Stereo Matching Network with Transformer-CNN Feature Fusion and ConvGRU Refinement for High-resolution Satellite Stereo Images

Mengran Yang [1], San Jiang [2] *, Wanshou Jiang [3], Qingquan Li [2]

[1] School of Computer Science, China University of Geosciences, Wuhan 430074, China - jiangsan@cug.edu.cn
[2] Guangdong Key Laboratory of Urban Informatics, Shenzhen University, Shenzhen 518060, China
[3] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China

KEY WORDS: Satellite stereo images, Disparity estimation, Transformer, Gate recurrent unit, Convolutional neural network

ABSTRACT:

In photogrammetry and remote sensing, disparity estimation of satellite images has been a significant and challenging task, holding crucial importance for research and applications in this domain. Recent years have seen substantial progress in stereo matching methods, but challenges remain significant in ill-posed regions. Although deep learning-based stereo matching methods outperform traditional approaches in terms of performance and speed, their limited receptive field makes it difficult for networks to establish long-distance dependencies. This poses challenges in ill-posed areas such as textureless regions, repetitive patterns, and occluded areas. This paper proposes an end-to-end model for high-resolution satellite remote sensing images. First, in the feature extraction stage, we use two independent Transformer and CNN modules to extract global and local features of stereo image pairs. Subsequently, by designing effective fusion strategies, we merge these two types of features to obtain richer and more accurate feature representations. Next, we utilize multi-scale features to construct multi-level cost volumes, supervising each level of cost volume from coarse to fine. This allows lower-level cost volumes to provide prior knowledge to higher-level cost volumes, guiding them to acquire richer and more accurate information. Finally, we employ a ConvGRU-based recurrent module in the refinement module on geometrically encoded cost volumes containing geometric and contextual information to iteratively update disparity maps with finer details and structures. In experiments, we validate our approach using publicly available datasets and compare it with traditional methods. Experimental results demonstrate significant performance improvements in stereo matching tasks, proving the effectiveness of our proposed method.

1. INTRODUCTION

The estimation of disparity in satellite stereo imagery has long been a significant and challenging task in photogrammetry and remote sensing. It refers to the process of determining the three-dimensional geometric relationships of corresponding points on the earth's surface using multi-angle image data obtained from satellites that have undergone epipolar rectification, by calculating the disparity information between the images. This process has broad applications in geographic information systems, urban planning, environmental monitoring, and can be used for digital map production, terrain measurement, resource management, and other areas. However, the large amount of data brought by high-resolution imagery and the presence of ill-posed areas such as common textureless, repetitive textures, and occlusions in satellite images pose numerous challenges for depth estimation in satellite stereo imagery.

Traditional methods typically use a four-step pipeline approach in stereo matching (Scharstein et al., 2001): matching cost computation, cost aggregation, disparity computation, and disparity optimization. Despite the progress made by traditional methods (Hosni et al., 2013; Scharstein and Szeliski, 2003)over the years, handcrafted features are often too simplistic and lack computational complexity, thus having limitations in handling complex situations. These limitations include a weak ability to handle textureless or repetitive texture regions, which often lead to matching difficulties, thereby affecting the accuracy of the matches. Additionally, traditional methods (Bleyer et al., 2011; Fife and Archibald, 2013) often struggle to effectively capture long-range dependencies, resulting in poor performance when dealing with large-scale and irregular surfaces. Traditional methods also face significant challenges in processing high-resolution satellite remote sensing imagery due to high computational complexity and time-consuming operations. Furthermore, their dependence on manual parameter adjustment or heuristic methods results in a lack of adaptability and intelligence. Therefore, addressing these limitations, deep learning-based methods (Chang and Chen, 2018; Kendall et al., 2017; Mayer et al., 2015) have gradually gained attention in recent years and have made significant progress in stereo matching.

GC-Net (Kendall et al., 2017)is the first model to apply an end-to-end method to stereo matching. It uses 3D convolutions to aggregate and regularize the 4D cost volume and employs soft argmin linear regression to generate the disparity map. PSM-net (Chang and Chen, 2018) utilizes a pyramid pooling module to extract multi-scale features and aggregates the cost volume using an encoder-decoder structure, supervised by the regression of multi-stage disparity maps. Although learning-based methods have made significant progress in stereo matching, most methods (Guo et al., 2019; Khamis et al., 2018) typically construct the 4D cost volume, requiring substantial computation and memory resources for 3D convolution aggregation and regularization. Additionally, their limited receptive field causes poor performance in textureless and weakly textured areas, as well as difficulty aligning the final disparity map with remote sensing images due to multiple downsampling operations during feature extraction.

Recent studies (Carion et al., 2020; Liu et al., 2021; Zheng et al., 2021) have demonstrated the advantages of Transformer in capturing long-range dependencies. STTR (Li et al., 2021) re-evaluates stereo matching from a sequence-to-sequence perspective and replaces the construction of cost volumes with dense pixel matching using positional information and attention mechanisms. The ultimate outcome confirms the feasibility and effectiveness of Transformer in the field of stereo matching.

The proposal of RAFT-Stereo (Lipson et al., 2021) has introduced a new perspective to stereo matching research.

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-X-G-2025-995-2025 | © Author(s) 2025. CC BY 4.0 License.

Inspired by the RAFT (Teed and Deng, 2020), it iteratively updates the disparity map along the coarse-to-fine pipeline and employs a ConvGRUs as the core unit. Utilizing a lookup operator to retrieve features from the correlation cost volume within the current disparity range, it effectively updates the disparity map, thereby contributing to advancements in stereo matching research.

