ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XI-2-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission II, 4–11 July 2026, Toronto, Canada

Diachronic Stereo Matching for Multi-Date Satellite Imagery

Elías Masquil[1], Luca Savant Aira[2], Roger Marí[3], Thibaud Ehret[4], Pablo Musé[1,5], Gabriele Facciolo[5,6]

[1] IIE, Facultad de Ingeniería, Universidad de la República, Uruguay
[2] Politecnico di Torino, Corso Duca degli Abruzzi, Torino TO, Italia
[3] Eurecat, Centre Tecnologic de Catalunya, Multimedia Technologies, Barcelona, Spain [4] AMIAD, Pôle Recherche, France
[5] Universite Paris-Saclay, CNRS, ENS Paris-Saclay, Centre Borelli, Gif-sur-Yvette, France [6] Institut Universitaire de France

Keywords: Stereo Matching, 3D Reconstruction, Diachronic Matching, Multi-Date Satellite Images

Abstract

Recent advances in image-based satellite 3D reconstruction have progressed along two complementary directions. On one hand, multi-date approaches using NeRF or Gaussian-splatting jointly model appearance and geometry across many acquisitions, achieving accurate reconstructions on opportunistic imagery with numerous observations. On the other hand, classical stereoscopic reconstruction pipelines deliver robust and scalable results for simultaneous or quasi-simultaneous image pairs. However, when the two images are captured months apart, strong seasonal, illumination, and shadow changes violate standard stereoscopic assumptions, causing existing pipelines to fail. This work presents the first Diachronic Stereo Matching method for satellite imagery, enabling reliable 3D reconstruction from temporally distant pairs. Two advances make this possible: (1) fine-tuning a state-of-the-art deep stereo network that leverages monocular depth priors, and (2) exposing it to a dataset specifically curated to include a diverse set of diachronic image pairs. In particular, we start from a pretrained MonSter model, trained initially on a mix of synthetic and real datasets such as SceneFlow and KITTI, and fine-tune it on a set of stereo pairs derived from the DFC2019 remote sensing challenge. This dataset contains both synchronic and diachronic pairs under diverse seasonal and illumination conditions. Experiments on multi-date WorldView-3 imagery demonstrate that our approach consistently surpasses classical pipelines and unadapted deep stereo models on both synchronic and diachronic settings. Fine-tuning on temporally diverse images, together with monocular priors, proves essential for enabling 3D reconstruction from previously incompatible acquisition dates.

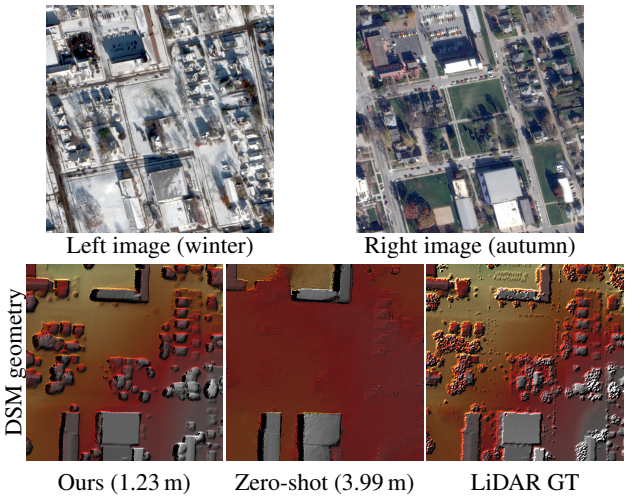

Figure 1. Output geometry for a winter-autumn image pair from Omaha (OMA_331 test scene). Our method recovers accurate geometry despite the diachronic nature of the pair, exhibiting strong appearance changes, which cause existing zero-shot methods to fail. Missing values due to perspective shown in black. Mean altitude error in parentheses; lower is better.

1. Introduction

Stereoscopic reconstruction models are becoming increasingly powerful, with the latest advances in the state of the art driven by the integration of monocular priors (Cheng et al., 2025, Bartolomei et al., 2025, Wen et al., 2025). In the remote sensing community, progress often builds on these general advances, with models adapted to the particular requirements of satellite imagery. After fine-tuning on the target domain, these models consistently achieve superior performance, surpassing classical techniques. Deep stereo matching architectures now outperform long-standing semi-global matching algorithms (Hirschmuller, 2007, Facciolo et al., 2015, Marí et al., 2022), commonly used in satellite stereoscopic pipelines, such as S2P (Amadei et al., 2025, De Franchis et al., 2014) and ASP (Beyer et al., 2018).

Despite this progress, significant challenges remain. As highlighted in (Tosi et al., 2025), one persistent difficulty arises in scenarios with challenging weather conditions. In Earth observation systems, diachronic stereoscopic pairs, i.e., images of the same geographical area captured at different dates with different viewing angles, are common. In contrast, true stereo acquisitions from the same date are more costly and less frequent. Consequently, addressing stereo reconstruction from diachronic images is of practical importance. As noted in (Facciolo et al., 2017, Marí et al., 2022), performance deteriorates as the temporal gap between image pairs increases. This degradation becomes particularly severe under substantial seasonal differences, such as snow versus no-snow conditions, where both traditional and learning-based methods fail dramatically. Moreover, as discussed in (Amadei et al., 2025), pipelines originally designed for near-simultaneous imagery struggle with the specific challenges of multi-temporal stereo, including variations in lighting, shadows, weather conditions, and moving objects, all of which adversely affect the accuracy of geometric reconstruction.

In this work, we present the first model capable of reliably performing diachronic satellite stereo matching, see Fig.1. To

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-483-2026 | © Author(s) 2026. CC BY 4.0 License.

enable this, we fine-tune a pretrained MonSter model (Cheng et al., 2025) on a stereoscopic dataset derived from Track 3 of the *DFC2019* remote sensing challenge (Bosch et al., 2019, Le Saux et al., 2019), specifically curated to include both synchronic and diachronic pairs selected according to temporal and geometric criteria. Each pair was rectified to ensure consistent disparity direction and alignment with reference Digital Surface Models (DSMs) derived from the LiDAR ground truth. We chose MonSter as it leverages monocular depth estimates, which are largely invariant to appearance changes unrelated to geometric structure.

