Satellite Foundation Stereo Model for DSM Reconstruction via Multi-Scale Cascaded Adaptation

Liupeng Su[a], Han Hu[a,*], Yuhao Ye[a], Zeyuan Dai[b,c], Junfan Wang[a], Zhihao Jia[a], Qianrui Guo[d], Heyi Li[d] and Qing Zhu[a]

[a]Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu, 611756, Sichuan, China
[b]Department of Military Oceanography and Hydrography and Cartography, Dalian Naval Academy, Dalian, 116018, China
[c]Key Laboratory of Hydrographic Surveying and Mapping of PLA, Dalian Naval Academy, Dalian, 116018, China
[d]Institute of Remote Sensing Satellite, China Academy of Space Technology, Beijing, 100094, China

ARTICLE INFO

Keywords:
Satellite Stereo Matching
Vision Foundation Models
Cascade Architecture
Bilateral Grid
Digital Surface Model

ABSTRACT

Foundation stereo models that leverage Vision Foundation Models (VFMs) have shown impressive performances in radiometric variations, structural boundary preservation, and occlusion handling. These properties are highly desirable for satellite stereo matching and Digital Surface Model (DSM) reconstruction. However, directly transferring foundation stereo models to satellite imagery remains challenging because satellite stereo involves extremely wide disparity ranges spanning both positive and negative values, together with a substantial domain gap caused by differences in imaging geometry, viewing conditions, and scene distributions, which together prevent zero-shot generalisation. In this paper, we propose SatFS, a satellite foundation stereo framework that adapts the single-scale FoundationStereo model into a cascade coarse-to-fine architecture. The proposed framework decomposes the computationally demanding global disparity search into a sequence of tractable local matching stages and dynamically modulates the search range using pixel-wise uncertainty. Specifically, to preserve high-fidelity structural boundaries during cost volume upsampling between cascade stages, we introduce PBU, a VFM-guided cost-volume upsampling module using a bilateral grid, which injects VFM-derived monocular depth estimates and feature maps as geometric guidance. Furthermore, we propose a Lightweight Geometry-Aware Encoding Volume (LGEV) that replaces the pre-computed full correlation volume with on-the-fly feature sampling, reducing the memory complexity of correlation volume construction from $O(H \cdot W^2)$ to $O(H \cdot W \cdot R)$ with negligible computational overhead. This design also supports negative disparities, overcoming the positive-only disparity limitation of FoundationStereo. Extensive experiments on four satellite stereo benchmarks demonstrate that SatFS achieves strong performance across diverse evaluation settings. It obtains the lowest D1 error of 9.57% and EPE of 1.33 px on WHU-Stereo, and achieves relative improvements of 10.0% in D1 and 19.0% in EPE under zero-shot cross-domain evaluation on WHU-SSIDE. For multi-sensor DSM reconstruction, SatFS attains the lowest RMSE of 2.47 m and 3.33 m on GF-7 and WorldView scenes, outperforming the best competing methods by 7.5% and 22.9%, respectively. Code and trained models will be made available at https://vrlab.org.cn/~hanhu/projects/satfs/.

1. Introduction

Satellite stereo reconstruction is entering a new regime in which robust matching can be driven not only by local image similarity, but also by structural cues encoded in features extracted by 3D Vision Foundation Models (VFMs) (Yang et al., 2024). Recent foundation stereo models show that these features provide strong invariance to radiometric variations, occlusions, structural boundaries, and fine-grained details in dense reconstruction (Wen et al., 2025; Cheng et al., 2025; Jiang et al., 2025a). In this work, we demonstrate that, when properly adapted to satellite imagery, 3D VFM-based foundation stereo provides a decisive solution to several long-standing bottlenecks of satellite reconstruction, including off-nadir occlusion, tile-boundary inconsistency, fine-structure degradation, severe shadow interference, and multi-temporal appearance variation. As illustrated in Fig. 1, our method translates these VFM features into visibly more coherent and detailed satellite reconstructions, sharply reducing edge artifacts and substantially improving the continuity of large-scale digital surface model (DSM) in occluded areas.

Despite this promise, adapting foundation stereo models (Wen et al., 2025; Cheng et al., 2025; Jiang et al., 2025a) to satellite imagery still remains a non-trivial problem rather than a direct deployment. The primary barrier lies in the inherent mismatch between the ground-view stereo geometry assumed by existing foundation models and the satellite imaging geometry targeted in this work. As shown in Fig. 2(a), satellite stereo pairs exhibit wide bidirectional disparity ranges that span both positive and negative values and often extend to hundreds of pixels, whereas existing foundation stereo models are designed for bounded, non-negative disparities. This geometric mismatch breaks a core assumption of current foundation stereo architectures. For example, FoundationStereo (Wen et al., 2025) restricts the disparity search to non-negative ranges and relies on a dense pre-computed correlation volume with $O(H \cdot W^2)$ memory complexity, making direct application both geometrically invalid and computationally impractical for high-resolution satellite inference. Moreover, even on positive-disparity satellite pairs that nominally satisfy this search constraint,

*Corresponding author
yueyuebird@my.swjtu.edu.cn (L. Su); han.hu@swjtu.edu.cn (H. Hu)

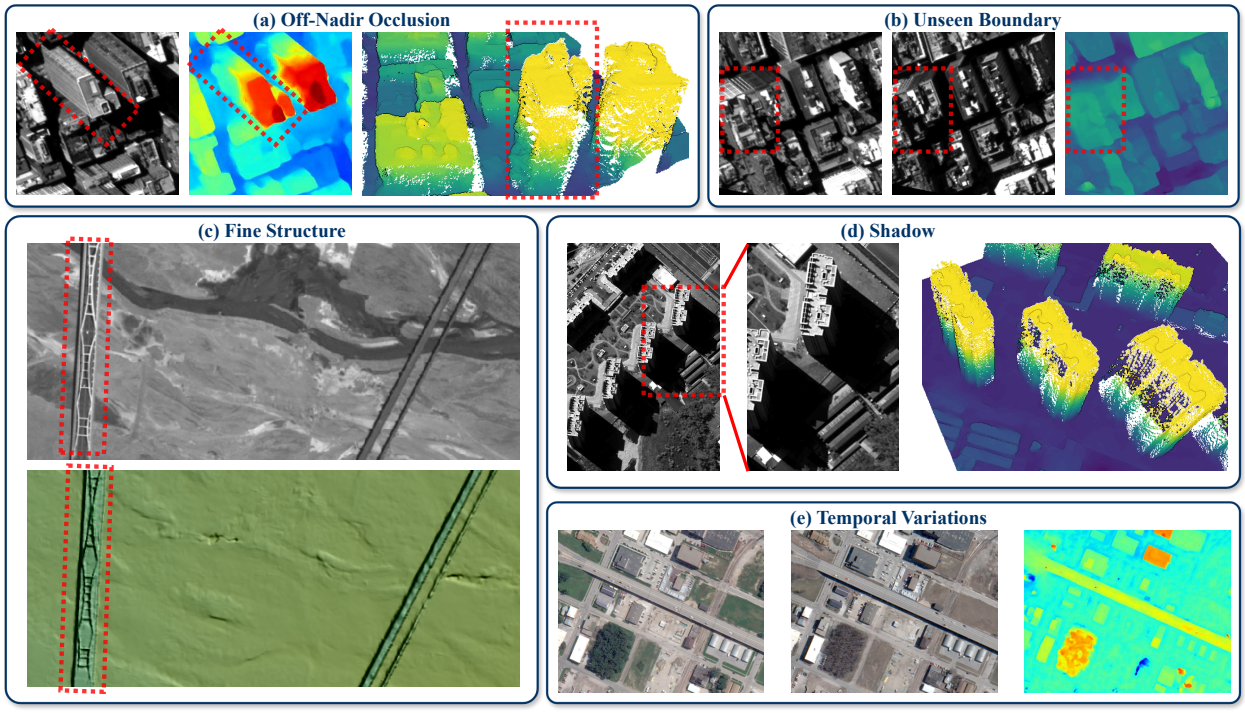

Figure 1: VFM-enhanced 3D satellite stereo reconstruction under challenging conditions, including off-nadir occlusions (a), unseen boundaries (b), fine structures (c), shadows (d), and temporal variations (e).

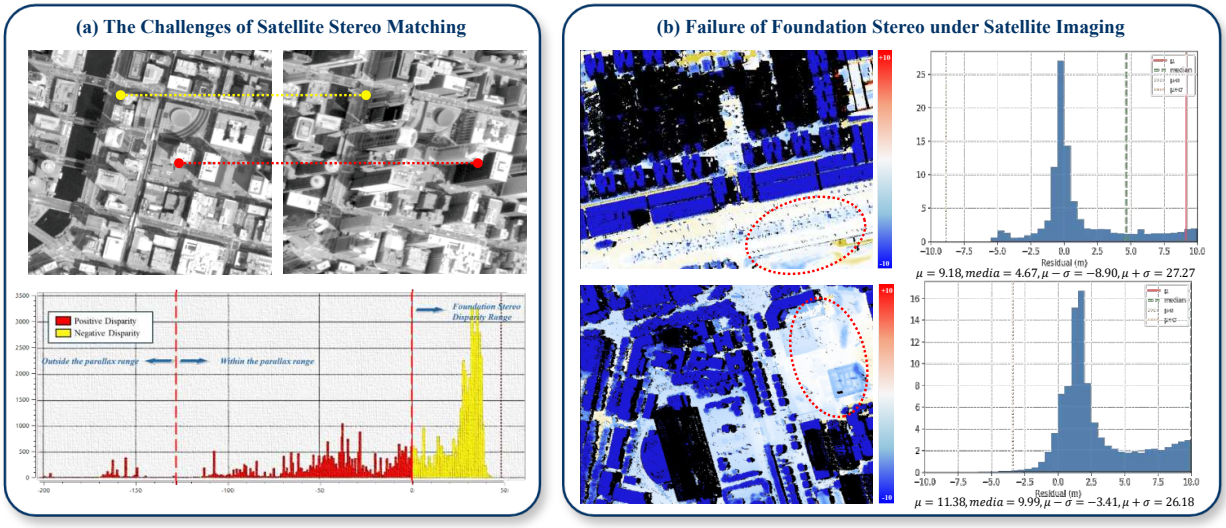

Figure 2: Wide disparity ranges in satellite stereo and limitations of FoundationStereo. (a) Satellite stereo pairs contain bidirectional disparities with both positive and negative values, beyond the bounded positive range assumed by ground-level stereo methods. (b) FoundationStereo still yields large off-ground errors and biased residuals on positive-disparity satellite pairs. Left: error map; right: residual histogram.

FoundationStereo still produces systematically biased error distributions, as shown in Fig. 2(b). This result indicates that the satellite domain gap is not merely a disparity-range issue, but also reflects differences in imaging geometry, radiometric conditions, and scene-structure distributions; consequently, direct zero-shot transfer becomes an invalid assumption.

On the other hand, existing satellite stereo methods remain constrained by the representation capability of the features. Traditional pipelines (Hirschmuller, 2005; Youssefi et al., 2020) rely on hand-crafted matching costs, making them vulnerable to radiometric variations. More recent deep stereo methods (Chang and Chen, 2018; Guo et al., 2019) and satellite-specific approaches (He et al., 2022; Rao et al., 2024; Zhang et al., 2026) learn local features and regularization regressors, but their representations are still shaped by limited scene diversity. Consequently, they lack the invariance needed

to generalize across regions, acquisition times, and imaging conditions. Recent analyses further show that the satellite domain gap arises not only from sensor differences, but also from discrepancies in scene semantics and structural distributions (He et al., 2023; Jiang et al., 2025b). Therefore, modifying cost volumes or network architectures alone cannot resolve the problem; robust satellite stereo requires stronger and more transferable feature representations.

To address these limitations, we propose SatFS, a cascade framework that adapts foundation stereo models (Wen et al., 2025; Cheng et al., 2025; Jiang et al., 2025a) to satellite imagery by replacing the original single-scale feature adaptation with multi-scale VFM side-tuning and enhanced lightweight stereo geometric reasoning. SatFS keeps the large VFM backbone frozen and learns only lightweight adapters, making the foundation features usable for satellite stereo without expensive full-model fine-tuning. These features are introduced at multiple image scales, so the model can first handle large satellite disparities at coarse resolution and then refine fine details at higher resolution.

Following the common coarse-to-fine strategy used in cascade MVS networks (Gu et al., 2020), the framework narrows the wide positive-and-negative disparity range step by step, instead of matching the full range at once. Multi-scale VFM guidance is further used during upsampling to keep object boundaries sharp and reduce the smoothing artifacts that commonly appear near edges. Finally, we replace the memory-heavy full correlation volume used by Iterative Geometry Encoding Volume (IGEV) (Xu et al., 2023) with a lightweight geometry-aware design, termed Lightweight Geometry-Aware Encoding Volume (LGEV), that samples only the needed disparity candidates. This reduces memory complexity from $O(H \cdot W^2)$ to $O(H \cdot W \cdot R)$, where R denotes the sampled disparity range, while naturally supporting both positive and negative disparities.

The main contributions of this work are summarized as follows:

- We propose SatFS, built upon FoundationStereo (Wen et al., 2025), as a unified cascade framework that brings frozen VFM features into each matching scale through lightweight side-tuning. Unlike existing methods that inject such features at a single scale, SatFS uses them throughout the matching process — from coarse-level wide-range ambiguity reduction to fine-level geometric refinement.
- We design the **Prior-guided Bilateral Upsampling (PBU)** module, which distinguishes itself from existing edge-aware upsampling methods by directly injecting VFM-derived monocular depth estimates and feature maps as geometric guidance, preserving high-fidelity structural boundaries without requiring explicit edge detection.
- We redesign IGEV into **LGEV**, which avoids storing the full pre-computed correlation volume and instead samples only the disparity candidates needed during inference. This reduces correlation memory from $O(H \cdot W^2)$ to $O(H \cdot W \cdot R)$, achieving up to an 86% reduction with negligible runtime overhead ($< 8\%$). LGEV also natively supports both positive and negative disparities, overcoming the positive-only limitation of FoundationStereo.

2. Related Work

2.1. Vision Foundation Models in 3D Vision

Self-supervised Vision Foundation Models (VFMs) trained on web-scale data provide robust visual representations with strong invariance to illumination, seasonal, and appearance changes. DINOv2 (Oquab et al., 2023) trains a billion-parameter Vision Transformer on curated large-scale datasets, yielding general-purpose features that surpass OpenCLIP on most benchmarks without fine-tuning. Its successor, DINOv3 (Siméoni et al., 2025), introduces Gram anchoring to prevent dense feature degradation during extended training and further scales both data and model capacity. These models therefore offer radiometrically robust features, which are highly desirable for satellite stereo matching but are difficult to obtain from conventionally trained CNNs.

This robustness has motivated increasing efforts to integrate VFMs into depth estimation and dense matching. Depth Anything V2 (DAv2) (Yang et al., 2024) trains monocular depth models on large-scale synthetic and pseudo-labelled data, achieving strong zero-shot depth prediction across diverse scenes. Building on such monocular depth priors, recent foundation stereo models have substantially improved ground-level dense matching. FoundationStereo (Wen et al., 2025) adapts VFM features through side-tuning and constructs a million-pair synthetic training set with automatic self-curation, establishing strong zero-shot stereo performance. MonSter (Cheng et al., 2025) and Stereo Anywhere (Bartolomei et al., 2025) use dual-branch architectures to iteratively fuse monocular depth estimates with stereo cues. DEFOM-Stereo (Jiang et al., 2025a) incorporates a depth foundation model into recurrent updates with scale-aware initialization, while PromptStereo (Wang et al., 2026) replaces conventional GRUs with a prompt-guided recurrent unit built on the monocular depth decoder. FormerStereo (Zhang et al., 2025) further transfers ViT-based foundation features to matching-specific representations through a reconstruction-constrained decoder.

However, existing foundation stereo methods are designed primarily for ground-level imagery, where disparities are typically bounded, positive, and estimated at moderate resolution. These assumptions do not hold in satellite stereophotogrammetry, which involves large bidirectional disparity ranges, sharp structural discontinuities, and large-scale inference under strict memory constraints. Thus, adapting foundation stereo models to satellite imagery remains a critical but unresolved problem.

