This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

s2p-hd: Gpu-Accelerated Binocular Stereo Pipeline for Large-Scale Same-Date Stereo

Tristan Amadei[1], *Enric Meinhardt-Llopis*[1], *Carlo de Franchis*[1,2], *Jérémy Anger*[1], *Thibaud Ehret*[3], *Gabriele Facciolo*[1]

[1]Universite Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, 91190, Gif-sur-Yvette, France
[2]Kayrros SAS, France
[3]AMIAD, Pôle Recherche, France
https://centreborelli.github.io/s2p-hd-earthvision

Abstract

The increasing availability of high-resolution satellite imagery has driven advances in 3D reconstruction techniques for the generation of Digital Elevation Models (DEM). Recent research focuses on opportunistic stereo, using sophisticated techniques like Neural Radiance Fields and Gaussian Splats, that are able to exploit a collection of multi-date images of the same site. These techniques give optimal results, but they are computationally expensive, so in practice, they are only used for small regions of interest. In contrast, quasi-simultaneous stereo products are routinely acquired for large-scale mapping, needing a focus on efficiency, robustness and scalability in their processing. This paper introduces s2p-hd, a binocular stereo pipeline designed for high-throughput processing of same-date satellite imagery. Building upon the open-source s2p pipeline, s2p-hd adds several key improvements that enhance its performance and robustness, tuned for same-date stereo imagery. These include a refined disparity range estimation leveraging reference models and multiscale analysis, the adaptation of a highly optimized GPU-based Semi-Global Matching (SGM) algorithm, and enhanced rectification and tiling strategies. We benchmark s2p-hd against standard stereo pipelines and show that it outperforms them both in accuracy and processing speed, making it a powerful tool for generating high-quality DEMs from large-scale optical satellite imagery, while balancing precision and computational efficiency.

1. Introduction

In the last few decades, high-resolution satellite imagery has evolved from being a rare resource to becoming a commodity, largely due to agile satellite constellations. This shift has fueled the development of applications across a wide range of fields [47], including agriculture, urban planning, environmental monitoring, and disaster response [2]. The availability of high-resolution data has also facilitated advanced processing techniques, such as the generation of 3D models from satellite images.

Satellite digital elevation models (DEMs) are traditionally generated using dedicated, quasi-simultaneous stereoscopic acquisitions, such as binocular or tristereo imagery [8, 20, 40]. In recent years, however, there has been increasing interest in leveraging opportunistic imagery acquired at different dates to create DEMs. The 2016 IARPA Multi-View Stereo 3D Mapping Challenge [4] and the 2019 IEEE GRSS Data Fusion Contest [5] both focused on assessing these capabilities. Early efforts in this direction applied traditional stereoscopic pipelines—designed for simultaneous imagery—to pairs of images selected from multi-temporal series, subsequently fusing the results [14, 19, 35]. Each image pair was used to produce an independent digital surface model (DSM), and when multiple pairs were available, the resulting DSMs were merged to create a more accurate model. Multi-View Stereo techniques have also been directly applied to satellite image time-series [44]. However, these methods have notable limitations. Pipelines originally designed for simultaneous imagery struggle with the unique challenges posed by multi-temporal imagery, such as variations in lighting, shadows, atmospheric conditions, and object movement, which affect the geometric accuracy.

Ultimately, these difficulties have led to the development of methods capable of addressing the radiometric and geometric variations present in multi-temporal imagery. Techniques based on inverse rendering, such as Neural Radiance Fields (NeRF) [27, 28, 38, 46] and Gaussian Splatting [1, 26, 41], have shown promise in producing more

