Advances in Stereo Matching for Disparity Estimation from Satellite Imagery: Traditional Scanline Aggregation Methods versus Deep Learning-Based RAFTStereo

Yazgı Nur Sayın [1, 2], Ali Özgün Ok [1, 3]

[1] Hacettepe University, Dept. of Geomatics Engineering, Ankara, Türkiye –
yazgi.sayin@hacettepe.edu.tr, ozgunok@hacettepe.edu.tr
[2] TUBITAK Space Technologies Research Institute, Ankara, Türkiye – yazgi.sayin@tubitak.gov.tr
[3] National Intelligence Academy, Dept. of Computer Engineering, Ankara, Türkiye – a.ok@mia.edu.tr

Keywords: Stereo Matching, Disparity Map, Satellite Imagery, More Global Matching, Semi Global Matching, RAFTStereo.

Abstract

Stereo image analysis plays a critical role in geospatial domains, enabling the generation of high-resolution disparity maps from satellite imagery, and these maps are essential for producing accurate 3D surface models used in applications such as terrain modeling, urban planning, and environmental monitoring. This study conducts an evaluation of traditional scanline aggregation stereo matching methods, including Semi-Global Matching (SGM) and More Global Matching (MGM), with a deep learning-based approach, i.e., RAFTStereo. For satellite images, traditional stereo matching methods are still popular due to their balance of efficiency and robustness, and SGM / MGM could provide reliable disparity maps. However, the maturity of deep learning and availability of high-quality benchmark datasets have been steadily shifting the process toward fully automatic, accurate, and scalable solutions. In this study, the performance of SGM, MGM and RAFTStereo methods were investigated for disparity estimation using stereo images of Gaofen-7 satellite using the WHU-Stereo satellite dataset. Experimental evaluations indicate that MGM consistently achieves the lowest numerical errors (≈ 3-5 pixels), while RAFTStereo produces more visually coherent disparity maps with reduced noise and improved surface continuity. Traditional methods such as SGM and MGM remain robust and require no training, yet deep learning-based approach RAFTStereo demonstrate superior performance in radiometrically and geometrically complex scenes.

1. Introduction

Stereo image analysis plays a critical role in geospatial domains, and despite the recent developments, disparity estimation still possesses significant challenges. Acquisitions over large textureless areas (smooth agricultural land, water bodies), repetitive patterns, transparent objects (e.g., glass and water), high-saturation scenes, or poorly lit / shadow areas still pose difficulties for disparity estimation. Additionally, atmospheric effects may naturally create differences between observation times and radiometric variations (e.g., across track stereo), further complicating stereo matching. These challenges highlight the importance of developing robust methods that can generalize to varying conditions for satellite images.

Stereo processing primarily consists of stereo rectification and dense disparity estimation. Stereo rectification transforms image planes such that corresponding points are aligned horizontally, significantly simplifying the matching problem. Dense stereo matching algorithms, either traditional or learning-based, are then used to estimate disparity maps that represent pixel shifts between stereo images. For satellite images, traditional stereo matching methods such as Semi Global Matching (SGM) (Hirschmüller, 2007) are still popular due to their balance of efficiency and robustness (d'Angelo and Reinartz, 2012; Rothermel et al., 2012). Through combining matching cost at pixel-wise level with smooth constraints at multiple path aggregation, the SGM could provide reliable disparity maps featured by remote sensing (Xia et al., 2020). Besides, More Global Matching (MGM) improves upon SGM by incorporating additional information from previously visited neighbouring pixels, enhancing disparity consistency and robustness, particularly in angular directions (Facciolo et al., 2015). MGM also involves computing local matching costs, directional aggregation to reduce ambiguity, selecting minimum cost disparity, and post-processing refinement. Although very successful, traditional approaches suffer from complex scenes containing low texture, and changing illuminations, which are very typical scene properties also observed from spaceborne imagery datasets (Treible et al., 2018). To overcome these limitations, some fusion models have been proposed, obtaining large performance improvements for different satellite datasets (Gómez et al., 2023).

The establishment of Convolutional Neural Networks (CNNs) aimed at changing the stereo matching thoroughly. Earlier attempts, for example Matching Cost CNN (MC-CNN), already demonstrated the successful utilization of matching cost computations (Zbontar and LeCun, 2016). Deep learning methods developed afterwards, like GANet (Zhang et al., 2019), which incorporated ideas from SGM into an end-to-end trainable technique, resulting in substantial improvements in both accuracy and efficiency (Xia et al., 2022; Gómez et al., 2022). However, recent developments in deep learning have brought forward methods like RAFTStereo (Lipson et al., 2021) which utilize feature learning and guided cost aggregation to deliver improved disparity estimation, even under challenging conditions. RAFTStereo is an extension of the Recurrent All-Pairs Field Transforms (RAFT) model (Teed and Deng, 2020), in which the disparity map is refined iteratively through a recurrent update module based on Long Short-Term Memory (LSTM) structure. Gómez et al. (2022) examined multiple methods, both classical and deep learning-based stereo approaches, and Xia et al. (2020) provided a detailed comparison between traditional SGM and learning-based methods like GANet. The studies indicated that deep learning methods have promising prospects, either as parts of hybrid chains or as end-to-end solutions of satellite stereo processing systems in the future.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-4-W18-2025-287-2026 | © Author(s) 2026. CC BY 4.0 License.