Our approach achieves state-of-the-art performance in terms of the mean absolute error (MAE) between the reference and reconstructed DSMs, derived from the estimated disparities. We demonstrate this through an extensive evaluation, using pairs from the *DFC2019 Track 3* data, covering areas in Jacksonville and Omaha (USA), and pairs from the *IARPA 2016* dataset (Bosch et al., 2016), covering areas near Buenos Aires (Argentina). All models are trained exclusively on a subset of *DFC2019 Track 3*, and tested on previously unseen sites, such as the Buenos Aires areas. The *IARPA 2016* data and a subset of *DFC2019* Jacksonville scenes (specifically JAX_004, JAX_068, JAX_214, and JAX_260) have become benchmark sites for satellite multi-view reconstruction, used to evaluate recent methods such as EO-NeRF (Marí et al., 2023) and EOGS (Savant Aira et al., 2025). In our evaluation, Jacksonville scenes serve as a reference for synchronic stereo, while Omaha and Buenos Aires sites provide more challenging diachronic conditions, characterized by large temporal gaps and substantial appearance variations.

Across all these datasets, we observe a clear performance hierarchy: zero-shot state-of-the-art models, such as MonSter (Cheng et al., 2025), FoundationStereo (Wen et al., 2025), and StereoAnywhere (Bartolomei et al., 2025), underperform relative to models fine-tuned on domain-specific data. Fine-tuning MonSter only on synchronic pairs leads to significant improvements, while fine-tuning on combined synchronic–diachronic pairs yields the best overall results. Our fine-tuned model also outperforms both the classical s2p-hd pipeline (Amadei et al., 2025) and previous learning-based methods such as RAFT-Stereo (Lipson et al., 2021), even when fine-tuned on our dataset. Furthermore, in aerial imagery benchmarks such as Enschede and EuroSDR-Vaihingen (Wu et al., 2024), our model performs on par with or surpasses competing methods.

In summary, this work introduces two key contributions:

- We present the first model capable of reliably performing diachronic stereo matching, enabling 3D reconstruction from pairs of satellite images acquired at distant dates and under strong appearance changes. We show that both the inclusion of diachronic data during training and the use of monocular depth priors are key for achieving these results.
- We release[1] a curated dataset for stereo matching in satellite imagery, including ground-truth disparities and DSMs, organized into synchronic and diachronic pairs. We also propose a simple heuristic based on metadata and image appearance to label pairs as diachronic or synchronic.

[1] https://centreborelli.github.io/diachronic-stereo

2. Related Work

2.1 Stereoscopic Reconstruction

We refer the reader to the comprehensive survey by (Tosi et al., 2025) for an in-depth review of the most recent advances in stereo matching. The past five years have seen remarkable progress, driven by key architectural designs and novel paradigms for addressing open challenges such as domain generalization, accuracy, over-smoothing of predictions, and efficiency.

Starting from RAFT-Stereo (Lipson et al., 2021), iterative optimization-based architectures have achieved impressive results in stereoscopic reconstruction. IGEV (Xu et al., 2023) further advanced RAFT's iterative cost-volume refinement paradigm, reaching state-of-the-art performance and later serving as the backbone of MonSter (Cheng et al., 2025).

Beyond architectural innovations, recent years have also been characterized by the extensive use of synthetic data. Synthetic datasets, e.g. (Mayer et al., 2016), have played a crucial role in the pre-training of stereo networks (Tosi et al., 2025). Although zero-shot adaptation to real imagery was initially challenging, recent works trained exclusively on synthetic data have demonstrated remarkable performance and strong generalization to real-world settings (Lipson et al., 2021, Xu et al., 2023, Tosi et al., 2023, Wen et al., 2025, Cheng et al., 2025, Bartolomei et al., 2025). It is also worth noting that the final boost in accuracy is commonly achieved when fine-tuning on the real target domain becomes feasible.

Recently, a new wave of stereo matching models has emerged that leverage monocular depth priors to improve performance in ill-posed scenarios such as occlusions and textureless regions (Cheng et al., 2025, Wen et al., 2025, Bartolomei et al., 2025). Of particular relevance to this work is MonSter (Cheng et al., 2025), which we adopt as our base model, as it is the only one that has released the complete training code.

As noted by (Tosi et al., 2025), despite the remarkable progress achieved in recent years, several challenges remain open. Among them are handling very high-resolution imagery, coping with challenging weather conditions, addressing ill-posed scenes, and recovering fine structural details.

2.2 Stereovision for Satellite Imagery

Recent advances in 3D reconstruction from multiple images have been driven by Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023). In the context of satellite imagery, this trend is no exception. Recent works have adapted these approaches to the satellite domain, including EO-NeRF (Marí et al., 2023) and EOGS (Savant Aira et al., 2025), based on NeRF and 3DGS, respectively. These methods achieve remarkably strong results, both in appearance modeling and 3D reconstruction, but also have limitations. While they can handle complex illumination changes, they fail to capture severe seasonal variations such as snow versus no-snow conditions. Moreover, their performance relies on the availability of multiple views, typically five or more, and degrades significantly when only a few images are available (Zhang and Rupnik, 2023, Masquil et al., 2025).

On the other hand, when only two images are available and captured within a short time span, scalable stereo pipelines based on semi-global matching (Hirschmuller, 2007) perform very

well (Amadei et al., 2025, De Franchis et al., 2014). These approaches can be extended to multi-date configurations by pairing temporally close images and fusing the reconstructed DSMs to form a consistent 3D model (Facciolo et al., 2017, Gómez et al., 2023). Such pairwise fusion often yields better results than traditional multi-view methods (Facciolo et al., 2017).

Recent studies, such as (Wu et al., 2024), have analyzed state-of-the-art learning-based stereo matching methods in the context of aerial imagery. They found that domain adaptation plays a crucial role, with substantial gains observed after fine-tuning models on the target domain. Similarly, (Marí et al., 2022) demonstrated that deep learning methods generally outperform classical matching-based algorithms in satellite stereo reconstruction under ideal conditions, and that pretrained aerial models can adapt well to satellite data. However, these models require careful input preprocessing to match training conditions and often produce incomplete reconstructions in complex or unusual scenarios (e.g., very distant acquisition dates or narrow baselines).

Existing datasets for stereo matching in satellite imagery have been reviewed in (Patil and Guo, 2023), which also introduces a new large-scale benchmark. They highlight several limitations of previous datasets, including limited geographic coverage, lack of diversity, and most critically, the presence of *bipolar* disparities, i.e., disparities with mixed signs, which are not compatible with the most recent approaches. Although their dataset effectively addresses many of these issues, it remains unsuitable for our purposes. In particular, their rectification process results in extremely wide disparity ranges, which must be constrained for deep stereo models to operate effectively and to enable manageable crop sizes, as their distributed image tiles are approximately 5000×5000 pixels. Moreover, the image pairs are provided without the corresponding camera models or rectification homographies, making it impossible to evaluate DSM reconstructions and restricting assessment to disparity errors alone.