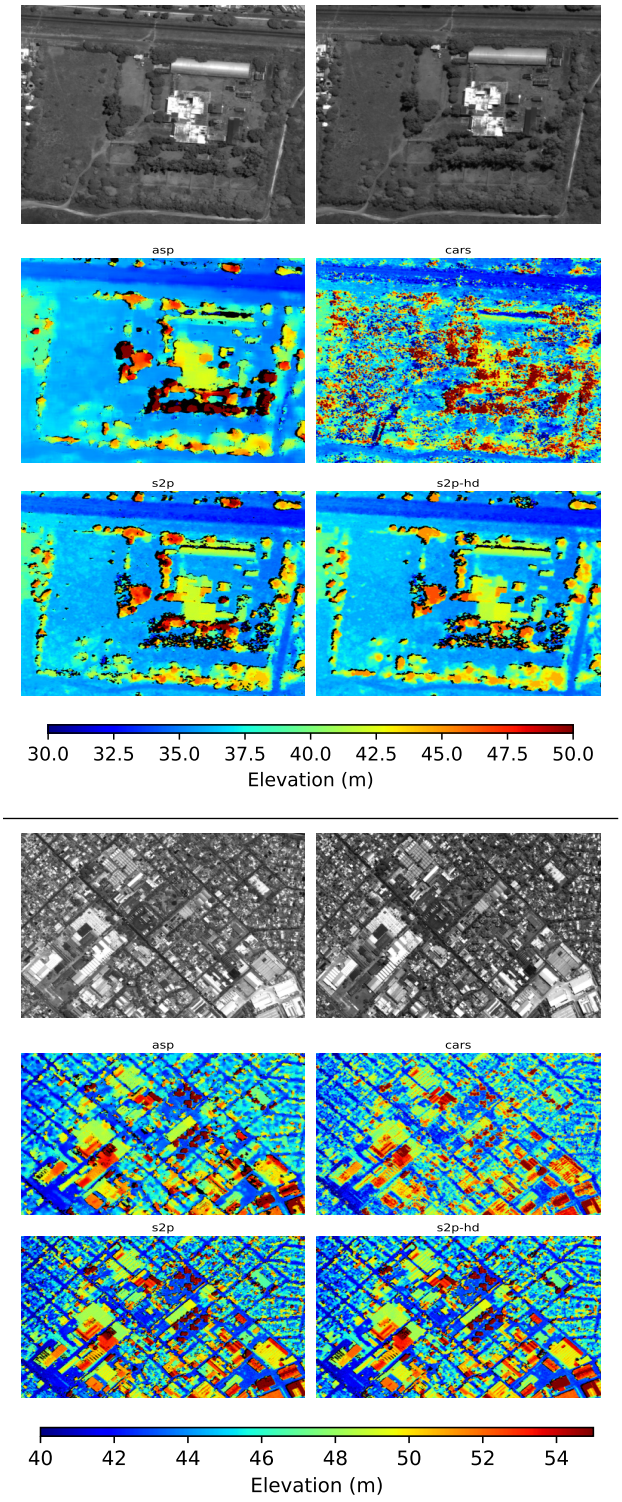

Figure 1. DSMs of s2p-hd and three other pipelines, on two small tiles of around 1000×1000 pixels from the IARPA challenge dataset [4], which are displayed above the DSMs. Pixels displayed in black represent points with no data, *i.e.* zones where the pipelines failed to predict a height.

dense and accurate DEMs. They can adapt to varying radiometric conditions, and produce accurate results despite the differences in acquisition dates. However, these methods often require a significantly larger number of input images compared to traditional simultaneous acquisitions, and an increased computational complexity. These two tasks—quasi-simultaneous and opportunistic stereo—are distinct, each requiring tailored solutions. While the former is primarily focused on large-scale mapping and should be highly efficient, opportunistic imaging aims to address more targeted sites. In the quasi-simultaneous setting, the goal is to process pairs of images quickly and accurately, ensuring that the resulting 3D models meet high standards without compromising performance. This approach should rely on algorithms designed to handle large images efficiently, often leveraging GPU acceleration to speed up processing times. This is the case of the NASA Ames stereo pipeline [3], MicMac [36] by the French National Geographic Institute (IGN), CARS [31] by the French Space Agency (CNES), s2p [10] by the Centre Borelli research center, or the CATENA [23] multi-stereo chain by the German Aerospace Center (DLR).

In summary, while both quasi-simultaneous and opportunistic stereo share common goals in terms of generating 3D models from satellite imagery, they differ significantly in their objectives, challenges, and required solutions. Quasi-simultaneous stereo focuses on large-scale efficiency, whereas opportunistic stereo addresses more targeted needs with greater computational demands.

In this paper, we will focus on the high-efficiency large-scale problem and describe the modifications proposed to the open source s2p package in order to meet these requirements. Unlike general-purpose pipelines like the original s2p, or CARS, which cater to a broad range of applications, the proposed s2p-hd pipeline is tailored for high-throughput processing of satellite imagery. It addresses the limitations of its predecessor, s2p, by improving disparity range estimation, integrating GPU-accelerated stereo matching [7], and enhancing post-processing techniques. Figure 1 shows an example of the DSM reconstructions of s2p-hd and three other pipelines. The main modifications in s2p-hd include:

- A robust and bounded disparity range estimation that leverages reference models, multiscale analysis, and acquisition parameters.
- The integration and adaptation of a highly optimized GPU-based SGM algorithm [7], significantly accelerating the stereo matching process.
- Enhanced rectification and tiling strategies to improve robustness and efficiency.

Comparisons with some openly-available satellite stereo reconstruction pipelines show that these modifications enable s2p-hd to generate high-quality DEMs, balancing accuracy and computational efficiency. Figure 2 shows how s2p-hd

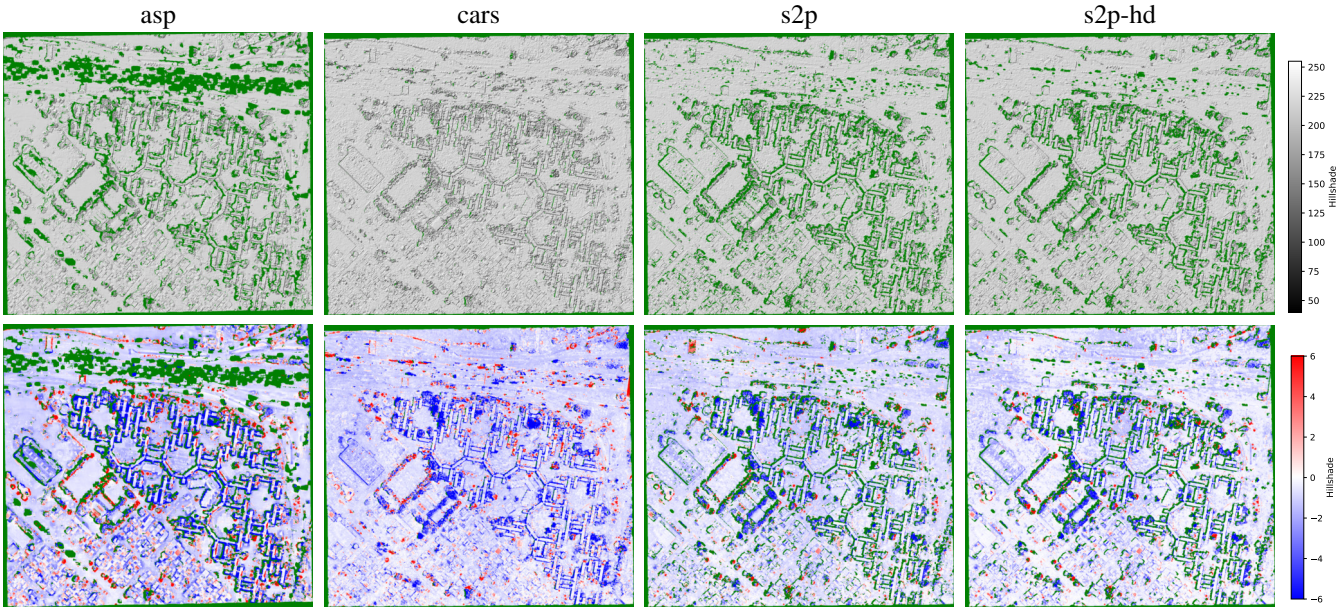

Figure 2. The first row displays hillshade visualizations of the DSMs computed by different pipelines on the training site of the IARPA challenge dataset [4]. The second row shows the elevation differences betweens the DSMs and the ground truth computed with a LiDAR. The green points represent the areas where the methods failed to predict a value. The results computed by s2p-hd are more accurate than those from other pipelines: s2p-hd gets an average RMSE of 0.84m over all the IARPA zones with a ground truth available, while the other pipelines go up to 1.80m. Furthermore, its computation is two to four times faster than the other pipelines.

compares favorably to other methods in terms of reconstruction accuracy.