This study assesses traditional scanline aggregation-based stereo matching methods, specifically Semi Global Matching (SGM) and More Global Matching (MGM), in relation to a deep learning-based method, RAFTStereo. The aim is to highlight their respective strengths, limitations, and applicability to geospatial tasks. Our emphasis is on evaluating the efficacy of the methods for disparity estimation utilizing high-resolution stereo images obtained from the Gaofen-7 (GF-7) satellite (Li et al., 2023).

The remainder of this paper is organized as follows: Section 2 introduces the three stereo matching methods under investigation, namely SGM, MGM, and RAFTStereo, and the pipeline framework utilised. Section 3 introduces the GF-7 dataset and pre-processing steps, including stereo rectification and ground truth. Section 4 presents the experimental setup, and discusses the comparative results, emphasizing the strengths and weaknesses of each approach. Finally, Section 5 concludes the paper with a summary of findings and directions for future research.

2. Methods Evaluated for Stereo Matching

Stereo matching, also referred to as correspondence matching, involves estimating a disparity map for rectified stereo image pairs that show positional shift of objects between two image views as a proxy for depth/elevation information. A multi-stage optimization pipeline that consists of matching cost computation, cost aggregation, disparity optimization and post-processing is typically used to mitigate the problem of stereo matching.
Traditional methods use low-level features extracted from local image patches around each pixel to estimate disparities (Hirshmuller and Scharstein, 2008). The effectiveness of traditional stereo matching methods is often limited by the use of manually generated features in matching cost functions. To get around these limitations, several deep learning-based methods have been introduced. The accuracy of depth estimation from stereo image pairs can be greatly enhanced by these novel techniques. However, because deep models require a lot of memory, they are frequently challenging to use in practice, especially when dealing with very-high-resolution satellite/aerial imagery. This is especially true when the graphics processing unit (GPU) memory needs to hold a large amount of 3D matching volumes. For this reason, traditional stereo algorithms continue to be useful, especially when dealing with large sized datasets. Even though deep learning is becoming more and more popular in many domains, this trend has not yet fully reached satellite stereo pipelines, which are still largely based on classical algorithms (Gómez et al., 2022).

Recent years have seen the emergence of deep learning-based techniques for determining matching costs; CNNs have demonstrated superior performance by improving their ability to assess patch similarity (Žbontar & LeCun, 2016). In general, deep learning models have been utilized either for aggregating matching costs using classical techniques like cost filtering and SGM or for separating the processes of cost computation and disparity estimation (Albanwan & Qin, 2022; Patil, 2022). Although these methods sometimes outperform traditional matching, they still encounter difficulties in accurately estimating disparity in ill-posed regions such as textureless or occluded areas. Despite reaching state-of-the-art performance, these hybrid frameworks are still limited by the drawbacks of conventional cost aggregation techniques, which frequently lead to imprecise disparity estimations, especially along object boundaries, in reflective regions, low-texture surfaces and occluded areas.

2.1 SGM

SGM was introduced to solve the computational cost of global methods (Hirschmüller, 2007). The method requires the generation of epipolar images as a prerequisite and consists of four main stages: matching cost computation, cost aggregation, disparity computation/optimization and refinement. To obtain a reliable depth map, a discontinuity preserving energy function is minimized. This energy is calculated along one-dimensional (1D) paths, not two-dimensional (2D) (e.g., left to right, top to bottom, etc.). In this way, data is collected from 8 or 16 different directions (Figure 1 - left). The information about each pixel from different directions is aggregated and, in the end, the lowest total cost is selected (winner-takes-all). This method is fast, but it can leave streaking artifacts. This is because aggregation for each line/column is seperated from each other.

2.2 MGM

MGM incorporates 2D contextual information into the 1D path-wise optimization framework of SGM (Facciolo et al., 2015). This is effectively accomplished by utilizing messages passed from previously visited pixels along the previous scanline (i.e., from pixels above). In the MGM algorithm, the matching decision at a given pixel is shaped not only by a few immediate neighbors but also by a broader region from the relevant quadrant (Figure 1 - right). This represents a major distinction from SGM, which only considers information propagated from specific directions.