3. Fine-Tuning MonSter for Diachronic Stereo Matching

We describe the methodology used to fine-tune MonSter for reliable diachronic stereo matching in satellite imagery. Sections 3.1-3.4 cover the main stages of the training process: (i) dataset curation and RPC-based rectification of image pairs to enforce unipolar positive horizontal disparities and derive disparity supervision from DSMs; and (ii) fine-tuning a state-of-the-art stereo matching network (MonSter) on a balanced mix of synchronic and diachronic pairs. The predicted disparities are triangulated using the RPC models and projected to produce a DSM (Section 3.5). The reconstructed DSMs are used for evaluation.

3.1 Dataset Curation

We use the WorldView-3 RGB crops from Track 3 of the *DFC2019* challenge (Bosch et al., 2019, Le Saux et al., 2019) and their corresponding ground-truth DSMs built from LiDAR to derive disparities. The dataset comprises 110 areas of interest (AOIs), consisting of 54 in Jacksonville (JAX) and 56 in Omaha (OMA). On average, each JAX scene comprises approximately 20 images, while each OMA site features around 32.

In this work, we label image pairs according to the temporal gap between the two images and their visual similarity, measured by the number of SIFT (Lowe, 2004) feature matches. This SIFT-based criterion is effective in practice, as these features are highly sensitive to appearance variations such as those caused by seasonal changes (Marí et al., 2019). Thus, we define

- **Diachronic pairs** as pairs of satellite images of the same geographic area acquired more than 30 days apart (modulo 1 year), and exhibiting fewer than 40 SIFT matches between them.
- **Synchronic pairs** as image pairs acquired within a 30-day interval and having at least 40 SIFT matches between them. We note that the 40-match threshold could be normalized with respect to the surface area to be image size agnostic.

To build a diverse dataset with an emphasis on diachronic conditions, we iterate over all AOIs and extract, for each one, at least 30 diachronic pairs in Omaha and 3 diachronic pairs in Jacksonville. Additionally, we randomly sample 5 synchronic pairs per site to balance the dataset. Diachronic effects are less frequent in Jacksonville, where vegetation and lighting remain more consistent throughout the year, resulting in more minor appearance differences even between acquisitions made several months apart. This *diachronic+synchronic* dataset is the main training resource used in this work. Additionally, we build a *synchronic-only* dataset with approximately 15 synchronic pairs per AOI, which is used for ablation studies.

For each selected pair, we rectify the images and compute a ground-truth disparity map using the reference DSM and RPC camera models, as detailed in Section 3.2. In total, we obtain 2,246 pairs in the diachronic+synchronic dataset and 1,567 pairs in the synchronic-only dataset.

3.2 Rectification of Multi-Date Satellite Imagery

This section describes the procedure used to rectify multi-date satellite image pairs into a geometry compatible with stereo networks (Algorithm 1) and the procedure for computing ground-truth disparities from DSMs for supervised training (Algorithm 2).

Deep stereo architectures, including MonSter, require disparities to be *unipolar*-that is, pixels in the right image should appear shifted to the left relative to those in the left image– and to *increase* with the underlying altitude, so that higher elevations correspond to larger disparity values. Enforcing these geometric constraints is nontrivial, as it requires identifying at least a few reliable image matches, *correspondences* across diachronic pairs, where substantial appearance changes due to season, illumination, or vegetation often make classical feature matching difficult.

To address this, we design a rectification strategy (Algorithm 1) that transforms arbitrary multi-date pairs into rectified stereo pairs compatible with MonSter's architecture and training distribution. The algorithm progressively refines a pair of rectifying homographies. It is initialized using RPC-based virtual correspondences following the approach of (De Franchis et al., 2014), then refined by fitting a horizontal shear transformation to reduce the disparity range. Because virtual correspondences alone are insufficient to guarantee *unipolar* disparities, a small number of actual matches are required across the diachronic pair. We employ DISK (Tyszkiewicz et al., 2020) for keypoint detection and LightGlue (Lindenberger et al., 2023) for feature matching with conservative matching settings, as this combination provides greater robustness to seasonal and illumination variations than

traditional descriptors such as SIFT (Figure 2). Similar to (Marí et al., 2022), these sparse matches are used to estimate a global horizontal translation (and, if necessary, a polarity swap) that enforces unipolar and altitude-increasing disparities. The resulting rectifying homographies are further refined in order to minimize any remaining vertical alignment error, yielding a final configuration that satisfies the geometric assumptions of modern deep stereo architectures even for strongly diachronic imagery.

For completeness, Algorithm 2 outlines how we derive the reference disparities from the ground-truth DSM. This step, required for fine-tuning and benchmarking, reprojects each pixel in the left rectified view through the DSM and camera models to compute the corresponding horizontal displacement in the right view, producing accurate disparity supervision in rectified coordinates.

Algorithm 1 Diachronic rectification

Require: Left image I_L with its camera model RPC_L, right image I_R with its camera model RPC_R, average altitude estimate z_{avg}
Ensure: Rectified images $\hat{I}_L$ and $\hat{I}_R$, rectifying homographies H_L and H_R
▷ Run (De Franchis et al., 2014) Algorithm 1.
$(H_L, H_R) \leftarrow \mathrm{RPC\_Rectification}(\mathrm{RPC}_L, \mathrm{RPC}_R)$
▷ Reduce disparity range with tilt, shear, and translation using RPC matches at z_{avg}
$H_R \leftarrow \mathrm{ReduceDispRange}(H_L, H_R, \mathrm{RPC}_L, \mathrm{RPC}_R, z_{\mathrm{avg}})$
if $\mathrm{DispDecreasesWithAltitude}(H_R, H_L, RPC_R, RPC_L)$ **then**
▷ Reverse the role of left and right images
$I_L, I_R \leftarrow I_R, I_L$
$H_L, L_R \leftarrow H_R, H_L$
end if
$(\hat{I}_L, \hat{I}_R) \leftarrow \mathrm{Warp}(I_L, H_L), \mathrm{Warp}(I_R, H_R)$
▷ Extract robust keypoint matches
$M \leftarrow \mathrm{LightGlueMatching}(\hat{I}_L, \hat{I}_R)$
▷ Enforce positive unipolar disparities
$t \leftarrow \min_{(u_L, v_L) \leftrightarrow (u_R, v_R) \in M}(u_L - u_R)$
▷ Correct residual vertical alignment errors
$s \leftarrow \mathrm{median}_{(u_L, v_L) \leftrightarrow (u_R, v_R) \in M}(v_L - v_R)$
$H_R \leftarrow \begin{bmatrix} 1 & 0 & t \\ 0 & 1 & s \\ 0 & 0 & 1 \end{bmatrix} H_R$
$(\hat{I}_L, \hat{I}_R) \leftarrow \mathrm{Warp}(I_L, H_L), \mathrm{Warp}(I_R, H_R)$
return $\hat{I}_L, \hat{I}_R, H_L, H_R$