While these studies (Teed and Deng, 2020; Zheng et al., 2021) have significantly advanced stereo matching, satellite images present unique challenges due to their different acquisition characteristics, leading to more prevalent ill-posed regions (mainly textureless areas and multi-scale objects, as well as the preservation of edge structure details). Satellite imagery (Li et al., 2023; Tao et al., 2020) often contains large buildings and objects of various scales, numerous areas with weak or repetitive textures, and disparities resulting from different shooting angles. Addressing these complex issues requires long-range dependency in stereo matching from remote sensing images and poses challenges for recovering fine details in disparity map edge structures.

To address these challenges, we propose TFGR-Stereo (Stereo matching network with Transformer-CNN Feature Fusion and ConvGRU refinement for high-resolution satellite stereo images) in this paper. Firstly, our network employs Transformer modules similar to the U-net (Ronneberger et al., 2015) structure to extract global features and fuse them with locally extracted features based on CNN. This fusion captures advantages in long-range dependencies and enriches the network with detailed information. Secondly, we design effective fusion strategies to merge different features to obtain richer and more accurate feature representations. Finally, we utilize an improved module based on ConvGRU to iteratively update the disparity map, thereby achieving a more refined result with better details.

In this paper, we will revisit and summarize related work in Chapter 2, provide a detailed description of our network structure in Chapter 3, and finally present and discuss our experiments in Chapter 4.

2. RELATED WORKS

2.1 CNN-based Methods

In recent years, significant advancements have been made in stereo matching through learning-based methods. GC-Net (Kendall et al., 2017), as a pioneering effort in deep learning for stereo matching, introduced an end-to-end approach by incorporating soft-argmin regression for generating the stereo disparity map. PSM-Net (Chang and Chen, 2018) integrated a spatial pyramid pooling module (Zhao et al., 2017) during feature extraction, effectively fusing multi-scale features and constructing a 4D cost volume. Subsequently, it employed a 3D hourglass convolutional block for cost aggregation learning. To alleviate computational and memory burdens, GWC-Net (Guo et al., 2019) employed a grouped correlation method for cost volume construction, grouping feature channels and computing their dot products sequentially. ACV-Net (Xu et al., 2024) addressed redundancy issues in existing methods' cost volumes, utilizing a subnetwork to generate attention weights for suppressing redundant information, thereby easing the regularization pressure in aggregation networks. ACV-Net (Xu et al., 2024) also serves as a lightweight module embeddable in most stereo matching frameworks. AA-Net (Xu and Zhang, 2020) replaced 3D convolutions with an adaptive cost aggregation method for enhanced aggregation, employing Intra-Scale Cost Aggregation (ISA) and Cross-scale Cost Aggregation (CSA) algorithms to mitigate edge-fattening in cases of disparity discontinuity and matching errors in textureless regions. CFNet (Shen et al., 2021) constructed a pyramid cost volume to narrow the disparity search range and refined the disparity map in a coarse-to-fine manner.

In the field of stereo matching in satellite remote sensing imagery, there have been significant advancements in recent years. To address common challenges such as textureless regions, disparity discontinuities, and occluded areas in remote sensing imagery, DSM-net (He et al., 2021) proposed a dual-scale learning network for stereo matching, with low-level capturing coarse-grained information and high-level capturing fine-grained information, aiding in matching objects at different scales. For simultaneous semantic segmentation and stereo matching tasks in remote sensing imagery, BGA-Net (Rao et al., 2021) introduced a multi-task architecture of a bidirectional guidance attention network, where both tasks share feature information to enhance overall performance. HMSM-Net (He et al., 2022), on the other hand, trained the network by hierarchically regressing multiple disparity maps, with the low-level cost volume providing prior knowledge to the high-level cost volume, guiding it to obtain richer and more accurate feature representations. Our approach shares conceptual similarities with the former methods but employs different strategies in feature extraction, feature fusion, and enhancement modules, resulting in further improvements in model performance.

2.2 Transformer-based Methods

In recent years, the exceptional performance of Transformer in computer vision (Wang et al., 2021; Zamir et al., 2021) has demonstrated their strong self-attention mechanisms' advantage in capturing long-range dependencies between pixels. STTR (Li et al., 2021) reexamined the depth estimation problem from a sequence-to-sequence perspective, replacing the construction of cost volumes with dense pixel matching using positional information and attention mechanisms, thereby advancing stereo matching. ELF-Net (Lou et al., 2023) introduced internal evidence fusion and external evidence fusion modules based on a mixed normal-inverse Gamma distribution (MoNIG) to simultaneously integrate multi-scale cost volume information and Transformer-extracted feature information, further enhancing disparity estimation accuracy and robustness. Although Transformers excel in extracting global features, they often overlook the detailed features needed for local regions, such as texture and other shallow information. To address this limitation, this paper introduces a novel feature extraction module that combines locally extracted features from CNNs with globally extracted features from Transformers. This fusion method integrates both local and global information to achieve a more comprehensive and accurate feature representation.