Furthermore, since evaluating each method requires a significant investment of time (from understanding installation procedures to identifying key hyperparameters and their effects) we provide a Docker sandbox to streamline future research in this domain. This environment enables easy deployment of all methods, simplifies DSM generation, and allows straightforward comparisons between predicted DSMs and ground truth data. The code of s2p-hd and the docker sandbox are available on the project webpage.

2. Related work

The increasing availability of high-resolution satellite imagery has contributed to the advance of the field of 3D reconstruction. A variety of methodologies for stereo matching and pipelines have emerged in pursuit of improved accuracy, robustness, and efficiency.

Stereo matching and disparity estimation. Early approaches to stereo matching predominantly relied on traditional block-matching techniques [17]. To improve consistency, optimization-based methods were introduced to regularize block-matching results [37]. The development of Semi-Global Matching (SGM) [22] marked a significant advancement, enabling robust disparity estimation by approximating global energy minimization while being computationally efficient. However, these methods can still struggle with complex scenes with many discontinuities, textureless regions and repetitive patterns.

Recent advances have seen the integration of deep learning to address these limitations. Convolutional Neural Networks (CNNs) and transformer-based stereo networks have demonstrated strong performance on disparity predictions [6, 11, 25]. These networks leverage large datasets to learn robust features, improving accuracy in challenging scenarios. For example, networks have been designed to directly regress disparity maps, or to refine disparity maps produced by traditional methods [39]. A review and comparison with traditional methods on satellite images are performed in [29].

Pipelines for satellite images. Several well-established stereo reconstruction pipelines cater to diverse applications. The Ames Stereo Pipeline (ASP) [3] is a comprehensive open-source toolkit developed by NASA, widely used for planetary and Earth observation applications. ASP integrates a variety of stereo matching algorithms, supporting a range of input modalities and sensor models. It is known for its robustness and flexibility, making it suitable for complex processing tasks. The Satellite Stereo Pipeline (s2p) [10] focuses on large-scale stereo reconstruction from high-resolution satellite images. It uses dense matching techniques optimized for computational efficiency, making it suitable for processing large datasets. CARS (CNES Al-

gorithms to Reconstruct Surface) [31] developed by CNES, is another efficient pipeline designed to optimize DSM generation. CARS balances computational efficiency and accuracy through optimized tiling and aggregation strategies. MicMac [36] from IGN is an open-source photogrammetry suite that offers a wide range of tools for stereo reconstruction, including dense matching and 3D reconstruction.

Optimization and refinement techniques. Post-processing and refinement techniques are crucial for improving the quality of reconstructed surfaces. Multi-View Stereo (MVS) methods, which integrate information from multiple viewpoints, are often used to enhance depth estimation consistency [16]. By looking at the scene from multiple viewpoints, these methods can see past occluded areas, giving a more complete and accurate estimation.

Deep learning-based refinement techniques, such as depth map fusion and uncertainty estimation, have further advanced reconstruction accuracy [34]. These methods can learn to correct errors and inconsistencies in disparity maps, leading to more accurate and robust reconstructions. Recent works have explored the integration of self-supervised learning and domain adaptation to improve generalization across different satellite sensors and terrains [18, 42]. Self-supervised learning allows models to learn from unlabeled data, reducing the need for large labeled datasets.

Challenges and future directions. Despite these advances, several challenges remain in satellite stereo reconstruction. The handling of occlusions, texture-less regions, and varying lighting conditions remains a major problem. Future research directions include the integration of multi-sensor fusion, neural implicit representations, and self-supervised learning paradigms to further enhance 3D reconstruction quality [32, 43]. Neural implicit representations, such as Neural Radiance Fields (NeRF), offer a promising approach for representing complex 3D scenes.

3. Method

Same-date stereo imaging is preferred for fast and reliable 3D reconstruction, as it avoids challenges linked to temporal changes like lighting, object motion, and atmospheric conditions. In contrast, opportunistic stereo—using images from different dates—is more ill-posed due to variations in shadows, scene dynamics, and seasonal effects, which can hinder accurate depth estimation. As illustrated in Figure 3, images from different dates require distinct processing and parametrization (in this case a stronger regularization would be needed) compared to same-date imagery. More examples are shown in Section 6.1 of the supplementary material.

It is worth noting that while opportunistic methods can also benefit from fast algorithms for initialization [27, 45], they generally require more complex processing to account for the changes over time, which limits their scalability to large-scale mapping.

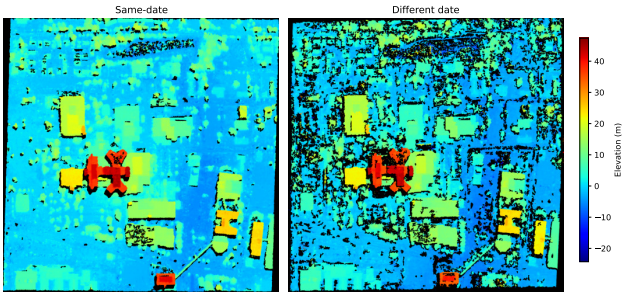

Figure 3. The two DSMs both correspond to the *JAX_068* zone from the GRSS dataset. The one on the left was created with two same-date images, while the one on the right was constructed from images taken at different dates. The same pipeline parameters were used for both results.