Algorithm 2 Ground truth disparities computation

Require: Rectified left image $\hat{I}_L$, rectifying homographies H_L and H_R, ground truth DSM, camera models RPC_L and RPC_R
Ensure: Ground truth disparity D
$\mathrm{DSM}_L \leftarrow \mathrm{ProjectDSM}(\mathrm{DSM}, \mathrm{RPC}_L)$
for each pixel (u, v) in $\hat{I}_L$ **do**
$(x, y) \leftarrow H_L^{-1}(u, v)$
$z \leftarrow \mathrm{DSM}_L(x, y)$
$(u_R, v_R) \leftarrow H_R \circ \mathrm{RPC}_R \circ \mathrm{RPC}_L^{-1}(x, y, z)$
$D(u, v) \leftarrow u - u_R$
end for
return D

3.3 Stereo Architecture

Our approach builds upon MonSter (Cheng et al., 2025), a recent state-of-the-art stereo model that couples monocular priors with stereo matching. MonSter makes extensive use of the monocular depth model Depth Anything V2 (Yang et al., 2024), employing it both as a backbone for feature extraction and as a depth estimator. The architecture consists of two main branches and a mutual refinement module. The stereo branch extracts features from the left and right images and follows the IGEV framework (Xu et al., 2023) to produce an initial disparity estimate. The monocular branch predicts an initial relative depth map, which is converted into a metric monocular disparity using the initial stereo prediction. Finally, both the metric monocular depth and stereo disparities are mutually refined through an iterative refinement stage to obtain the final disparity estimate.

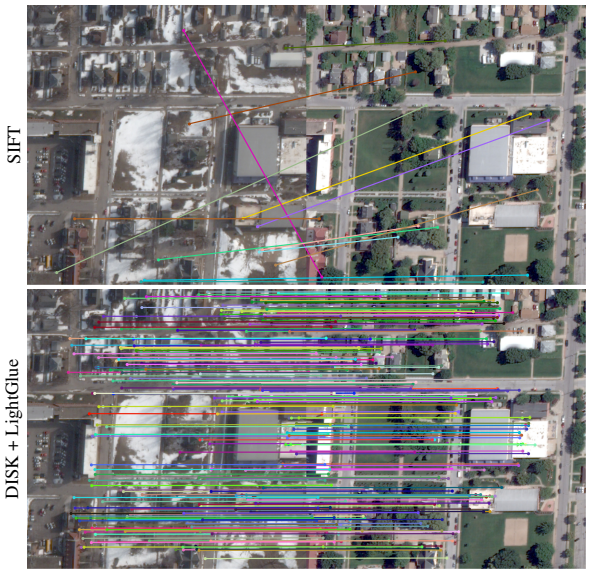

Figure 2. Comparison of SIFT and DISK+LightGlue matches for the diachronic pair OMA_331_017 – OMA_331_036.

A key motivation for using monocular priors in diachronic stereo is their invariance to image appearance changes unrelated to geometric structure. Seasonal or illumination variations, such as those between winter and summer conditions or different shadows, can significantly alter pixel intensities corresponding to the same underlying geometry. While monocular depth models cannot produce metrically consistent values across images, they preserve the geometric structure of the scene. As a result, the features extracted by such models remain stable under varying conditions, providing strong and complementary cues for reliable correspondence estimation in diachronic matching. In our experiments, we observed that the monocular predictions of the same area across different seasons exhibit similar geometric structure despite large radiometric differences between the input images (cf. Figure 3).

3.4 Fine-Tuning Details

We fine-tune the MonSter model starting from its publicly available *mix_all* checkpoint, pretrained on a combination of the KITTI 2012 (Geiger et al., 2012), KITTI 2015 (Menze and Geiger, 2015), Middlebury (Scharstein et al., 2014), ETH3D (Schops et al., 2017), and SceneFlow (Mayer et al., 2016) datasets. The model is optimized for 50,000 iterations using the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 5×10^{-4} and a weight decay of 1×10^{-5}. Fine-tuning was performed on a single NVIDIA RTX 6000 GPU with a total training time of approximately 24 hours.

Input images are kept in the original dynamic range of $[0, 255]$ and randomly cropped to 512×512 pixels for training. To prevent edge effects, a 32-pixel border is excluded from the loss computation. The fine-tuning configuration follows the default

Method type	Model (checkpoint) {fine-tuning}	Jacksonville	Buenos Aires	Omaha Synchronic	Omaha Diachronic
classical pipeline	s2p-hd	2.04 ± 0.69	2.64 ± 0.78	1.33 ± 0.65	7.90 ± 3.68
stereo networks (without monocular priors)	RAFT-Stereo (sceneflow)	6.01 ± 3.64	5.07 ± 1.36	1.44 ± 0.81	2.24 ± 0.99
	RAFT-Stereo {fine-tuned}	1.73 ± 0.61	2.08 ± 0.32	0.89 ± 0.35	1.04 ± 0.44
stereo networks (with monocular priors)	FoundationStereo (23-51-11)	2.33 ± 0.89	7.18 ± 4.41	1.42 ± 1.45	1.30 ± 0.95
	StereoAnywhere (sceneflow)	5.06 ± 2.04	3.53 ± 0.41	1.04 ± 0.44	1.38 ± 0.59
	MonSter (mix all)	2.15 ± 0.63	2.55 ± 0.45	0.92 ± 0.34	1.61 ± 0.63
	Ours {*synchronic-only*}	1.29 ± 0.52	1.65 ± 0.10	$\mathbf{0.76 \pm 0.37}$	0.97 ± 0.44
	Ours {*diachronic+synchronic*}	$\mathbf{1.20 \pm 0.52}$	$\mathbf{1.55 \pm 0.22}$	0.77 ± 0.33	$\mathbf{0.84 \pm 0.34}$

Table 1. Evaluation on all satellite test sets for all methods: Altitude MAE in meters.

