Stereo Matching of High-Resolution Satellite Images via Hierarchical ViT and Self-Supervised DINO

Xu He [1], Mengran Yang [1], San Jiang [2, 4 *], 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 Univeristy, Shenzhen 518060, China

[3] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China

[4] Engineering Research Center of Natural Resource Information Management and Digital Twin Engineering Software, Ministry of Education, Wuhan 430074, China

KEY WORDS: Satellite Image, Dense Matching, Deep Learning, Semi-global Matching

ABSTRACT:

Dense matching plays an important role in 3D modeling from satellite images. Its purpose is to establish pixel-by-pixel correspondences between two stereo images. This study presents a learning-based dense matching approach that integrates self-supervised learning with a multi-head attention mechanism to achieve feature fusion. Since stereo matching in satellite datasets is restricted by the disparity range, the pixel-by-pixel method can reduce the limitation. In the feature extraction module, we have performed attention-based in-depth learning on the smallest-scale feature using the self-supervised DINO. In addition, a CEP (Context-Enhanced Path) module is added outside the main matching path, and continuously enhanced position embedding is used to improve relative position encoding. The effectiveness of this method has been demonstrated through experiments on the US3D and WHU-Stereo datasets.

1. INTRODUCTION

Dense matching of stereo images is a classic problem in the field of photogrammetry and computer vision (Ji et al., 2019, Jiang et al., 2021). Its core task is to establish the pixel-by-pixel correspondences between two images to recover the 3D information of the target (Geiger et al., 2010). Stereo-dense matching has become the most crucial component in many tasks that range from localization tracking to 3D reconstruction (Geiger et al., 2011). As the popularity and quality of satellite images continue to improve, stereo matching based on high-resolution satellite images has been widely used in various applications, such as 3D modeling of large-scale cities (Facciolo et al., 2017, Huang et al., 2017). Thus, efficient and robust stereo matching becomes the key to applying high-resolution satellite images.

Given a pair of rectified stereo images, the first step in stereo-density matching is to compute the disparity of each pixel in the reference image, which is further used to recover depth and 3D information (Gu et al., 2020). For common close-range datasets, scholars had proposed various dense matching algorithms, which greatly promoted deep learning progress in this field (Kendall et al., 2017, Shen et al., 2021, He et al., 2023). These algorithms learned the correspondence between pixels through neural networks to achieve high-precision disparity prediction for close-range datasets (Lin et al., 2024). Early applications of convolutional neural networks (CNN) to stereo matching aimed to improve individual steps of an established process (Poggi et al., 2021). With the proposal of DispNet (Mayer et al., 2016) and GC-Net (Kendall et al., 2017), some end-to-end deep learning models have further improved in accuracy and efficiency, and occupy a dominant position in commonly used benchmark tests, such as KITTI (Geiger et al., 2012, Menze and Geiger, 2015), Middlebury (Scharstein et al., 2014), SceneFlow (Mayer et al., 2018), and other remote sensing datasets (Patil et al., 2019).

The range of the disparity in parallax estimation for satellite images differs from that of close-up datasets (He et al., 2021). Current methods trained with a fixed disparity range cannot cover the whole disparity range of other datasets without significant modification due to this imbalanced disparity distribution. The 3D cost volume (Gu et al., 2020) cascade formula handles this problem by first building the cost volume using larger-scale semantic 2D features and a sparsely sampled disparity range assumption. Later, the early estimated disparity map is used to modify the sampling range of the disparity assumption, resulting in the construction of a new cost volume that applies finer semantic features for more accurate disparity prediction. CFNet implemented a related concept. Uncertainty estimate was utilized to assess the pixel-level confidence of disparity computation and enhance the disparity search range after merely fusing several low-resolution dense cost volumes to increase the receptive field (Shen et al., 2021).

To increase the accuracy of disparity estimation, deep learning architectures with multi-head attention methods, such as Transformer, have been introduced into the dense matching field recently (Zuo et al., 2016, Fei et al., n.d.). Using the sequential nature and geometric properties of stereo matching, the STTR network inputs the image features acquired using CNN into the Transformer to capture the long-range correlation between pixels (Li et al., 2021). This technique imposes uniqueness restrictions in the epipolar direction by densely and precisely computing the correlation between pixels in the left and right pictures during stereo matching. The feature representation is updated with the use of image context and location, and the self-attention and cross-attention processes are integrated to execute alternating actions inside and across images.

*Corresponding author

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-2025-357-2025 | © Author(s) 2025. CC BY 4.0 License.

In order to describe distant context information and enhance the network's robustness and generalization capacity, CSTR incorporates a CEP (Context Enhanced Path) module based on STTR (Guo et al., 2022). Furthermore, STransMNet substitutes the Transformer structure in STTR with Swin Transformer and proposes feature differentiation loss to strengthen the algorithm's focus on feature details and optimize the loss function (Ding et al., 2022). For use to manage 3D perception tasks including optical flow, stereo matching, and depth estimation, GMFlow presents a unified global matching framework that compares feature similarities to establish the connection between pixels (Xu et al., 2022). It does this by using a cross-attention method. GOAT proposes to use a parallel attention mechanism to calculate the initial disparity map and occlusion mask and designs a parallel disparity and occlusion estimation module, as well as an occlusion-aware full aggregation module to refine the disparity in the occlusion area (Liu et al., 2024).