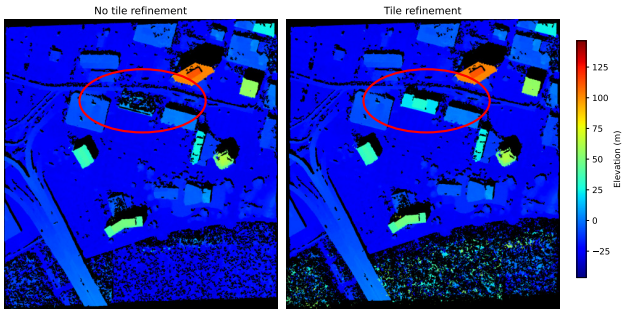

Figure 4. The left DSM was generated without tile-wise disparity range refinement, while the right DSM includes refined disparity ranges per tile. As illustrated, the central building is missing in the reconstruction without refinement, whereas it is accurately captured when refinement is applied. The lower part of the image corresponds to water, which poses challenges for stereo matching; as a result, both methods struggle to reconstruct accurate elevation in those areas.

The s2p-hd pipeline is designed for the efficient generation of high-resolution Digital Surface Models from same-date, large-scale satellite stereo imagery. It builds upon the open-source s2p pipeline, addressing its limitations and significantly enhancing its performance and robustness. With s2p-hd, we focus on quasi-simultaneous stereo, prioritizing speed and scalability, and distinguishing it from more computationally intensive opportunistic stereo methods.

3.1. Disparity range estimation

A critical improvement in s2p-hd is the refined disparity range estimation. The original s2p pipeline relied on SIFT keypoint matches [24] extracted for each tile as a way to estimate the disparity range to be explored (*i.e.* the max and min disparity of the matched keypoints). However, this could lead to computational inefficiencies and, in some cases, artifacts in the reconstructed DSM, such as clipped

Method	Matching algorithm	IARPA Dataset				GRSS Dataset				San Diego			
		RMSE	NMAD	Percentile 90	Time (s)	RMSE	NMAD	Percentile 90	Time (s)	RMSE	NMAD	Percentile 90	Time (s)
ASP	SGM	1.80	0.87	2.85	39.08	2.76	1.07	4.60	140.55	3.15	0.51	20.24	54.32
CARS	SGM	1.23	0.59	2.10	5948.48	3.84	1.59	6.34	66.65	4.84	1.31	7.06	12,355.65
s2p	MGM	0.91	0.43	1.42	32.03	2.72	1.10	4.53	36.09	2.79	0.44	2.36	38.77
s2p-hd	SGM	0.89	0.45	1.28	**18.52**	2.70	1.05	4.37	**23.38**	2.74	0.42	**1.69**	**17.53**
	MGM	0.86	0.41	1.26	20.32	2.70	1.08	4.51	28.33	2.85	0.44	2.41	34.03
	SGM + MGM pp	**0.84**	**0.40**	**1.25**	20.56	**2.62**	**1.01**	**4.23**	34.00	**2.66**	**0.40**	2.19	23.50

Table 1. Comparison of different pipelines on the IARPA [4], GRSS [5] and San Diego [30] datasets. Best values in each column are in **bold**, second best are underlined. SGM + MGM pp (last row) means that SGM was used for the stereo matching and MGM was used for the subpixel refinement post-processing.

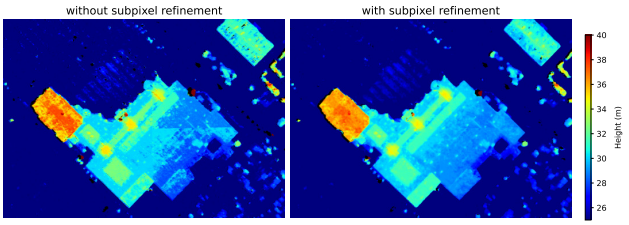

Figure 5. DSM of a building without (left) and with (right) subpixel refinement. We can see the effect of tile discontinuity on the left DSM, where the building is split in its predicted height. However, the right DSM displays a smooth and continuous prediction of the building's height.

buildings due to incorrect range bounds. Indeed, on one hand, sparse keypoints do not reliably capture the full extent of scene depth variations as keypoints may fail to pick up tall structures, leading to an underestimated disparity range and missing height information. On the other hand, sparse keypoints can also result in spurious matches, causing an overestimated disparity range that increases noise and computation time. To address these issues, s2p-hd introduces a bounded range estimation approach, incorporating:

- **Reference model integration:** When available, a coarse reference DEM (*e.g.* SRTM [15]) is used to establish an initial altitude range.
- **Multiscale estimation:** A multiscale approach refines the disparity range by progressively analyzing image features at different resolutions.
- **Acquisition configuration analysis:** The baseline-to-height (b/H) ratio and other acquisition parameters are used to constrain the disparity range based on geometric considerations.

Furthermore, s2p-hd incorporates a heuristic to avoid problematic tiles during the disparity range estimation, preventing outliers from throwing off the overall estimation. This improved estimation significantly enhances the accuracy and efficiency of the subsequent stereo matching process.

In s2p, the disparity is computed on a tile-by-tile basis, without considering neighboring tiles. This can result in large structures, such as buildings, being clipped or entirely missed. If a tile requires a large disparity range, its neighboring tiles often do as well; otherwise, some features may be inaccurately reconstructed. To ensure consistency across neighboring tiles, s2p-hd estimates the disparity range by analyzing keypoint matches from both the current tile and its eight neighboring tiles. First, altitudes are computed for these putative matches via triangulation. If the resulting set of altitude spans a small range (set to less than 300m in our experiments), then all the matches are considered valid and used to determine the disparity range. However, if the range of altitudes is too large, some matches are likely to be incorrect. Therefore, the matches yielding extreme altitudes are discarded. In practice, the altitude thresholds are computed as: $\text{median}(\{\text{altitudes}\}) \pm 250m$. Figure 4 shows that this disparity refinement allows for a smoother reconstruction. Furthermore, its computation overhead is negligible.