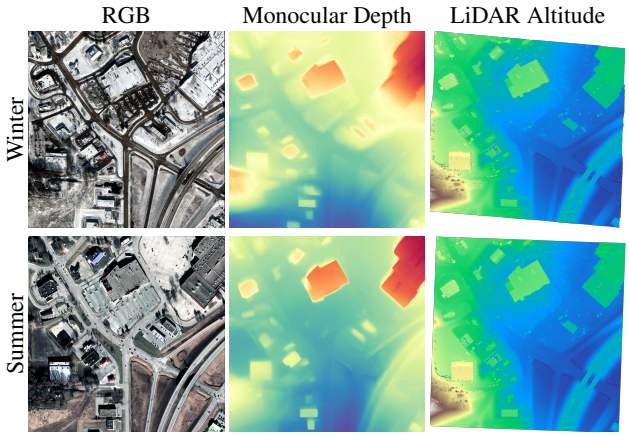

Figure 3. Monocular depth consistency across seasons. Despite large radiometric differences between seasonal images from the Omaha OMA_893 scene, the monocular Depth Anything V2 predictions exhibit coherent geometric structure.

MonSter setup, including the data augmentation strategies derived from RAFT-Stereo. Model performance is monitored using the absolute disparity error (end-point error) on the validation set, and the best checkpoint is selected based on this metric.

3.5 DSM Reconstruction

Given a predicted disparity map, we want to reconstruct a DSM. First, we estimate an altitude value for each pixel in the rectified left view, using the iterative triangulation strategy described in Algorithm 2 of (De Franchis et al., 2014). This algorithm searches for the height that best satisfies the epipolar constraint between the two RPC projections. We refer to this collection of all the estimated altitude values as an *altitude image*. We then reproject the altitude image into a uniform ground-aligned grid using the left RPC model, obtaining a DSM.

4. Experiments

We evaluate our approach on four test sets derived from DFC2019 Track 3 (Bosch et al., 2019, Le Saux et al., 2019) and the IARPA2016 dataset (Bosch et al., 2016), commonly used in satellite multi-view reconstruction benchmarks:

- Jacksonville (DFC2019): Four AOIs from Jacksonville, characterized by minimal seasonal or shadow changes between image pairs. This subset represents a synchronic test scenario.
- Buenos Aires (IARPA2016): Three AOIs from Buenos Aires with illumination and shadow differences between acquisitions. This represents a soft diachronic setting.
- Omaha Synchronic (DFC2019): Five AOIs from Omaha (OMA_084, OMA_134, OMA_230, OMA_247, OMA_331) using only synchronic image pairs.
- Omaha Diachronic (DFC2019): The same five AOIs from Omaha, but using diachronic image pairs, showing strong seasonal contrasts between distant acquisitions, most notably snow versus no-snow conditions. This represents a strong diachronic scenario.

All models are trained on the Jacksonville and Omaha AOIs, excluding those reserved for testing. For each test AOI, we evaluate on twenty stereo pairs. For Jacksonville and Buenos Aires, pairs are randomly selected from all available combinations. For Omaha, we randomly pick 20 pairs for each test set according to the definitions from Section 3.1. All data and splits are shared to ensure reproducibility and to support future research.

Our primary metric is the pixelwise mean absolute error (MAE) between the reconstructed DSM and the LiDAR reference DSM. For each AOI, we evaluate it across multiple stereo pairs and calculate the median MAE as the AOI score. Since the difficulty varies significantly across regions, this aggregation produces a single, stable value that is robust to outliers and enables fair comparison across them. We then report the mean and standard deviation of these AOI scores as the score of each dataset.

Since vegetation height varies seasonally and may differ between image acquisitions and LiDAR capture dates, we also exclude vegetation regions. These ignored pixels are obtained from classification masks of the DSM provided in the original datasets. The LiDAR DSMs have a spatial resolution of 30–50 cm/pixel. To avoid boundary artifacts, a 32-pixel margin is cropped from all DSMs before error and loss computations, during training and testing.

4.1 Diachronic Fine-Tuning of MonSter

Table 1 reports the quantitative results on all test sets using the classical s2p-hd pipeline, different state-of-the-art stereo models (Masquil et al., 2026) (RAFT-Stereo (Lipson et al., 2021), FoundationStereo (Wen et al., 2025), StereoAnywhere (Bartolomei et al., 2025), MonSter (Cheng et al., 2025)), and our MonSter model fine-tuned with and without diachronic pairs (*diachronic+synchronic* vs *synchronic-only*, respectively). All third-party methods are executed with their default configurations. Across all test sets, our model consistently achieves the lowest MAE, demonstrating its robustness to diachronic conditions. Reduced variance indicates improved reliability and robustness, with no catastrophic errors.

Figures 4 and 5 show qualitative results comparing our fine-tuned model against other zero-shot and fine-tuned baselines and

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-483-2026

AOI	Pair	#SIFT	FoundationStereo	StereoAnywhere	MonSter	s2p-hd	Ours (Sync)	**Ours (Dia+Sync)**
IARPA_003	12JUN15-01SEP15	8	9.84	2.24	2.21	1.66	1.59	**1.38**
JAX_260	018-005	7	1.41	8.63	3.63	3.90	1.17	**0.79**
OMA_084	042-016	0	4.87	11.92	4.53	8.98	1.69	**1.30**
OMA_134	029-013	0	**0.89***	2.82	3.58	16.97	2.39	1.33
OMA_230	013-031	0	0.68	8.51	1.21	12.33	0.46	**0.34**
OMA_247	009-023	7	1.10	1.29	3.08	12.21	0.86	**0.66**
OMA_331	017-036	7	1.75	3.10	2.68	4.14	1.22	**0.88**

Table 2. Altitude MAE (m) for challenging diachronic test pairs exhibiting few SIFT matches. * FoundationStereo’s prediction for OMA_134 is nearly flat, producing a deceptively small error despite a visually incorrect reconstruction, as shown in Figure 4.