2.2. General Stereo Matching

Classical photogrammetric pipelines approach stereo matching through path-cost aggregation. Semi-Global Matching (SGM) (Hirschmuller, 2005) and its variants approximate global energy minimization by aggregating one-dimensional path costs along multiple directions. Integrated systems such as S2P (De Franchis et al., 2014), ASP (Shean et al., 2016), and CARS (Youssefi et al., 2020) extend this principle with robust outlier filtering and multi-view fusion, while learning-enhanced variants (Seki and Pollefeys, 2017; Schönberger et al., 2018) replace heuristic parameters with learned alternatives. Although these methods remain effective in many photogrammetric settings, their hand-crafted matching costs are vulnerable to severe radiometric variations.

Deep stereo methods replaced hand-crafted costs with learned cost volumes and neural regularization. The use of 4D cost volumes with 3D CNN regularization (Kendall et al., 2017; Chang and Chen, 2018; Guo et al., 2019) established the dominant learning-based paradigm, which has been further improved by multi-scale context aggregation, group-wise correlation, uncertainty-guided sampling (Mao et al., 2021), and hybrid attention-based aggregation (Cheng et al., 2023). However, dense cost-volume regularization incurs memory costs that grow rapidly with image resolution and disparity range, making it impractical for large satellite tiles with wide search spaces. In addition, most methods estimate only bounded positive disparities, which conflicts with the bidirectional geometry of satellite stereo.

Cascade and iterative refinement paradigms have been developed to reduce the computational burden of dense cost-volume methods. Cascade methods (Gu et al., 2020; Shen et al., 2021) progressively narrow the disparity search space across stages; CFNet introduces variance-based uncertainty for adaptive search, DFPN (Hu et al., 2023) proposes adaptive region aggregation, and SEDNet (Chen et al., 2023) jointly estimates disparity and uncertainty via KL divergence. Iterative methods replace 3D CNN aggregation with recurrent update operators. RAFT-Stereo (Lipson et al., 2021) iteratively refines disparity using multi-level ConvGRUs on all-pairs correlations, IGEV-Stereo (Xu et al., 2023) and IGEV++ (Xu et al., 2025b) encode geometry and context into iterative geometry encoding volumes, and ICGNet (Gong et al., 2024) incorporates both intra-view and cross-view geometric features. Nevertheless, cascade methods often rely on bilinear interpolation for cost-volume upsampling, which discards structural information and blurs depth boundaries, whereas iterative methods still construct dense feature correlations with quadratic memory growth. Edge-aware bilateral grid upsampling (Richardt et al., 2010; Barron et al., 2015; Xu et al., 2021, 2025a) can alleviate boundary blur, but existing approaches mainly use photometric guidance and lack explicit geometric guidance from monocular depth estimates.

2.3. Satellite Stereo Matching

Recent satellite stereo methods address domain-specific imaging challenges through architectures tailored to multi-scale geometry and disparity refinement. Most approaches follow a pipeline of multi-scale feature extraction, hierarchical cost-volume construction, and refinement, while differing in how they adapt to satellite imagery. HMSM-Net (He et al., 2022) and MSCA-Net (Wang et al., 2025) leverage multi-scale cost volumes with edge-aware refinement; Zheng et al. (2024) propose hybrid feature fusion for DSM generation; MaskCRNet (Rao et al., 2024) uses masked pre-training to improve feature quality; SRCV-Net (Kim et al., 2025) introduces self-refining cost volumes for textureless regions; Yang et al. (2025) combine Transformer and CNN modules with iterative refinement; and MetaMRGE (Zhang et al., 2026) incorporates metadata-guided geometric encoding for large-disparity scenarios.

Despite these advances, satellite-specific methods remain limited by feature representations learned from relatively small and heterogeneous satellite datasets. Comparative evaluations (Albanwan and Qin, 2022; He et al., 2023) show that models trained on close-range data generalize poorly to satellite imagery. Jiang et al. (2025b) further demonstrate that this domain gap arises not only from sensor differences, but also from discrepancies in scene semantics and structural distributions. RSMT (Sun et al., 2026) additionally identifies non-negligible label errors in widely used datasets, which further constrains fully supervised learning. As a result, existing satellite-specific models remain susceptible to representation drift across geographic regions, acquisition conditions, and scene structures.

The above studies reveal a gap between two complementary research directions. Satellite-specific stereo methods incorporate relevant geometric assumptions but lack robust cross-domain representations, whereas foundation stereo models provide stronger visual priors but assume ground-level disparity geometry. SatFS bridges this gap by coupling VFM structural features with satellite-specific geometric reasoning through multi-scale side-tuning. It further introduces a cascade coarse-to-fine architecture for large bidirectional disparities, PBU for edge-preserving stage transitions, and a Lightweight Geometry-Aware Encoding Volume for memory-efficient large-scale inference.

3. Method

3.1. Overview

Given a rectified satellite stereo pair, our goal is to estimate a dense disparity map at the highest feasible resolution under extremely wide disparity ranges, while maintaining robustness to severe radiometric and geometric variations. As illustrated in Fig. 3, SatFS builds upon Foundation-Stereo (Wen et al., 2025) and adapts foundation stereo models to satellite imagery by addressing both representation and efficiency bottlenecks. SatFS introduces three key changes to make foundation stereo practical for satellite imagery: multi-scale VFM side-tuning, which uses VFM features at every pyramid level; cascade disparity estimation, which estimates the wide positive-and-negative disparity range from coarse to fine; and lightweight geometry-aware encoding, which avoids

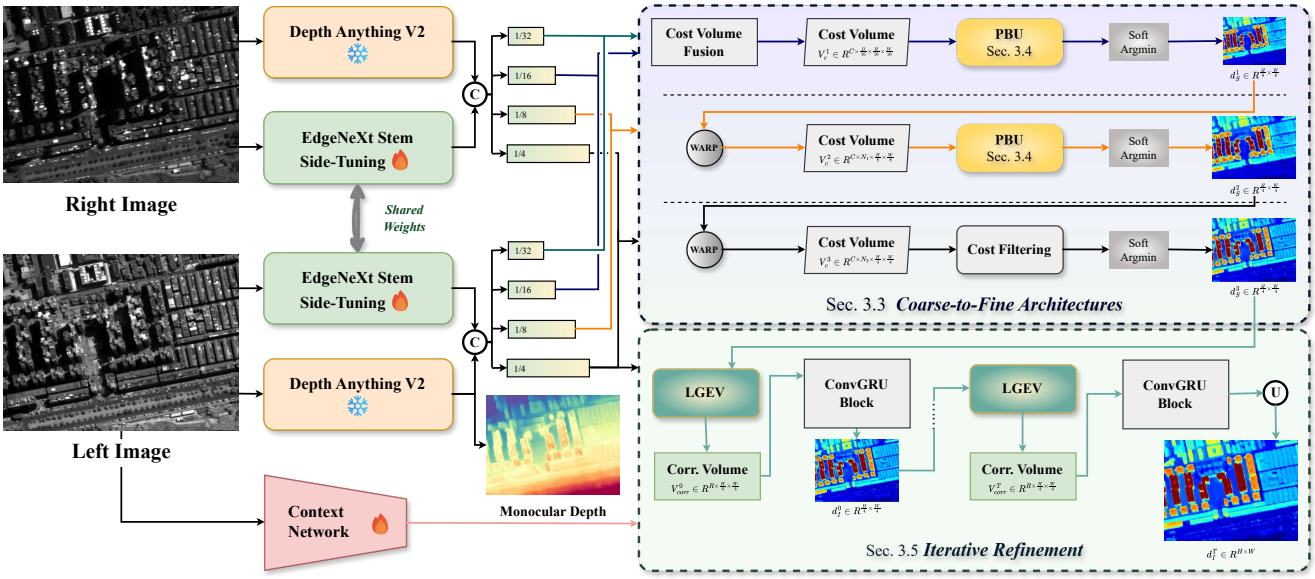

Figure 3: Overview of the proposed SatFS architecture. SatFS consists of multi-scale VFM-enhanced feature extraction, cascade coarse-to-fine disparity estimation, and LGEV-based lightweight iterative refinement for scalable satellite stereo matching.

storing a full correlation volume by sampling only the needed candidates during inference.

Multi-scale VFM-enhanced representation. The frontend contains two complementary branches. The matching branch extracts hierarchical stereo features with a shared CNN backbone at 1/4–1/32 resolutions and enriches each level with frozen DAv2 features through multi-scale side-tuning, providing robust cues for coarse ambiguity resolution and fine geometric refinement (Section 3.2). The context branch follows FoundationStereo (Wen et al., 2025) and progressively downsamples the input to produce structural context at 1/4, 1/8, and 1/16 scales. These context features initialize the hidden states of multi-level ConvGRUs (Lipson et al., 2021) and condition the recurrent disparity updates throughout optimization (Section 3.2).

Cascade coarse-to-fine matching pipeline. Driven by the multi-scale VFM side-tuned features, our coarse-to-fine matching pipeline progressively densify the disparities, as illustrated in Fig. 4. In the coarse stage (Section 3.3.1), cost volumes at 1/32 and 1/16 scales are aggregated using a standard encoder–decoder Cost Volume Fusion module commonly used in deep stereo networks (Chang and Chen, 2018; Guo et al., 2019; Wen et al., 2025) and upsampled to 1/8 resolution using the proposed PBU module (Section 3.4). PBU leverages VFM-derived monocular depth estimates and feature maps for edge-preserving reconstruction, after which a soft-argmin decoder yields an initial disparity map d_S^1. Subsequently, two successive cascade stages at 1/8 and 1/4 resolutions progressively refine this estimate (Section 3.3.2). Each stage constructs a compact local cost volume by sampling N_1 and N_2 disparity hypotheses around the previous stage's prediction, with the search range dynamically modulated by per-pixel matching uncertainty. While the 1/8 stage regularizes its cost volume via PBU to output d_S^2, the 1/4 stage employs Attentive Hybrid Cost Filtering (Wen et al., 2025) for joint spatial-disparity aggregation, producing the refined disparity d_S^3.

Iterative refinement with LGEV. To bypass the memory bottleneck of conventional correlation structures, a Lightweight Geometry-aware Encoding Volume (LGEV, Section 3.5) replaces the pre-computed full correlation volume with on-demand feature sampling. This mechanism reduces the correlation memory complexity from $O(H \cdot W^2)$ to $O(H \cdot W \cdot R)$ while natively accommodating negative disparities. Starting from d_S^3, LGEV dynamically samples correlation features around the current estimate d_I^{t-1} at each iteration t. A motion encoder and a selective ConvGRU (Lipson et al., 2021) then regress a residual update to yield d_I^t. After T iterations, the final full-resolution disparity map d_I^T is recovered via learned convex upsampling.

3.2. Multi-scale VFM-enhanced Representation

Given a rectified stereo pair $\mathbf{I}_l, \mathbf{I}_r \in \mathbb{R}^{H\times W\times 3}$, SatFS first extracts stereo matching features with a shared EdgeNeXt-S backbone (Maaz et al., 2022), following the feature design of FoundationStereo (Wen et al., 2025). Unlike FoundationStereo, which injects VFM features at a single scale, SatFS applies DAv2 (Yang et al., 2024) side-tuning at all cascade levels. The output is a VFM-enhanced feature pyramid $\{\mathbf{f}_l^{(i)}, \mathbf{f}_r^{(i)}\}_{i\in\{4,8,16,32\}}$, where $\mathbf{f}^{(i)} \in \mathbb{R}^{C_i\times H_i\times W_i}$ denotes the feature map at scale $1/i$. These features are used to construct cost volumes and provide matching cues from coarse disparity search to fine geometric refinement.

SatFS also adopts the context network from FoundationStereo (Wen et al., 2025) to provide recurrent update guidance.

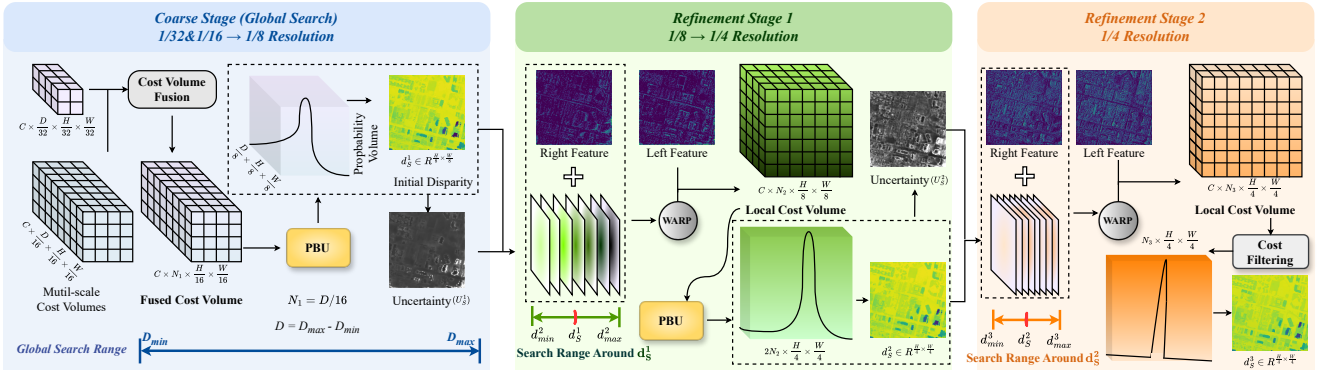

Figure 4: Overview of the cascade coarse-to-fine architecture. An initial disparity is estimated at the coarsest resolution using multi-scale cost volume fusion over a global search range. Subsequent refinement stages progressively narrow the search space by constructing local cost volumes centred on the disparity estimate from the preceding stage. The search range is adaptively modulated according to matching uncertainty, enabling efficient and robust coarse-to-fine refinement. PBU is employed for cost regularisation at intermediate resolutions, while the attentive hybrid cost filtering module is applied at finer resolutions to further enhance local matching reliability.

The context branch follows a standard CNN encoder with progressive downsampling and fuses DAv2 features at the $1/4$ scale before producing context features at $1/4$, $1/8$, and $1/16$ resolutions. The resulting context features initialise the hidden states of the multi-level ConvGRUs (Lipson et al., 2021) and condition the iterative disparity updates in LGEV (Section 3.5).

3.3. Cascade Coarse-to-Fine Architecture

This section details the design of the cascade architecture, including the coarse stage and cascade refinement stages. The detailed cascade pipeline is illustrated in Fig. 4.

3.3.1. Coarse Stage

Single-scale cost volumes are constrained by a resolution-robustness trade-off: high-resolution features provide necessary geometric detail but are often susceptible to radiometric fluctuations, whereas low-resolution features are more stable but often lose fine-grained geometric fidelity (He et al., 2022; Zhang et al., 2026). To address this trade-off and handle the large disparity variations in satellite stereo pairs, we build hierarchical cost volumes and fuse them across scales.

Cost Volume Construction. Given the extracted features $\mathbf{f}_l^{(i)}$ and $\mathbf{f}_r^{(i)}$ at scale $1/i$, we construct a hybrid cost volume combining feature concatenation and group-wise correlation (Guo et al., 2019). For brevity, we omit the scale superscript (i) in the following formulation when no ambiguity arises:

$$\begin{aligned} \mathbf{V}_{cat} &= [\mathrm{Conv}(\mathbf{f}_l)(h,w), \mathrm{Conv}(\mathbf{f}_r)(h,w-d)], \\ \mathbf{V}_{gwc} &= \langle \widehat{\mathbf{f}}_{l,g}(h,w), \widehat{\mathbf{f}}_{r,g}(h,w-d) \rangle, \\ \mathbf{V} &= \mathrm{Conv}([\mathbf{V}_{cat}, \mathbf{V}_{gwc}]) \end{aligned} \tag{1}$$

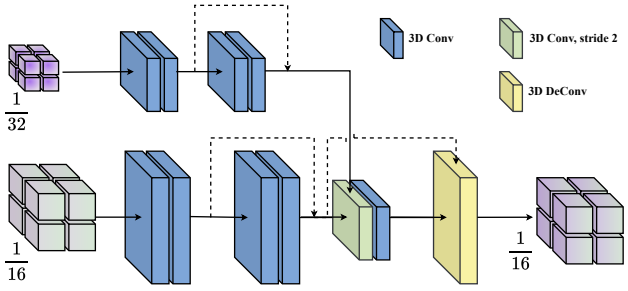

Figure 5: Architecture of the Cost Volume Fusion module. The module aggregates cost volumes from multiple coarse scales and produces a fused cost volume for subsequent upsampling and disparity regression.

where d denotes the disparity hypothesis at scale $1/i$, and $\widehat{\mathbf{f}}_{l,g}$ and $\widehat{\mathbf{f}}_{r,g}$ represent the features divided into G groups for group-wise correlation. To control channel dimensionality, we first apply a 1×1 convolution to project the features into a compact 14-dimensional embedding space before constructing $\mathbf{V}_{cat}$. In parallel, $\mathbf{V}_{gwc}$ is computed via inner-product matching within each feature group, capturing structured similarity cues with lower computational overhead. The final cost representation $\mathbf{V}$ concatenates both branches followed by a convolutional fusion layer.