3.2. GPU-accelerated stereo matching

The stereo matching process, the computational bottleneck in many stereo pipelines, is accelerated in s2p-hd using a GPU-based Semi-Global Matching (SGM) algorithm. Based on the open-source `libSGM` [7], the implementation in s2p-hd includes several key modifications:

- **Extend libSGM disparity range and improve its left-right consistency check:** Modifications to the CUDA code of libSGM to extend the disparity range from 256 to 512 pixels, accommodating a wider range of scene depths, and correcting the left-right consistency check to handle negative disparities.
- **Variable window sizes:** The implementation supports variable window sizes for the census transform, allowing for flexibility in matching different image features.
- **Adjusted regularity parameters:** The regularity parameters of SGM are tuned to optimize performance for same-date satellite imagery.
- **GPU memory scheduling:** A GPU memory scheduler is implemented to prevent out-of-memory errors, allowing parallel processing of multiple tiles on GPUs with limited memory.

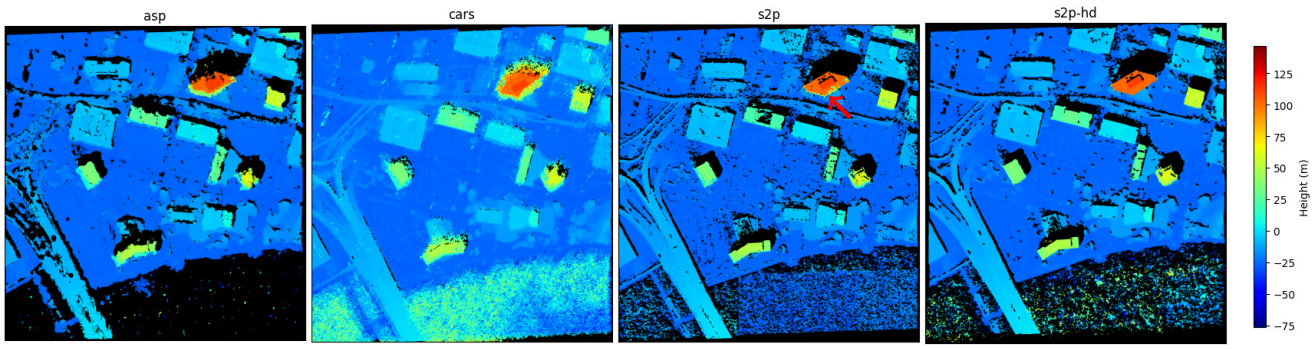

Figure 6. DSM visualizations computed by each method for the *JAX_166* area from the GRSS dataset [5]. As shown on Table 1, s2p-hd achieves more accurate and dense representations, while being much faster than other pipelines. s2p-hd achieves an RMSE of 2.70m compared to ASP's 2.76m or CARS' 3.84m. Furthermore, s2p-hd computes this DSM in only 23s, two to four times faster than the other methods. The red arrow highlights an overlying artifact in s2p resulting from averaging the altitudes of the projected DSM, rather than taking the maximum as done in s2p-hd.

- **Subpixel refinement:** An optional subpixel disparity refinement step using the More Global Matching (MGM) correlator [13] is added to improve disparity map accuracy.

These modifications result in a significant speedup in the stereo matching process, enabling the processing of large-scale datasets in a fraction of the time required by CPU-based methods.

3.3. Rectification and tiling

Rectification speedup. By replacing the keypoint matching routine with an optimized OpenCV implementation, s2p-hd achieves a 5x acceleration in the rectification process. Furthermore, it improves robustness by skipping tiles where rectification fails.

Disparity quantization. For short-baseline images (common in native stereo), small disparity estimation errors translate into larger altitude differences. Since the disparity is estimated with limited discrete precision, elevation values become quantized. When processing images in tiles, adjacent tiles can have different quantization offsets, leading to visible discontinuities between tiles (see Figure 5). To mitigate the effect of tile discontinuity, a subpixel refinement of the disparity is computed by applying the MGM [13] correlator restricted to a narrow band of disparities around the result computed by SGM; we will denote this refinement as SGM + MGM pp.

Tile margin management. When computing a large-scale DSM, tiling the images is essential to handle the push-broom geometry [9] and allows to manage memory consumption and processing efficiency. However, selecting the appropriate tile size is crucial. Tiles that are too large can cause memory overload or prevent multiprocessing, significantly increasing computation time. Conversely, tiles that are too small can impact the quality of the result. First, small tiles may compromise epipolar rectification, especially if SIFT-based feature matching is used, as insufficient features may be detected. Additionally, small tiles can lead to large structures—such as tall buildings—being split across multiple tiles, which disrupts the matching process and affects reconstruction accuracy.

To mitigate these issues, s2p-hd applies an overlap margin when tiling images, ensuring continuity across tiles. A left and right margin is added to each tile, defined as

$$\text{margin_size} = b/H \times h_{max} \times \text{GSD}, \tag{1}$$

where b/H represents the baseline-to-altitude ratio of the satellite stereo pair [12], h_{max} is the maximum altitude in the scene and GSD denotes the ground sampling distance (*i.e.*, the pixel resolution in meters).

The optimal tile size is then determined based on the available VRAM. It is obtained via the formula

$$\text{tile_size} \times (2 \cdot \text{margin_size} + \text{tile_size}) \times 512 \times B = \frac{\text{VRAM}}{k}, \tag{2}$$

where $k \geq 2$ is the number of processes that should run concurrently at this tile size, 512 corresponds to the disparity range and B is the number of bytes per pixel (4 bytes in our case).

3.4. DSM generation and post-processing

The generation of the DSM from the disparity maps is optimized in s2p-hd by employing efficient triangulation, aggregation, and filtering methods. Specifically, we replaced the default averaging of point clouds with a maximum value approach that prevents overlay errors such as the ones highlighted in Figure 6. Post-processing steps, such as 3D filtering of small disconnected components, and hole filling, are streamlined with simplified parameters and improved parallel handling.