2.3 Iterative Methods

The impressive performance of iterative methods in stereo matching is remarkable. Inspired by the RAFT (Teed and Deng, 2020) optical flow estimation model, RAFT-Stereo (Lipson et al., 2021) extends its framework to stereo matching. Initially, the model constructs a 4D cost volume by computing similarities between all pairs of pixels. Subsequently, it utilizes a lookup operator to retrieve iterative features from the correlation volume within the current disparity range, and finally updates disparities using a ConvGRU-based update operator. However, DLNR (Zhao et al., 2023) pointed out the potential loss of high-frequency information during the iteration process and overly tight coupling in the ConvGRU update operator modules, proposing decoupled modules to mitigate these issues. On the other hand, IGEV-Stereo (Xu et al., 2023) argued that algorithms similar to RAFT-Stereo (Lipson et al., 2021) lack non-local

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-X-G-2025-995-2025 | © Author(s) 2025. CC BY 4.0 License.

geometric knowledge when building global correlations, thus introducing a method that combines encoding of geometric cost volumes integrating geometric, contextual, and local information. Our study adopts the concept of geometric encoded cost volumes in the improved ConvGRU module, utilizing a 1/4 resolution cost volume regularized by an hourglass aggregation module, facilitating iterative updates of disparities and generating a weighted combination necessary for final full-resolution disparity.

3. METHOD

This chapter provides a detailed description of the network architecture of TFGR-Stereo (see Figure 1), and key components including a feature extractor based on Transformer and CNN for extracting and integrating feature information from images. Additionally, a feature fusion module is designed to effectively integrate multi-scale information from different levels to enhance overall performance. Furthermore, an improved module based on ConvGRU (Zamir et al., 2021) aims to address the loss of detailed information during downsampling, catering to the alignment requirements between the disparity map and reference image structures in depth estimation. Finally, the loss function is defined and optimized to guide objective optimization during the training process. Through a comprehensive discussion of these key components, we elucidate the methods and mechanisms by which TFGR-Stereo achieves efficient and precise depth estimation in stereo matching tasks for satellite remote sensing images.

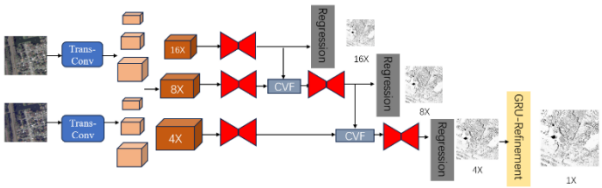


Figure1. Overview of our proposed TFGR -Stereo. TFGR-Stereo firstly passes the satellite remote sensing images which have undergone epipolar rectification into a dual-branch feature extraction module. This module is capable of modeling long-range dependencies and integrating rich detailed information. The extracted multi-scale feature information is organized into multi-level cost volumes, which are then regularized in a hierarchical manner from coarse to fine. Finally, a refinement module based on ConvGRU is employed to produce the ultimate full-resolution disparity.

3.1 Feature Extraction

3.1.1 Channel-Attention Transformer extractor: In the task of stereo matching for satellite remote sensing imagery, accurate and comprehensive feature representation is crucial for subsequent steps. Most learning-based networks (He et al., 2022; Khamis et al., 2018) employ feature extraction modules similar to residual networks (He et al., 2016), but the limited receptive field prevents the network from capturing global information, which is particularly disadvantageous for satellite remote sensing images with numerous textureless or weakly-textured areas. The success of Transformer in computer vision (Carion et al., 2020; Liu et al., 2021) has demonstrated their advantage in long-range modeling, yet their computational complexity grows quadratically due to their self-attention mechanism. Inspired by Restormer (Zamir et al., 2021) and DLNR (Zhao et al., 2023), we have designed a channel-attention Transformer feature extraction module resembling a U-net (Ronneberger et al., 2015) structure (as shown in Figure 2) to output only the final 1/4 resolution feature map.

3.1.2 Channel Attention Mechanism: Because of the high-resolution nature of satellite remote sensing imagery, traditional Transformer modules (Carion et al., 2020; Zheng et al., 2021) impose significant computational burdens when processing such image data. To address this challenge, we introduce the CWSA module, derived from Restromer's MDTA module. The CWSA module enhances computational efficiency by linearly handling self-attention across channel dimensions, thereby significantly improving model efficiency.

3.1.3 Dual-branch Feature extractor architecture: In the task of stereo matching for satellite remote sensing imagery, Transformer offers advantages in capturing global contextual information. However, they may overlook detailed information required for local regions. Maintaining sharp edges of objects of varying sizes and ensuring alignment between disparity maps and reference image structures are crucial for the final results. A hybrid deep neural network (Zhang et al., 2022) based on Transformer and CNN has been proposed for semantic segmentation of Very High-Resolution remote sensing images. In this network architecture, the encoder utilizes Transformer modules for global modeling, while the decoder employs CNN modules to restore detailed information. It is posited that Transformers and CNNs, due to their inherent computational mechanisms, each have specific strengths in information extraction. When integrated into a single network, these modules might not fully exploit their respective advantages in feature extraction. Additionally, satellite imagery contains rich high-frequency information, including textures and other details. Integrating Transformer and CNN modules within a similar U-net-like network structure could lead to one module failing to extract necessary high-frequency information. Consequently, in this paper we use a separate CNN feature extraction module (He et al., 2022) specifically for capturing local information while the Transformer module is used for global information extraction. The features extracted by each module are then fused and fed into a pyramid pooling module, providing a rich and accurate multi-level feature representation for subsequent processing steps.