MGM defines a dedicated scanning order for each direction (e.g. top-right, bottom-left) and accordingly computes directed cost accumulation values for each direction. The costs accumulated from all directions are then merged, correcting for redundant information. This merging is performed according to formula and the final matching is determined using the Winner-Takes-All method, selecting the disparity with the minimum total cost.

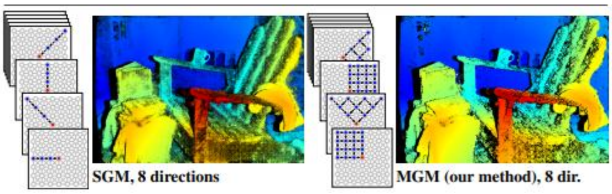


Figure 1. SGM and MGM approaches (Facciolo et al., 2015)

2.3 RAFTStereo

RAFTStereo is a deep learning model used for stereo depth estimation (Lipson et al., 2021). It is derived from the RAFT optical flow algorithm (Teed and Deng, 2020), where optical flow refers to the technique of detecting motion between two images. RAFTStereo computes the disparity (i.e., the horizontal shift) between two images to generate a depth map. Its key contributions include the use of multi-level Gated Recurrent Units (GRUs), cost volume optimization, and improved real-time performance.

RAFTStereo utilizes multi-level GRUs to enable more efficient information propagation across different resolutions. Unlike previous methods that employed computationally expensive 3D convolutional networks to process stereo cost volumes, RAFTStereo adopts a lightweight approach using 2D convolutions and simpler cost volume computations. Besides, its architecture is specifically designed for speed, enabling real-time inference, which makes RAFTStereo suitable for practical applications.

RAFTStereo employs a dual-encoder architecture consisting of a feature encoder and a context encoder for independent feature extraction (Figure 2). The feature encoder processes both the left and right input images, generating dense feature maps that are subsequently used to build the correlation volume. It is composed of a sequence of residual blocks and downsampling layers that generate feature maps at a resolution of 1/4 or 1/8 of the original image, contingent upon the number of downsampling layers. The feature encoder utilizes instance normalization throughout its layers. In contrast, the context encoder shares an identical architectural design but substitutes instance normalization with batch normalization and is exclusively applied to the left input image. The hidden state of the update module is initialized using context features, which are also fed into the GRU at every time step.

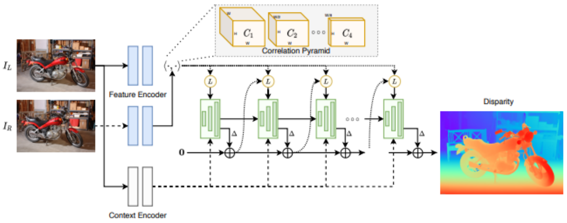


Figure 2. RAFTStereo Architecture (Lipson et al., 2021)

The correlation volume captures the matching information that the model uses to find correspondences between pixels in the left and right images. For each pixel, visual similarity to pixels in the opposite image is measured, generating a volume that helps the model predict disparity. To improve efficiency, a correlation pyramid is constructed by repeatedly applying average pooling to the last dimension, creating four levels of correlation volumes at different resolutions. Analyzing images at multiple scales enables the model to capture both large structures at lower resolutions and finer details at higher resolutions. Updates are performed across these different scales to enhance accuracy. During correlation lookup, a one-dimensional grid of integer offsets is generated around the current disparity estimate. The grid serves as an index for sampling features across multiple levels of the correlation pyramid. Due to the continuous nature of the grid coordinates, bilinear interpolation is employed during sampling. The resulting values are aggregated into a single feature map. The network iteratively updates disparity maps at multiple resolutions (e.g., 1/8, 1/16, and 1/32 of the input resolution). This allows the network to better handle areas with large textures and regions containing fine details.

RAFTStereo employs separate backbones for extracting correlation features and context features to support GRU updates. It was demonstrated that employing a unified backbone for extracting both correlation and context features enables faster inference while maintaining predictive accuracy. This single-backbone architecture is utilized in RAFTStereo.

2.4 Stereo Processing Pipelines Utilised

Recently, stereo processing pipelines including Satellite Stereo Pipeline (S2P) and GPU-Accelerated Binocular Stereo Pipeline (S2P-HD) have been designed focusing exclusively on satellite data, incorporating automation of numerous pre-analysis, matching, and triangulation steps with domain-specific corrections. S2P is an open-source, modular framework developed by École Normale Supérieure Paris-Saclay with CNES for automatic Digital Surface Model (DSM) generation from high-resolution optical satellite stereo imagery (De Franchis et al., 2014). Optimized for pushbroom sensors, it comprises geometric pre-processing, epipolar rectification, disparity estimation, and 3D triangulation. The framework addresses complex epipolar surfaces of pushbroom sensors by approximating Rational Function Model (RFM) with local affine models from first-order Taylor expansions, followed by dense stereo matching using SGM/MGM with Census-based cost metrics.