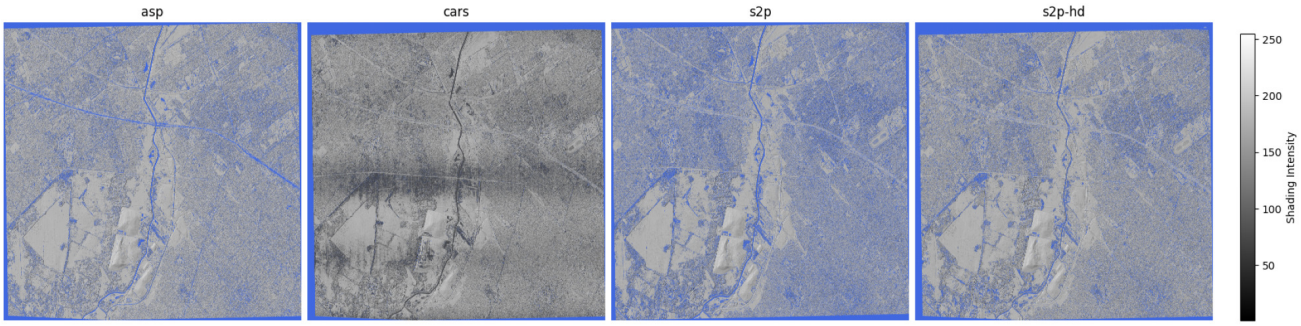

Figure 7. Hillshade visualizations of the DSMs computed by the different pipelines on a full-size Worldview-3 image pair ($36,864 \times 43,008$ pixels at a resolution of 37.5cm/pixel). The results computed by s2p-hd are slightly more dense and more precise than the others, for a much smaller running time, between two and four times faster than other pipelines.

4. Experiments & results

All the experiments were run on an Intel Core Ultra 9 185H CPU with 22 cores and 32GB of RAM, and an NVIDIA RTX 1000 Ada GPU with 6GB of VRAM. We analyze here the experiments conducted on the first two datasets shown in Table 1. The results on the San Diego dataset [21, 30], corresponding to a landscape with larger natural terrain variability, are discussed in Section 7.3 of the supplementary material.

4.1. 2016 IARPA Multi-View Stereo 3D Mapping Challenge

We evaluated s2p-hd using the 2016 IARPA Multi-View Stereo 3D Mapping Challenge dataset [4], which comprises 47 panchromatic WorldView-3 images (30 cm nadir resolution), accompanied by a 20 cm resolution airborne LiDAR ground truth. Most images were acquired on different dates; however, since s2p-hd is designed for same-date stereo, we selected a single same-date pair based on the recommendations in [14] (see Figure 1 and Figure 12 in supplementary material). Ground truth data consists of 20 cm resolution airborne LiDAR covering a 20 square kilometer subset.

We perform our evaluation by computing DSMs on small areas of interest with available LiDAR ground truth, comparing the DSMs generated by ASP, CARS, s2p, and s2p-hd. To mitigate the impact of georeferencing errors and sensor biases, we first coregistered the predicted DSMs to the LiDAR ground truth using the Nuth & Kääb algorithm [33] (details in supplementary material). We then assessed performance using **Root Mean Square Error (RMSE)**, 90^{th} **percentile error**, and **Normalized Median Absolute Deviation (NMAD)**, calculated as: $nmad = 1.486 \times median(|data - median(data)|)$.

Figure 2 illustrates the differences between DSMs predicted by each pipeline relative to the LiDAR ground truth. As shown in Table 1, s2p-hd achieves lower RMSE than other methods while maintaining competitive computation times. CARS, which always performs epipolar rectification on full-resolution images (approximately $40{,}000 \times 40{,}000$ pixels), exhibits significantly higher processing costs. The other methods extract the Region of Interest (ROI) before rectification, allowing for faster processing. The integration of a GPU-accelerated SGM algorithm significantly reduces the computation time of s2p-hd compared to s2p (see Table 2). However, we observed a slight trade-off: while more efficient, the GPU-accelerated SGM correlator resulted in marginally lower accuracy compared to the CPU-based MGM. The best compromise on the dataset is achieved by combining SGM for stereo matching with MGM for subpixel refinement. This setup delivers the best performance while adding only a small amount of computation time, remaining faster than the other pipelines.

Note that the results presented in Table 1 are averages calculated across all zones with available ground truth. In Figure 1 and Figure 2, we observe that CARS employs a different approach to handle occluded or texture-less regions. While the other methods leave these areas blank, CARS tends to fill them with smooth depth estimates. This can lead to unreliable predictions in regions such as water and forests. To compensate for this, CARS can also output an ambiguity map, which highlights pixels where the algorithm struggles to confidently select a single depth value. As shown in Figure 8, which corresponds to the DSM for the *JAX_166* area from the GRSS dataset (Figure 6), the map reveals that while CARS fills regions left blank by other pipelines, these areas show high ambiguity, signaling that the reconstruction in those regions is uncertain.

s2p-hd leverages GPU acceleration for faster matching, however this advantage cannot be observed by working on small areas of interest. To better evaluate performance, we computed the DSM for the entire IARPA image pair, rather than restricting it to a specific region. Table 2 highlights the speed improvement of s2p-hd, whether using MGM (CPU-based) or SGM (which benefits from GPU acceleration); or

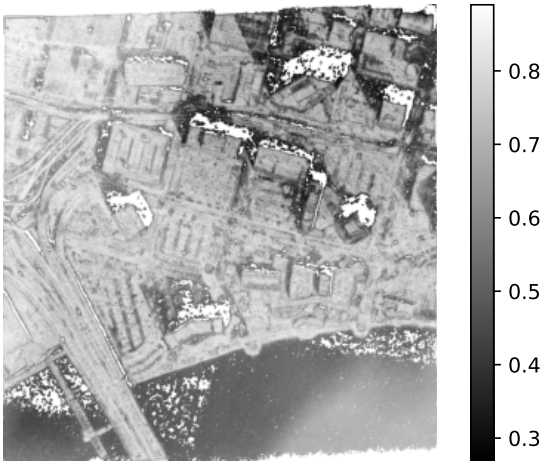

Figure 8. Ambiguity map produced by CARS, during the DSM reconstruction for the *JAX_166* area from the GRSS dataset.