Cost Volume Fusion. As illustrated in Fig. 5, we adopt a standard encoder–decoder cost-volume fusion design following common deep stereo practice (Chang and Chen, 2018; Guo et al., 2019; Wen et al., 2025). The module takes the cost volumes built at $1/32$ and $1/16$ scales as input, aggregates them with 3D convolutions and skip connections, and outputs a fused cost volume at $1/16$ scale.

The fused cost volume is then upsampled via the PBU module (Section 3.4) to reconstruct the probability volume

at 1/8 scale. A soft-argmin operation extracts the initial disparity map d_S^1:

$$d_S^1(h,w) = \sum_{d=\frac{D_{\min}}{8}}^{\frac{D_{\max}}{8}-1} d \cdot \mathbf{P}_S^1(h,w,d), \tag{2}$$

where $D_{\max}$ and $D_{\min}$ bound the global disparity search range at the original image scale, and $\mathbf{P}_S^1(h,w,d)$ is the matching probability of hypothesis d at position (h,w). This probability-weighted average enables sub-pixel disparity estimation, providing a reliable foundation for subsequent cascade refinement.

3.3.2. Cascade Refinement

Building upon the initial coarse estimate d_S^1, we progressively refine the disparity through two cascade stages at 1/8 and 1/4 resolutions. At each stage, a local cost volume is constructed around the disparity estimate from the previous stage, where the search range is adaptively modulated by the estimation uncertainty. At the 1/8 stage, the local cost volume is regularized and upsampled via the PBU module (Section 3.4), followed by a soft-argmin operation to obtain an updated disparity map d_S^2. At the 1/4 stage, the local cost volume is regularized by the Attentive Hybrid Cost Filtering module from FoundationStereo (Wen et al., 2025), which jointly aggregates features along both spatial and disparity dimensions, producing a refined disparity map d_S^3.

Uncertainty-Aware Range Modulation. After obtaining the single-channel probability volume, well-matched pixels tend to exhibit a concentrated probability distribution, whereas ambiguous pixels display a dispersed one. We therefore quantify the matching uncertainty of each pixel via the variance of its probability distribution. Given a position (h,w) with probability distribution $\mathbf{P}_S^n(h,w,d)$ over disparity hypotheses d, we define the uncertainty $\mathbf{U}_S^n(h,w)$ as follows:

$$\begin{aligned} d_S^n(h,w) &= \sum_d d \cdot \mathbf{P}_S^n(h,w,d), \\ \mathbf{U}_S^n(h,w) &= \sum_d \left(d - d_S^n(h,w)\right)^2 \cdot \mathbf{P}_S^n(h,w,d), \end{aligned} \tag{3}$$

where $d_S^n(h,w)$ is the expected disparity at position (h,w). The uncertainty $\mathbf{U}_S^n(h,w)$ captures the dispersion of the probability distribution, with higher values indicating greater ambiguity in the matching. We then modulate the local search range at the next stage proportionally to $\mathbf{U}_S^n$. Thus, the next stage's disparity searching range is defined as:

$$\begin{aligned} d_{\max}^{n+1} &= d_S^n + (\alpha^n + 1)\sqrt{\mathbf{U}_S^n} + \beta^n, \\ d_{\min}^{n+1} &= d_S^n - (\alpha^n + 1)\sqrt{\mathbf{U}_S^n} - \beta^n, \end{aligned} \tag{4}$$

where α^n and β^n are learnable scale and offset parameters that control the sensitivity of the search range to the uncertainty.

We then uniformly sample N_{n+1} discrete disparity hypotheses within the defined range:

$$d_m^{n+1} = d_{\min}^{n+1} + m \cdot \max\left(\frac{d_{\max}^{n+1} - d_{\min}^{n+1}}{N_{n+1}-1},\ \Delta_{\min}\right), \quad m \in \{0,1,\ldots,N_{n+1}-1\}, \tag{5}$$

where N_{n+1} is the number of disparity hypotheses at stage $n+1$, and $\Delta_{\min}$ is a preset minimum sampling interval that prevents overly dense hypotheses when the adaptive range is narrow. Given the sampled hypotheses, we index into the left and right feature maps to retrieve the corresponding features, forming a 4D volume. The hybrid cost volume at stage $n+1$ is then constructed following the same concatenation and group-wise correlation fusion strategy described in Section 3.3.1.

Cost Filtering. While the 1/8 stage employs PBU for edge-preserving upsampling across resolutions, the 1/4 stage operates at the highest cascade resolution where the primary challenge shifts from boundary preservation to fine-grained spatial-disparity discrimination. We therefore adopt the Attentive Hybrid Cost Filtering module from FoundationStereo (Wen et al., 2025), which performs joint spatial-disparity aggregation via decoupled convolutions and disparity-level attention, yielding a well-regularized probability volume. The filtered volume is then decoded via soft-argmin to produce the refined disparity map d_S^3.

3.4. Prior Bilateral Upsampling

Standard cost volume upsampling between cascade stages tends to oversmooth disparity discontinuities at structural boundaries. To preserve edge fidelity during upsampling, we propose PBU, which incorporates VFM-derived monocular depth estimates and feature maps as geometric guidance through bilateral-grid slicing. Unlike standard bilateral grid methods (Xu et al., 2021) that rely solely on learnable feature-based guidance, our PBU uses both high-level VFM features and a normalised monocular depth map to steer the slicing process, producing edge-preserving upsampled cost volumes with structurally coherent disparity distributions. The module is illustrated in Fig. 6.

Multi-Modal Guidance. We concatenate the high-resolution image feature maps with a normalised monocular depth map, both at resolution $H^{(o)} \times W^{(o)}$. The depth map serves as a geometric constraint, providing coarse but essential structural information. The concatenated features are then passed through a series of convolutions to produce a unified guidance map $\mathbf{G} \in \mathbb{R}^{B\times 1\times H^{(o)}\times W^{(o)}}$. This guidance map determines the lookup coordinates along the guidance dimension of the bilateral grid, ensuring that disparity discontinuities align with both radiometric edges and geometric depth boundaries.

Grid Learning Branch. The low-resolution input cost volume is processed by a 3D hourglass network with skip connections. This encoder-decoder architecture effectively aggregates local and global cost information, transforming the raw matching costs into a structured bilateral grid $\mathcal{B} \in$

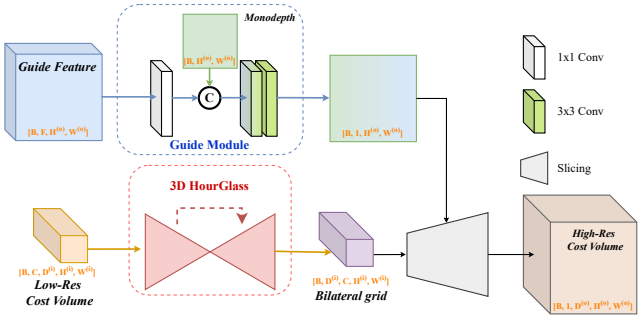

Figure 6: Architecture of the PBU module. PBU performs edge-aware cost volume upsampling in a bilateral grid from the input resolution $H^{(i)}\times W^{(i)}$ to the output resolution $H^{(o)}\times W^{(o)}$. It integrates monocular depth cues with feature-based guidance to mitigate oversmoothing and preserve structural boundaries during disparity estimation.

$\mathbb{R}^{B\times N_d^{(i)}\times C_g\times H^{(i)}\times W^{(i)}}$, where $N_d^{(i)}$ is the number of disparity hypotheses and C_g denotes the guidance dimension. The resulting grid encodes the matching probability distribution in a learned high-dimensional space.

Slicing and Reconstruction. The high-resolution cost volume C_H is obtained via a differentiable slicing layer. Using the guidance map $\mathbf{G}$, this layer performs data-dependent lookups in the low-resolution bilateral grid through multi-linear interpolation. Formally, the slicing operation is defined as:

$$C_H(x,y,d) = \mathcal{B}(s_x x,\ s_y y,\ s_d d,\ s_g \mathbf{G}(x,y)), \tag{6}$$

where $s_x, s_y \in (0,1)$ are the spatial scale ratios of the grid relative to the high-resolution cost volume, s_d is the disparity scale ratio, and $s_g \in (0,1)$ is the ratio of the grid's guidance range to the guidance map range. The slicing layer is parameter-free and can be implemented efficiently. After slicing, the output cost volume is passed through a softmax layer to produce the final probability distribution for disparity estimation.

3.5. Iterative Refinement with LGEV

The cascade stages produce disparity estimates at 1/4 resolution. To further refine these estimates with sub-pixel accuracy, iterative update operators have proven effective, yet existing approaches either pre-compute dense correlation volumes with prohibitive memory cost or assume non-negative disparities, both unsuitable for satellite stereo imagery. We address these limitations by modifying IGEV (Xu et al., 2023) into LGEV coupled with a selective ConvGRU update operator, which provides on-demand correlation features at each iteration while natively supporting negative disparities; the refined 1/4-resolution disparity is finally upsampled to full resolution via learned convex upsampling.

3.5.1. Lightweight Geometry-aware Encoding Volume

Iterative stereo methods such as RAFT-Stereo (Lipson et al., 2021) and IGEV-Stereo (Xu et al., 2023, 2025b) perform on-the-fly correlation lookups at each iteration. FoundationStereo (Wen et al., 2025) adopts a similar strategy but pre-computes a dense geometry encoding volume at 1/4 scale, consuming $O(H\cdot W^2)$ memory and limiting inference resolution. Moreover, these methods assume non-negative disparities with a fixed search range starting from zero, precluding satellite stereo pairs that exhibit negative disparities due to orbital geometry.

To address these limitations, we propose LGEV that replaces pre-computed correlation with on-demand feature sampling—caching only the L2-normalised left feature $\bar{\mathbf{f}}_L$ and a multi-level right feature pyramid $\{\bar{\mathbf{f}}_R^l\}_{l=0}^{L-1}$—while introducing configurable disparity range support. At iteration t, given the current disparity estimate d_I^{t-1}, we extract features at R corresponding positions from the right feature pyramid via bilinear sampling, and compute the inner product with the left features to obtain the correlation feature $\mathbf{V}_{\text{corr}}^t$.

This design offers three advantages: (1) memory consumption is reduced from $O(H\cdot W^2)$ to $O(H\cdot W\cdot R)$, where $R \ll W$ is the number of sampled positions, enabling high-resolution satellite inference on commodity GPUs; (2) by introducing a configurable minimum disparity $d_{\min}$, LGEV supports negative disparity search to accommodate the orbital geometry of satellite stereo pairs; (3) per-pixel search start s_t and step δ_t enable cascade disparity sampling that progressively focuses on high-probability intervals during iterative refinement.

Specifically, at iteration t, the sampling coordinate at pyramid level l with offset Δ is defined as:

$$\xi_t^l(h,w;\Delta) = \frac{w - \tilde{d}_t(h,w) + \Delta}{2^l}, \tag{7}$$

where $\Delta \in \{-r,\dots,+r\}$ enumerates $R = 2r+1$ uniformly spaced integer offsets centred around the current disparity estimate, $l = 0,\dots,L-1$ indexes the pyramid level, and $\tilde{d}_t$ is the normalised disparity. Depending on the stage configuration, $\tilde{d}_t$ is defined as: $d_I^{t-1} - d_{\min}$ when negative disparity support is required (satellite stereo pairs with orbital geometry), $(d_I^{t-1} - s_t)/\delta_t$ for cascade refinement with a local search centre s_t and step δ_t, or simply d_I^{t-1} in the basic mode. The on-the-fly correlation feature is then computed as:

$$\mathbf{V}_{\text{corr}}^t(h,w,\Delta,l) = \left\langle \bar{\mathbf{f}}_L(h,w),\ \ \hat{\mathbf{f}}_R^l(h,\xi_t^l) \right\rangle, \tag{8}$$

where $\bar{\mathbf{f}}_L$ is the L2-normalised left feature and $\hat{\mathbf{f}}_R^l$ denotes the bilinearly sampled feature from the l-th level of the right pyramid. The final encoding feature is obtained by concatenating the geometry encoding volume lookup with the real-time correlation features along all offsets and pyramid levels:

$$\mathbf{FV}_t(h,w) = \left[\mathbf{V}_C(\tilde{d}_t,h,w),\ \ \{\mathbf{V}_{\text{corr}}^t\}_{\Delta,l} \right], \tag{9}$$

where $\mathbf{V}_C$ is the geometry encoding volume—a lightweight 3D convolution-based volume pre-computed from the 1/4-scale features (following (Xu et al., 2023))—queried at the normalised disparity $\tilde{d}_t$.

3.5.2. Iterative Refinement

Given the encoding feature $\mathbf{FV}_t$ constructed by LGEV, we adopt a GRU-based iterative framework to progressively refine the disparity. At each iteration, the update proceeds in three steps: motion encoding, selective ConvGRU update, and residual disparity regression.

Motion Encoding. A motion encoder $\mathcal{E}_m$ encodes the current cost feature $\mathbf{FV}_t$ and disparity d_I^{t-1} into a motion feature:

$$\mathbf{x}_t = \left[\mathcal{E}_m(\mathbf{FV}_t, d_I^{t-1}),\ d_I^{t-1},\ \mathbf{c}\right], \tag{10}$$

where $\mathcal{E}_m$ comprises two convolutional branches that independently process $\mathbf{FV}_t$ and d_I^{t-1} before concatenation and projection. The context feature $\mathbf{c} = \mathrm{ReLU}(\mathbf{f}_c)$ is extracted from the left image and fused with VFM features, providing monocular geometric guidance.

Selective ConvGRU. We employ a Selective ConvGRU (Lipson et al., 2021) for hidden state update. Unlike standard GRU with a single convolution kernel, it maintains two parallel branches with a small kernel (3×3) and a large kernel ($1\times5+5\times1$), adaptively weighted by a spatial attention map $\mathbf{a}$ generated via cascade channel and spatial attention on the context features. This allows the model to favour large receptive fields in flat textureless regions for low-frequency structure and small receptive fields near edges for high-frequency detail. Hidden states are synchronously updated at three resolution levels (1/4, 1/8, 1/16) in a coarse-to-fine order, with initial states $h_0^{(i)} = \tanh(\mathbf{f}_c^{(i)})$. Coarse levels provide global context to fine levels via bilinear upsampling, while fine levels supply local detail via downsampling, enabling bidirectional information exchange.

Disparity Update. The disparity is updated residually via $d_I^t = d_I^{t-1} + \mathrm{Conv}_\Delta(h_t^{(4)})$. After T iterations, the 1/4-resolution prediction d_I^T is upsampled to full resolution via convex upsampling with learned 3×3 per-pixel weights.

3.6. Progressive Satellite Domain Adaptation

We start from the FoundationStereo (Wen et al., 2025) checkpoint, which is pre-trained on 1M synthetic stereo pairs. Direct end-to-end fine-tuning on satellite data is not ideal: it can weaken the useful representations learned during large-scale pre-training, and the pretrained model has not seen the negative disparities caused by satellite orbital geometry. We therefore use a four-stage progressive fine-tuning pipeline that gradually introduces satellite-specific geometry and radiometry, moving from synthetic augmentation to real-domain alignment.

Stereo-Pair Inversion for Negative Disparity. To enable the pretrained model to reason about negative disparities, we apply a stereo-pair inversion augmentation to the 190K synthetic images in the FoundationStereo dataset. By randomly swapping the left and right images during training, the model learns cost aggregation and regression over the full range $[-D_{\max}, D_{\max}]$, breaking the original non-negative disparity assumption.

Adaptive Cost Filtering Adaptation. We freeze the pretrained backbone and fine-tune only the AHCF module. This allows the model to adapt its cost volume distribution to the negative-disparity geometry introduced by stereo-pair inversion, while preserving the pretrained VFM features.

Iterative Refinement Warm-up. We unfreeze the key update components: the motion encoder $\mathcal{E}_m$ and the 1/4-resolution Selective ConvGRU, while keeping the backbone and most of AHCF frozen. This stage focuses on learning the residual update logic for satellite-specific disparity adjustments.

