applied sciences

MDPI

Article

EWT-HA-Net: An Efficient Wavelet-Convolution-Enhanced Hybrid Attention Network for Sensorless Freehand 3D Ultrasound Reconstruction

Yuqing Yin, Yaoxian Zhang, Zhongxu Bao * and Qiang Niu

School of Computer Science and Technology/School of Artificial Intelligence, China University of Mining and Technology, Xuzhou 221116, China; yinyuqing@cumt.edu.cn (Y.Y.); zhangyaoxian@cumt.edu.cn (Y.Z.); niuq@cumt.edu.cn (Q.N.)

* Correspondence: baozx@cumt.edu.cn

Abstract

Sensorless freehand three-dimensional (3D) ultrasound reconstruction eliminates the need for external tracking devices but remains challenging due to speckle noise, weak textures, ambiguous anatomical boundaries, and the computational cost of existing high-performance methods. To address these issues, this paper proposes EWT-HA-Net, an efficient wavelet-convolution-enhanced hybrid attention network for sensorless freehand 3D ultrasound reconstruction. The network estimates inter-frame transformations by exploiting the selected frame pair and sequence contextual information, and recovers the probe trajectory through sequential transformation accumulation. An Enhanced Wavelet Transform Convolution (EWTConv) module integrates learnable wavelet decomposition and Dynamic Frequency Fusion Gating (DFFG) for multi-scale spatial-frequency feature extraction, while a complementary CoordAtt-SE attention strategy enhances position-sensitive and channel-wise feature representation. Experimental results demonstrate that EWT-HA-Net achieves improved reconstruction performance with low computational complexity, providing a favorable balance between accuracy and efficiency.

Keywords: 3D ultrasound reconstruction; Wavelet Transform Convolution; attention mechanisms; deep learning; sensorless

1. Introduction

Ultrasound is widely used in clinical imaging because it provides real-time, non-ionizing, and cost-effective visualization of tissues and organs. Three-dimensional (3D) ultrasound further provides volumetric anatomical information for visualization, measurement, and clinical assessment [1]. Compared with dedicated 3D array probes or mechanically swept systems, freehand 3D ultrasound reconstructs volumetric data by compounding two-dimensional (2D) ultrasound slices acquired along a manually guided probe trajectory. This approach is compatible with conventional ultrasound systems and offers a flexible field of view, but its reconstruction quality strongly depends on accurate estimation of the spatial position and orientation of each frame.

Accurate estimation of the probe trajectory remains one of the principal challenges in freehand 3D ultrasound reconstruction. External optical and electromagnetic tracking systems can provide reliable probe pose measurements, but their clinical deployment requires additional hardware, system calibration, and modifications to the conventional scanning workflow. Optical trackers depend on an unobstructed line of sight, whereas

Academic Editor: Christos Bouras

Received: 5 August 2026
Revised: 11 September 2026
Accepted: 14 September 2026
Published: 17 September 2026

electromagnetic trackers may be affected by nearby metallic objects and electromagnetic interference [2]. In contrast, sensorless reconstruction methods instead estimate probe motion directly from ultrasound image sequences, eliminating external hardware and tracker-to-probe calibration while improving portability and system compatibility.

Deep learning has been increasingly applied to sensorless probe trajectory estimation. Early methods mainly employed Convolutional Neural Networks (CNNs) or ResNet-based architectures to extract local image features and regress inter-frame transformations [3]. However, their reliance on local or pairwise information often leads to cumulative errors in long sequences with complex, non-linear trajectories. To incorporate sequential context, Luo et al. employed Long Short-Term Memory (LSTM) networks to aggregate global information and capture temporal dependencies [4], but recurrent neural networks limit computational parallelism and may struggle to preserve long-range dependencies [5]. Similarly, 3D CNNs can capture spatio-temporal information but generally introduce substantial computational and memory costs [6]. Qi Li et al. combined ResNet and transformer architectures to extract local and global features from ultrasound sequences and formulated transformation estimation as a multi-task learning problem [7]. More recent studies have explored fine-grained spatio-temporal modeling and physics-guided learning to improve trajectory recovery under complex scanning conditions [8,9]. Despite these advances, existing methods still face a fundamental trade-off between motion estimation accuracy and computational efficiency.

To address this challenge, this paper proposes an Efficient Wavelet-Convolution-Enhanced Hybrid Attention Network, termed EWT-HA-Net, for sensorless freehand 3D ultrasound reconstruction. EWT-HA-Net is designed to estimate inter-frame transformations from paired ultrasound images, from which the probe scanning trajectory is subsequently recovered by sequential transformation accumulation.