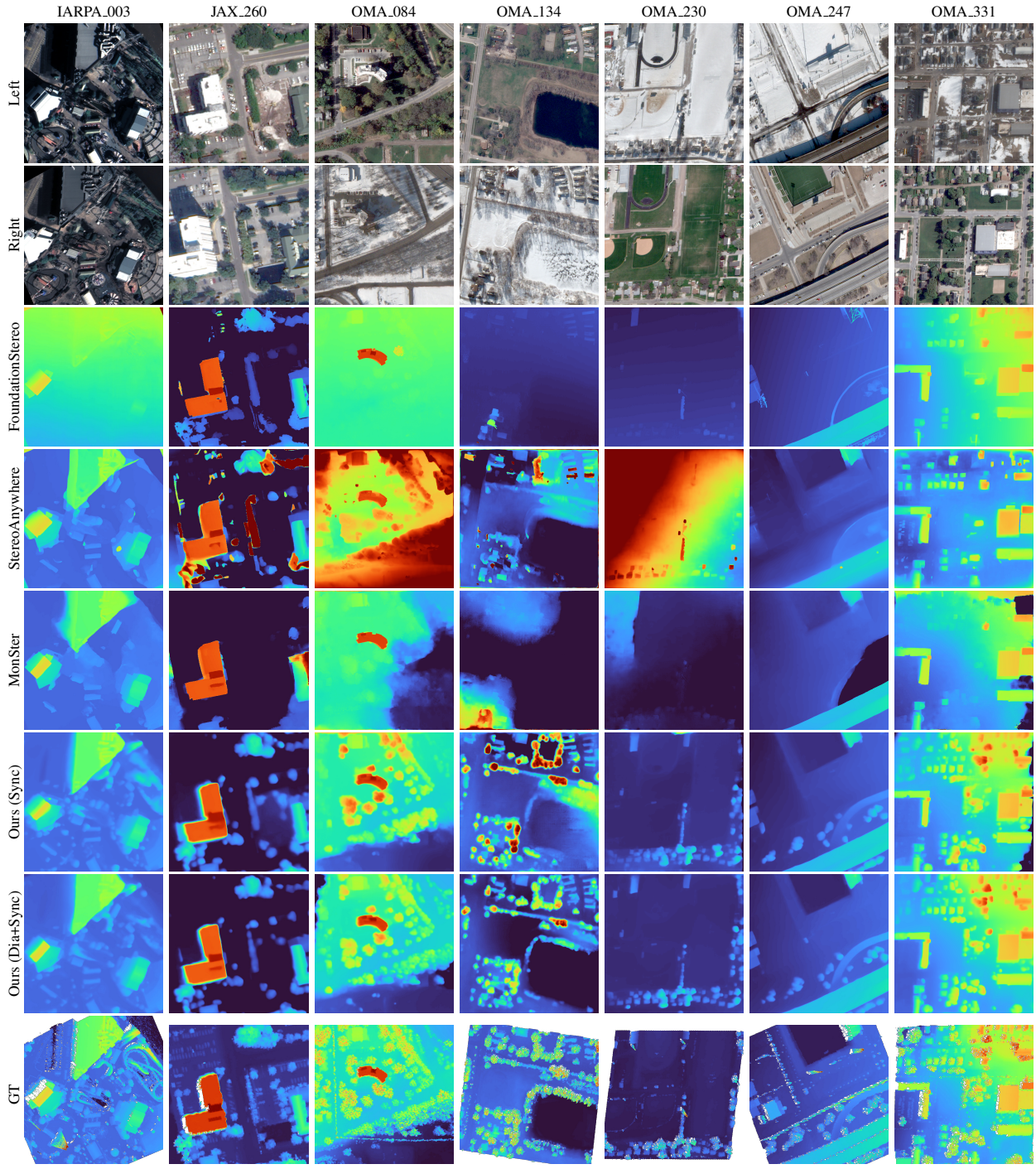

Figure 4. Qualitative results of disparity predictions on a selection of hard diachronic image pairs from the test set, listed in Table 2.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-483-2026 | © Author(s) 2026. CC BY 4.0 License.

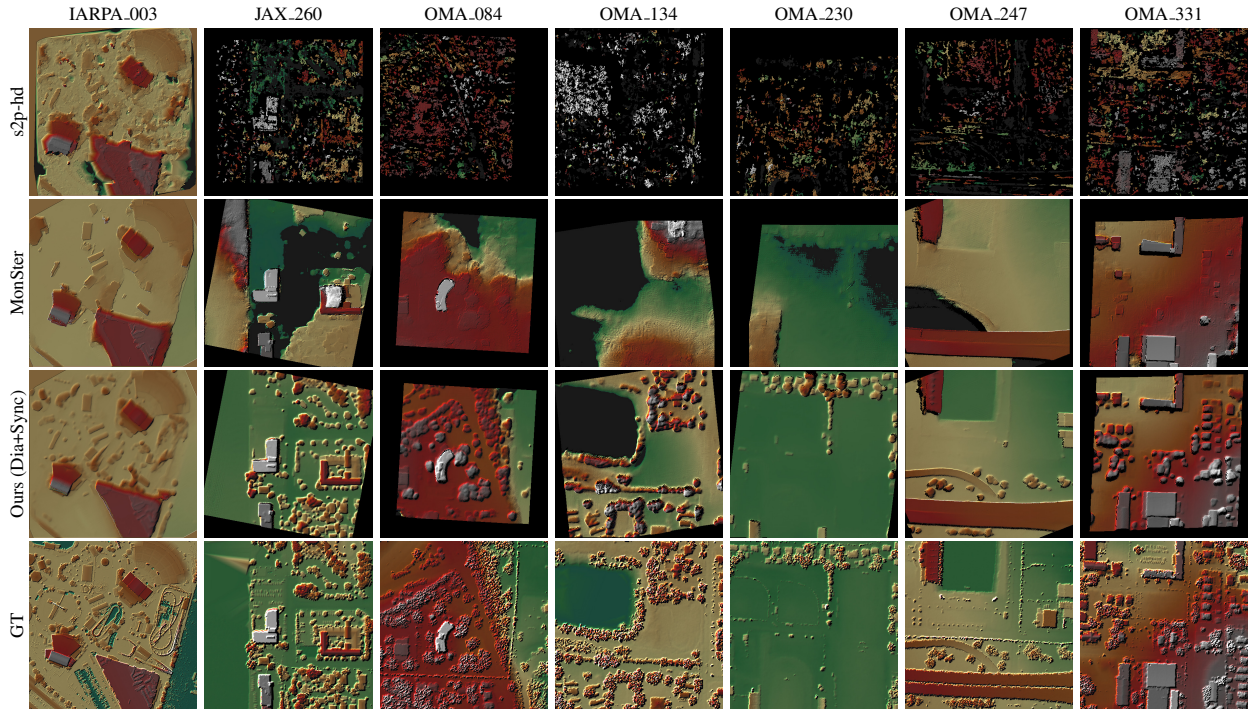

Figure 5. Top to bottom: DSMs produced by the classical s2p-hd pipeline, zero-shot MonSter (mix all), our method, and the ground-truth DSM on a selection of hard diachronic image pairs from the test set, listed in Table 2. Missing values shown in black.

s2p-hd, on a selection of test pairs. For each Omaha and Buenos Aires AOI, we deliberately select pairs with minimal SIFT feature matches, which correlate with severe appearance changes and thus represent the most challenging diachronic scenarios. The zero-shot stereo networks and s2p-hd fail under substantial seasonal variations, producing noisy (s2p-hd), incomplete (MonSter, StereoAnywhere) or over-smoothed surfaces (FoundationStereo) whereas our approach maintains consistent accuracy. The quantitative metrics for these pairs are provided in Table 2.