Joint Satellite Domain Alignment. Finally, we perform full-pipeline adaptation on real satellite imagery. The side-tuning branch, AHCF, motion encoder, and 1/4-resolution Selective ConvGRU are jointly optimised, while the VFM backbone remains frozen. Through exposure to real remote sensing scenes with complex illumination and radiometric variations, the model completes the domain alignment from synthetic to satellite data.

3.7. Loss Function

The model is trained with a cascade-supervised objective. The cascade predictions are supervised with manually assigned stage weights, while the iterative refinement outputs are supervised with exponentially increasing weights following (Wen et al., 2025):

$$\mathcal{L} = \sum_{i=1}^{N} \lambda_i \, |d_S^i - \bar{d}|_{\text{smooth}} + \sum_{t=1}^{T} \gamma^{T-t} |d_I^t - \bar{d}|_{\text{smooth}}, \tag{11}$$

where $N = 3$ cascade stages with manually set weights $\lambda_i \in \{0.6, 0.8, 1.0\}$, $\bar{d}$ denotes the ground-truth disparity (downsampled to the corresponding resolution), d_I^t is the iterative refinement output at iteration t, $|\cdot|_{\text{smooth}}$ denotes the smooth L1 loss, and $\gamma = 0.9$.

4. Experiments

We evaluate SatFS from five complementary perspectives:

- **In-distribution accuracy.** Section 4.2 compares SatFS with state-of-the-art stereo matching methods on satellite benchmarks and quantifies the gains from VFM-guided matching and cascade cost volume refinement.
- **Component analysis.** Section 4.3 isolates the contribution of multi-scale VFM side-tuning, cascade cost volume construction, PBU, multi-level VFM injection, and GRU-based iterative refinement, together with a memory–runtime analysis of LGEV.
- **Generalization analysis.** Section 4.4 evaluates cross-dataset transfer on WHU-SSIDE (Zhang et al., 2026) and multi-temporal robustness on SatStereo (Patil et al., 2019).

- **Geographically diverse real-world application.** Section 4.5 evaluates DSM reconstruction across unseen regions and multi-sensor satellite imagery, with quantitative LiDAR assessment and comparisons against commercial photogrammetry pipelines including Inpho (Trimble Inc., 2026), Metashape (Agisoft LLC, 2022), and G3D-SAT (DASpatial, 2026).
- **Limitations.** Section 4.6 discusses remaining limitations and directions for future work.

4.1. Experimental Setup

4.1.1. Datasets

Table 1 summarises the four benchmark datasets used in our experiments and their roles in different evaluation settings. US3D and WHU-Stereo are used for supervised training, in-distribution benchmark evaluation, and component analysis. WHU-SSIDE is used to evaluate cross-dataset transfer under wider disparity ranges, while SatStereo is used to assess multi-temporal robustness on independent WorldView imagery.

US3D. The Urban Semantic 3D (US3D) dataset (Bosch et al., 2019) provides multi-view WorldView-3 satellite imagery over two US cities—Jacksonville, FL (26 scenes) and Omaha, NE (43 scenes)—covering approximately 100 km^2 at 0.3 m panchromatic GSD. Ground truth disparity maps and four-class semantic labels (building, ground, tree, water) are derived from airborne LiDAR at 0.8 m aggregate pulse spacing. Imagery collected between 2014 and 2016 introduces substantial seasonal appearance differences, making US3D a challenging benchmark for robust stereo matching. Following (He et al., 2022), we use 1,600 epipolar-rectified stereo pairs of 1024×1024 pixels from Jacksonville for training and evaluate cross-region generalisation on 2,153 pairs from Omaha.

WHU-Stereo. As a second training source, we incorporate the WHU-Stereo dataset (He et al., 2022), which provides near-synchronous along-track panchromatic stereo imagery from the Chinese GaoFen-7 satellite (GSD< 0.8 m) over six Chinese cities: Shaoguan, Kunming, Yingde, Qichun, Wuhan, and Hengyang. Ground truth disparity maps are generated by projecting airborne LiDAR point clouds into the epipolar-rectified image planes. Following the official split, the dataset comprises 1,220 training pairs and 415 testing pairs at 1024×1024 pixels, covering diverse urban and rural landscapes. Compared with US3D, WHU-Stereo introduces additional challenges due to lower spatial resolution, single-channel panchromatic input, and the presence of both positive and negative disparities. We adopt a mixed-training strategy, jointly training on US3D and WHU-Stereo to encourage cross-sensor feature invariance.

WHU-SSIDE. To evaluate cross-dataset transfer under wider disparity ranges, we further test on the recently released WHU-SSIDE dataset (Zhang et al., 2026). WHU-SSIDE extends WHU-Stereo to 3,737 high-quality stereo pairs over the same six cities, with a disparity coverage of approximately 384 pixels. Each sample includes epipolar-rectified stereo images, disparity ground truth, and structured metadata, including geographic coordinates, satellite angles, and solar angles. This dataset is used exclusively for generalization evaluation without fine-tuning, and we report results on the official test split of 376 pairs.

Table 1
Summary of datasets. US3D and WHU-Stereo are used for supervised training, in-distribution evaluation, and component analysis. WHU-SSIDE and SatStereo are used for cross-dataset transfer and multi-temporal robustness evaluation, respectively.

Dataset	Sensor / GSD	Pairs	Role
US3D	WV-3 / 0.3 m	1600 / 2153	Train & Test
WHU-Stereo	GF-7 / <0.8 m	1220 / 415	Train & Test
WHU-SSIDE	GF-7 / <0.8 m	376	Validation
SatStereo	WV-2, 3 / 0.3–0.5 m	7258	Validation

SatStereo. To assess multi-temporal robustness on an independent benchmark, we additionally evaluate on the SatStereo dataset (Patil et al., 2019). SatStereo provides stereo-rectified multi-date WorldView-3 imagery, with partial WorldView-2 coverage, over 10 areas of interest in the United States at 0.3–0.5 m GSD, with 53–505 stereo pairs per area. Ground truth disparity maps are generated by fusing multi-view DSMs and aligning them with 30 cm airborne LiDAR; building masks are provided to indicate regions where the disparity ground truth is reliable. This dataset is used exclusively for multi-temporal robustness evaluation without fine-tuning, and we evaluate on all 7,258 available stereo pairs.

4.1.2. Evaluation Metrics

Following (Bosch et al., 2019; He et al., 2022), we adopt two standard metrics to evaluate disparity accuracy. The **end-point error** (EPE) measures the mean absolute disparity error over all valid pixels:

$$\mathrm{EPE} = \frac{1}{|\mathcal{V}|} \sum_{(h,w)\in\mathcal{V}} \left| \hat{d}(h,w) - \tilde{d}(h,w) \right|, \tag{12}$$

where $\mathcal{V}$ denotes the set of valid labelled pixels, $\hat{d}$ is the predicted disparity, and $\tilde{d}$ is the ground-truth disparity.

The **D1 error rate** measures the percentage of valid pixels whose absolute disparity error exceeds a predefined threshold:

$$\mathrm{D1} = \frac{1}{|\mathcal{V}|} \sum_{(h,w)\in\mathcal{V}} \mathcal{I}\left(\left| \hat{d}(h,w) - \tilde{d}(h,w) \right| > \varepsilon \right) \times 100\%, \tag{13}$$

where ε is the error threshold, set to 3 pixels in this study. $\mathcal{I}(\cdot)$ denotes the Iverson bracket, which returns 1 if the condition is satisfied and 0 otherwise. For both disparity metrics, lower values indicate better performance (↓).

For DSM reconstruction evaluation, we report the **root mean square error** (RMSE) and **mean absolute error** (MAE) of the height error, where lower values indicate better reconstruction accuracy (↓). We also report the **percentage**

of absolute error within a threshold (PAE@t), which measures the percentage of reconstructed points whose absolute height error falls within a practical tolerance t:

$$\mathrm{PAE@t} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}\left(|e_i| \le t\right) \times 100\%, \tag{14}$$

where e_i denotes the height error of the i-th valid reconstructed point, and n is the number of valid points. A higher PAE@t value indicates better DSM reconstruction quality (↑).

4.1.3. Implementation Details

We implement SatFS in PyTorch and train it on 8 NVIDIA RTX 4090 GPUs for 100 epochs, with a per-GPU batch size of 2 and an effective batch size of 16. We use the AdamW optimiser with an initial learning rate of 10^{-4}, a weight decay of 10^{-5}, and $\epsilon = 10^{-8}$. A OneCycleLR scheduler is adopted with a peak learning rate of 10^{-4} and a 5% warmup phase. Gradient clipping is applied with a maximum norm of 0.01, and mixed-precision training with AMP is enabled.

During training, images are randomly cropped to 384×768 pixels. We apply colour jittering with brightness, contrast, and saturation factors sampled from $[0.6, 1.4]$, and hue perturbation sampled from $[-0.5, 0.5]$. We also use random erasing, horizontal flipping with a probability of 0.7, and random RGB-to-grayscale conversion with a probability of 0.8 to simulate single-channel panchromatic input. During inference, images are padded so that both spatial dimensions are divisible by 32.

The VFM backbone, namely DAv2 ViT-L, remains frozen throughout training. Only the side-tuning branches, cascade cost volume modules, and GRU refinement components are updated. The global disparity search range is set to $[-512, 512]$ pixels to accommodate both positive and negative disparities. The two cascade refinement stages sample N_2=32 and N_3=16 disparity hypotheses at the $1/8$ and $1/4$ scales, respectively. For group-wise correlation, we use G=8 feature groups. The LGEV module uses R=9 sampling positions, corresponding to a radius of r=4, across L=2 pyramid levels, and performs K=22 GRU iterations during inference.

The model is initialised from a FoundationStereo checkpoint that has been fine-tuned with stereo-pair inversion augmentation to support negative disparities, as described in Section 3.6.

4.2. Comparative Evaluations

We compare SatFS with both classical and learning-based baselines. For classical stereo matching, we include SGM (Hirschmuller, 2005), using the implementation adopted in (Sun et al., 2026). For learning-based methods, we evaluate CFNet (Shen et al., 2021), HMSMNet (He et al., 2022), IGEV++ (Xu et al., 2025b), FoundationStereo (FS) (Wen et al., 2025), and MetaMRGE (Zhang et al., 2026). Hereafter, FS denotes FoundationStereo in all tables and discussions. To ensure fair comparisons, all deep learning baselines are retrained on the same mixed training set, i.e., US3D and WHU-Stereo, using identical data augmentation and comparable hyperparameter settings where applicable. All methods are then evaluated on the same test sets using the same metrics.

Table 2
Quantitative comparison on the US3D and WHU-Stereo datasets. * denotes results reported in the original papers. † denotes SatFS without the iterative GRU refinement module. The best and second-best results are highlighted in **red** and blue, respectively.

Year	Method	US3D		WHU Stereo	
		D1↓	EPE↓	D1↓	EPE↓
2005	SGM	19.36	3.67	19.02	4.07
2021	CFNet	6.37	1.09	10.73	1.45
2022	HMSMNet*	-	-	14.74	1.76
2022	HMSMNet	6.72	1.14	11.69	1.53
2025	IGEV++	6.15	1.06	10.62	1.42
2026	MetaMRGE*	-	-	11.84	1.67
2026	MetaMRGE	6.24	1.09	10.56	1.44
2025	FS	**5.43**	**0.97**	10.55	1.40
2026	SatFS†	5.77	1.05	9.83	1.34
2026	SatFS	5.75	1.01	**9.57**	**1.33**

We first evaluate all methods on the in-distribution test sets of US3D (Omaha) and WHU-Stereo, where the test data come from sensor domains included in the mixed training set. The quantitative results are reported in Tab. 2. On WHU-Stereo, SatFS achieves the best performance, with a D1 error rate of 9.57% and an EPE of 1.33 pixels. It outperforms the second-best method, FS, which obtains 10.55% D1 and 1.40 EPE. The advantage is also evident over satellite-specific baselines: the retrained HMSMNet and MetaMRGE obtain D1 error rates of 11.69% and 10.56%, respectively, both higher than that of SatFS. For reference, their original-paper results, marked with * in the table, are 14.74% and 11.84% D1, respectively.

On US3D, SatFS achieves the second-best performance, with a D1 error rate of 5.75% and an EPE of 1.01 pixels, closely following FS, which obtains 5.43% D1 and 0.97 EPE. As further analysed in the ablation study (Section 4.3), the relatively small margin on US3D may be attributed to the strong alignment between DAv2's RGB-oriented pre-training and the multispectral imagery used in US3D. In this setting, FS already benefits substantially from VFM features, leaving less room for further improvement through multi-scale VFM-enhanced representation.

Figures 7 and 8 present qualitative comparisons on WHU-Stereo and US3D, respectively. On WHU-Stereo (Fig. 7), the first two rows show that SatFS† already recovers richer rooftop structures than the baselines, while the full SatFS further sharpens fine geometric details through GRU-based iterative correction. The third row demonstrates the robustness of SatFS in scenes with large elevation variations, where it preserves more complete disparity estimates and avoids the severe over-smoothing observed in competing methods.

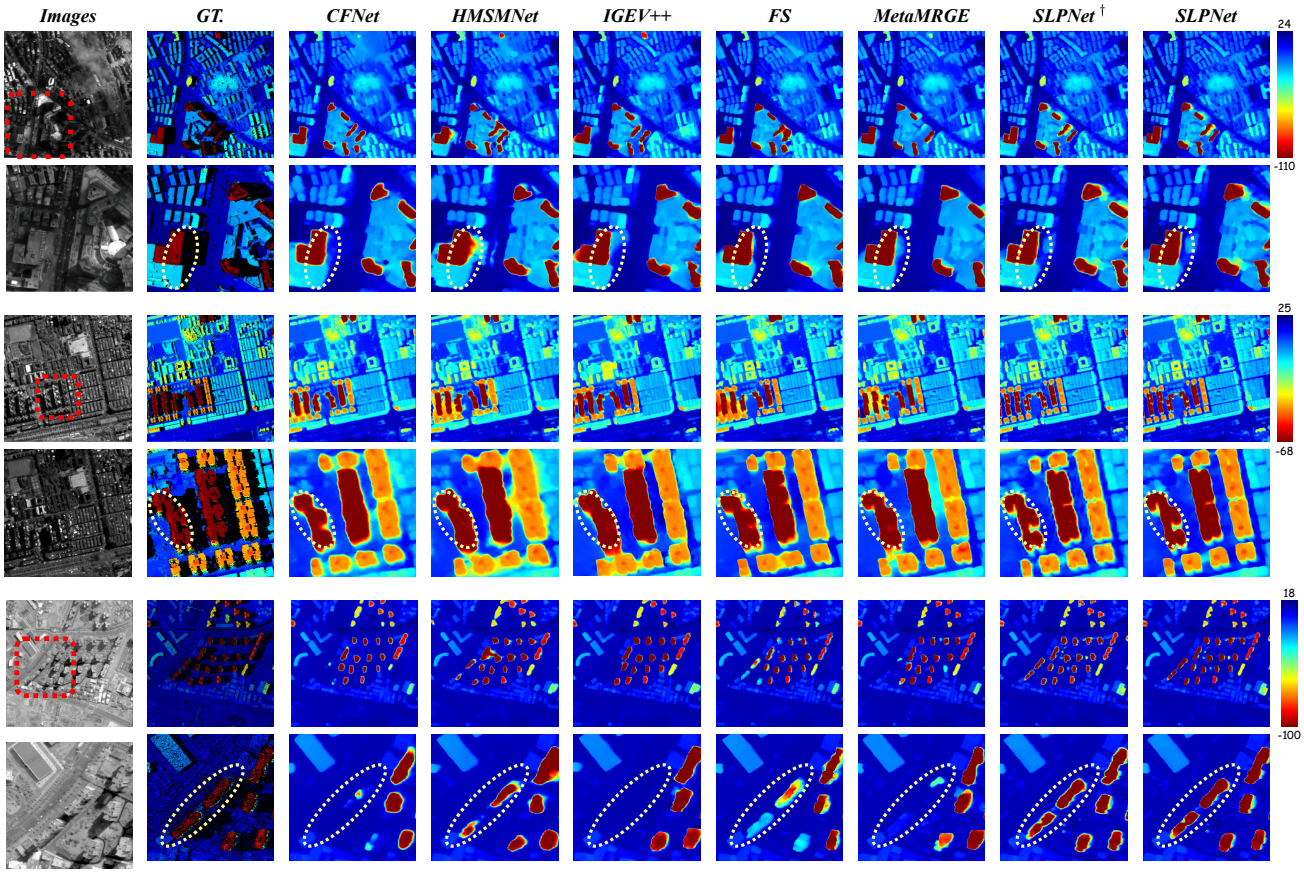

Figure 7: Qualitative comparison on WHU-Stereo. †: SatFS without the iterative GRU refinement module.