S2P-HD represents a very recent sophisticated evolution of S2P, incorporating specialized modifications for concurrent optical satellite data processing (Amadei et al., 2025). Key improvements include refined disparity range estimation leveraging reference models and multi-resolution analysis, highly optimized GPU-enhanced SGM techniques, improved rectification methods, and strategic tiling strategies for large geographic areas.

3. Dataset and Framework

The WHU-Stereo satellite dataset (Li et al., 2023) is built using high-resolution stereo imagery captured by the Chinese GF-7 satellite, which is equipped with a dual-line array stereoscopic camera system. High-quality in-track stereo pairs are ensured by the simultaneous capture of the panchromatic images from forward and backward viewing angles (−5° and +26°, respectively) and a ground sampling distance (GSD) of better than 0.8 meters. With the help of the dual-line array camera system aboard the GF-7 satellite, there is only a brief temporal offset between the forward- and backward-looking images, which are taken during the same orbital pass. This temporal alignment allows for accurate disparity estimation.

All computational experiments and evaluations presented in this study were carried out on a high-performance computing system. equipped with an Intel® Xeon® Platinum 8380 CPU running at 2.30 GHz, featuring 160 logical processors (40 cores per socket with 2 threads per core). The system also includes 503 GB of RAM and NVIDIA L4 24 GB GPUs, which provided significant acceleration for deep learning-based stereo matching and large-scale DSM generation tasks.

The S2P pipeline provides built-in support for the SGM algorithm and its extensions, the MGM framework. In addition to these, it incorporates several algorithmic variations (e.g., tvl1, msmw, sgbm, mgm_multi) that allow flexible adaptation of stereo processing to diverse terrain and radiometric conditions (De Franchis et al., 2014). S2P also offers configurable options for cost functions, regularization strength, and consistency filtering, enabling fine-tuning of disparity estimation accuracy and robustness without modifying the core algorithmic structure.

RAFTStereo includes iterative updates to the disparity map through a multi-level recurrent GRU based architecture. It utilizes correlation pyramids to keep the global context and works directly with high-resolution images. Training sessions (including fine tuning operations) were performed for RAFTStereo based on the WHU datasets. In the production of disparity maps, 1,222 left and right stereo images and disparity images with a resolution of 0.8 meters in the pan band, each measuring 1,024 x 1,024 pixels, were used and shared with ground truth information for training. In this study, the Shaoguan, Kunming, Yingde, and Qichun regions were provided for training. For testing, two 1024x1024 images from the Hengyang, and Shaoguan regions that were not included in the training process were utilised.

The WHU dataset consisting of 1024×1024 pixels epipolar rectified stereo pairs was used without additional preprocessing. Configuration was selected to preserve spatial resolution while keeping memory consumption within the limits of available hardware. Mixed precision training was utilized to accelerate computation without degrading numerical stability or performance.

4. Results and Discussion

The performance of stereo matching algorithms has been evaluated using both visual analysis and numerical measures. Table 1 presents the overall performance of the RAFTStereo algorithm on the generated disparity maps.

The comparative analysis between traditional stereo matching algorithms (SGM and MGM) and deep learning-based method (RAFTStereo) reveals several differences in their performance when applied to satellite stereo imagery. Traditional methods, particularly MGM, demonstrate robust and consistent performance across various datasets, offering reliable disparity maps and digital surface models (DSMs) with relatively lower computational requirements. MGM's structured cost aggregation approach provides smoother disparity outputs than SGM, effectively reducing streaking artefacts and improving disparity completeness in moderately challenging scenes.

Figures 2 and 3 present the right and left stereo images (for SG_0 and HY_6 regions), ground truth and output disparity maps produced by traditional and deep learning methods. According to the numerical results presented, the RMS distance is computed as ≈ 3-5 pixels for both the traditional and deep learning methods.