Impact of Monocular Priors To assess the contribution of monocular priors, we compare our method with a similarly fine-tuned RAFT-Stereo baseline using the *diachronic+synchronic* training data. RAFT-Stereo (Lipson et al., 2021) serves as the foundation for several recent architectures (Cheng et al., 2025, Bartolomei et al., 2025). However, unlike MonSter, it does not incorporate any monocular depth prior. Both models are fine-tuned using identical settings, data splits, and training duration to ensure a fair comparison. As shown in Table 1, our fine-tuned MonSter significantly outperforms the fine-tuned version of RAFT-Stereo.

Impact of Diachronic Training Data We further analyze the role of diachronic training data by comparing two variants of our model: one fine-tuned exclusively on the synchronic-only dataset and another fine-tuned on the combined diachronic+synchronic dataset. Results reported in Table 1 indicate that incorporating diachronic pairs into the training set has a twofold effect. (i) It is non-harmful when evaluating on synchronic pairs, as demonstrated by the Jacksonville split and the Omaha synchronic split. (ii) It is essential when evaluating on diachronic pairs, as evidenced by the Omaha diachronic split.

4.2 Generalization to Aerial Data

In this section, we evaluate whether fine-tuning on diachronic satellite data improves generalization to a different domain: aerial imagery. We compare MonSter and our fine-tuned model on two datasets from the benchmark proposed by (Wu et al., 2024): Enschede and EuroSDR-Vaihingen.

For these datasets, we report errors in pixels for the disparity estimation, including both MAE and root mean square error (RMSE). The EuroSDR-Vaihingen dataset represents a relatively simple aerial stereo benchmark. Conversely, Enschede poses a more complex challenge, as noted by the benchmark authors (Wu et al., 2024), with all methods showing higher error levels.

As shown in Table 3, MonSter performance is comparable to our fine-tuned model in EuroSDR-Vaihingen. However, in the Enschede dataset, fine-tuning on diachronic satellite data reduces the domain gap, yielding improvements in MAE and RMSE.

Model	EuroSDR-Vaihingen		Enschede	
	MAE (px)	RMSE (px)	MAE (px)	RMSE (px)
MonSter (zero-shot)	**1.32**	2.81	5.04	12.64
Ours (DFC fine-tuned)	1.57	**2.74**	**3.93**	**9.35**

Table 3. Evaluation on aerial datasets: MAE and RMSE in pixels.

4.3 Limitations

Despite the strong performance achieved by fine-tuning, our model still exhibits certain limitations inherent to learning-based methods. Once stereo matching becomes data-driven, the model inevitably learns to reproduce patterns present in the training data. Any systematic bias in the ground truth disparity maps is likely to appear at inference time.

Vegetation Bias We trained an alternative version of our model using DSMs where vegetation pixels were replaced by the minimum height in their neighborhood, removing trees from the ground truth disparities. When comparing the two models, we observe that the model trained on the original, tree-inclusive data consistently predicts higher disparities corresponding to

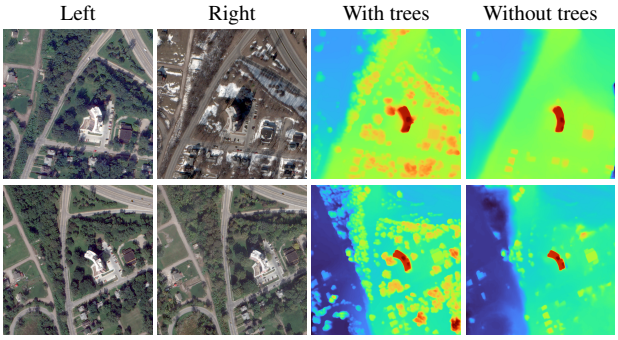

Figure 6. Disparity predictions after fine-tuning on ground-truth disparity maps with vs without trees. Tree-inclusive fine-tuning is biased toward tree-inclusive predictions, while the opposite is biased toward vegetation-free predictions, regardless of whether trees are visible in one or both input images (top vs bottom). Tree-inclusive ground-truth disparities are shown in Figure 4.

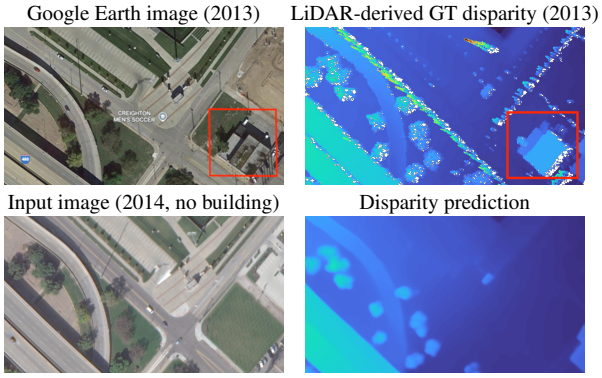

Figure 7. Temporal mismatch in OMA_247 between images and LiDAR-derived ground-truth disparity. The ground-truth (2013) includes a building absent from the input images (2014-2015). These inconsistencies add noise to both training and evaluation.

trees, even when they are not visible in both images. Conversely, the model trained without trees produces flat disparities in those regions, regardless of the trees appearing in both images. As shown in Figure 6, this behavior is invariant to whether trees appear in one or both images.

LiDAR–Imagery Temporal Discordance Using LiDAR-derived disparities as ground truth can introduce temporal noise that affects both training and evaluation. In Figure 4, the GT disparity for OMA_247 contains a building that is absent in the images. This mismatch likely stems from LiDAR captured around 2013, before the structure was removed, while images were acquired between 2014 and 2015 (Bosch et al., 2019, Le Saux et al., 2019). Figure 7 illustrates the discrepancy: the 2013 image (Google Earth) shows the building, while the 2014–2015 image does not. Although our model is generally robust to such inconsistencies during training, evaluation may be unfairly penalized, resulting in an increased reported error.

Reduced Sharpness Another limitation caused by the training data is the reduced sharpness of the predicted disparities. As shown in Figure 8, zero-shot predictions are visually sharper, although they are less accurate overall. In contrast, our fine-tuned model produces more geometrically accurate results but with smoother edges. We attribute this loss of sharpness to the training disparities, which are derived from DSMs from LiDAR measurements. Although accurate, they may have limited details

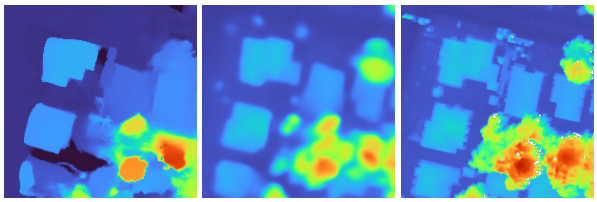

Figure 8. Comparison of disparity sharpness (detail from JAX_004). Fine-tuning improves the accuracy of the predicted disparities but produces smoother, less sharp boundaries.

due to their spatial resolution. This contrasts with the original training data from MonSter, which consists of perfect disparities with sharp boundaries from purely synthetic data.