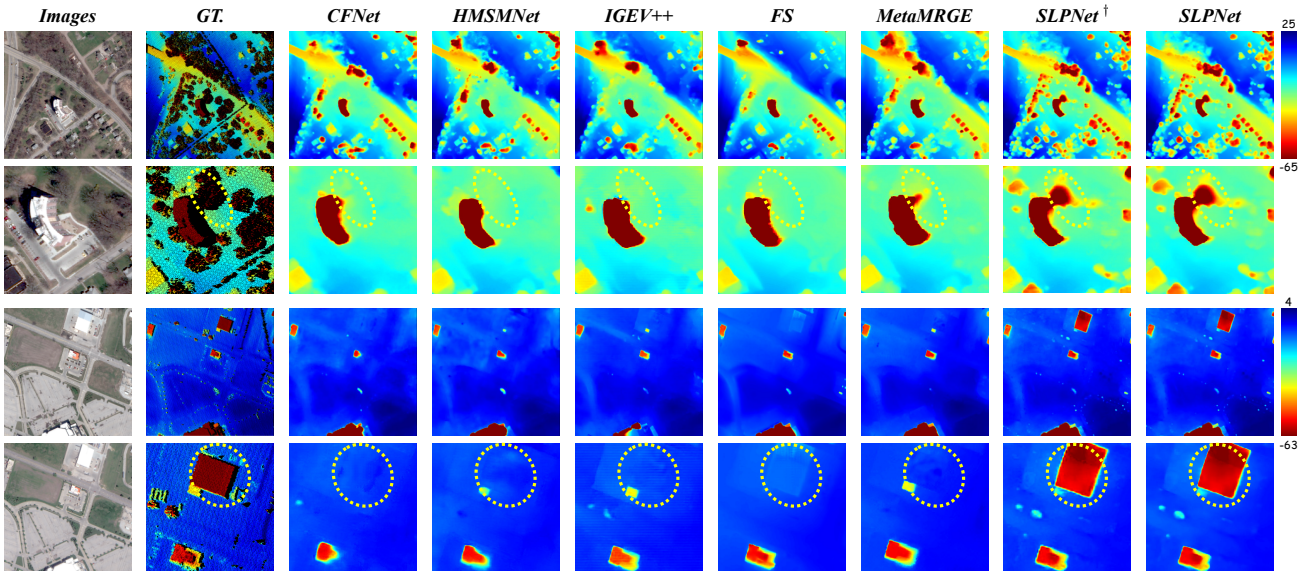

Figure 8: Qualitative comparison on US3D. †: SatFS without the iterative GRU refinement module.

On US3D (Fig. 8), the first scene shows that SatFS produces sharper object boundaries and better preserves the disparities of narrow structures, such as leafless trees. The second scene further confirms its robustness in large textureless rooftop regions, where SatFS maintains coherent disparity estimates while the baselines struggle to recover plausible values.

Table 3
Ablation study on cascade cost-volume stages, PBU, and GRU-based iterative refinement on the US3D and WHU-Stereo datasets. • indicates that the corresponding module incorporates VFM features, whereas ∘ indicates that the module is used without VFM enhancement. FS# denotes the FoundationStereo variant adapted to negative disparities through left–right image swapping and trained on a 190 K-pair subset.

Settings	Coarse		Refine		PBU	GRU	US3D		WHU-Stereo	
	1/32	1/16	1/8	1/4			D1↓	EPE↓	D1↓	EPE↓
FS#				•		✓	14.18	2.14	18.11	3.39
FS				•		✓	**5.43**	**0.97**	10.55	1.40
1				•			6.36	1.12	14.36	2.23
2			∘				6.89	1.19	17.51	2.61
3	∘		∘	•			6.18	1.08	11.74	1.56
4		∘	∘	•			5.87	1.04	11.40	1.51
5	∘	∘	∘	•			5.85	1.02	11.16	1.46
6	∘	∘	∘	•	✓		5.75	1.01	10.79	1.43
7	•	•	•	•	✓		5.77	1.05	9.83	1.34
Ours	•	•	•	•	✓	✓	5.75	1.01	**9.57**	**1.33**

4.3. Ablation Studies

To isolate the contribution of each proposed module, we conduct progressive ablation experiments on the WHU-Stereo and US3D test sets. Tab. 3 traces the progression from the FoundationStereo baseline to the full SatFS by successively adding cascade cost volume stages, PBU, multi-level VFM injection, and GRU-based iterative refinement. In addition, Tab. 4 analyses the memory efficiency of the lightweight geo-encoding volume, and Fig. 9 provides qualitative comparisons for representative configurations.

Domain adaptation baseline. As shown in Tab. 3, FS# denotes a FoundationStereo variant adapted to negative disparities through left–right image swapping and trained exclusively on a 190 K-pair subset of synthetic data. It obtains D1 errors of 18.11 on WHU-Stereo and 14.18 on US3D, revealing a substantial domain gap when transferring ground-level foundation stereo models directly to satellite imagery. Joint fine-tuning on satellite data, denoted as FS, substantially mitigates this gap, achieving the best accuracy on US3D (D1: 5.43, EPE: 0.97) and competitive results on WHU-Stereo (D1: 10.55, EPE: 1.40). Compared with FS#, this corresponds to relative D1 reductions of 61.7% on US3D and 41.7% on WHU-Stereo.

Cascade cost volume stages. Single-scale cost volumes are insufficient for satellite stereo, where disparity ranges can span hundreds of pixels. As shown in Tab. 3, the single-stage 1/4-scale variant (Setting 1) yields a WHU-Stereo D1 of 14.36, while the single-stage 1/8-scale variant (Setting 2) obtains 17.51. Both are substantially worse than the four-stage cascade result of 11.16 (Setting 5), confirming that a single scale cannot effectively handle the wide disparity range in satellite stereo matching. Since Setting 1 and Setting 2 each contain only one refinement stage, their comparison also shows that higher-resolution refinement is more important for final accuracy, as the 1/4-scale variant clearly outperforms the 1/8-scale variant.

Introducing a 1/32 coarse stage in addition to the two refinement stages reduces the WHU-Stereo D1 from 14.36 to 11.74 (Setting 3), corresponding to an 18.2% relative improvement. This indicates that coarse-level geometry provides useful constraints for large-displacement matching. Replacing the 1/32 stage with a 1/16 stage yields a comparable improvement (Setting 4: WHU-Stereo D1 of 11.40). The full four-stage cascade, which combines both coarse scales, achieves the best result among variants without PBU (Setting 5: WHU-Stereo D1 of 11.16 and US3D D1 of 5.85), suggesting that the two coarse scales provide complementary geometric constraints.

PBU. Equipping the full cascade with PBU (Setting 6) further reduces the WHU-Stereo D1 from 11.16 to 10.79, corresponding to a 3.3% relative improvement. It also reduces the US3D D1 from 5.85 to 5.75. These consistent gains demonstrate that VFM-guided cost volume upsampling through bilateral-grid slicing improves the multi-scale cascade by better preserving structural boundaries during probability-volume reconstruction. The qualitative results in Fig. 9 further show that Setting 6 produces sharper disparity boundaries near building edges and more complete estimates in textureless regions than Setting 5.

Multi-level VFM injection. Extending VFM feature injection from the 1/4 scale only (Setting 6) to the full feature pyramid (Setting 7) brings substantial improvements on WHU-Stereo. The D1 error decreases from 10.79 to 9.83, corresponding to an 8.9% relative improvement, and the EPE decreases from 1.43 to 1.34. On US3D, the results remain comparable, with D1 changing from 5.75 to 5.77 and EPE from 1.01 to 1.05. This asymmetric effect is consistent with the VFM pre-training domain: DAv2 is pre-trained on natural RGB images, which aligns better with the multispectral imagery in US3D, whereas the single-channel panchromatic imagery in WHU-Stereo deviates more from the pre-training distribution and therefore benefits more from multi-level

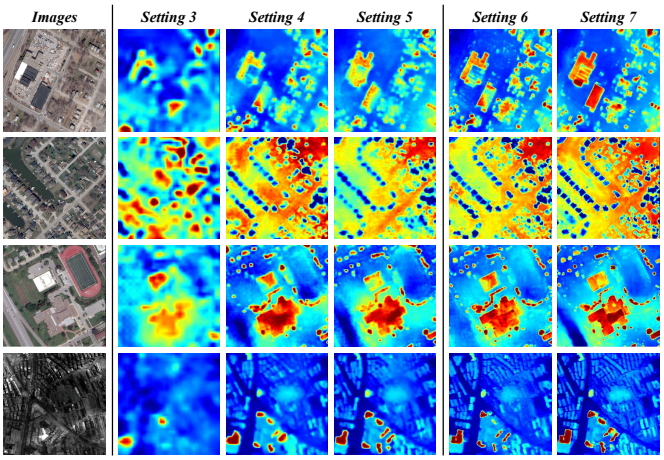

Figure 9: Coarse-stage disparity (d_S^1) visualization under progressive module addition. Each column (Settings 3–7) corresponds to an experimental configuration in Table 3. Progressive module addition recovers increasingly fine-grained disparity details, particularly in textureless regions and near object boundaries.

VFM injection. The qualitative results in Fig. 9 further confirm that Setting 7 produces disparity maps with better structural completeness and sharper edges than Setting 6.

GRU refinement. Adding the GRU-based iterative refinement module on top of full-pyramid VFM injection further improves the final accuracy. Compared with Setting 7, the full SatFS reduces the WHU-Stereo D1 from 9.83 to 9.57, corresponding to a 2.6% relative improvement, and slightly lowers the EPE from 1.34 to 1.33. On US3D, the D1 decreases from 5.77 to 5.75, and the EPE decreases from 1.05 to 1.01.

Notably, even without GRU refinement, Setting 7 already outperforms all compared methods on WHU-Stereo, with a D1 of 9.83 as reported in Tab. 2. This indicates that multi-scale VFM feature injection alone provides a strong matching representation. The GRU-based refinement further consolidates this advantage by applying geometry-aware iterative correction, enabling the full SatFS to achieve the best overall performance on WHU-Stereo, with a D1 of 9.57 and an EPE of 1.33.

Complexity Analysis of LGEV. We evaluate the proposed LGEV by replacing the pre-computed dense $H\times W\times W$ correlation volume with on-demand local sampling at $R \ll W$ positions during each iteration. This reduces the memory complexity from $O(HW^2)$ to $O(HWR)$.

As reported in Tab. 4, the lightweight formulation substantially reduces GPU memory consumption across different resolutions, sampling ranges, and batch sizes. At a resolution of 128×256 with R=9, memory usage decreases from 74 MB to 31 MB, corresponding to a 58.1% reduction. When the resolution increases to 256×512, memory consumption decreases from 488 MB to 124 MB under R=9, achieving a 74.6% reduction. With a larger sampling range of R=18, memory

Table 4
Memory and runtime comparison between dense all-pairs correlation and LGEV. Full denotes the pre-computed dense all-pairs correlation volume, whereas Light denotes the proposed lightweight geometry-aware encoding strategy. Larger memory savings are observed at higher spatial resolutions, with only marginal runtime overhead. R denotes the half-width of the local sampling window.

Resolution	B	R	Full (MB) ↓	Light (MB) ↓	Reduction (%) ↑	Overhead (%) ↓
128×256	1	9	74	31	58.1	0.6
256×512	1	9	488	124	74.6	6.0
256×512	1	18	600	236	60.7	7.8
256×512	2	9	976	248	74.6	7.8
512×1024	1	9	3488	496	85.8	7.8
1024×1024	1	9	6976	992	85.8	7.9

usage is reduced from 600 MB to 236 MB, corresponding to a 60.7% reduction. For high-resolution satellite imagery at 512×1024, LGEV further reduces memory usage from 3,488 MB to 496 MB, achieving an 85.8% reduction while introducing only 7.8% runtime overhead. Even at 1024×1024 resolution, its memory consumption remains below 1 GB, requiring 992 MB compared with 6,976 MB for dense correlation construction.

The memory consumption of LGEV scales linearly with the local sampling range R and batch size B, while remaining substantially lower than dense all-pairs correlation across all tested configurations. As illustrated in Fig. 10, this memory advantage becomes increasingly pronounced at higher spatial resolutions, which is consistent with the theoretical reduction from dense correlation construction to localised feature sampling.

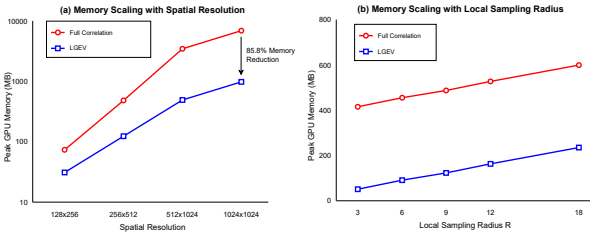

Figure 10: Empirical memory scaling analysis of LGEV. (a) Memory consumption under increasing spatial resolutions. Dense all-pairs correlation grows rapidly with image size, whereas LGEV achieves substantially better scalability through on-demand local sampling. (b) Memory consumption under different local sampling ranges R. Unlike dense correlation construction, LGEV scales linearly with the local sampling range, providing controllable memory complexity for large-range satellite stereo matching.

Despite this substantial memory reduction, LGEV preserves numerical equivalence to dense correlation lookup at the queried sampling locations. The maximum element-wise difference remains below 10^{-4}, indicating negligible numerical discrepancy, while the additional runtime overhead remains below 8%.

4.4. Generalisation Analysis

A robust satellite stereo matching method should generalise beyond the training distribution. We evaluate this ability along two complementary axes: *cross-dataset transferability* on WHU-SSIDE, which introduces wider disparity ranges and unseen test data from the same sensor family, and *multi-temporal robustness* on SatStereo, which contains substantial illumination and seasonal appearance variations across acquisition dates. In both settings, the trained model is directly evaluated without any fine-tuning or adaptation.

4.4.1. Cross-dataset Generalisation

To evaluate cross-dataset transferability under wider disparity ranges, we apply the model trained on US3D and WHU-Stereo directly to WHU-SSIDE. Although WHU-SSIDE shares the GF-7 sensor with WHU-Stereo, it provides a substantially wider disparity range of up to 384 pixels and contains unseen test data, making it a challenging benchmark for cross-dataset generalisation under increased matching difficulty.

As reported in Tab. 5, SatFS achieves the best performance on WHU-SSIDE across both metrics, with a D1 error rate of 10.30% and an EPE of 1.63 pixels. Compared with the second-best methods, namely FS in terms of D1 (11.45%) and MetaMRGE in terms of EPE (2.01), SatFS yields relative improvements of 10.0% and 19.0%, respectively.

The qualitative comparisons in Fig. 11 further demonstrate the robustness of SatFS in challenging urban scenes with severe occlusions and large disparity variations. Compared with the baselines, SatFS reconstructs more complete rooftop structures, handles occluded regions more consistently, and produces disparity predictions that are visually closer to the ground truth.

4.4.2. Multi-temporal Generalisation

To evaluate robustness to temporal appearance variations, we further test the trained model on SatStereo without fine-tuning or adaptation. SatStereo contains multi-date WorldView-3 and WorldView-2 imagery, introducing significant illumination and seasonal changes across acquisition dates.

As reported in Tab. 5, SatFS achieves the best EPE of 2.35 pixels and the second-best D1 error rate of 13.23%, slightly behind FS, which obtains a D1 of 12.42%. This indicates that SatFS reduces the average disparity error most effectively, although FS produces a slightly lower proportion of large-error pixels under the D1 threshold. In contrast, MetaMRGE and IGEV++ exhibit substantial performance degradation, with D1 error rates of 30.72% and 24.92%, respectively. This suggests that their strong performance on in-domain benchmarks does not directly transfer to multi-temporal satellite imagery.

Representative qualitative results are shown in Fig. 12. Although the quantitative metrics are computed only within the provided rooftop mask regions, SatFS also produces cleaner and more spatially consistent disparity predictions around building boundaries and surrounding non-building areas. These results demonstrate its stronger robustness to temporal appearance variations.

