results of the conventional iterative process. This confirms that the core limitation of disparity iteration methods is indeed the failure of binocular matching caused by factors such as occlusion or physical scene changes—not the alignment between depth and disparity per se. Although the aligned depth may exhibit minor deviations due to erroneous local details in the data, which could cause the performance of the monocular disparity optimization branch to drop significantly in the short term, the macro-structural information that is independent of binocular vision can still effectively improve the standard disparity. In this way, the limitations caused by binocular matching failures in remote sensing scenes—such as occlusions or physical scene changes—can be overcome. Ultimately, the feature differences between the left and right images, computed after warping based on the disparity derived from monocular–stereo joint iteration, guide the positive refinement of both types of disparity.

6. Conclusions

This paper presents a novel disparity estimation method for remote sensing that leverages a depth foundation model and incorporates non-local geometric and contextual information via a cyclic iterative framework. Furthermore, the proposed multi-scale fusion enables learning correspondences across levels, improving adaptability in challenging regions. To improve prediction accuracy for ill-posed scenes, we introduce a joint reconstruction approach that integrates the depth estimation foundation model at both the feature and image levels, utilizing its generalized representations. Within the iterative framework, non-local geometric and contextual cues are encoded into recurrent features to progressively refine disparities, advancing stereo matching accuracy. Experiments on US3D and GaoFen-7 datasets show significant gains in texture-less, disparity-discontinuous, repetitive-pattern, and spatio-temporally heterogeneous areas. Despite its superior performance, our approach exhibits certain limitations in inference time due to the use of high-dimensional information and the iterative refinement process. In the future, we plan to further investigate the trade-off between accuracy and efficiency across modules, aiming to develop remote sensing stereo matching methods that achieve both high precision and computational efficiency.

Author Contributions: Conceptualization, K.H.; methodology, K.H. and W.Z.; software, K.H.; validation, K.H.; formal analysis, W.Z.; investigation, K.H.; resources, W.Z.; data curation, K.H.; writing—original draft preparation, K.H.; writing—review and editing, K.H. and W.Z.; visualization, K.H.; supervision, W.Z.; project administration, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation of this manuscript, the authors used DeepSeek-V4 for the purposes of language polishing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

References

1. Zenati, N.; Zerhouni, N. Dense Stereo Matching with Application to Augmented Reality. In *Proceedings of the 2007 IEEE International Conference on Signal Processing and Communications*; IEEE: New York, NY, USA, 2007; pp. 1503–1506.
2. Dong, N.; Ren, X.; Sun, M.; Jiang, C.; Zheng, H. Fast Stereo Aerial Image Construction and Measurement for Emergency Rescue. In *Proceedings of the 2013 Fifth International Conference on Geo-Information Technologies for Natural Disaster Management*; IEEE: New York, NY, USA, 2013; pp. 119–123.
3. Kitayama, D.; Touma, Y.; Hagiwara, H.; Asami, K.; Komori, M. 3D Map Construction Based on Structure from Motion Using Stereo Vision. In *Proceedings of the 2015 International Conference on Informatics, Electronics & Vision (ICIEV)*; IEEE: New York, NY, USA, 2015; pp. 1–5.
4. Madake, J.; Patwardhan, S.; Shah, P.; Bhatlawande, S. Stereo Matching for Terrain Mapping Using Aerial Imagery. In *Proceedings of the 2024 4th Asian Conference on Innovation in Technology (ASIANCON)*; IEEE: New York, NY, USA, 2024; pp. 1–8.
5. 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. [CrossRef]
6. Hirschmüller, H. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *30*, 328–341. [CrossRef]
7. Lipson, L.; Teed, Z.; Deng, J. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. In *Proceedings of the 2021 International Conference on 3D Vision (3DV)*; IEEE: New York, NY, USA, 2021; pp. 218–227.
8. Xu, G.; Wang, X.; Ding, X.; Yang, X. Iterative Geometry Encoding Volume for Stereo Matching. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 21919–21928.
9. Wang, X.; Xu, G.; Jia, H.; Yang, X. Selective-Stereo: Adaptive Frequency Information Selection for Stereo Matching. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 19701–19710.
10. Wu, Z.; Zhu, H.; He, L.; Liu, Y.; Du, S. Towards Accurate Binocular Vision of Satellites: A Cascaded Multi-Scale Pyramid Network for Stereo Matching on Satellite Imagery. *Expert Syst. Appl.* **2024**, *253*, 124276. [CrossRef]
11. 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]
12. Mayer, N.; Ilg, E.; Häusser, 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 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2016; pp. 4040–4048.
13. Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Westling, P. High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. In *German Conference on Pattern Recognition*; Springer International Publishing: Cham, Switzerland, 2014.
14. Menze, M.; Geiger, A. Object Scene Flow for Autonomous Vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2015.
15. Geiger, A.; Lenz, P.; Urtasun, R. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*; IEEE: New York, NY, USA, 2012.
16. Bosch, M.; Foster, K.; Christie, G.; Wang, S.; Hager, G.D.; Brown, M. Semantic Stereo for Incidental Satellite Images. In *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2019; pp. 1524–1532.
17. 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.
18. 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]
19. Liao, P.; Zhang, X.; Chen, G.; Wang, T.; Li, X.; Yang, H.; Zhou, W.; He, C.; Wang, Q. S2Net: A Multitask Learning Network for Semantic Stereo of Satellite Image Pairs. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5601313. [CrossRef]
20. 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]

21. Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; Zhao, H. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 10371–10381.
22. Barron, J.T.; Poole, B. *The Fast Bilateral Solver*; Springer: Cham, Switzerland, 2016.
23. Teed, Z.; Deng, J. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *European Conference on Computer Vision*; Springer Science and Business Media Deutschland GmbH: Cham, Switzerland, 2020.
24. Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Bry, A. End-to-End Learning of Geometry and Context for Deep Stereo Regression. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2017.
25. Khamis, S.; Fanello, S.R.; Rhemann, C.; 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)*; IEEE: New York, NY, USA, 2018.
26. Chang, J.R.; Chen, Y.S. Pyramid Stereo Matching Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018.
27. Foroosh, H.; Zerubia, J.B.; Berthod, M. Extension of Phase Correlation to Subpixel Registration. *IEEE Trans. Image Process.* **2002**, *11*, 188–200. [CrossRef] [PubMed]
28. Lucas, B.D.; Kanade, T. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the Proceedings of the 7th International Joint Conference on Artificial Intelligence—Volume 2*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1981; pp. 674–679.
29. Delon, J.; Rougé, B. Small Baseline Stereovision. *J. Math. Imaging Vis.* **2007**, *28*, 209–223. [CrossRef]
30. Sabater, N.; Morel, J.M.; Almansa, A. Sub-Pixel Stereo Matching. In *Proceedings of the 2010 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: New York, NY, USA, 2010; pp. 3182–3185.
31. Chao, W.; Wang, X.; Wang, Y.; Wang, G.; Duan, F. Learning Sub-Pixel Disparity Distribution for Light Field Depth Estimation. *IEEE Trans. Comput. Imaging* **2023**, *9*, 1126–1138. [CrossRef]
32. Cho, K.; Merrieenboer, B.V.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Doha, Qatar, 2014.
33. Schps, T.; Schnberger, J.L.; Galliani, S.; Sattler, T.; Geiger, A. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017.
34. Tao, R.; Xiang, Y.; You, H. An Edge-Sense Bidirectional Pyramid Network for Stereo Matching of VHR Remote Sensing Images. *Remote Sens.* **2020**, *12*, 4025. [CrossRef]
35. He, S.; Zhou, R.; Li, S.; Jiang, S.; Jiang, W. Disparity Estimation of High-Resolution Remote Sensing Images with Dual-Scale Matching Network. *Remote Sens.* **2021**, *13*, 5050. [CrossRef]
36. 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]
37. Huang, T.; Pan, H.; Zhou, N.; Zou, S.; Zhou, S. A Mutual-Structure Weighted Sub-Pixel Multimodal Optical Remote Sensing Image Matching Method. *Remote Sens.* **2026**, *18*, 1137. [CrossRef]
38. Jeong, W.; Park, S.-Y. UGC-Net: Uncertainty-Guided Cost Volume Optimization with Contextual Features for Satellite Stereo Matching. *Remote Sens.* **2025**, *17*, 1772. [CrossRef]
39. Caron, M.; Touvron, H.; Misra, I.; Jegou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 9630–9640.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
41. Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R.C.; Schindler, K. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 9492–9502.
42. Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; Zhao, H. Depth Anything V2. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*; Neural Information Processing Systems Foundation, Inc.: Vancouver, BC, Canada, 10–15 December 2024; pp. 21875–21911.
43. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2019.
44. 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 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021.

45. Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperGlue: Learning Feature Matching with Graph Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2020.
46. 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*; IEEE: New York, NY, USA, 2024.

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.