Nonetheless, there are several ill-posed regions, such as weak texturing, and a wide disparity range in satellite datasets (He et al., 2022). So the prediction accuracy of widely-used algorithms is not as high as it is on other close-range datasets. It is necessary to carefully tune and enhance dense matching algorithms for satellite images in order to accommodate specific data features. To minimize computation and guarantee efficient information transfer between windows, this study develops a feature extraction method that combines hierarchical ViT(Vision Transformer) (Dosovitskiy et al., 2020) and DINO (Caron et al., 2021) and implements a dense matching network for high resolution satellite images.

2. METHODOLOGY

This study proposes a hierarchical ViT-Stereo model. First, the overall architecture of the model is outlined, and then the working principles of the feature extraction module based on hierarchical ViT and DINO, the global information aggregation module of the context enhancement path, and the position information improvement module based on relative position encoding offset are described in detail. Finally, through a series of comparative experiments, this chapter discusses and analyzes the experimental results to verify the effect and advantages of the proposed method.

2.1 Network Architecture

As shown in Figure 1, the hierarchical ViT-Stereo model is roughly divided into three modules: feature extraction, disparity calculation, and disparity optimization.

(1) Benefiting from pyramid architecture and efficient attention mechanism, for a given pair of left and right images I_l, $I_r \in R^{H\times W\times C}$, the model in this study improves the learning of feature information by combining hierarchical ViT and DINO. The feature extraction module extracts feature vectors from the image, which are then converted into feature descriptors and used to calculate pixel relative distance codes. The specific details of the feature extraction module are shown in Section 2.2.

(2) Afterwards, the attention weight is calculated by fusing the main matching path of the contextual cross-polar line features, and the most likely matching position is found from the optimal transmission allocation matrix T, and a three-pixel window is constructed around it $N_3(k)$, renormalize the matching probability t_l within the window to obtain $\widetilde{t_l}$, so that its sum is 1:

$$\widetilde{t_l} = \frac{t_l}{\sum_{l\in N_3(k)} t_l}, l \in N_3(k) \tag{1}$$

According to probability, the original disparity $\widetilde{d}_{raw}$ is calculated by the weighted sum of candidate disparities:

$$\widetilde{d}_{raw} = \sum_{l\in N_3(k)} d_l \widetilde{t_l} \tag{2}$$

In addition, within the three-pixel window, the sum of the probabilities of each pixel is used to represent the network's reflection of the current estimated confidence. This representation is measured by the inverse occlusion probability. The occlusion probability p_{occ} is calculated as:

$$p_{occ}(k) = 1 - \sum_{l\in N_3(k)} t_l \tag{3}$$

(3) To further aggregate transpolar information, based on the initial disparity $\widetilde{d}_{raw}$, a convolutional layer is introduced to adjust the conditional input image to include disparity estimation of transpolar information. The structure of the disparity optimization module is shown in detail in Figure 2. This process connects the original disparity map and occlusion map with the original left image along the channel dimension, then aggregates the occlusion information through two convolution blocks and the ReLU layer, and uses the Sigmoid layer to generate the final occlusion map O_{final}.

For $\widetilde{d}_{raw}$, it is refined through a residual block, which consists of expanding the channel dimensions before passing it through the ReLU activation function and reducing it back to the original channel dimensions afterwards. Better results are obtained by repeatedly connecting the original disparity map with the residual block, and finally the output of the residual block is added back to the original disparity through long skip connections to achieve more accurate disparity refinement.

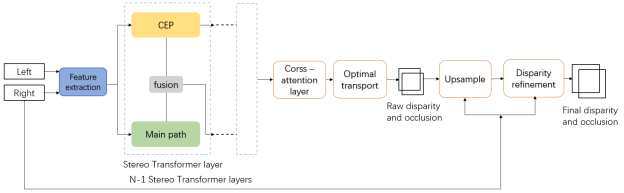

Figure 1: Network architecture.

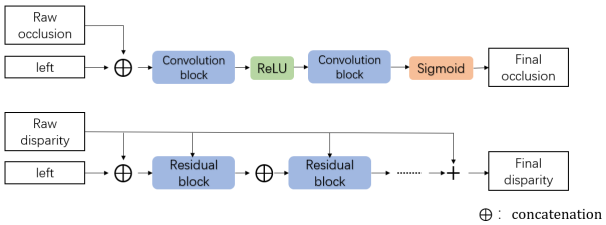

Figure 2: Disparity refinement module.