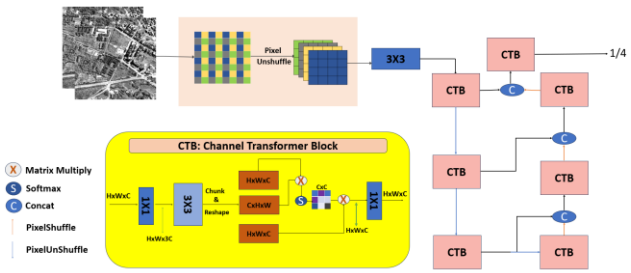


Figure2. Channel-Attention Transformer extractor. We construct a Transformer-based feature extraction module whose architecture is like U-net (Ronneberger et al., 2015). This module is capable of capturing global contextual information, providing significant advantages for satellite remote sensing images that contain abundant textureless or weakly textured regions. Additionally, it outputs feature maps at a 1/4 resolution.

3.2 Feature Fusion

In our approach, whether it is the early-stage feature fusion or the subsequent cost volume fusion, our design philosophy is to utilize the representation of global information as prior knowledge for local information representation and guide the network fusion (as shown in Figure 3). The specific structure is illustrated in Figure 3. Regarding the fusion of global features extracted by the Transformer and local features extracted by

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-X-G-2025-995-2025 | © Author(s) 2025. CC BY 4.0 License.

CNN, we first concatenate these two types of features along the channel, then generate channel attention vectors through a series of processing layers. These vectors are softmax-normalized to ensure their sum is equal to 1. Next, we multiply the global features extracted by the Transformer with the front half of the channel attention vectors, and multiply the local features extracted by CNN with the latter half of the channel attention vectors. Finally, these two parts are added together to obtain a rich and accurate feature representation, which is then fed into the subsequent pyramid pooling module for further processing. As for the cost volume fusion, we initially perform bilinear interpolation on the low-resolution cost volume to upsample it to a higher scale consistent with adjacent cost volumes. Then, similar to the feature fusion operation mentioned above, we ensure that the fused cost volume retains rich information and provides precise multi-scale representations for subsequent steps.

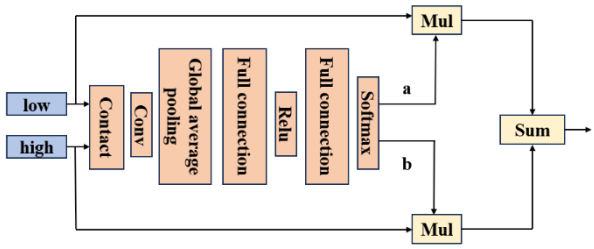


Figure3. First, the low-level features are concatenated with the high-level features and then fed into a convolutional processing layer for mapping. After undergoing a series of operations, such as global pooling, attention weights a and b are generated whose sum is 1. Subsequently, these two types of features are multiplied by their corresponding attention weights and then added together to produce a more accurate and detailed feature representation.

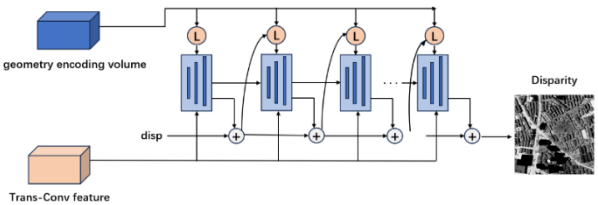


Figure4. We use the regularized 1/4 resolution cost volume as the geometric encoding volume and take its generated disparity as the initial disparity. Subsequently, we apply the refinement module based on ConvGRU for iterative updates and utilize the generated weights to output the full-resolution disparity map.

3.3 ConvGRU-based Refinement Module

Following multiple layers of hourglass-shaped cost aggregation (Chang and Chen, 2018) and fusion modules, initial resolution is achieved. Due to the high resolution of satellite remote sensing imagery, to alleviate computational and memory burdens, we process feature information fused from Transformer and CNN through a series of operations, mapping it to hidden layer information required for iterative modules. The normalized 1/4-resolution cost volume, containing both geometric and contextual information, serves as the geometric encoding cost volume (Xu et al., 2023). The resulting disparity map, derived from this, serves as the initial disparity map, both of which are fed into an iterative module based on ConvGRU (Lipson et al., 2021). Unlike iterative methods such as RAFT-Stereo (Lipson et al., 2021), where ConvGRU-based update operators iteratively refine disparities to generate weighted combinations needed for the final full resolution, our design is more streamlined and lightweight. During each iteration, search operators retrieve iterative features from the geometric encoding cost volume based on the current disparity, enriched by regularization and fusion. This ensures that the cost volume not only contains geometric information but also richer global context, enabling search operators to precisely supplement the necessary details for the final disparity map alignment with the reference image structure.

3.4 Loss Function

The disparity regressed by the hierarchical multi-scale cost volumes are optimized using Smooth L1 Loss:

$$L = \frac{1}{N}\sum_{i=1}^{N} smooth_{L1}(d_{gt} - d_i) \quad (1)$$

where N represents the total number of valid pixels, d_{gt} denotes the ground truth, and d_i signifies the predicted disparity value. The hierarchical total loss is the weighted sum of the disparity losses at each branch:

$$L_h = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3 \quad (2)$$

For the disparity generated by the refinement module based on ConvGRU, we utilize the following equation to compute the loss (Lipson et al., 2021):

$$L_{refine} = \sum_{i=1}^{N} \gamma^{N-i} \left\| d_{gt} - d_{refine} \right\|, \quad where\ \gamma = 0.9 \quad (3)$$

Where N denotes the total number of valid pixels, and, according to empirical experience, γ is set to 0.9, d_{gt} represents the ground truth, and d_{refine} signifies the predicted disparity value by ConvGRU-based refinement module.
The total loss is defined as:

$$L_{total} = L_h + \lambda L_2 \quad (4)$$

Where λ represents the loss weight, used to balance the learning between the refinement module and the hierarchical branches.