Table 5
Generalisation evaluation on SatStereo and WHU-SSIDE. SatStereo is used for multi-temporal evaluation, while WHU-SSIDE is used for cross-dataset evaluation. The best and second-best results are highlighted in **red** and blue, respectively.

Method	SatStereo		WHU-SSIDE	
	D1↓	EPE↓	D1↓	EPE↓
CFNet	14.13	3.59	14.66	2.34
HMSMNet	18.11	4.00	14.61	2.75
IGEV++	24.92	13.70	18.07	2.68
FS	**12.42**	2.47	11.45	2.20
MetaMRGE	30.72	15.21	13.27	2.01
SatFS	13.23	**2.35**	**10.30**	**1.63**

Table 6
Real satellite imagery used for real-world application evaluation. Scenes with LiDAR-derived ground truth (GT) are used for quantitative DSM accuracy assessment, while the remaining scenes are used for qualitative evaluation.

Sensor	Region	GSD (m)	GT
GF-7	Hawaii, USA	0.8	LiDAR
WV	UCSD, San Diego, USA	0.5	LiDAR
BJ-3	Harbin, China	0.3	—
BJ-3	Urumqi, China	0.3	—
WV-3	New York, USA	0.3	—
WV-3	Nepal	0.3	—
WV-3	Nigeria	0.3	—
GeoEye-1	Düsseldorf, Germany	0.4	—

4.5. Real-world Application Analysis

To assess real-world applicability beyond curated benchmarks, we evaluate SatFS on a diverse collection of operational satellite stereo imagery spanning multiple unseen geographic regions, sensors, and scene types, as summarised in Tab. 6. The evaluation is organised into two tiers. First, we conduct quantitative DSM reconstruction assessment on scenes with LiDAR-derived ground truth, where the imaging sensors are consistent with those used during training. Second, we perform qualitative evaluation on scenes without ground truth, which further introduces unseen sensors and broader geographic diversity.

Following the standard satellite stereo reconstruction pipeline, each predicted disparity map is transformed into a 3D point cloud using the rational polynomial camera (RPC) model and the corresponding epipolar rectification parameters. The resulting point cloud is then interpolated to generate a gridded DSM. In addition to the deep learning baselines compared in Section 4.2, we include three commercial photogrammetry packages—Inpho (Trimble Inc., 2026), Metashape (Agisoft LLC, 2022), and G3D-SAT (DASpatial,

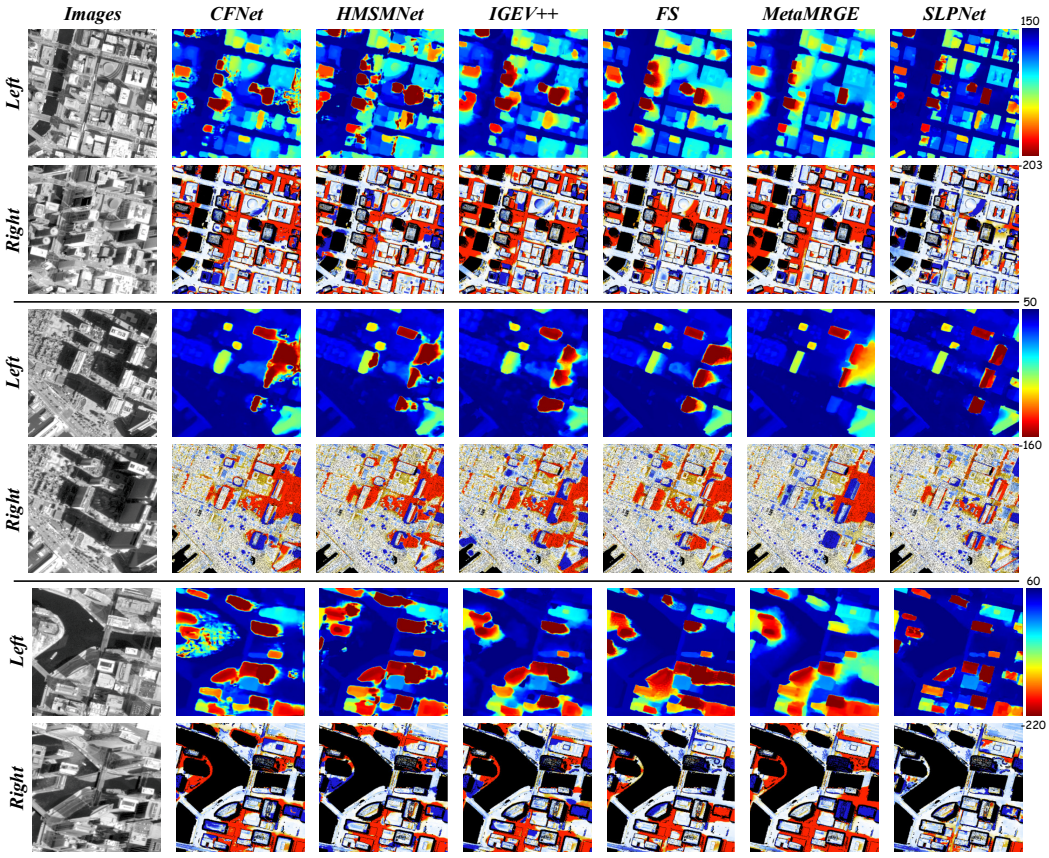

Figure 11: Qualitative comparison on WHU-SSIDE. SatFS reconstructs more complete rooftop structures and handles occluded regions more consistently in challenging urban scenes with large disparity variations. For each scene, the second row shows the disparity error map, computed as $\mathrm{pred} - \mathrm{GT}$ and colour-coded from red ($+10$ pixels) to blue (-10 pixels). Values outside $[-10, 10]$ pixels are clipped for visualisation.

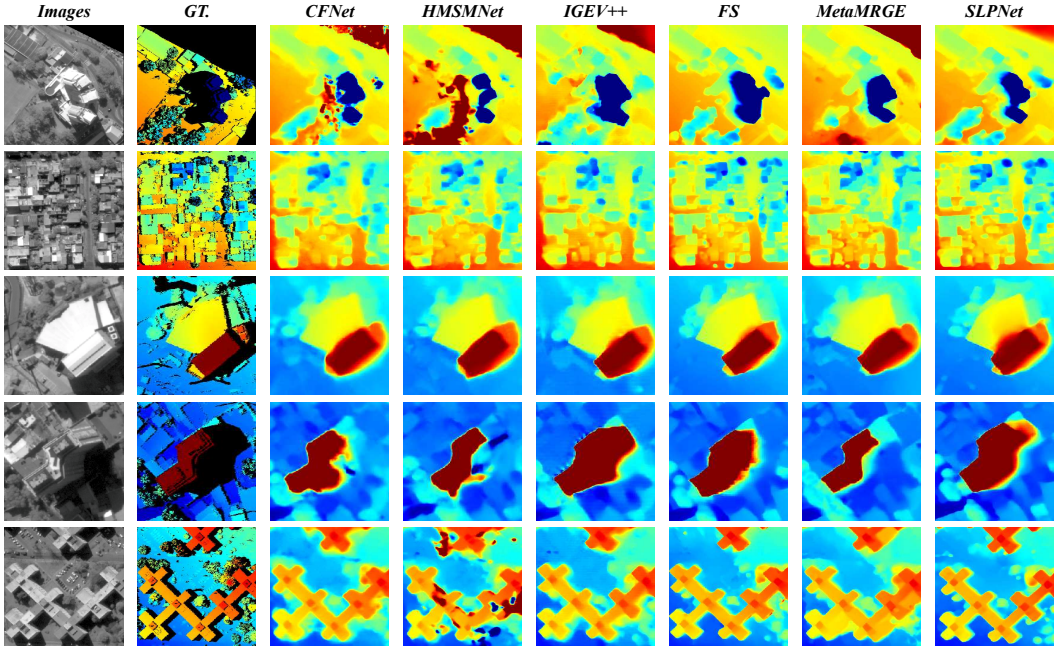

Figure 12: Qualitative comparison on SatStereo. SatStereo contains multi-date WorldView stereo pairs with significant cross-temporal appearance variations. Under these challenging acquisition conditions, SatFS produces more spatially consistent disparity predictions around building boundaries and within rooftop regions.

2026)—as production-grade reference pipelines widely used in operational satellite mapping.

4.5.1. LiDAR-based Quantitative Evaluation

We evaluate DSM reconstruction accuracy on two real-world scenes with available LiDAR-derived reference data. The first scene is a GF-7 stereo pair over Hawaii, with ground truth derived from USGS airborne LiDAR (OpenTopography and U.S. Geological Survey, 2017). The second scene is a WorldView stereo pair over the UCSD campus in San Diego, with ground truth obtained from the Merrick & Company LiDAR survey (Merrick, Inc. and OpenTopography, 2005). Both scenes are acquired by sensors consistent with those used in the training datasets, but they cover geographic regions and terrain types that are entirely unseen during training. They therefore provide a controlled setting for quantitatively assessing cross-geographic transferability while minimising the confounding effect of sensor-domain shifts.

Tab. 7 reports the overall and per-ROI DSM reconstruction accuracy on both scenes. SatFS achieves the lowest overall RMSE on both scenes and ranks first in RMSE across all five individual ROIs. On the GF-7 Hawaii scene, SatFS attains an RMSE of 2.47 m and an MAE of 1.26 m, outperforming the second-best G3D-SAT result (2.67 m / 1.33 m) by 7.5% and 5.3%, respectively. It also achieves the highest overall PAE values across the reported thresholds. On the WV UCSD scene, the advantage becomes more pronounced: SatFS reduces the RMSE to 3.33 m, which is 22.9% lower than CFNet, and reduces the MAE to 2.37 m, corresponding to a 12.5% improvement. In contrast, traditional photogrammetry pipelines such as Inpho and Metashape produce substantially larger errors, highlighting the advantage of learning-based methods for high-resolution satellite DSM reconstruction.

For the GF-7 Hawaii scene, we further analyse three ROIs with increasing scene complexity. SatFS consistently achieves the best RMSE, and its margin over the runner-up increases as the scene becomes more challenging. In the dense residential ROI-2, SatFS outperforms G3D-SAT by 8.5% in RMSE and 7.0% in MAE. In the textureless airport ROI-3, SatFS retains the best RMSE and PAE@2.5, indicating better robustness in weakly textured regions. The qualitative results in Fig. 13 further support these observations: SatFS preserves sharper structural boundaries and produces more coherent surfaces, whereas the competing methods tend to generate over-smoothed or fragmented reconstructions.

On the WV UCSD scene, a complementary pattern is observed. In ROI-1, where low-rise dense urban areas contain weak texture, most methods over-smooth rooftops, whereas SatFS better preserves fine-grained elevation variations. In the more challenging ROI-2, large buildings introduce occlusions and shadows, making complete reconstruction more difficult. SatFS not only recovers detailed rooftop structures but also generates more complete terrain in shadowed regions, achieving the largest performance gain in this evaluation, with an RMSE of 4.21 m and a 22.8% relative improvement. This trend is consistent with the visual comparison in Fig. 14, where other methods are more prone to structural gaps or geometric distortions under occlusion.

Overall, the advantage of SatFS mainly stems from its VFM-enhanced representation and coarse-to-fine matching strategy, which provide more stable correspondence cues under weak texture, degraded illumination, and complex structural conditions. Cost-volume-based learning methods such as CFNet and IGEV++ rely more heavily on local appearance similarity, leading to notable performance drops in low-texture or shadowed regions. Traditional photogrammetry pipelines such as Inpho and Metashape tend to produce noisier or fragmented surfaces, while G3D-SAT, although effective at noise suppression, suffers from over-smoothing and detail loss in vegetated and rooftop areas. It is worth noting that SatFS does not consistently lead at the strictest tolerance, PAE@1. In addition, off-nadir imagery may still produce local reconstruction gaps in regions occluded by tall buildings, suggesting that stronger geometric constraints or post-processing strategies could further improve DSM completeness and accuracy.

4.5.2. Cross-sensor Qualitative Evaluation

For satellite imagery without LiDAR-derived ground truth, we conduct a qualitative assessment focusing on structural completeness, edge fidelity, and fine-grained detail recovery. Fig. 15 provides a full-scene overview of SatFS DSM reconstruction across all cross-sensor test areas. As shown in Fig. 16, SatFS preserves richer structural details without post-processing while maintaining geometric smoothness comparable to G3D-SAT. Among the three commercial packages, G3D-SAT produces the smoothest DSMs but tends to over-smooth vegetated regions, Inpho often generates fragmented surfaces, and Metashape provides a more balanced result between smoothness and detail preservation.

On BJ-3 imagery (Fig. 16(a)), SatFS reconstructs fine-scale building structures in region a1 with fidelity comparable to G3D-SAT while retaining more rooftop details. In region a2, although severe off-nadir viewing occasionally leads to local holes, the surrounding disparity estimates remain consistent and geometrically plausible. In the vegetated region a3, SatFS and Metashape preserve vegetation details more faithfully, whereas G3D-SAT exhibits clear over-smoothing.

On WV-3 (Fig. 16(b)) and GeoEye-1 imagery (Fig. 16(c)), SatFS achieves building reconstruction quality comparable to G3D-SAT while producing noticeably better results in vegetated areas, where G3D-SAT tends to degrade due to over-smoothing. Specifically, in region b1, SatFS recovers herringbone rooftop structures and surrounding vegetation. In region b2, fine vegetation details are reconstructed with high fidelity to the original imagery. In region b3, high-rise rooftop structures are recovered more completely. In regions b4 and c1, SatFS preserves finer structural details around building interior edges and occlusion-induced gaps, maintaining good reconstruction quality even under cloud occlusion in scene c1.

Table 7
Quantitative DSM reconstruction comparison on the GF-7 Hawaii and WorldView UCSD scenes. ^{T} denotes the traditional stereo matching mode of G3D-SAT. The best and second-best results are highlighted in **red** and blue, respectively.

Site	ROI	Metric	Traditional Methods			Deep Learning						
			Inpho	Metashape	G3D-SAT^{T}	G3D-SAT	CFNet	FS	HMSMNet	IGEV++	MetaMRGE	SatFS
GF-7, Hawaii	All Sites	RMSE↓	6.98	7.00	2.98	2.67	2.75	3.15	2.88	3.10	3.02	**2.47**
		MAE↓	2.23	2.01	1.55	1.33	1.42	1.50	1.66	1.74	1.60	**1.26**
		PAE@1 ↑	58.64	55.30	60.41	67.04	62.81	63.01	53.13	53.69	60.49	**67.43**
		PAE@2.5 ↑	80.80	79.59	82.52	85.24	84.50	84.02	80.10	79.34	81.02	**86.65**
		PAE@7.5 ↑	95.18	96.20	97.26	97.91	97.82	97.53	97.59	96.84	97.10	**98.27**
	ROI-1	RMSE↓	6.37	6.36	3.81	3.44	3.30	4.34	3.31	3.34	3.38	**3.15**
		MAE↓	2.45	2.31	1.76	1.49	1.52	1.64	1.70	1.64	1.52	**1.41**
		PAE@1 ↑	56.01	54.18	58.86	65.68	64.25	64.06	55.84	59.10	65.02	**66.31**
		PAE@2.5 ↑	77.34	76.96	80.61	83.87	83.33	83.89	80.51	81.57	82.45	**84.48**
		PAE@7.5 ↑	93.35	94.60	96.23	97.15	97.10	96.89	97.03	96.93	97.14	**97.60**
	ROI-2	RMSE↓	11.01	10.37	2.15	1.99	2.02	2.13	2.40	2.74	2.77	**1.82**
		MAE↓	2.57	1.90	1.42	1.28	1.35	1.43	1.72	1.93	1.86	**1.19**
		PAE@1 ↑	52.08	50.44	54.60	58.49	55.09	53.22	42.20	40.33	45.59	**60.31**
		PAE@2.5 ↑	80.79	79.13	81.91	83.93	83.40	81.68	76.10	72.97	74.23	**86.55**
		PAE@7.5 ↑	97.50	98.67	99.22	99.41	99.38	99.26	98.97	97.89	97.54	**99.57**
	ROI-3	RMSE↓	3.56	4.27	2.98	2.57	2.93	2.98	2.92	3.23	2.93	**2.44**
		MAE↓	1.67	1.80	1.47	1.23	1.39	1.43	1.55	1.65	1.42	**1.17**
		PAE@1 ↑	67.83	61.27	67.78	**76.94**	69.09	71.75	61.35	61.63	70.85	75.67
		PAE@2.5 ↑	84.26	82.68	85.06	87.92	86.79	86.49	83.69	83.47	86.37	**88.93**
		PAE@7.5 ↑	94.70	95.31	96.32	97.18	96.99	96.43	96.76	95.71	96.64	**97.64**
WV, UCSD San Diego	All Sites	RMSE↓	5.99	13.87	4.60	4.41	4.32	5.24	4.76	4.87	5.41	**3.33**
		MAE↓	3.61	5.71	3.11	3.22	2.71	2.98	3.01	2.96	3.24	**2.37**
		PAE@1 ↑	14.94	14.29	11.72	7.41	16.73	18.18	16.74	**18.89**	13.05	15.54
		PAE@2.5 ↑	62.60	61.55	65.41	56.14	73.26	**74.85**	67.43	70.18	65.91	74.31
		PAE@7.5 ↑	89.60	85.66	92.36	93.96	94.39	92.92	92.64	93.21	92.45	**97.05**
	ROI-1	RMSE↓	3.54	10.67	3.08	3.37	2.72	2.51	2.79	2.57	3.19	**2.46**
		MAE↓	2.49	3.78	2.36	2.80	2.02	**1.89**	2.13	1.96	2.32	2.03
		PAE@1 ↑	17.48	17.18	14.17	7.26	19.78	22.96	19.93	**23.00**	15.90	16.82
		PAE@2.5 ↑	71.15	70.39	73.47	58.48	80.72	**83.94**	75.72	79.64	74.13	78.35
		PAE@7.5 ↑	95.86	93.18	96.99	97.12	98.16	98.22	97.74	98.24	97.46	**99.08**
	ROI-2	RMSE↓	8.44	17.08	6.13	5.45	5.91	7.97	6.72	7.17	7.63	**4.21**
		MAE↓	4.73	7.65	3.85	3.64	3.40	4.06	3.89	3.95	4.16	**2.72**
		PAE@1 ↑	12.40	11.41	9.28	7.57	13.68	13.40	13.56	**14.78**	10.19	14.27
		PAE@2.5 ↑	54.04	52.71	57.35	53.79	65.80	65.76	59.14	60.72	57.69	**70.27**
		PAE@7.5 ↑	83.35	78.14	87.73	90.80	90.63	87.62	87.55	88.19	87.44	**95.01**