2.2 Feature Extraction via Hierarchical ViT and DINO

Considering the correlation between the left and right images, input images I_l, $I_r \in R^{B\times C\times H\times W}$ are first merged into $I_s \in R^{2B\times C\times H\times W}$ in the batch dimension, which is then downsampled to half the original resolution to reduce computational and memory overhead. In order to adapt to the fine-tuning needs of different resolutions, hierarchical ViT adopts an efficient attention mechanism and combines robust position encoding for different scales. As shown on the left side of Figure 3, hierarchical ViT uses multi-scale features $\left\{F^{(h,s)}\right\}_{s=1}^{S=4}$ relative to the initial resolution (1/8, 1/16, 1/32, 1/64) size, encoding the original image. In addition to leveraging the pyramid architecture, this model replaces the absolute position encoding in ViT with a conditional position encoding generated by a Position Encoding Generator (PEG), enabling it to learn position cues from zero padding and passing them through an appropriate convolutional neural network (CNN) inductive bias to break the permutation equivalence of ViT. As shown in Figure 4, PEG is placed after the first encoder block of each level. In order to capture local neighborhood information, the flattened input sequence $X \in R^{B\times N\times C}$ is first reshaped in two-dimensional space into $X\prime \in R^{B\times H\times W\times C}$. Afterwards, the function F is repeatedly applied to the local patches in $X\prime$ to generate the conditional position encoding $E^{B\times H\times W\times C}$. PEG is efficiently implemented by a zero-padded 2D convolution with kernel k and (k-1)/2.

In addition, to further enhance the feature learning ability, this study utilizes the attention map generated by the last layer of CLS tokens in self-supervised DINO. Considering the importance of feature matching in dense matching tasks, leveraging prior knowledge of object or scene segmentation helps distinguish foreground and background, thus avoiding confusion in depth prediction. Given that the input size of hierarchical ViT has been halved, irregular embedding using a convolution kernel with size and stride of 16 enables DINO to perform on feature maps with resolutions of (H/32, W/32) Attention-based learning. To better utilize its segmentation capabilities, trainable gated linear units (GLU) are used to reduce feature dimensions. Assume that $A \in R^{\frac{H}{32}\times\frac{W}{32}\times 1}$ represents the attention map of the last layer of CLS tokens in DINO, and h represents the number of attention heads. $\hat{A} \in R^{\frac{H}{32}\times\frac{W}{32}\times 1}$ means averaging along h heads of A, then GLU can be expressed as:

$$\widetilde{F}^{(p)} = Swish(ConvBN_l([F_{dino}; A]) \odot Swish(ConvBN_r(ConvBN_r([F_{dino} \odot \hat{A}]))) \quad (4)$$

Among them, Swish(x)=x·sigmoid(x); $\odot$ means element-wise multiplication; [•; •] means splicing operation. GLU helps protect important features that benefit from segmentation attention maps during the dimensionality reduction process, thereby effectively improving model performance. After that, $\widetilde{F}^{(p)}$ is upsampled to $F^{(p)}\in \mathrm{R}^{\frac{H}{8}\times\frac{W}{8}\times C}$ using two transposed convolutions, used to add features to the FPN encoder with channel C, and finally fuse features of different scales:

$$F^{(h)} = FPN\left(F^{(h,1)}, F^{(h,2)}, F^{(h,3)}, F^{(h,4)}\right) \quad (5)$$

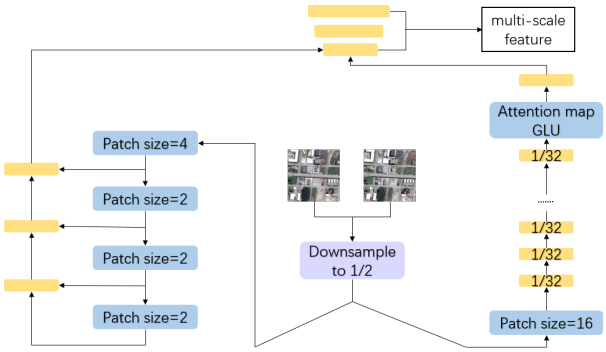

Figure 3: Feature extraction.

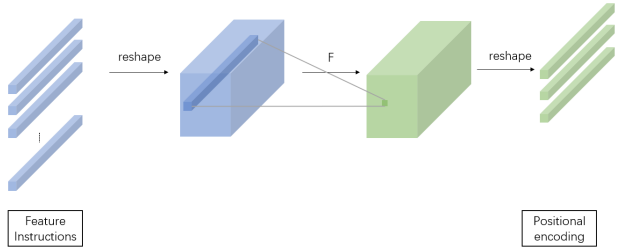

Figure 4: Position encoding.

2.3 Global Information Fusion via Context-Enhanced Path

When calculating disparity, relying solely on the correlation between pixels on the image epipolar lines is often insufficient to capture sufficient information, especially because this method does not fully utilize the information across the epipolar lines, which is crucial for processing global information. In the case of large textureless areas, specular reflections, or low transparency, features in the left and right images may be similar and misleading, causing feature matching based solely on a single epipolar line to become unclear in these areas. Therefore, in order to accurately predict disparity, leveraging long-range contextual information becomes a critical task. This chapter introduces a plug-in module called Context Enhancement Path (CEP), which is designed to enhance the aggregation of global context information and provide comprehensive geometric information for the model.