4. EXPERIMENTS

4.1 Evaluation Metrics

We employ the average endpoint error and the fraction of erroneous pixels as the evaluation criteria for this experiment. The specific evaluation equations are presented as follows:

$$EPE = \frac{1}{N}\sum_{k\in T} \left| d_{gt} - d \right| \quad (5)$$

$$D_1 = \frac{1}{N}\sum_{k\in T} \left[\left| d_{gt} - d \right| \right] > t \quad (6)$$

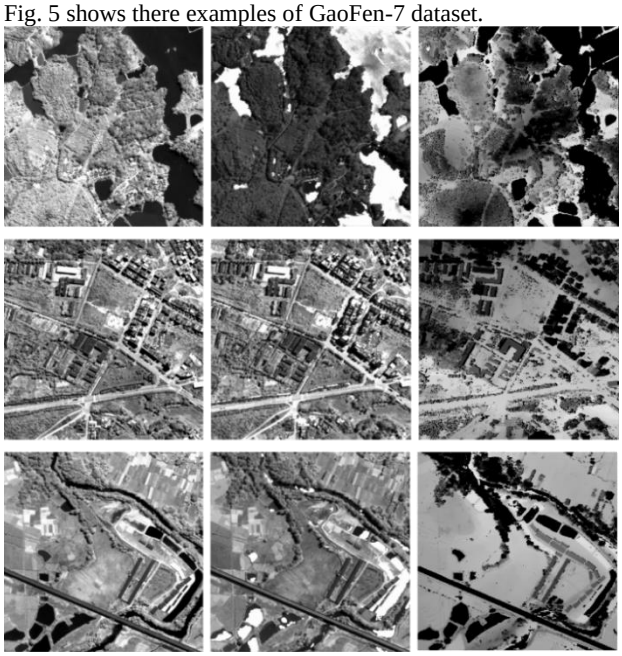
The number N represents the number of valid pixels, and T denotes the collection of pixels within a given threshold. The symbols d_{gt} and d correspond to the ground truth and network-predicted disparities, respectively, while t stands for the disparity error threshold, typically set at 3.

4.2 Dataset

GaoFen-7 (He et al., 2022) is an open dataset designed for high-resolution satellite image stereo matching, comprising 490 image pairs that have been rectified using ground control points and annotated with ground truth labels. Among these, 400 image pairs are allocated for training, while the remainder are used for validation and testing. Each image in this dataset is a single-channel image with a depth of 16 bits and a resolution of 1024

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-X-G-2025-995-2025 | © Author(s) 2025. CC BY 4.0 License.

×1024 pixels. Disparity maps are stored as 16-bit floating-point values, covering scenes from various Chinese cities and surrounding areas. These images depict diverse and complex environments with objects of different sizes (as illustrated in Figure 5), including buildings, roads, water bodies, and forests. Due to variations in lighting and angles during acquisition, regions of ambiguity are prevalent, posing challenges for models. Fig. 5 shows there examples of GaoFen-7 dataset.



(a) (b) (c)

Figure 5.The examples of the GaoFen-7 dataset. (a) left images; (b) right images; (c) ground-truth disparity maps

4.3 Implementation Details

We implemented TFGR-Stereo using PyTorch and conducted experiments on an NVIDIA RTX 4090 GPU. Due to the iterative nature of our model compared to other methods, during the training of the Transformer module, we employed an end-to-end Adam optimizer ($\beta 1 = 0.9$, $\beta 2 = 0.999$). The network was initialized with pre-trained weights from HMSM-Net (He et al., 2022) and trained for 100 epochs with an initial learning rate of 0.001, halving it every 10 epochs. For training the ConvGRU-based refinement module, we adopted a strategy of freezing the model parameters initially, utilizing the AdamW optimizer with an initial learning rate of 0.0002, and performing a total of 2000K training steps. Subsequently, we unfroze the parameters to involve the entire model in training for an additional 2000K steps. For the GaoFen-7 dataset, we normalized input images to a pixel intensity range of -1 to 1 for data preprocessing. To preserve high-frequency information in images and maintain object integrity, normalized image pairs of size 1024×1024 were directly input into the network without cropping or resizing during training, and no data augmentation was applied. The disparity range for GaoFen-7, based on disparity distribution, was set to [-112, 16). Empirically, the loss weights $\lambda 1$, $\lambda 2$, $\lambda 3$, and λ were set to 0.5, 0.7, 1.0, and 0.6, respectively.

4.4 Results

Table 1 presents a comparison of EPE and D1 metrics on the GaoFen-7 test set. It is evident that our method outperforms traditional approaches by a large margin, and achieves state-of-the-art performance among learning-based methods. Compared to HMSM-Net (He et al., 2022), our proposed method surpasses it by 8.41% in terms of the D1-all metric. The examples shown in Figure 6 and Figure 7 fully demonstrate the superiority of our proposed network. Intuitively, we can observe that our network predicts excellent disparity results in textureless areas such as building rooftops and forests. Moreover, in the case of multiple objects, edge predictions are also remarkably smooth. In satellite remote sensing imagery, the presence of abundant common urban objects such as roads, buildings, and forests often leads to the problem of thickening edges. However, as indicated in Figure 6 and Figure 7, the road marked in the image shows that the edges predicted by our network are very sharp, resulting in a well-aligned predicted disparity map with the original image structure. All these results strongly validate the effectiveness of our proposed module.