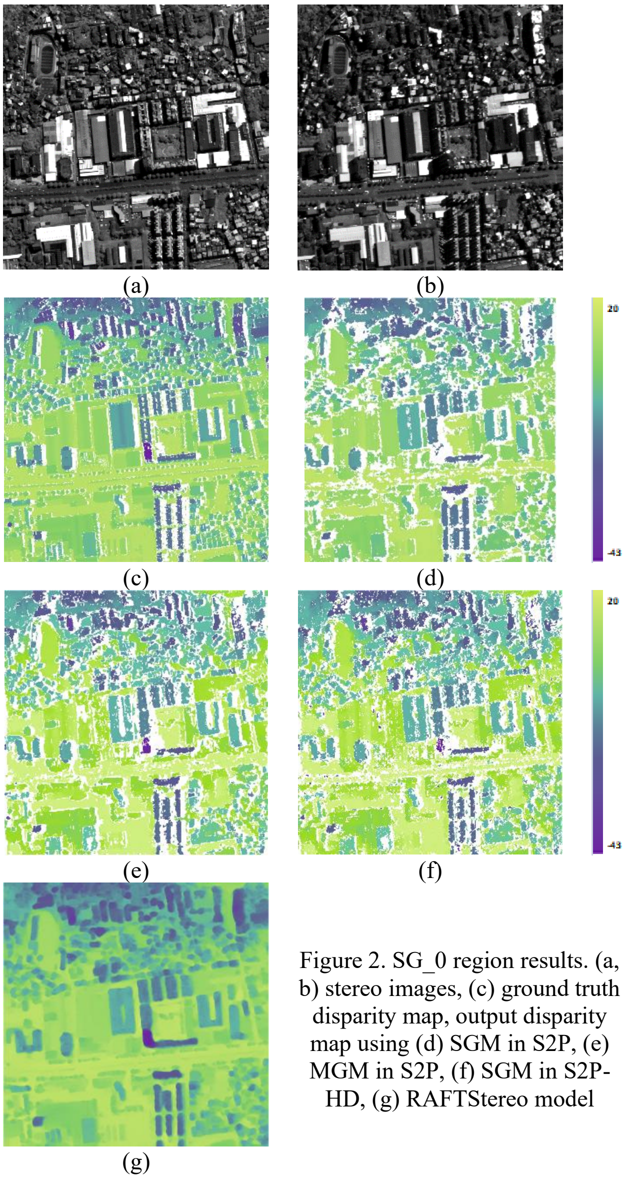
Data	*SGM (S2P)*			*MGM (S2P)*		
	COMP (%)	RMSE	MAE	COM P (%)	RMSE	MAE
GF-7 HY_6	53.40	4.78	2.20	64.76	**4.34**	1.98
GF-7 SG_0	73.59	3.60	1.93	75.51	**3.29**	1.73
Data	*SGM-GPU (S2P-HD)*			*RAFTStereo*		
	COMP (%)	RMSE	MAE	COM P (%)	RMSE	MAE
GF-7 HY_6	63.69	6.71	2.83	100	4.93	2.54
GF-7 SG_0	76.03	4.02	2.03	100	3.31	1.96

Table 1. The overall performance of the different methods. Best RMSE computed are given in bold.

Our experimental results reveal that MGM achieves high geometric accuracy, particularly in moderately textured areas, while RAFTStereo demonstrates performance in handling radiometric inconsistencies and textureless surfaces, offering computational efficiency during the inference stage.

As shown in Table 1, MGM method achieved the lowest error values, thus standing out in terms of numerical accuracy. This can be associated with MGM producing fewer systematic errors in disparity estimation. However, visual inspection revealed that the MGM outputs still contained localized noise and discontinuities. SGM produced slightly higher error than MGM. Nevertheless, it provided a more balanced result in terms of noise reduction and surface continuity. SGM was observed to generate more stable results in homogeneous regions, although it tended to cause detail loss around object boundaries.

Despite having the highest error among the four methods, S2P-HD SGM stands out with its ability to capture fine details and distinguish complex surface transitions. However, this detail-preserving property comes along with significant noise and reduced surface continuity. In other words, while the algorithm makes small structures more distinguishable, it also introduces instability and patchy results in homogeneous areas. This indicates that S2P-HD has a strong tendency toward detail preservation but is weaker in noise suppression. RAFTStereo



(a) (b)

(c) (d)

(e) (f)

(g)

Figure 2. SG_0 region results. (a, b) stereo images, (c) ground truth disparity map, output disparity map using (d) SGM in S2P, (e) MGM in S2P, (f) SGM in S2P-HD, (g) RAFTStereo model