Two consistent patterns emerge across all qualitative scenes. First, SatFS preserves structural details, including building edges, narrow roads, and vegetation, that traditional commercial pipelines often either over-smooth or fragment. Second, these advantages remain stable across different unseen sensors, including BJ-3 at 0.3 m, WV-3 at 0.3 m, and GeoEye-1 at 0.4 m GSD, without any fine-tuning. This demonstrates the strong cross-sensor generalisation capability of SatFS in real-world satellite DSM reconstruction.

Taken together, the LiDAR-based quantitative evaluation and the cross-sensor qualitative assessment show that SatFS predictions can be reliably converted into metrically accurate 3D reconstructions, while also maintaining visually coherent and geometrically consistent DSM quality across unseen sensors and geographic regions without ground truth.

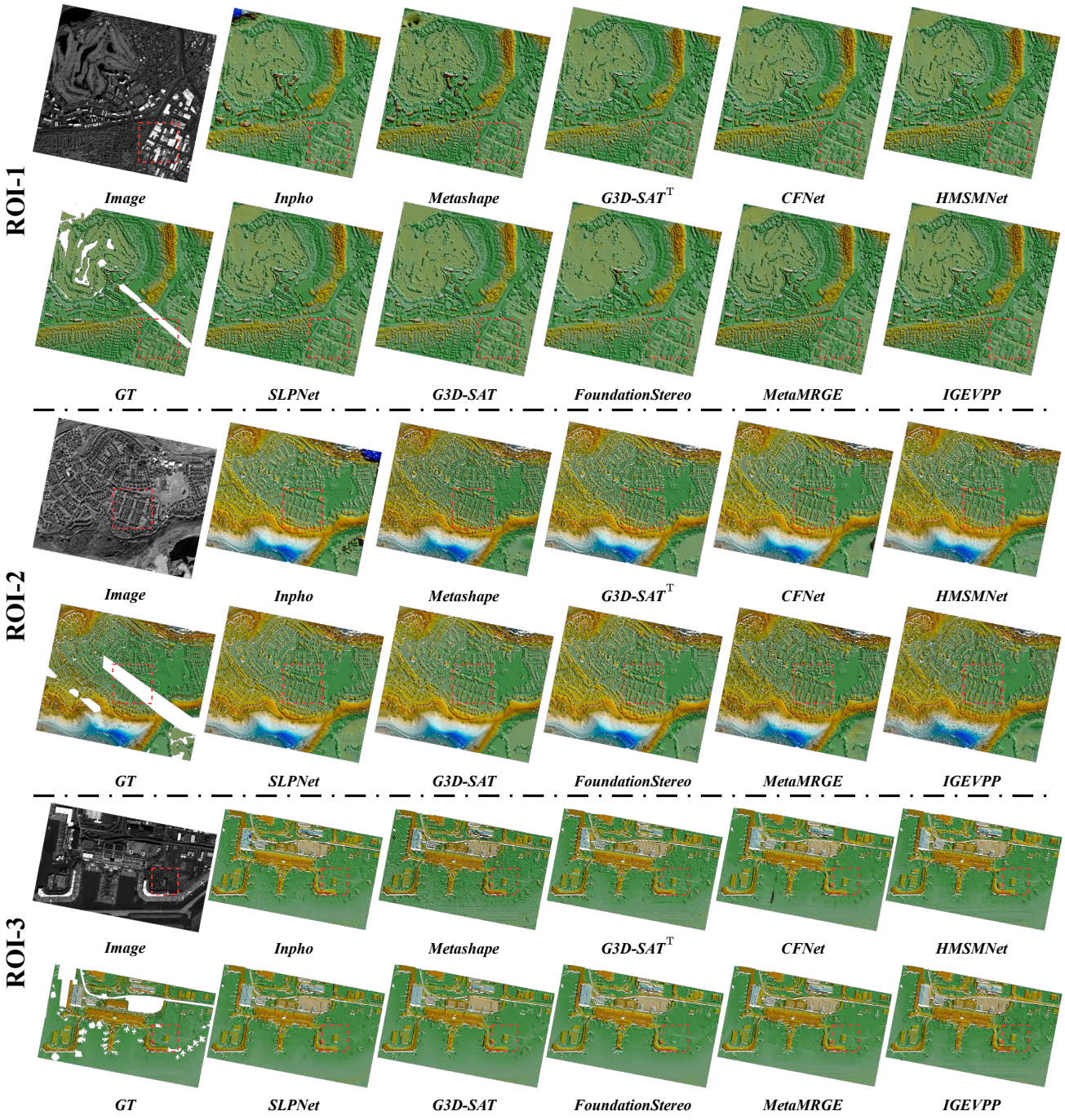

Figure 13: Qualitative DSM comparison on the GF-7 Hawaii scene.

4.6. Discussion and Limitations

Despite the promising results demonstrated above, several limitations of the proposed approach warrant discussion.

First, in regions with severe off-nadir viewing geometry, SatFS occasionally produces incomplete DSMs with localised reconstruction gaps, as illustrated in the WV-3 New York scene of Fig. 15. These gaps mainly arise from inherent occlusions caused by tall buildings under large incidence angles. In such cases, facade or ground regions behind buildings may be invisible in one or both stereo views, leading to missing or unreliable 3D points during DSM fusion. Although spatial interpolation can fill these voids to some extent, it often produces artificially smooth elevation transitions near building boundaries, which are inconsistent with the sharp vertical discontinuities of man-made urban structures.

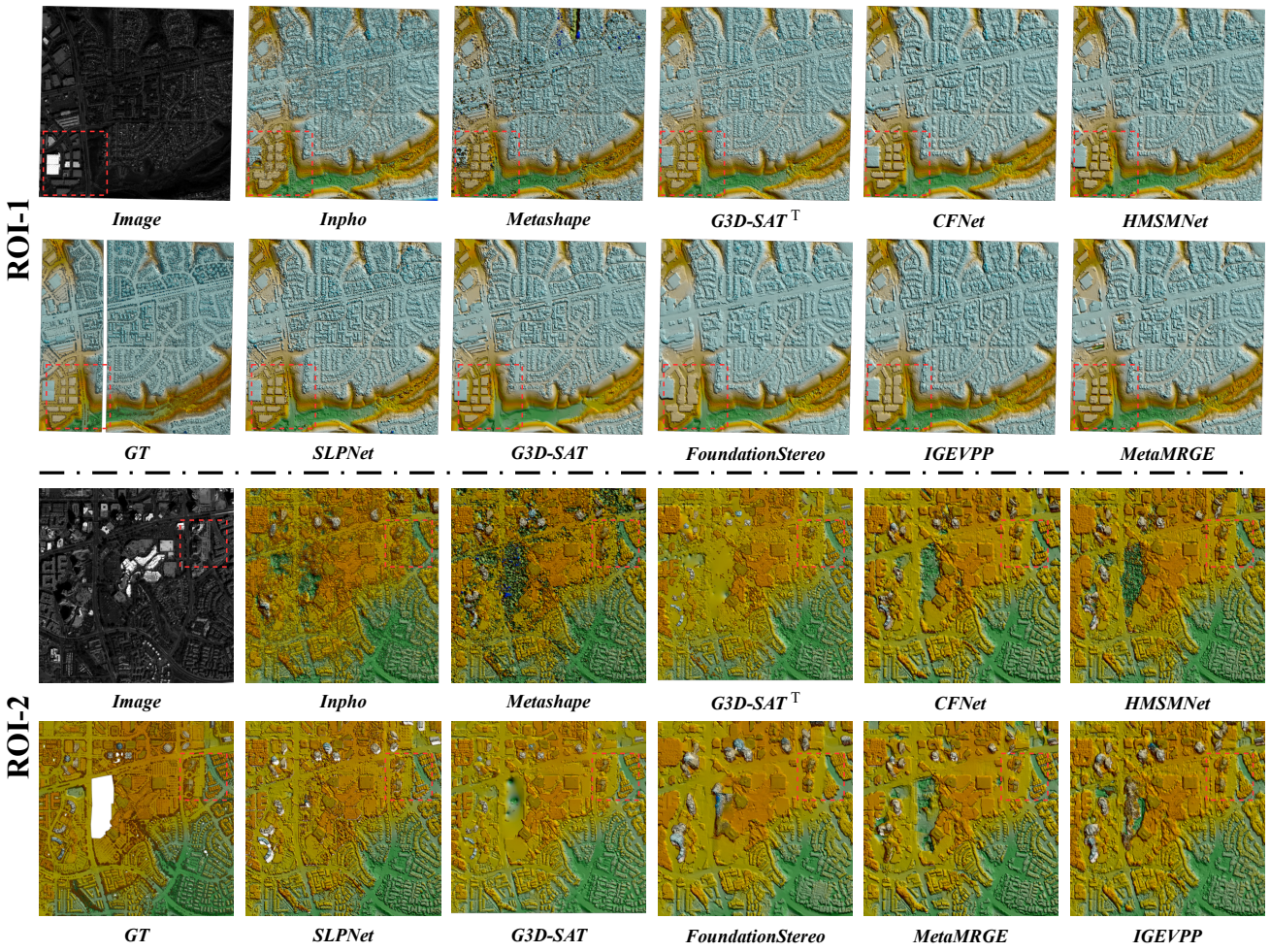

Figure 14: Qualitative DSM comparison on the WorldView UCSD scene.

Second, DSMs produced by SatFS exhibit slightly higher noise over building rooftops than those generated by the best-performing commercial baseline, G3D-SAT. This difference likely stems from the stronger geometric regularisation embedded in commercial photogrammetry pipelines, where rooftop planar constraints and multi-scale surface refinement are explicitly enforced to improve structural consistency. In contrast, SatFS reconstructs DSMs directly from predicted disparities without task-specific surface regularisation or post-processing. This limitation is also common to learning-based stereo methods, which are primarily optimised for pixel-wise disparity accuracy rather than surface-level regularity. Incorporating building-aware planar constraints or differentiable surface refinement into the reconstruction stage is therefore a promising direction for reducing rooftop noise.

Third, the quality and consistency of existing satellite stereo benchmarks also limit both training supervision and quantitative evaluation. Most benchmarks are constructed by co-registering airborne LiDAR acquisitions with satellite imagery, a process that is vulnerable to geometric inconsistencies caused by temporal offsets between the two modalities. These inconsistencies may arise from urban development, vegetation change, seasonal variation, or transient scene elements, making it difficult to distinguish true model errors from label noise. This issue complicates the assessment of model generalisation and highlights the need for temporally co-registered, multi-sensor satellite stereo benchmarks.

These limitations suggest several directions for future research. Occlusion-induced reconstruction gaps could be mitigated by incorporating point cloud classification to distinguish ground from non-ground regions, allowing interpolation to be applied selectively to ground surfaces while preserving sharp building boundaries. Rooftop noise could be reduced by introducing geometry-aware regularisation, such as building planar constraints or piecewise-smooth surface constraints, either as a post-processing step or within a differentiable end-to-end training objective. Finally, constructing benchmarks with tighter temporal co-registration between reference data and satellite observations would provide more reliable supervision and evaluation, enabling a more rigorous assessment of generalisation across diverse sensors and geographic regions.

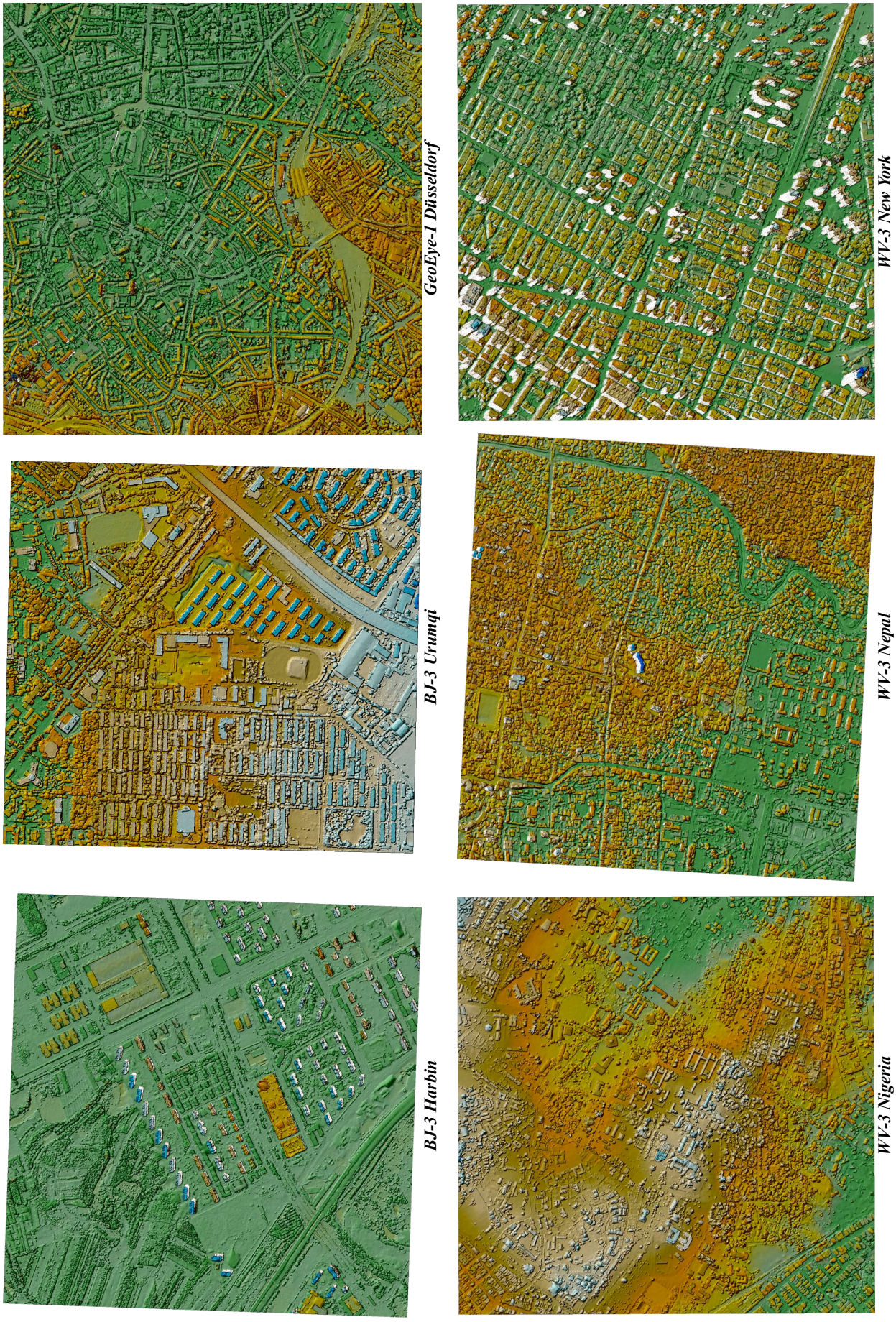

Figure 15: **Full-scene DSM reconstruction overview on cross-sensor satellite imagery.** SatFS produces structurally complete and geometrically consistent DSMs across all evaluated scenes without fine-tuning or post-processing.