Network	EPE	D1
DenseMapNet	3.066	35.01
StereoNet	2.091	21.07
PSMNet	1.971	18.62
HMSM-Net	1.762	14.74
TFGR-stereo	1.702 ↓	13.40 ↓

Table 1. Quantitative comparison of different methods on the GaoFen-7 testing set.

4.5 Ablation Study

To validate and better comprehend the effectiveness of the proposed modules, extensive ablation experiments were conducted on the GaoFen-7 dataset, maintaining consistency in the setting of all hyperparameters with the aforementioned training. Detailed results are presented in Tables 2 and 3.

Model	Channel-Attention Transformer Extractor	ConvGRU-based Refinement Module	GaoFen7	
			EPE	D1
baseline			1.762	14.7
Net-V1	√		1.719	13.4
Net-V2		√	1.747	13.6
Net-full	√	√	1.702 ↓	13.4 ↓

Table 2. Ablations on different modules of our proposed model

4.5.1 Channel-Attention Transformer extractor: The Channel-Attention Transformer feature extractor (Zamir et al., 2021) alleviates the challenge of large textureless regions present in satellite remote sensing images due to buildings, water surfaces, and flat areas. It demonstrates significant improvements in ablation experiments (as shown in Table 2), which validate the advantage of Transformers in long-range modeling. By applying this module only to our baseline, a decrease of 8.41% in D1 error is observed on the GaoFen-7 dataset, showcasing the effectiveness of our proposed Transformer module in enhancing prediction accuracy.

4.5.2 Dual-branch Feature extractor architecture: To validate the effectiveness of the dual-branch network, we conducted three sets of experiments (as shown in Figure 8). In the first set, features were extracted solely using the Transformer module. The second set employed a U-net (Ronneberger et al., 2015) structure, where the encoder part used the Transformer module and the decoder part used the CNN module. The final set of experiments involved the proposed dual-branch network. According to the results in Table 3, comparing the first and second sets of experiments confirms the effectiveness of feature extraction with the CNN module, suggesting that despite the

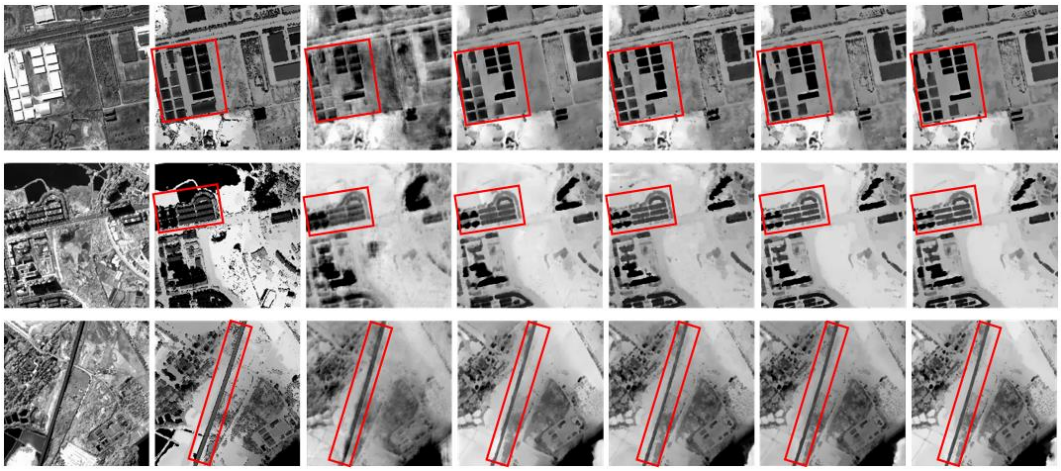


Figure 6. Visualization examples of disparity estimation results on the GaoFen-7 testing set generated by different methods. From left to right: left image, ground truth, DenseMApNet, StereoNet, PSMNet, HMSM-Net, our TFGR-Stereo.

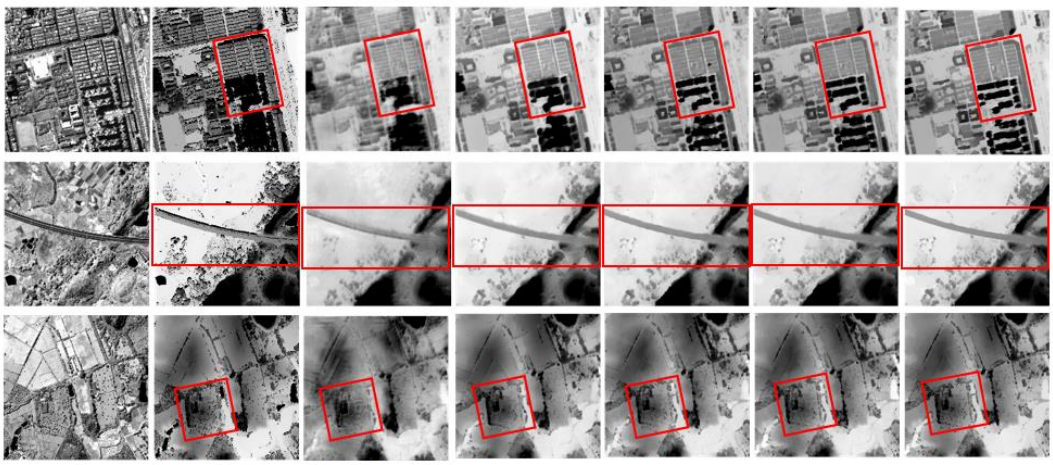


Figure 7. Visualization examples of disparity estimation results on the typical regions generated by different methods. From left to right: buildings, roads, forests. From top to bottom: left image, ground truth, DenseMapNet, StereoNet, PSMNet, HMSM-Net, and our TFGR-Srereo.