In order to take into account efficient calculation and aggregation of global information, Wang et al. [77] proposed Axial-Attention, using two consecutive axial attention layers for the height axis and width axis respectively. Assume that $N_{(W\times 1)}$(i) is the relative position code represented by the $w \times h$ scale area around pixel i, q, k, v, respectively represent the query, key, and value, and S represents the Softmax function, then the width axis attention layer y_i can be described as:

$$y_i = \sum_{j\in N_{W\times 1}(i)} S(q_i^T k_i + q_i^T r_{j-1})(v_j) \quad (6)$$

Compared to locally constrained attention, y_i computes attention line-by-line through weight sharing. For height-axis attention, the attention calculations on the vertical and horizontal axes are the same except that the attention is calculated column-wise.

The goal of CEP is to maintain the contextual features of the left and right images and provide the contextual features as ad-

ditional supplementary information to the main matching path. Its detailed structure is shown in Figure 5. In addition, the self-attention layer in the matching path is replaced with an axial attention layer, including horizontal and vertical directions, to collect contextual information from the horizontal and vertical axes. As a layer-by-layer module, CEP first obtains contextual features from the previous CEP layer, and then applies axial attention layers and cross-attention layers to further process contextual features. These processed contextual features are then used as supplementary information to be fused with the information on the main matching path.

After obtaining the contextual information features, the features in the CEP are fused into the main matching path through the path fusion module. This step allows the main matching path to capture long-range contextual information from low-resolution features, thereby enhancing its overall effect on the scene. The structure is shown in Figure 6. Specifically, the context feature $f_C \in R^{H\times W/n\times C}$ is upsampled and spliced with the main matching feature $f_M \in R^{H\times W\times C}$ in the channel dimension. Two 3 × 3 convolutional layers are then applied to the connected features, and finally the fused features are used as input to the next main matching path module.

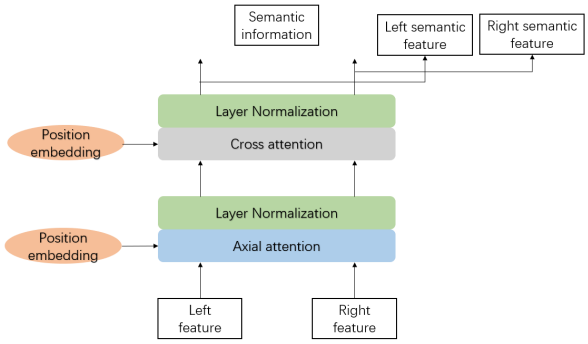

Figure 5: Contextual Enhancement Path (CEP).

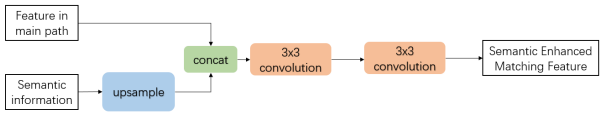

Figure 6: Fusion of CEP and main matching path.

2.4 Relative Position Encoding based on Continuous Enhanced Position Embedding

In the Transformer model, the introduction of absolute position encoding (APE) is to compensate for the model's own inability to process the order of elements in the sequence, so that the model can identify the order of elements in the input sequence. To enhance model flexibility, relative position encoding (RPE) is introduced. Different from APE, RPE not only captures the local dependencies within the sequence by encoding the relative distance between elements in the sequence, but also enables the model to dynamically adjust its focus based on the relative positions between elements, thereby processing sequence data more effectively. Therefore, this study combines APE and RPE in Transformer to enable the model to better understand and adapt to different contextual situations. APE provides a stable sequential framework, and RPE provides the model with the ability to dynamically adjust according to context, making the model more flexible and effective in processing high-resolution satellite images containing complex global and contextual information. APE is used in the original attention mechanism of the Transformer framework, and the absolute position embedding $P = (p_1, p_2, ..., p_L)$ is added to the input embedding x:

$$x_i = x_i + p_i \tag{7}$$

For the mark at position n, each component of its position encoding is calculated by the sine function and the cosine function, where the index of the component is k=1,2,...,K/2.

$$E_{2k}(n) = cos\omega_k n E_{2k+1}(n) = sin\omega_k n \tag{8}$$

Among them $\omega_k = 10000^{-2k/K}$. This method generates a unique encoding for each position in the sequence through a periodic function, allowing the model to recognize and exploit positional information. On a two-dimensional image with horizontal and vertical coordinates, for each position (x, y), its K-dimensional absolute sine position embedding into equation 8 is:

$$\begin{aligned} E_{2k}(x,y) &= cos\pi(\omega_{k,x}x + \omega_{k,y}y) \\ E_{2k+1}(x,y) &= sin\pi(\omega_{k,x}x + \omega_{k,y}y) \end{aligned} \tag{9}$$