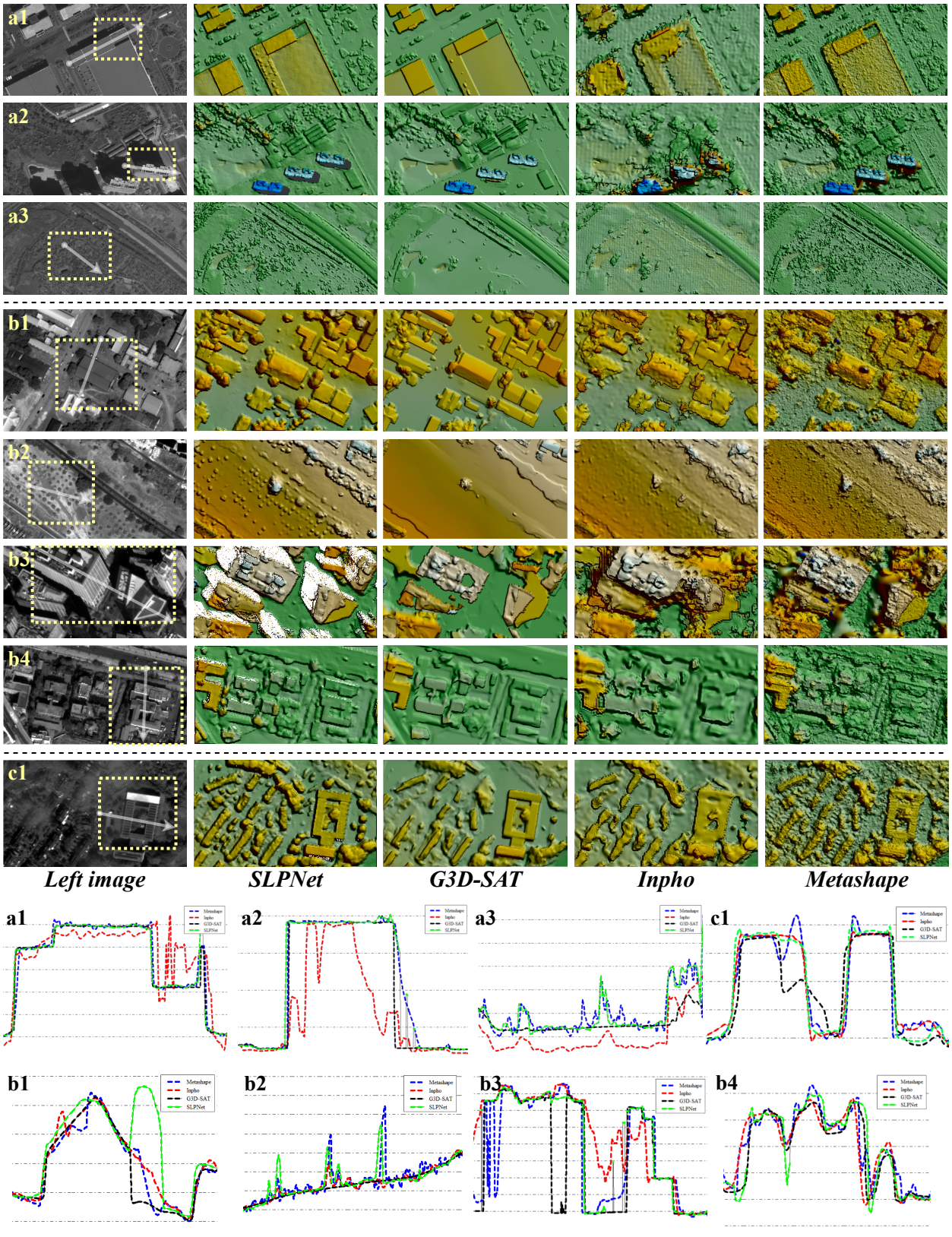

Figure 16: Cross-sensor qualitative DSM comparison on BJ-3, WV-3, and GeoEye-1 imagery. Panels (a), (b), and (c) show representative results on BJ-3, WV-3, and GeoEye-1 scenes, respectively. We visualise local reconstruction details and elevation profiles to compare structural completeness, edge fidelity, and fine-scale geometric preservation across different sensors without LiDAR supervision.

5. Conclusion

This paper presented SatFS, a satellite foundation stereo framework that adapts VFM-based stereo matching to high-resolution satellite photogrammetry. Rather than directly transferring a ground-level foundation stereo model, SatFS combines frozen VFM representations with satellite-specific geometric reasoning. It introduces multi-scale VFM side-tuning for robust feature adaptation, a cascade coarse-to-fine architecture for large bidirectional disparity search, Prior-guided Bilateral Upsampling for structure-preserving cost volume propagation, and LGEV for memory-efficient iterative refinement. Together, these components address the main challenges of satellite stereo matching: wide positive and negative disparity ranges, radiometric and temporal appearance variations, and fragile correspondence around sharp urban structures.

Extensive experiments confirm the effectiveness of SatFS across both disparity estimation and DSM reconstruction tasks. On in-distribution benchmarks, SatFS achieves the best results on WHU-Stereo, with a D1 error rate of 9.57% and an EPE of 1.33 px, and obtains competitive performance on US3D, with a D1 error rate of 5.75% and an EPE of 1.01 px. Under zero-shot transfer to WHU-SSIDE, SatFS improves D1 and EPE by 10.0% and 19.0% over the closest competitors, demonstrating stronger generalisation to wider disparity ranges. On SatStereo, SatFS obtains the best EPE under multi-temporal WorldView imagery, indicating improved robustness to cross-date appearance changes. For DSM reconstruction, SatFS achieves the lowest RMSE values of 2.47 m on GF-7 and 3.33 m on WorldView scenes, outperforming the best competing methods by 7.5% and 22.9%, respectively. Ablation studies further show that multi-scale VFM injection, PBU, GRU-based iterative refinement, and LGEV contribute to accuracy, structural preservation, and memory efficiency.

Several limitations remain. Severe off-nadir viewing can still produce local reconstruction gaps in regions occluded by tall buildings, and SatFS may retain more rooftop noise than commercial photogrammetry pipelines that explicitly enforce surface regularisation. In addition, current satellite stereo benchmarks are affected by temporal misalignment between LiDAR reference data and satellite imagery, which limits both training supervision and quantitative evaluation. Future work should therefore explore occlusion-aware point cloud completion, building-aware planar or piecewise-smooth surface constraints, and temporally co-registered multi-sensor benchmarks. These extensions would further improve DSM completeness and provide a more rigorous basis for evaluating cross-sensor and cross-region generalisation.

Acknowledgement

This study was supported in part by the National Natural Science Foundation of China (Project No. U25A20772, 42230102) and the Natural Science Foundation of Sichuan Province under Grant 2026NSFSCZY0054.

References

Agisoft LLC, 2022. Agisoft metashape – intelligent photogrammetry enhanced with lidar data processing, version 1.8.4.

Albanwan, H., Qin, R., 2022. A comparative study on deep-learning methods for dense image matching of multi-angle and multi-date remote sensing stereo-images. The Photogrammetric Record 37, 385–409.

Barron, J.T., Adams, A., Shih, Y., Hernandez, C., 2015. Fast bilateral-space stereo for synthetic defocus, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4466–4474.

Bartolomei, L., Tosi, F., Poggi, M., Mattoccia, S., 2025. Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail, in: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1013–1027.

Bosch, M., Foster, K., Christie, G., Wang, S., Hager, G.D., Brown, M., 2019. Semantic stereo for incidental satellite images, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1524–1532.

Chang, J.R., Chen, Y.S., 2018. Pyramid stereo matching network, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE. pp. 5410–5418.

Chen, L., Wang, W., Mordohai, P., 2023. Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 17235–17244.

Cheng, J., Liu, L., Xu, G., Wang, X., Zhang, Z., Deng, Y., Zang, J., Chen, Y., Cai, Z., Yang, X., 2025. Monster: Marry monodepth to stereo unleashes power, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6273–6282.

Cheng, J., Xu, G., Guo, P., Yang, X., 2023. Coatrsnet: Fully exploiting convolution and attention for stereo matching by region separation. International Journal of Computer Vision 132, 56–73.

DASpatial, 2026. Get3d cluster satellite.

De Franchis, C., Meinhardt-Llopis, E., Michel, J., Morel, J.M., Facciolo, G., 2014. Automatic sensor orientation refinement of pléiades stereo images, in: 2014 IEEE Geoscience and Remote Sensing Symposium, pp. 1639–1642.

Gong, R., Liu, W., Gu, Z., Yang, X., Cheng, J., 2024. Learning intra-view and cross-view geometric knowledge for stereo matching, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 20752–20762.

Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P., 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 2492–2501.

Guo, X., Yang, K., Yang, W., Wang, X., Li, H., 2019. Group-wise correlation stereo network, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 3268–3277.

He, S., Li, S., Jiang, S., Jiang, W., 2022. HMSM-net: Hierarchical multi-scale matching network for disparity estimation of high-resolution satellite stereo images. ISPRS Journal of Photogrammetry and Remote Sensing 188, 314–330.

He, X., Jiang, S., He, S., Li, Q., Jiang, W., Wang, L., 2023. Deep learning-based stereo matching for high-resolution satellite images: A comparative evaluation. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLVIII-1-W2-2023, 1635–1642.

Hirschmuller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutual information, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 807–814 vol. 2.

Hu, H., Su, L., Mao, S., Chen, M., Pan, G., Xu, B., Zhu, Q., 2023. Adaptive region aggregation for multi-view stereo matching using deformable convolutional networks. The Photogrammetric Record .

Jiang, H., Lou, Z., Ding, L., Xu, R., Tan, M., Jiang, W., Huang, R., 2025a. Defom-stereo: Depth foundation model based stereo matching, in: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21857–21867.

Jiang, L., Wang, F., Zhang, W., Li, P., You, H., Xiang, Y., 2025b. Rethinking the key factors for the generalization of remote sensing stereo matching networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 18, 4936–4948.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A., 2017. End-to-End Learning of Geometry and Context for Deep Stereo Regression, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE. pp. 66–75.

Kim, J., Cho, S., Chung, M., Kim, Y., 2025. Improving disparity consistency with self-refined cost volumes for deep learning-based satellite stereo matching. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 18, 9262–9278.

Lipson, L., Teed, Z., Deng, J., 2021. Raft-stereo: Multilevel recurrent field transforms for stereo matching, in: 2021 International conference on 3D vision (3DV), IEEE. pp. 218–227.

Maaz, M., Shaker, A., Cholakkal, H., Khan, S., Zamir, S.W., Anwer, R.M., Shahbaz Khan, F., 2022. Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications, in: Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII, Springer-Verlag, Berlin, Heidelberg. p. 3–20.

Mao, Y., Liu, Z., Li, W., Dai, Y., Wang, Q., Kim, Y.T., Lee, H.S., 2021. UASNet: Uncertainty adaptive sampling network for deep stereo matching, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE. pp. 6291–6299.

Merrick, Inc., OpenTopography, 2005. 2005 san diego urban region lidar. Funded by City of San Diego; airborne LiDAR, 1.41 pts/m^2, survey area 1,190 km^2.

OpenTopography, U.S. Geological Survey, 2017. Usgs lidar point cloud hi hawaii island 2017.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2023. Dinov2: Learning robust visual features without supervision.

Patil, S., Comandur, B., Prakash, T., Kak, A.C., 2019. A new stereo benchmarking dataset for satellite images. arXiv:1907.04404.

Rao, Z., Li, X., Xiong, B., Dai, Y., Shen, Z., Li, H., Lou, Y., 2024. Cascaded recurrent networks with masked representation learning for stereo matching of high-resolution satellite images. ISPRS Journal of Photogrammetry and Remote Sensing 218, 151–165.

Richardt, C., Orr, D., Davies, I., Criminisi, A., Dodgson, N.A., 2010. Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid, in: Daniilidis, K., Maragos, P., Paragios, N. (Eds.), Computer Vision – ECCV 2010, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 510–523.

Schönberger, J.L., Sinha, S.N., Pollefeys, M., 2018. Learning to fuse proposals from multiple scanline optimizations in semi-global matching, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. Springer International Publishing. volume 11217, pp. 758–775.

Seki, A., Pollefeys, M., 2017. SGM-nets: Semi-global matching with neural networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 6640–6649.

Shean, D.E., Alexandrov, O., Moratto, Z.M., Smith, B.E., Joughin, I.R., Porter, C., Morin, P., 2016. An automated, open-source pipeline for mass production of digital elevation models (dems) from very-high-resolution commercial stereo satellite imagery. ISPRS Journal of Photogrammetry and Remote Sensing 116, 101–117.

Shen, Z., Dai, Y., Rao, Z., 2021. Cfnet: Cascade and fused cost volume for robust stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13906–13915.

Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P., 2025. DINOv3. arXiv:2508.10104.

Sun, H., Wang, T., Cheng, Q., Huang, J., 2026. RSMT: Robust stereo matching training with geometric correction, clean pixel selection and loss weighting. ISPRS Journal of Photogrammetry and Remote Sensing 232, 421–436.

Trimble Inc., 2026. Trimble inpho photogrammetry software.

Wang, X., Yang, H., Wang, H., Cheng, J., Xu, G., Lin, M., Yang, X., 2026. PromptStereo: Zero-shot stereo matching via structure and motion prompts. arXiv:2603.01650.

Wang, Y., Wen, Z., Huang, X., 2025. MSCA-net: Multiscale chunked attention network for high-resolution satellite stereo matching. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 18, 27745–27763.

Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S., 2025. Foundationstereo: Zero-shot stereo matching, in: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 5249–5260.

Xu, B., Xu, Y., Yang, X., Jia, W., Guo, Y., 2021. Bilateral grid learning for stereo matching networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1–10.

Xu, G., Liu, J., Wang, X., Cheng, J., Deng, Y., Zang, J., Chen, Y., Yang, X., 2025a. BANet: Bilateral Aggregation Network for Mobile Stereo Matching, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 28870–28880.

Xu, G., Wang, X., Ding, X., Yang, X., 2023. Iterative geometry encoding volume for stereo matching, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21919–21928.

Xu, G., Wang, X., Zhang, Z., Cheng, J., Liao, C., Yang, X., 2025b. Igev++: Iterative multi-range geometry encoding volumes for stereo matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 47, 7108–7122.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth anything v2, in: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 21875–21911.

Yang, M., Jiang, S., Jiang, W., Li, Q., 2025. Stereo matching network with transformer-CNN feature fusion and ConvGRU refinement for high-resolution satellite stereo images. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences X-G-2025, 995–1002.

Youssefi, D., Michel, J., Sarrazin, E., Buffe, F., Cournet, M., Delvit, J.M., L'Helguen, C., Melet, O., Emilien, A., Bosman, J., 2020. CARS: A photogrammetry pipeline using dask graphs to construct a global 3D model, in: IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, pp. 453–456.

Zhang, G., Jiang, Y., Wei, S., Shen, X., Wu, K., 2026. Towards robust disparity estimation in satellite stereo imagery: A new high-quality benchmark dataset and a metadata-informed multi-range geometric encoding network. ISPRS Journal of Photogrammetry and Remote Sensing 235, 511–535.

Zhang, Y., Wang, L., Li, K., Wang, Y., Guo, Y., 2025. Learning representations from foundation models for domain generalized stereo matching, in: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (Eds.), Computer Vision – ECCV 2024. Springer Nature Switzerland. volume 15100, pp. 146–162.

Zheng, Z., Wan, Y., Zhang, Y., Hu, Z., Wei, D., Yao, Y., Zhu, C., Yang, K., Xiao, R., 2024. Digital surface model generation from high-resolution satellite stereos based on hybrid feature fusion network. The Photogrammetric Record 39, 36–66.