Method	ASP	CARS	s2p	s2p-hd (MGM)	s2p-hd (SGM)	s2p-hd (SGM + MGM pp)
Time (s)	7359	10,817	10,732	6483	**4522**	6837

Table 2. Computation time (in seconds) required by different pipelines to predict a full DSM. The time repartition of each step for s2p-hd is detailed in the supplementary material.

the slightly more computationally expensive but more accurate SGM with MGM subpixel refinement post-processing. However, since ground truth data is unavailable for such a large area, we cannot compute error metrics for these full-image DSMs. Figure 7 shows the reconstruction quality on the full-size images.

4.2. 2019 IEEE GRSS Data Fusion Contest

The 2019 IEEE GRSS Data Fusion Contest [5] provided a dataset focused on urban 3D reconstruction, featuring multi-view satellite imagery from Maxar's WorldView-3 and WorldView-4 satellites over Jacksonville, Florida, USA. The dataset includes high-resolution panchromatic and multispectral images acquired on different dates, with varying illumination and atmospheric conditions. It also provides a LiDAR-derived DSM at 50 cm resolution as ground truth for a subset of the area.

We picked images from *Track 3: Multi-view semantic stereo* of this challenge. As we are interested in same-date stereo images, we needed to select the best image pair from each zone to consider. Following [14], we filter pairs to retain only those with baseline angles between $5°$ and $45°$, while ensuring that the incidence angle of both images remains below $40°$ to avoid excessive distortions. Within this filtered set, pairs are sorted by increasing acquisition date difference, favoring those captured closest in time to minimize variations in lighting and seasonal changes. Figure 13 (in the supplementary material) shows the image pairs of two areas of this dataset, selected with this heuristic. To ensure the same-date, quasi-simultaneous nature of s2p-hd, we carefully selected image pairs with closely matching acquisition times from the already filtered set. The final pairs have an average acquisition time difference of 11 seconds.

Similarly to the previous dataset, s2p-hd processes DSMs significantly faster than other pipelines while maintaining high accuracy. In fact, it achieves a lower RMSE compared to the other methods. Our method produces DSMs that better preserve scene details and align more closely with the LiDAR ground truth. To illustrate this, Figure 6 displays DSMs generated by each method over a region from the GRSS dataset. Compared to the IARPA dataset, the LiDAR ground truths in the GRSS dataset have a lower resolution and cover much smaller areas. This explains why the RMSE difference between s2p-hd and the other methods is less pronounced.

5. Conclusion

We present s2p-hd, a GPU-accelerated binocular stereo pipeline designed for the efficient generation of high-resolution Digital Surface Models from same-date satellite imagery. Building upon the open-source s2p pipeline, s2p-hd introduces several key improvements to enhance performance and robustness, addressing the specific demands of large-scale, quasi-simultaneous stereo reconstruction.

The integration of a highly optimized GPU-based Semi-Global Matching (SGM) algorithm significantly accelerates the stereo matching process, a computational bottleneck in many pipelines. Modifications such as extended disparity range, improved left-right consistency checks, variable window sizes, and adjusted regularity parameters contribute to the algorithm's performance.

Experimental results on the IARPA Multi-View Stereo 3D Mapping Challenge dataset and the IEEE GRSS Data Fusion Contest demonstrate the effectiveness of s2p-hd. Compared to established pipelines, s2p-hd achieves competitive accuracy with significantly reduced computation time. Specifically, the GPU-accelerated SGM algorithm in s2p-hd achieves lower RMSE values while requiring less processing time, showcasing the benefits of our optimized implementation. Furthermore, the ability to process full-size images highlights the pipeline's scalability and robustness.

Acknowledgments

This work was performed using HPC resources from GENCI-IDRIS (grant 2024-AD011012453R4). Centre Borelli is also with Université Paris Cité, SSA and INSERM.

References

[1] Luca Savant Aira, Gabriele Facciolo, and Thibaud Ehret. EOGS: Gaussian splatting for earth observation. *arXiv preprint arXiv:2412.13047*, 2024. 1

[2] Olalekan Mumin Bello and Yusuf Adedoyin Aina. Satellite remote sensing as a tool in disaster management and sustainable development: Towards a synergistic approach. *Procedia - Social and Behavioral Sciences*, 120:365–373, 2014. 3rd International Geography Symposium, GEOMED2013, 10-13 June 2013, Antalya, Turkey. 1

[3] Ross A. Beyer, Oleg Alexandrov, and Scott McMichael. The ames stereo pipeline: Nasa's open source software for deriving and processing terrain data. *Earth and Space Science*, 5 (9):537–548, 2018. 2, 3

[4] Marc Bosch, Zachary Kurtz, Shea Hagstrom, and Myron Brown. A multiple view stereo benchmark for satellite imagery. In *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–9. IEEE, 2016. 1, 2, 3, 5, 7

[5] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D Hager, and Myron Brown. Semantic stereo for incidental satellite images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1524–1532. IEEE, 2019. 1, 5, 6, 8

[6] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 3

[7] Fixstars Corporation and Fixstars Autonomous Technologies. libsgm: A cuda implementation of the semi-global matching algorithm, 2023. https://github.com/fixstars/libSGM, Accessed: 2025-03-03. 2, 5

[8] A Cuartero, AM Felicísimo, and FJ Ariza. Accuracy of dem generation from terra-aster stereo data. *International Archives of Photogrammetry and Remote Sensing*, 35(B2): 559–563, 2004. 1

[9] C. de Franchis, E. Meinhardt-Llopis, J. Michel, J.-M. Morel, and G. Facciolo. On stereo-rectification of pushbroom images. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5447–5451, 2014. 6

[10] Carlo De Franchis, Enric Meinhardt-Llopis, Julien Michel, Jean-Michel Morel, and Gabriele Facciolo. An automatic and modular stereo pipeline for pushbroom images. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2014. 2, 3

[11] Véronique Defonte, Loïc Dumas, Myriam Cournet, and Emmanuelle Sarrazin. Evaluation of mc-cnn based stereo matching pipeline for the co3d earth observation program. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 7670–7673. IEEE, 2021. 3