Among them, $\omega_{k,x} = 10^{2k/K} cos\ k$, $\omega_{k,y} = 10^{2k/K} sin\ k$. Since the dense matching model needs to take advantage of the continuity of the matched images, the traditional sinusoidal position encoding of the APE in the Transformer model is now changed from a discrete position-based encoding to a continuous position-based encoding. Furthermore, information about the relative positions between tokens can be preserved by employing specific position embedding enhancement techniques during training. First, each embedding in the sequence is transformed using the S^δ operator with a global random shift from a zero-mean uniform distribution, $\delta \in \mathcal{U}(-\delta_{max}, \delta_{max})$:

$$S^\delta X_i = X_k + e^{i\omega_k \delta} \tag{10}$$

Substituting equation 9, we obtain $(x_i', y_i') = (x_i + \delta_x, y_i + \delta_y)$ after global random offset. To further enhance and prevent spontaneous correlations from being captured, a zero-mean uniformly distributed local offset is introduced, $\epsilon_i \in \mathcal{U}(-\epsilon_{max}, \epsilon_{max})$, when$(x_i', y_i') = (x_i + \epsilon_{x,i}, y_i + \epsilon_{y,i})$. Finally, to prevent distance memory, a random global scale λ is introduced, $log\lambda \sim \mathcal{U}(-log\lambda_{max}, log\lambda_{max})$, get the final $(x_i', y_i') = (\lambda x_i, \lambda y_i)$.

3. EXPERIMENTS

3.1 Datasets

The US3D dataset is a large-scale remote sensing image dataset proposed for multiple tasks. For stereo matching, 4292 RGB image pairs and publicly available ground truth disparity maps are provided, and the image size is 1024*1024 pixels. These images were collected from the WorldView-3 satellite and cover the two cities of Jacksonville (JAX) and Omaha (OMA) in the United States. In the experiments of this study, 1600 image pairs of Jacksonville were used for training, while the remaining image pairs of the city were used for validation and testing. All image pairs from Omaha are used to evaluate the generalization ability of the network. Figure 7 is an example image of the US3D data set. From left to right, they are the left image, the right image, and the ground truth disparity.

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-2025-357-2025 | © Author(s) 2025. CC BY 4.0 License.

WHU-Stereo is an open source dataset for stereo matching of high-resolution satellite images, containing more than 1,700 epipolar-corrected image pairs (Li et al., 2023). Similar to the US3D dataset, there are a total of 1981 epipolar-corrected stereo image pairs in WHU-Stereo, of which 1757 image pairs provide ground truth data generated from aerial LiDAR point clouds. The dataset consists of panchromatic band images with 16-bit depth, the size is 1024*1024 pixels, and the disparity map is stored on 16-bit floating point values, covering six cities in China. Among them, Shaoguan, Kunming, Yingde and Qichun are used to evaluate the geographical generalization ability of the deep learning model within cities; Wuhan and Hengyang are used to evaluate the geographical extrapolation ability of the model between cities. Figure 8 is an example image of the WHU-Stereo data set. From left to right, they are the left image, the right image, and the ground truth disparity.

Figure 7: Sample images in US3D.

Figure 8: Sample images in WHU-Stereo.

3.2 Evaluation metrics

We choose endpoint error (EPE) and 3-pixel error ratio (D1) as the evaluation indicators of the comparison method. EPE is the average of the Euclidean distances between the predicted value and the true value. D1 refers to the percentage of error points in all effective pixels on the basis that the difference between the predicted value and the real disparity value exceeds 3 pixels, which is considered an error.

$$EPE = \frac{1}{N} \sum_{k \in T} |\hat{d}_k - \widetilde{d}_k| \tag{11}$$

$$D1 = \frac{1}{N} \sum_{k \in T} [|\hat{d}_k - \widetilde{d}_k| > t] \tag{12}$$

where $\widetilde{d}_k$ = ground-truth disparity, $\hat{d}_k$ = estimated disparity, N, T = number and set of labelled pixels in the image, t = threshold of erroneous disparity.

3.3 Environment and parameter

In this experiment, all deep learning model training and testing were completed on an NVIDIA GeForce RTX 4090 graphics card with 24GB of video memory. All models are implemented based on the Pytorch framework. In addition, the inputs of the semantic disparity optimization layer are set to 4 and 2 respectively according to the number of channels of the images in the US3D and WHU-Stereo data sets. The embedding dimension and patch size in layered ViT and DINO were adjusted according to the image size to ensure that the model can effectively handle inputs of different sizes. The entire model was trained from scratch on two data sets for 200 epochs. In addition, the initial learning rate of the model was set to 0.0001, and as the training progressed, a weight decay strategy of 0.99 was adopted when using the AdamW optimizer to optimize the training process.

3.4 Results and Discussions

In order to fully verify the dense matching performance of the hierarchical ViT-Stereo model, in the experiment, the dense matching algorithm STTR based on position information and attention mechanism was used as the benchmark to compare the impact of the improvement of three key modules on the performance: Hierarchical ViT and DINO improve feature extraction module (FE), aggregate context enhancement path to main matching path (CEP), and continuous enhancement position embedding to relative position encoding (PE). Table 1 and Table 2 show the hierarchical ViT-Stereo results of the fusion of STTR and CEP modules, the fusion of STTR and PE modules, and the fusion of the three modules, covering the two indicators of EPE error and D1 error.