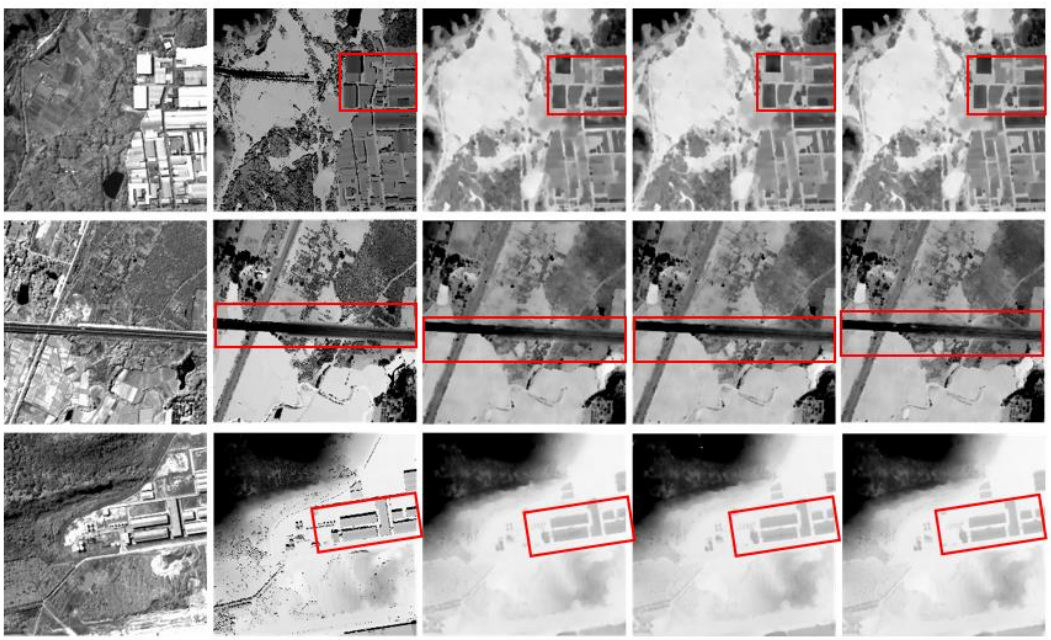


Figure 8. Visualization examples of disparity estimation results on the GaoFen-7 testing set. From Top to bottom: left image, ground truth, V1 (only use Transformer modules), V2(structure like U-net where encoder uses Transformer modules and decoder uses CNN modules), V3(dual-branch network we propose in this paper).

https://doi.org/10.5194/isprs-annals-X-G-2025-995-2025

advantage of Transformers in extracting global features, they may overlook local regions. In contrast, the CNN module, based on window-based computation, can compensate for the drawbacks of Transformers in handling local regions, resulting in better performance when processing areas rich in texture details. By comparing the results of the second and third sets, the effectiveness of our proposed dual-branch network is fully demonstrated, confirming the respective strengths of the Transformer and CNN modules in extracting global and local features. The design of the dual-branch network allows both the Transformer and CNN modules to utilize the high-frequency information from the reference image, independently extracting effective and accurate feature representations. These representations are then fused efficiently in the feature fusion module to obtain feature representations rich in global and local contextual relationships, further enhancing network performance.

Structure	GaoFen7	
	EPE	D1
V1	1.857	15.7
V2	1.800	15.0
V3	1.702 ↓	13.4 ↓

Table 3. Ablations on different structures of our proposed feature extractor Where V1 represents the structure only by Transformer, V2 represents the structure like U-net where the encoder utilizes Transformer modules while the decoder employs CNN modules, and V3 is the dual-branch network proposed in this paper.

4.5.3 ConvGRU-based Refinement Module: The effectiveness of the ConvGRU module is demonstrated through ablation studies in this paper. As shown in Table 2, the refinement module based on ConvGRU (Lipson et al., 2021) significantly improves the accuracy and precision of disparity estimation(Compared to the baseline network, our proposed module achieved a 7.48% reduction in the D1 metric on the Gaofen-7 dataset). Additionally, it addresses the issue of disparity inconsistency caused by multiple objects in satellite imagery and ensures alignment between the disparity map and the reference image structure. We utilize the regularized 1/4 resolution cost volume as the geometric encoding cost volume (Xu et al., 2023), which extracts effective and accurate non-local information and scene prior knowledge. These feature representations effectively enhance the accuracy of disparity prediction. Furthermore, during training, the initial disparity predicted by the geometric encoding cost volume is used as supervision for the ConvGRU refinement module. This strategy accelerates module training and optimization while reducing the training pressure on the network, resulting in a more concise and lightweight overall structure. At each iteration, the lookup operator retrieves iterative features from the geometric encoding cost volume based on the current disparity within a predefined range. With this mechanism, the lookup operator (Lipson et al., 2021) accurately supplements the detailed information required for the final disparity, ensuring perfect alignment between the final disparity map and the reference image structure.