5. Conclusion

This work addressed the problem of diachronic stereo matching, i.e., recovering 3D geometry from multi-date satellite pairs affected by strong seasonal, illumination, and shadow changes. Our experiments reveal that deep stereo models, when carefully adapted, can solve this problem.

A key finding is that robustness to diachronic variations in this case does not necessitate architectural changes, but rather emerges from the data used during fine-tuning. We observe a clear performance hierarchy: zero-shot models perform worst, followed by models fine-tuned on synchronic satellite data only, while those trained on a combination of synchronic and diachronic pairs achieve the best results. Another key factor is the use of monocular priors, which help maintain geometric consistency when stereo cues might be weak due to appearance changes.

Nonetheless, our findings also expose persistent challenges common to learning-based stereo models. These models tend to reproduce the biases of their supervision data, e.g., predicting tree-like structures even when vegetation is not visible, or producing overly smooth results. While synthetic data could mitigate these effects by providing sharp and accurate supervision, capturing the full complexity of real-world settings remains an open challenge.

Acknowledgments

This work was partially supported by Agencia Nacional de Investigación e Innovación (ANII, Uruguay) under the graduate scholarship POS_NAC_2023_1_177798, and by the ECOS-Sud French-Uruguayan cooperation program under grant U25E01. The project was provided with computing HPC and storage resources by GENCI at CINES, thanks to grant AD011016256 on the supercomputer Adastra’s MI250x partition.

References

Amadei, T., Meinhardt-Llopis, E., de Franchis, C., Anger, J., Ehret, T., Facciolo, G., 2025. s2p-hd: Gpu-accelerated binocular stereo pipeline for large-scale same-date stereo. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2339–2348.

Bartolomei, L., Tosi, F., Poggi, M., Mattoccia, S., 2025. Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1013–1027.

Beyer, R. A., Alexandrov, O., McMichael, S., 2018. The Ames Stereo Pipeline: NASA's open source software for deriving and processing terrain data. *Earth and Space Science*, 5(9), 537–548.

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

Bosch, M., Kurtz, Z., Hagstrom, S., Brown, M., 2016. A multiple view stereo benchmark for satellite imagery. *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, IEEE, 1–9.

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. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6273–6282.

De Franchis, C., Meinhardt-Llopis, E., Michel, J., Morel, J.-M., Facciolo, G., 2014. An automatic and modular stereo pipeline for pushbroom images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Facciolo, G., De Franchis, C., Meinhardt-Llopis, E., 2017. Automatic 3d reconstruction from multi-date satellite images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 57–66.

Facciolo, G., Franchis, C. d., Meinhardt, E., 2015. MGM: A Significantly More Global Matching for Stereovision. *Procedings of the British Machine Vision Conference 2015*, British Machine Vision Association, 1–90.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 3354–3361.

Gómez, A., Randall, G., Facciolo, G., von Gioi, R. G., 2023. Improving the pair selection and the model fusion steps of satellite multi-view stereo pipelines. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6344–6353.

Hirschmuller, H., 2007. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2), 328–341.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 139–1.

Le Saux, B., Yokoya, N., Hänsch, R., Brown, M., 2019. Data Fusion Contest 2019 (DFC2019).

Lindenberger, P., Sarlin, P.-E., Pollefeys, M., 2023. Lightglue: Local feature matching at light speed. *Proceedings of the IEEE/CVF international conference on computer vision*, 17627–17638.

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

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91–110.

Marí, R., de Franchis, C., Meinhardt-Llopis, E., Facciolo, G., 2019. To bundle adjust or not: A comparison of relative geolocation correction strategies for satellite multi-view stereo. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

Marí, R., Ehret, T., Facciolo, G., 2022. Disparity estimation networks for aerial and high-resolution satellite images: A review. *Image Processing On Line*, 12, 501–526.

Marí, R., Facciolo, G., Ehret, T., 2023. Multi-date earth observation NeRF: The detail is in the shadows. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2035–2045.

Masquil, E., Ehret, T., Musé, P., Facciolo, G., 2026. Deep S2P: Integrating Learning-Based Stereo Matching into the Satellite Stereo Pipeline. *arXiv preprint arXiv:2603.21882*.

Masquil, E., Marí, R., Ehret, T., Meinhardt-Llopis, E., Musé, P., Facciolo, G., 2025. S-EO: A large-scale dataset for geometry-aware shadow detection in remote sensing applications. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2383–2393.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T., 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4040–4048.

Menze, M., Geiger, A., 2015. Object scene flow for autonomous vehicles. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3061–3070.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99–106.

Patil, S., Guo, Q., 2023. Stellar: A Large Satellite Stereo Dataset for Digital Surface Model Generation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 433–440.

Savant Aira, L., Facciolo, G., Ehret, T., 2025. Gaussian splatting for efficient satellite image photogrammetry. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5959–5969.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P., 2014. High-resolution stereo datasets with subpixel-accurate ground truth. *German conference on pattern recognition*, Springer, 31–42.

Schops, T., Schonberger, J. L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A., 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3260–3269.

Tosi, F., Bartolomei, L., Poggi, M., 2025. A survey on deep stereo matching in the twenties. *International Journal of Computer Vision*, 133(7), 4245–4276.

Tosi, F., Tonioni, A., De Gregorio, D., Poggi, M., 2023. Nerf-supervised deep stereo. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 855–866.

Tyszkiewicz, M., Fua, P., Trulls, E., 2020. Disk: Learning local features with policy gradient. *Advances in neural information processing systems*, 33, 14254–14265.

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

Wu, T., Vallet, B., Pierrot-Deseilligny, M., Rupnik, E., 2024. An evaluation of Deep Learning based stereo dense matching dataset shift from aerial images and a large scale stereo dataset. *International Journal of Applied Earth Observation and Geoinformation*, 128, 103715.

Xu, G., Wang, X., Ding, X., Yang, X., 2023. Iterative geometry encoding volume for stereo matching. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21919–21928.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth anything v2. *Advances in Neural Information Processing Systems*, 37, 21875–21911.

Zhang, L., Rupnik, E., 2023. SparseSat-NeRF: Dense Depth Supervised Neural Radiance Fields for Sparse Satellite Images. *ISPRS Annals*.