The first challenge addressed by EWT-HA-Net is the extraction of reliable motion-sensitive features from noisy, low-contrast, and weakly textured ultrasound images [10]. To address this issue, we develop an Enhanced Wavelet Transform Convolution module, termed EWTConv. Inspired by wavelet transform convolution [11], EWTConv combines cascaded wavelet decomposition with lightweight small-kernel convolutions to capture multi-scale spatial-frequency information with low computational overhead. Learnable wavelet bases improve the adaptability of feature decomposition, while dynamic frequency fusion selectively integrates informative frequency components. This design enhances motion-sensitive structural representations while suppressing redundant responses. The second challenge is to identify features that are both anatomically salient and informative for probe motion estimation. To this end, EWT-HA-Net introduces a lightweight hybrid attention module integrating Coordinate Attention (CoordAtt) and Squeeze-and-Excitation (SE). CoordAtt captures positional dependencies, whereas SE recalibrates channel-wise responses. Their complementary integration enables the network to jointly model spatial and channel information and emphasize anatomically salient and motion-informative regions. Experimental results on a large-scale dataset demonstrate that EWT-HA-Net achieves superior reconstruction accuracy and computational efficiency compared with representative state-of-the-art methods. By jointly improving reconstruction accuracy and reducing computational complexity, the proposed method aims to provide a more efficient solution for sensorless freehand 3D ultrasound reconstruction and to facilitate its further evaluation in practical ultrasound imaging scenarios.

To summarize, the main contributions of this paper are as follows:

- We propose EWT-HA-Net, a lightweight framework for sensorless freehand 3D ultrasound reconstruction that estimates inter-frame spatial transformations by exploiting both the selected frame pair and contextual information from the ultrasound sequence, and recovers the global probe trajectory through sequential transformation accumulation.

- We develop an EWTConv-based spatial-frequency feature extraction mechanism by integrating learnable wavelet decomposition and Dynamic Frequency Fusion Gating (DFFG) into the lightweight MBConv architecture. This task-oriented design enables multi-level spatial-frequency representation and adaptive fusion of low- and high-frequency information while maintaining low computational cost.
- We design a stage-wise complementary CoordAtt-SE attention strategy that combines position-sensitive spatial modeling with channel-wise feature recalibration, thereby enhancing anatomically salient and motion-informative feature representations without introducing substantial computational overhead.

2. Methods

The primary objective of sensorless freehand 3D ultrasound reconstruction is to estimate the relative spatial transformations among ultrasound frames and recover their spatial arrangement along the probe scanning trajectory. Let $\{X_1, X_2, \ldots, X_m\}$ denote an ultrasound sequence containing m frames. For two selected frames X_i and X_j, where $1 \leq i < j \leq m$, the relative spatial transformation T_{i-j} describes the change in translation and rotation from frame i to frame j. The selected frames are not necessarily adjacent, and contextual information from other available frames in the sequence can be used to assist transformation estimation, as described in Section 2.1. For complete trajectory reconstruction, the predicted transformations between successive frames are sequentially accumulated with the first frame taken as the reference coordinate system. The resulting frame poses are then used to spatially position the ultrasound images in 3D space and form the reconstructed 3D frame stack.

Figure 1 illustrates the overall framework of EWT-HA-Net for sensorless freehand 3D ultrasound reconstruction. Given a sequence of 2D ultrasound frames, the network predicts the relative transformations between consecutive frames. Taking the first frame as the reference coordinate system, the predicted inter-frame transformations are sequentially accumulated to recover the global pose of each subsequent frame. Specifically, the global transformation of frame k is calculated as

$$\hat{T}_{0,k} = \hat{T}_{0,k-1}\hat{T}_{k-1,k}, \qquad \hat{T}_{0,0} = I, \tag{1}$$

where $\hat{T}_{k-1,k}$ denotes the predicted transformation between two consecutive frames. Using the calibration information provided with the dataset, the pixel coordinates of each ultrasound frame are converted to metric coordinates and transformed into the common coordinate system defined by the first frame. In this way, the spatially positioned 2D ultrasound frames form the reconstructed 3D ultrasound frame stack and the corresponding probe trajectory.

In the present reconstruction pipeline, the calibration parameters provided with the dataset are kept fixed and are used to convert image-pixel coordinates into metric coordinates and to establish the spatial relationship between the ultrasound image plane and the acquisition coordinate system. The same calibration procedure is applied to both the predicted and ground-truth transformations. It should be noted that the reconstruction evaluated in this study is a geometrically reconstructed 3D frame stack rather than a dense Cartesian voxel volume. Therefore, no additional voxel-grid resampling, voxel interpolation, intensity compounding of overlapping voxels, or empty-voxel filling is performed. Reconstruction accuracy is evaluated directly from the spatial coordinates of the transformed ultrasound frames using MDE, FD, and DR.

The core feature encoder is constructed using modified MBConv blocks based on the EfficientNet architecture [12]. Within each block, the input features are first expanded through a 1×1 convolution, followed by EWTConv for efficient extraction of multi-scale

spatial-frequency information. A hybrid attention module is then employed to jointly model positional dependencies and channel interactions, enabling the network to emphasize anatomically salient and motion-informative features. Finally, a 1×1 pointwise convolution projects the features to the desired dimension, while dropout and a shortcut connection improve regularization and feature propagation. These modifications enhance inter-frame transformation estimation and probe trajectory recovery while maintaining low computational complexity.