[12] Julie Delon and Bernard Rougé. Small baseline stereovision. *Journal of Mathematical Imaging and Vision*, 28:209–223, 2007. 6

[13] Gabriele Facciolo, Carlo De Franchis, and Enric Meinhardt. Mgm: A significantly more global matching for stereovision. In *BMVC 2015*, 2015. 6

[14] Gabriele Facciolo, Carlo De Franchis, and Enric Meinhardt-Llopis. Automatic 3d reconstruction from multi-date satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 57–66, 2017. 1, 7, 8

[15] Tom G. Farr, Paul A. Rosen, Edward Caro, Robert Crippen, Riley Duren, Scott Hensley, Michael Kobrick, Mimi Paller, Ernesto Rodriguez, Ladislav Roth, David Seal, Scott Shaffer, Joanne Shimada, Jeffrey Umland, Marian Werner, Michael Oskin, Douglas Burbank, and Douglas Alsdorf. The shuttle radar topography mission. *Reviews of Geophysics*, 45(2), 2007. 5

[16] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 4

[17] Mohammed Ghanbari. The cross-search algorithm for motion estimation (image coding). *IEEE Transactions on Communications*, 38(7):950–953, 1990. 3

[18] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 4

[19] Alvaro Gómez, Gregory Randall, Gabriele Facciolo, and Rafael Grompone von Gioi. An experimental comparison of multi-view stereo approaches on satellite images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 844–853, 2022. 1

[20] Ayman Habib, Eui-Myoung Kim, M Morgan, and Isabelle Couloigner. Dem generation from high resolution satellite imagery using parallel projection model. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 35(1), 2004. 1

[21] Shea Hagstrom, Hee Won Pak, Stephanie Ku, Sean Wang, Gregory Hager, and Myron Brown. Cumulative assessment for urban 3d modeling. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 3261–3264, 2021. 7

[22] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007. 3

[23] Thomas Krauß, Pablo d'Angelo, Mathias Schneider, and Veronika Gstaiger. The fully automatic optical processing system catena at dlr. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40:177–183, 2013. 2

[24] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 4

[25] Chuanhua Lu, Hideaki Uchiyama, Diego Thomas, Atsushi Shimada, and Rin-ichiro Taniguchi. Sparse cost volume for efficient stereo matching. *Remote sensing*, 10(11):1844, 2018. 3

[26] Xiaoyang Lyu, Yang-Tian Sun, Yi-Hua Huang, Xiuzhe Wu, Ziyi Yang, Yilun Chen, Jiangmiao Pang, and Xiaojuan Qi. 3dgsr: Implicit surface reconstruction with 3d gaussian splatting. *ACM Transactions on Graphics (TOG)*, 43(6):1–12, 2024. 1

[27] Roger Marí, Gabriele Facciolo, and Thibaud Ehret. Sat-nerf: Learning multi-view satellite photogrammetry with transient

objects and shadow modeling using rpc cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1311–1321, 2022. 1, 4

[28] Roger Marí, Gabriele Facciolo, and Thibaud Ehret. Multi-date earth observation NeRF: The detail is in the shadows. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2035–2045, 2023. 1

[29] Roger Marí, Thibaud Ehret, and Gabriele Facciolo. Disparity Estimation Networks for Aerial and High-Resolution Satellite Images: A Review. *Image Processing On Line*, 12:501–526, 2022. 3

[30] Elías Masquil, Roger Marí, Thibaud Ehret, Enric Meinhardt-Llopis, Pablo Musé, and Gabriele Facciolo. S-EO: A large-scale dataset for geometry-aware shadow detection in remote sensing applications. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. 5, 7, 3

[31] Julien Michel, Emmanuelle Sarrazin, David Youssefi, Myriam Cournet, Fabrice Buffe, Jean-Marc Delvit, Aurélie Emilien, Julien Bosman, Olivier Melet, and Céline L'Helguen. A new satellite imagery stereo pipeline designed for scalability, robustness and performance. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2:171–178, 2020. 2, 4

[32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 4

[33] Christopher Nuth and Andreas Kääb. Co-registration and bias corrections of satellite elevation data sets for quantifying glacier thickness change. *The Cryosphere*, 5(1):271–290, 2011. 7

[34] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3227–3237, 2020. 4

[35] Rongjun Qin. Automated 3d recovery from very high resolution multi-view satellite images. *arXiv preprint arXiv:1905.07475*, 2019. 1

[36] Ewelina Rupnik, Mehdi Daakir, and Marc Pierrot Deseilligny. Micmac–a free, open-source solution for photogrammetry. *Open geospatial data, software and standards*, 2:1–9, 2017. 2, 4

[37] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47:7–42, 2002. 3

[38] Federico Semeraro, Yi Zhang, Wenying Wu, and Patrick Carroll. Nerf applied to satellite imagery for surface reconstruction. *arXiv preprint arXiv:2304.04133*, 2023. 1

[39] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi DI Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 195–204, 2019. 3

[40] Thierry Toutin. Generation of dsms from spot-5 in-track hrs and across-track hrg stereo data using spatiotriangulation and autocalibration. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60(3):170–181, 2006. 1

[41] YuanZheng Wu, Jin Liu, and Shunping Ji. 3d gaussian splatting for large-scale surface reconstruction from aerial images. *arXiv preprint arXiv:2409.00381*, 2024. 1

[42] Jiayu Yang, Jose M Alvarez, and Miaomiao Liu. Self-supervised learning of depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7526–7534, 2021. 4

[43] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 4

[44] Kai Zhang, Jin Sun, and Noah Snavely. Leveraging vision reconstruction pipelines for satellite imagery. In *IEEE International Conference on Computer Vision Workshops*, 2019. 1

[45] L. Zhang and E. Rupnik. Sparsesat-nerf: Dense depth supervised neural radiance fields for sparse satellite images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-1/W1-2023:895–902, 2023. 4

[46] Tongtong Zhang, Yu Zhou, Yuanxiang Li, and Xian Wei. Satensorf: Fast satellite tensorial radiance field for multi-date satellite imagery of large size. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[47] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017. 1