On the US3D dataset, accuracy is first measured by the EPE error. The results show that although the improvement is small, the fusion of PE and CEP modules can effectively improve the accuracy of the STTR model. On the basis of these two modules, the hierarchical ViT-Stereo model further added FE, showing a more significant accuracy improvement compared to STTR. From the perspective of D1 index, the improvements of PE and CEP modules are similar. After adding the FE module, the accuracy of the hierarchical ViT-Stereo model has been significantly improved. In the JAX part of the data set, there are two evaluation indicators: EPE error and D1 error. It was reduced by 1.815 and 24.7% respectively, and it was reduced by 1.61 and 23.3% respectively in the OMA part of US3D.

Results of the WHU-Stereo showed that although the average values of the hierarchical ViT-Stereo model in the EPE error and D1 error indicators reached 3.31 and 29.0%, the accuracy is significantly improved compared to the original STTR model, which was reduced by 5.525 and 29.0% respectively, but the improvement effect of PE and CEP modules on D1 error index is not obvious. The EPE error index measures the average error between the predicted disparity and the actual disparity, and the D1 error index mainly measures the proportion of pixels in the disparity map with an error exceeding 3 pixels. The PE module mainly improves the pixel position accuracy of the disparity map, but this reduction of small errors may not be enough to reduce the error below the threshold of the D1 error. In addition, although the CEP module improves accuracy by aggregating contextual information beyond the main matching path, which may be more accurate in image predictions containing large natural landscapes such as mountains and rivers, in urban scenes due to complex buildings and road structures, its performance improvement is limited.

To more intuitively observe the difference in the performance of different modules in improving the STTR model in the satellite image dense matching task, Figure 9 shows multiple randomly

Method	US3D		WHU-Stereo
	JAX	OMA	
STTR	3.657	3.948	8.835
STTR+PE	3.354	3.479	7.287
STTR+CEP	3.434	3.401	6.341
Ours	1.842	2.338	3.310

Table 1: EPE error on US3D and WHU-Stereo (Pixels)

Method	US3D		WHU-Stereo
	JAX	OMA	
STTR	41.4	43.8	43.3
STTR+PE	36.5	35.6	43.3
STTR+CEP	37.1	34.2	46.9
Ours	16.7	20.5	29.0

Table 2: D1 error on US3D and WHU-Stereo (%)

selected predicted parallax images in the US3D data set, including the test results of JAX and OMA generalization results. From top to bottom are the experimental results of the original left image, ground truth parallax, STTR model, STTR+PE module, STTR+CEP module, and layered ViT-Stereo model. The corresponding image numbers from left to right are JAX-416-004-011, JAX-427-006-011, OMA-248-025-027, OMA132-031-028, OMA-315-003-006. According to the experimental results in Figure 9.

Compared with other models, there are some lateral textures in the disparity images predicted by the STTR model. It is initially believed that this phenomenon may be due to the insufficient integration of transpolar information and sufficient position information in the matching path of the STTR model, resulting in a lack of sufficient context information for image pairs in pixel-by-pixel matching. In addition, the results of the STTR+PE module and STTR+CEP show that the PE module can more accurately maintain pixel position information when predicting buildings and roads through improved continuous position embedding technology. The CEP module based on context enhancement pays more attention to the integration of global information, especially showing better performance in predicting continuous trees near buildings.

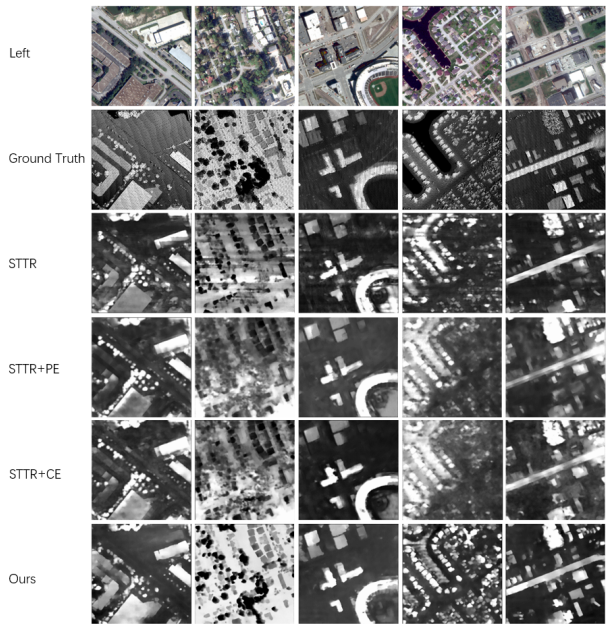

Figure 9: Results on US3D.

As for the hierarchical ViT-Stereo model that combined the PE module, the FE module, and the feature extraction module based on hierarchical ViT and DINO, its performance in disparity prediction is better than other models. From the perspective of JAX-416-004-011, which contains large and complete buildings, that is, the first column of images in Figure 9, the results predicted by the hierarchical ViT-Stereo model are 1.130 and 8.4% in EPE error and D1 error respectively. The results of the STTR model are 2.429 and 34.2%, the results of the STTR+PE module are 1.984 and 21.6%, and the results of the STTR+CEP module are 2.106 and 22.5%. Figure 10 enlarged the original image and the experimental results of each model. It is easy to observe that only the hierarchical ViT-Stereo model predicts the buildings in the red and blue boxes more completely and clearly.