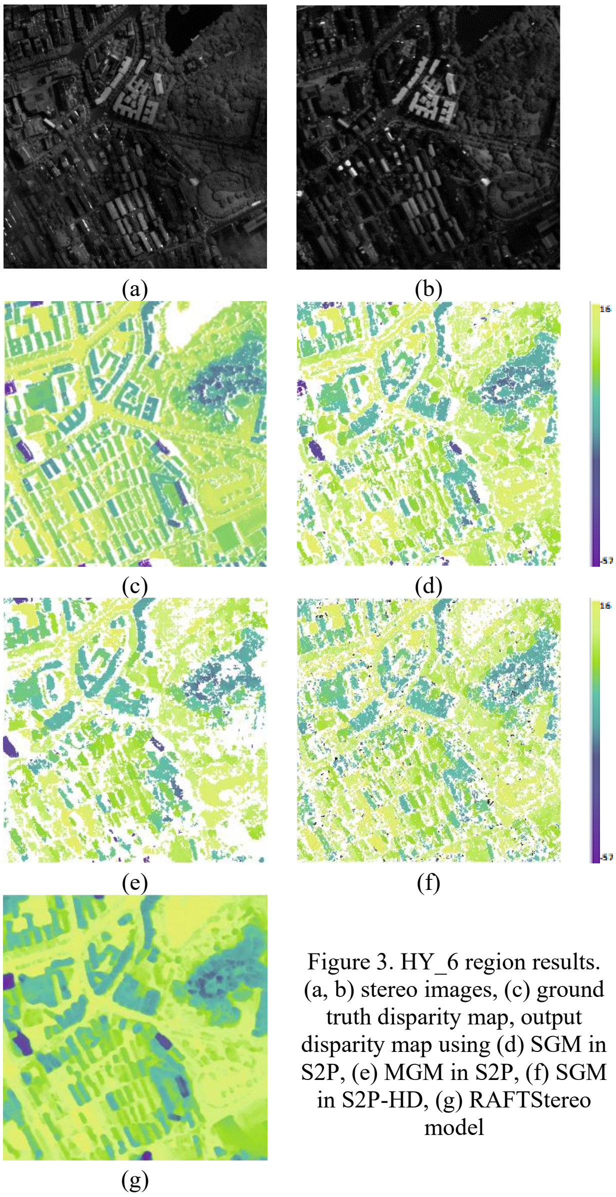


Figure 3. HY_6 region results. (a, b) stereo images, (c) ground truth disparity map, output disparity map using (d) SGM in S2P, (e) MGM in S2P, (f) SGM in S2P-HD, (g) RAFTStereo model

ranked second in terms of RMSE, but in visual analysis it produced one of the least noisy results with the highest surface continuity. The lower spatial variance observed in RAFTStereo outputs suggests that the method is advantageous in producing consistent disparity estimations in homogeneous regions. On the other hand, the slightly higher RMSE compared to MGM and SGM indicates that the method may introduce systematic shifts (bias) in elevation for certain parts.

MGM produced relatively stable results over water surfaces but exhibited considerable noise in forested areas. This indicates that the stability of MGM may decrease in textured regions. RAFTStereo yielded error values close to MGM but stood out in terms of visual quality. It produced smoother and more continuous surfaces in forested regions, with lower noise levels. Over water surfaces, its performance was comparable to MGM. This suggests that RAFTStereo can provide more consistent performance across both homogeneous and complex regions. Visual analysis of SGM revealed fragmented results in forested areas, while the method remained more stable in residential zones and along roads. Therefore, the performance of SGM appears to vary depending on the structural characteristics of the scene. S2P-HD SGM obtained the highest error among the methods, yet it demonstrated advantages in capturing small structural details and road boundaries. However, its tendency to generate significant noise in forested regions and lack of continuity over water surfaces highlight its weaknesses. This indicates that S2P-HD is strong in detail preservation but less effective on complex natural surfaces.

Overall, MGM and RAFTStereo produced comparable results in terms of numerical accuracy, while RAFTStereo showed superior visual quality, particularly in natural surfaces. SGM provided a balanced but scene-dependent performance, whereas S2P-HD SGM performed well in preserving small details but remained sensitive to noise in natural regions.

Figures 4 and 5 present the right and left stereo images (for SG_0 and HY_6 regions), difference output disparity maps produced by traditional and deep learning methods. Note that disparity errors exceeding 10 pixels with respect to the ground truth were masked out to facilitate clearer visualization and comparison.

In the first dataset (Figure 4 - SG_0 region), which predominantly consists of structured urban areas, roads, and rectangular agricultural parcels, the difference maps reveal the distinct characteristics of each method. SGM demonstrates low error magnitudes in an overall sense, yet systematic deviations are clearly concentrated along building edges and road boundaries, highlighting its limitations in capturing sharp structural discontinuities. Conversely, disparities over homogeneous agricultural fields are achieved with greater consistency. MGM provides a more uniform distribution of errors compared to SGM, with reduced concentrations in built-up regions and along roads, although residual noise persists in irregular structures and less-defined areas. S2P-HD SGM shows a stronger ability to preserve fine details, such as field boundaries and smaller objects, but this advantage comes at the expense of elevated error levels surrounding those regions, particularly around structural edges. RAFTStereo, on the other hand, produces the most coherent and visually smooth difference map, with very low errors across homogeneous surfaces and only localized deviations around dense building clusters and road intersections, confirming its superior capability in maintaining surface continuity.

The second dataset (Figure 5 - HY-6 region) presents a more complex scenario, comprising urban settlements, a water body, and dense forested regions. SGM achieves generally low error levels, yet suffers from significant deviations in forested areas and fragmented inconsistencies over the water surface, reflecting its difficulty in textureless and highly irregular regions. Within residential areas, however, its performance remains comparatively stable. MGM exhibits a more homogeneous error distribution and performs reliably over the water surface, though noise remains prominent within forested regions; error patterns in structured areas are more localized compared to SGM. S2P-HD SGM continues to preserve fine structural details and road boundaries, yet generates higher error magnitudes in natural regions, particularly forests and water surfaces. This behavior underscores its strength in detail preservation but also its issues in homogeneous or low-texture surfaces. RAFTStereo delivers the most balanced performance, with minimal errors over the water body and noticeably reduced noise in forested areas relative to the other methods. The remaining errors are largely confined to object boundaries and forest edges, highlighting RAFTStereo's robustness across both homogeneous and structurally complex natural environments.