5. CONCLUSIONS

In this paper, we propose a novel end-to-end deep learning network named TFGR-Stereo for disparity estimation in high-resolution satellite stereo images. The Transformer-based feature extractor captures global features, while a dual-branch architecture and effective fusion strategies provide the network with rich pyramid feature representations. Additionally, the network hierarchically learns stereo matching from coarse to fine scales through supervision applied at each scale. Finally, a refinement module based on ConvGRU is employed for fine-tuning the disparities. Leveraging these strategies, HMSM-Net demonstrates outstanding disparity estimation capabilities. We evaluate our network on the GauFen-7 dataset. Experimental results, along with comparisons to several state-of-the-art methods, demonstrate the superior performance of our approach and improved quality of disparity estimation in challenging regions. Furthermore, comprehensive ablation studies validate the effectiveness of the proposed modules.

ACKNOWLEDGMENTS

This research was funded by the National Natural Science Foundation of China (Grant No. 42371442) and the Hubei Provincial Natural Science Foundation of China (Grant No. 2023AFB568).

REFERENCES

Bleyer, M., Rhemann, C., Rother, C., 2011. PatchMatch Stereo - Stereo Matching with Slanted Support Windows, British Machine Vision Conference. In Bmvc (Vol. 11, No. 2011, pp. 1-11).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-End Object Detection with Transformers, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), Computer Vision – ECCV 2020. Springer International Publishing, Cham, pp. 213-229.

Chang, J.R., Chen, Y.S., 2018. Pyramid Stereo Matching Network, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5410-5418.

Fife, W.S., Archibald, J.K., 2013. Improved Census Transforms for Resource-Optimized Stereo Vision. IEEE Transactions on Circuits and Systems for Video Technology 23, 60-73.

Guo, X., Yang, K., Yang, W., Wang, X., Li, H., 2019. Group-Wise Correlation Stereo Network. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3268-3277.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778.

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

He, S., Zhou, R., Li, S., Jiang, S., Jiang, W., 2021. Disparity Estimation of High-Resolution Remote Sensing Images with Dual-Scale Matching Network, Remote Sensing, 13(24), 5050.

Hosni, A., Rhemann, C., Bleyer, M., Rother, C., Gelautz, M., 2013. Fast Cost-Volume Filtering for Visual Correspondence and Beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 504-511.

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

Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J.P.C., Izadi, S., 2018. StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction. In Proceedings of the European conference on computer vision (ECCV), pp. 573-590.

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.

Li, Z., Liu, X., Drenkow, N., Ding, A., Creighton, F., Taylor, R., Unberath, M., 2021. Revisiting Stereo Depth Estimation From a Sequence-to-Sequence Perspective with Transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6197-6206).

Lipson, L., Teed, Z., Deng, J., 2021. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching, 2021 International Conference on 3D Vision (3DV), pp. 218-227.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 9992-10002.

Lou, J., Liu, W., Chen, Z., Liu, F., Cheng, J., 2023. ELFNet: Evidential Local-global Fusion for Stereo Matching, 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 17738-17747.

Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T., 2015. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4040-4048.

Rao, Z., He, M., Zhu, Z., Dai, Y., He, R., 2021. Bidirectional Guided Attention Network for 3-D Semantic Detection of Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing 59, 6138-6153.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234-241.

Scharstein, D., Szeliski, R., 2003. High-accuracy stereo depth maps using structured light, 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. (Vol. 1, pp. I-I).

Scharstein, D., Szeliski, R., Zabih, R., 2001. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001), pp. 131-140.

Shen, Z., Dai, Y., Rao, Z., 2021. CFNet: Cascade and Fused Cost Volume for Robust Stereo Matching, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13901-13910.

Tao, R., Xiang, Y., You, H., 2020. An Edge-Sense Bidirectional Pyramid Network for Stereo Matching of VHR Remote Sensing Images, Remote Sensing, 12(24), 4025.

Teed, Z., Deng, J., 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), Computer Vision – ECCV 2020. Springer International Publishing, Cham, pp. 402-419.

Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2021. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 548-558.

Xu, G., Wang, X., Ding, X., Yang, X., 2023. Iterative Geometry Encoding Volume for Stereo Matching, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21919-21928.

Xu, G., Wang, Y., Cheng, J., Tang, J., Yang, X., 2024. Accurate and Efficient Stereo Matching via Attention Concatenation Volume. IEEE Transactions on Pattern Analysis and Machine Intelligence 46, 2461-2474.

Xu, H., Zhang, J., 2020. AANet: Adaptive Aggregation Network for Efficient Stereo Matching, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1956-1965.

Zamir, S.W., Arora, A., Khan, S.H., Hayat, M., Khan, F.S., Yang, M.-H., 2021. Restormer: Efficient Transformer for High-Resolution Image Restoration. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5718-5729.

Zhang, C., Jiang, W., Zhang, Y., Wang, W., Zhao, Q., Wang, C., 2022. Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing 60, 1-20.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6230-6239.

Zhao, H., Zhou, H., Zhang, Y., Chen, J., Yang, Y., Zhao, Y., 2023. High-Frequency Stereo Matching Network. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1327-1336.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S., Zhang, L., 2021. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6877-6886.