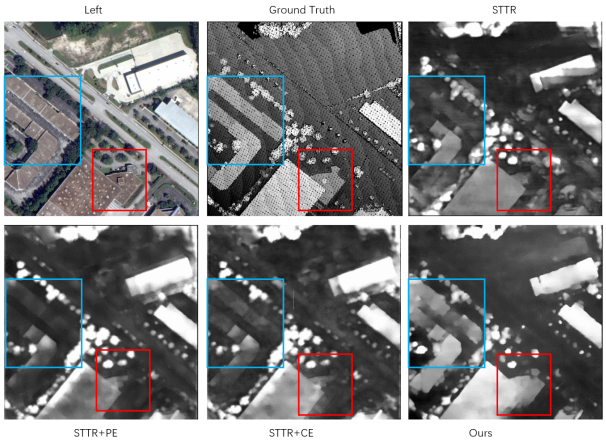

Figure 10: Disparity maps on JAX-416-004-011.

Judging from the OMA132-031-028 image that contains more continuous small buildings and trees, it is the fourth column of images in Figure 9. The prediction results of the hierarchical ViT-Stereo model are 1.536 and 12.7% in EPE error and D1 error respectively, results of the STTR model are 2.420 and 37.4%, the results of the STTR+PE module are 2.514 and 31%, and the results of the STTR+CEP module are 2.361 and 23.4%. In Figure 11, by zooming in on the original image and the experimental results of different models, the area marked by the red box in the lower right corner of the images, including small buildings and trees. Only the layered ViT-Stereo model can more completely display these details. However, for the river part in the upper left corner of the original image, almost all deep learning models are hard to accurately predict disparity.

Figure 12 shows the visualization results of each module in the WHU-Stereo data set test set. From top to bottom are the experimental results of the original left image, ground truth parallax, STTR model, STTR+PE module, STTR+CEP module, and layered ViT-Stereo model. The corresponding image numbers from left to right are KM-22, QC-398, and their partial enlargement image respectively. It can be observed that similar to the results on the US3D dataset, the results predicted by the STTR model also have lateral texture, especially near the invalid parallax area.

For instance, practically all of the tiny structures near the invalid parallax area in Figure 12’s first and second columns cannot be predicted with any degree of accuracy. The invalid region has a major impact on the prediction accuracy of the STTR+PE and STTR+CEP modules on the KM-22 picture. Utilizing its

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-2025-357-2025 | © Author(s) 2025. CC BY 4.0 License.

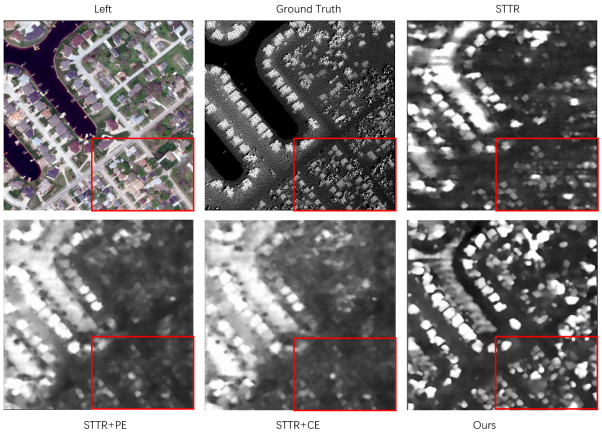

Figure 11: Disparity maps on OMA132-031-028.

feature extraction module, the hierarchical ViT-Stereo model accurately predicts the tiny building groups in the top section while minimizing the effects of invalid regions. Its D1 error and EPE error indicators, which are 10.3% and 1.481, respectively, are noticeably better.

For image QC-398, as shown in the third and fourth columns of Figure 12, each model successfully displays the parallax of each building in the red box. In addition, even in the ground-truth disparity map, the outline of the large land mass below the middle is not completely displayed, the disparity map predicted by the deep learning model can still calculate the results of this area relatively accurately. Comprehensive analysis showed that

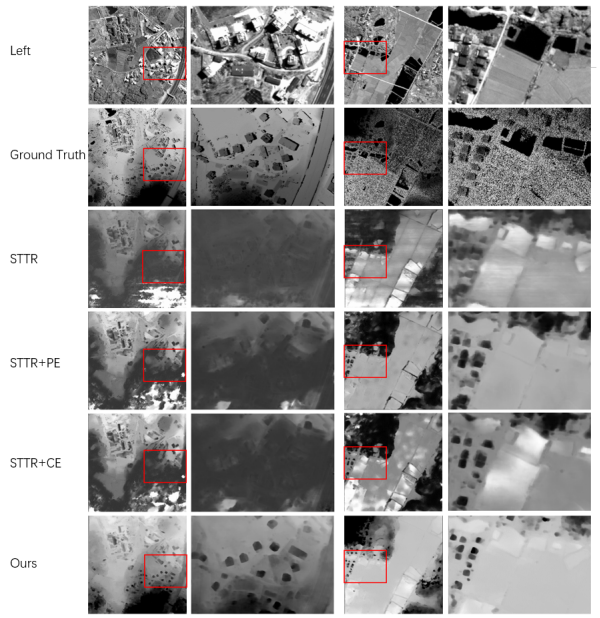

Figure 12: Disparity maps on WHU-Stereo.