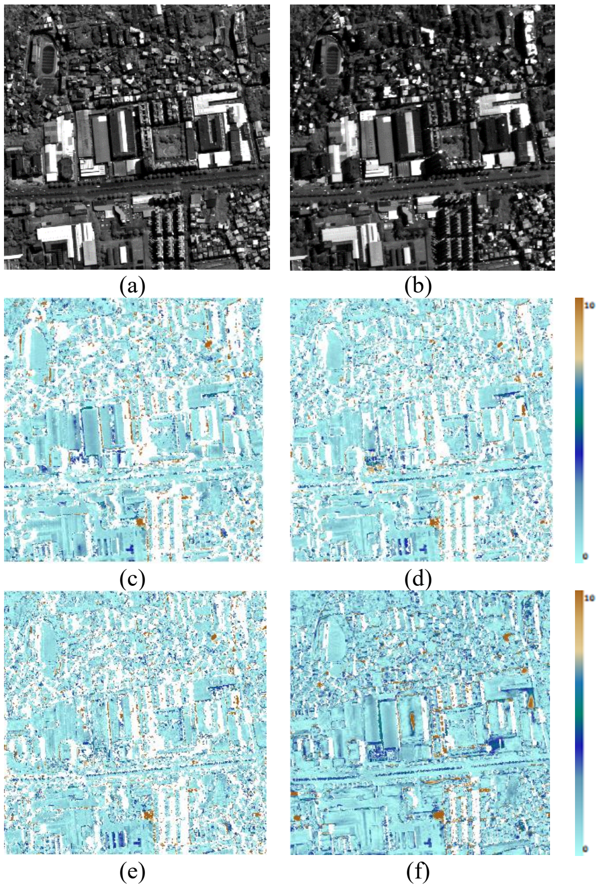


Figure 4. SG_0 region disparity difference maps. (a, b) stereo images, (c) difference map between GT and SGM output, (d) difference map between GT and MGM output, (e) difference map between GT and SGM (GPU acceralated) output, and (f) difference map between GT and RAFTStereo model output.

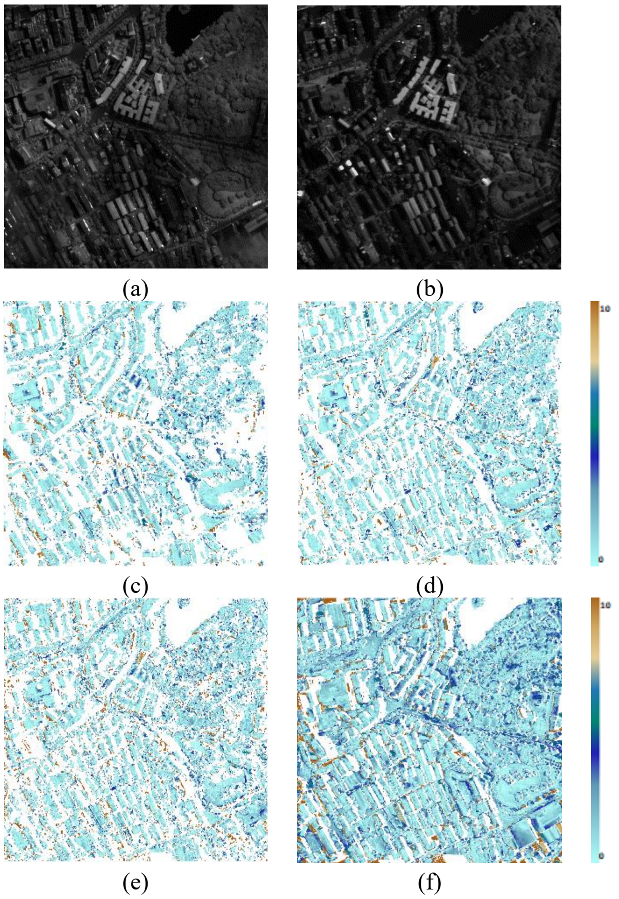


Figure 5. HY_6 region disparity difference maps. (a, b) stereo images, (c) difference map between GT and SGM output, (d) difference map between GT and MGM output, (e) difference map between GT and SGM (GPU acceralated) output, and (f) difference map between GT and RAFTStereo model output.

5. Conclusion

In this study, the process of generating disparity maps from high-resolution satellite stereo images using traditional stereo matching methods and a modern deep learning approach is examined using the stereo images of Gaofen-7 satellite available within the WHU-Stereo satellite dataset.

Traditional stereo matching methods, particularly SGM and MGM, continue to provide reliable performance in structured urban environments and textured terrains, especially under limited computational resources. However, their ability to generalize to challenging scenarios such as homogeneous surfaces, seasonal variations, and significant radiometric inconsistencies remains limited. Recent advancements in deep learning-based stereo matching methods, such as RAFTStereo, have demonstrated notable performance in disparity estimation accuracy, generalization and robustness across varying satellite imagery conditions. Although deep learning methods outperform classical techniques in most conditions, their dependency on extensive annotated training data and their generalization across different satellite platforms without fine-tuning remains a big concern.

The RMSE evaluations and difference map analyses presented in this study revealed complementary strengths and weaknesses among the tested methods. MGM consistently produced the lowest numerical errors, whereas RAFTStereo generated more visually coherent disparity maps with reduced noise and improved surface continuity, especially across homogeneous areas and natural surfaces such as water bodies and forests. SGM maintained stable performance in structured regions but showed systematic deviations along object boundaries, while S2P-HD SGM preserved fine structural details at the expense of increased noise and error concentrations in textureless regions. These results highlight the importance of jointly considering both numerical accuracy and spatial error distribution in the evaluation of stereo matching algorithms.

In summary, our analysis suggests that while traditional methods remain indispensable for specific conditions, the future of satellite stereo matching strongly leans towards deep learning-based and hybrid approaches, particularly for large-scale, high-precision disparity generation. Drawing upon the findings of this research and recent literature, future investigations should focus on designing memory-efficient deep architectures to efficiently process large satellite images and using hybrid methods. In addition, fine-tuning pre-trained models on satellite-specific datasets can significantly improve their ability to handle the unique characteristics of satellite imagery, including large disparity ranges, pushbroom sensor geometries, and multi-temporal variations. The availability of publicly available datasets remains essential for unbiased evaluation, and careful benchmark design is fundamental to driving progress in satellite-based disparity generation.

Acknowledgements

The presented work was supported by the Secretariat of Defence Industries (SSB) within the scope of the Defence Industry Artificial Intelligence Talent Cluster (SAYZEK) project.

References

Albanwan, H., Qin, R., 2022. A comparative study on deep-learning methods for dense image matching of multi-angle and multi-date remote sensing stereo-images. *Photogramm. Rec.*, 37, 385–409. doi.org/10.1111/phor.12425

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. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2025.

d'Angelo, P., Reinartz, P., 2012. Semiglobal matching results on the ISPRS stereo matching benchmark. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XXXVIII-4/C22, 79–84. doi.org/10.5194/isprsarchives-XXXVIII-4-W19-79-2011

De Franchis, C., Meinhardt-Llopis, E., Michel, J., Morel, J.-M., Facciolo, G., 2014. An automatic and modular stereo pipeline for pushbroom images. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, II-3, 49–56. doi.org/10.5194/isprsannals-II-3-49-2014

Facciolo, G., De Franchis, C., Meinhardt, E., 2015. MGM: A significantly more global matching for stereovision. Proc. Brit. Mach. Vis. Conf. (BMVC), 1–12. doi.org/10.5244/C.29.90

Gómez, A., Randall, G., Facciolo, G., von Gioi, R.G., 2022. An experimental comparison of multi-view stereo approaches on satellite images. Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2937–2946. doi.org/10.1109/WACV 51458.2022.00078

Hirschmüller, H., 2007. Stereo processing by semi-global matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2), 328–341. doi.org/10.1109/TPAMI.2007.1166

Hirschmüller, H., Scharstein, D., 2008. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(9), 1582–1599. doi.org/10.1109/TPAMI.2008.221

Li, S., He, S.,Jiang, S., Jiang, W., Zhang, L., 2023. WHU-Stereo: A challenging benchmark for stereo matching of high-resolution satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1-14. doi.org/10.1109/TGRS.2023.3245205

Lipson, L., Teed, Z., Deng, J., 2021. RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. International Conference on 3D Vision (3DV), London, United Kingdom, 2021, pp. 218-227, doi.org/10.1109/3DV53792.2021.00032

Patil, S.D., 2022. Automatic 3D Earth landscape reconstruction by satellite stereo. Doctoral dissertation, Purdue University, USA.

Rothermel, M., Wenzel, K., Fritsch, D. and Haala, N., 2012, December. SURE: Photogrammetric surface reconstruction from imagery. Proceedings LC3D workshop, Berlin 8(2).

Teed, Z. and Deng, J., 2020, August. RAFT: Recurrent all-pairs field transforms for optical flow. European conference on computer vision, pp. 402-419. Cham: Springer International Publishing, 2020. doi.org/10.24963/ijcai.2021/662

Žbontar, J., LeCun, Y., 2016. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(65), 1–32.