although the integration of the three modules has achieved an overall improvement in EPE error and D1 error, it proves the effectiveness of this research method. However, through detailed analysis of the ablation experimental results, it was found that different improvement strategies have different effects on different types of images. This discovery offered a crucial reference for subsequent optimization of single-channel images.

4. CONCLUSION

We presented the hierarchical ViT-Stereo model to perform dense pixel matching based on location information and attention mechanisms, given the wide variations in the disparity range of different datasets. Encoding information based on the relative positions of pixels in an image more accurately identifies occluded areas and provides confidence estimates. At the same time, during the matching process, the context enhancement path that integrates cross-polar features is integrated into the main matching path, which helps to improve the global understanding of the model. Furthermore, it has been demonstrated through comparative tests conducted on the US3D and WHU-Stereo datasets that the method has significantly improved the accuracy and robustness of disparity estimates.

ACKNOWLEDGEMENTS

This research was funded by the National Natural Science Foundation of China (Grant No. 42371442, 42371452).

REFERENCES

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Ding, Y., Yuan, W., Zhu, Q., Zhang, H., Liu, X., Wang, Y., Liu, X., 2022. Transmvsnet: Global context-aware multi-view stereo network with transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8585–8594.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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

Fei, J., Kai, G., Zhi, L., Jiarong, H., Jie, R., Qinggao, L., n.d. A performing analysis of unsupervised dense matching feature extraction networks. *Acta Geodaetica et Cartographica Sinica*, 51(3), 426.

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

Geiger, A., Roser, M., Urtasun, R., 2010. Efficient large-scale stereo matching. *ACCV (1)*, 25–38.

Geiger, A., Ziegler, J., Stiller, C., 2011. Stereoscan: Dense 3d reconstruction in real-time. *2011 IEEE intelligent vehicles symposium (IV)*, Ieee, 963–968.

Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P., 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2495–2504.

Guo, W., Li, Z., Yang, Y., Wang, Z., Taylor, R. H., Unberath, M., Yuille, A., Li, Y., 2022. Context-enhanced stereo transformer. *European Conference on Computer Vision*, Springer, 263–279.

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.

He, X., Jiang, S., He, S., Li, Q., Jiang, W., Wang, L., 2023. Deep Learning-Based Stereo Matching for High-Resolution Satellite Images: a Comparative Evaluation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 1635–1642.

Huang, X., Wen, D., Li, J., Qin, R., 2017. Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery. *Remote sensing of environment*, 196, 56–75.

Ji, S., Liu, J., Lu, M., 2019. CNN-based dense image matching for aerial remote sensing images. *Photogramm. Eng. Remote Sens*, 85, 415–424.

Jiang, S., Jiang, W., Wang, L., 2021. Unmanned Aerial Vehicle-Based Photogrammetric 3D Mapping: A survey of techniques, applications, and challenges. *IEEE Geoscience and Remote Sensing Magazine*, 10(2), 135–171.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A., 2017. End-to-end learning of geometry and context for deep stereo regression. *Proceedings of the IEEE international conference on computer vision*, 66–75.

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*.

Li, Z., Liu, X., Drenkow, N., Ding, A., Creighton, F. X., Taylor, R. H., Unberath, M., 2021. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. *Proceedings of the IEEE/CVF international conference on computer vision*, 6197–6206.

Lin, S., Zhuo, X., Qi, B., 2024. Accuracy and efficiency stereo matching network with adaptive feature modulation. *Plos one*, 19(4), e0301093.

Liu, Z., Li, Y., Okutomi, M., 2024. Global occlusion-aware transformer for robust stereo matching. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3535–3544.

Mayer, N., Ilg, E., Fischer, P., Hazirbas, C., Cremers, D., Dosovitskiy, A., Brox, T., 2018. What makes good synthetic training data for learning disparity and optical flow estimation? *International Journal of Computer Vision*, 126, 942–960.

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

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

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

Poggi, M., Tosi, F., Batsos, K., Mordohai, P., Mattoccia, S., 2021. On the synergies between machine learning and binocular stereo for depth estimation from images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9), 5314–5334.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P., 2014. High-resolution stereo datasets with subpixel-accurate ground truth. *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, Springer, 31–42.

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

Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Tao, D., 2022. Gmflow: Learning optical flow via global matching. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8121–8130.

Zuo, Z., Shuai, B., Wang, G., Liu, X., Wang, X., Wang, B., Chen, Y., 2016. Learning contextual dependence with convolutional hierarchical recurrent neural networks. *IEEE Transactions on Image Processing*, 25(7), 2983–2996.