The backbone follows an EfficientNet-B1-based configuration with seven MBConv stages. The output channel dimensions of the seven stages are 16, 24, 40, 80, 112, 192, and 320, respectively, comprising 23 MBConv blocks in total. The stride of the first block in each stage is 1, 2, 2, 2, 1, 2, and 1, respectively, while the remaining blocks within each stage use a stride of 1. No dilated convolution is employed, and the dilation factor is fixed to 1. After the final feature stage, global average pooling is applied, followed by a single fully connected regression head that outputs the inter-frame transformation representation. The network does not contain multiple task heads; the global probe trajectory is reconstructed subsequently by accumulating the predicted inter-frame transformations.

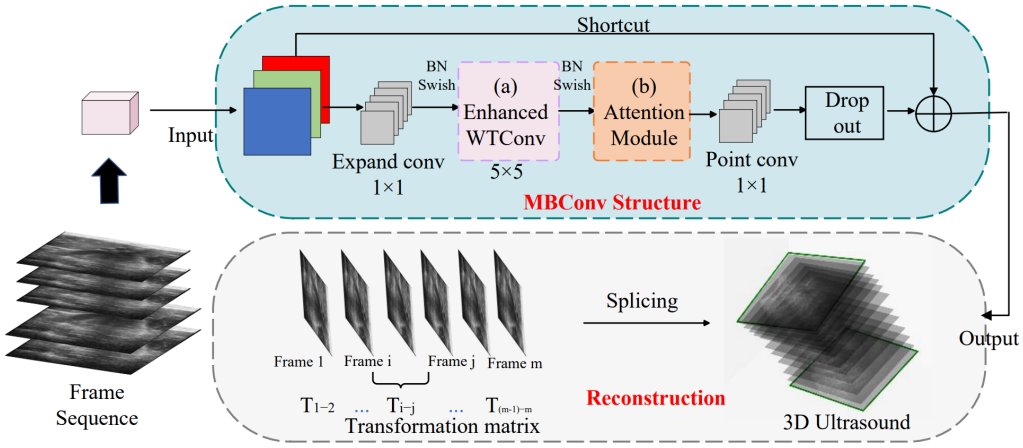

Figure 1. Overview of the EWT-HA-Net framework and the modified MBConv block.

2.1. Encoding of Ultrasound Sequences

Efficient sequence encoding is important for 3D ultrasound reconstruction because the estimation of inter-frame spatial transformations can benefit from the contextual relationships within an ultrasound sequence. Let $S_m = \{X_1, X_2, \ldots, X_m\}$ denote an ultrasound sequence containing m frames, where X_i and X_j denote the two selected frames whose relative spatial transformation is to be estimated, with $1 \leq i < j \leq m$. The selected frames are not restricted to being adjacent. In addition to X_i and X_j, contextual information may be obtained from other available frames in the sequence, including frames preceding X_i and following X_j [13]. Therefore, the selected frame pair and its surrounding sequence context are jointly represented by S_m and are used for spatial transformation prediction [7].

Based on this sequence representation, the recursive encoding function f, parameterized by θ, updates the contextual representation and predicts the relative transformation matrix T_{i-j} between the selected frames:

$$T_{i-j} = f(S_m, I_{(m-1)}; \theta), \tag{2}$$

where T_{i-j} denotes the relative spatial transformation from frame X_i to frame X_j, including relative translation and rotation, and $I_{(m-1)}$ denotes the contextual representation propagated from the preceding recursive update. Here, $I_{(m-1)}$ represents an intermediate sequence representation rather than an ultrasound image frame. Although X_i and X_j do

not appear explicitly as separate arguments in Equation (2), their information is contained in the sequence representation S_m.

The contextual representation is recursively updated as the sequence is processed:

$$I_m = f(S_m, I_{(m-1)}; \theta), \tag{3}$$

where I_m denotes the updated contextual representation at the current recursive step. Thus, the same recursive formulation describes both the propagation of sequence context and the estimation of the target inter-frame transformation. This notation should not be confused with the raw ultrasound frames, which are denoted by $X_1, \ldots, X_m$ in this subsection.

Through this sequence encoding strategy, spatial transformation estimation can exploit contextual information from the selected frame pair as well as other available frames in the acquired sequence. In the present formulation, contextual information may include both preceding and future frames. Therefore, the current sequence encoding is non-causal and is primarily intended for offline reconstruction. A strictly real-time implementation would require restricting the contextual information to the current and preceding frames and evaluating the corresponding inference latency.

2.2. Enhanced Wavelet Transform Convolutions

To enable more effective spatial-frequency feature extraction for ultrasound images, we incorporate an Enhanced Wavelet Transform Convolution (EWTConv) module into the lightweight MBConv architecture. EWTConv integrates two task-oriented design elements, namely learnable wavelet decomposition [14] and Dynamic Frequency Fusion Gating (DFFG) [15], to adapt frequency-domain feature extraction to the characteristics of ultrasound images.

Instead of standard convolutional operations, EWTConv decomposes the input features into multi-scale frequency components using learnable wavelet filters. For clarity, Equation (4) presents the basic one-dimensional filtering operation using the trainable low-pass and high-pass filters $g_\theta[n]$ and $h_\theta[n]$. In the actual two-dimensional implementation, this filtering operation is applied separably along the two spatial dimensions, resulting in the conventional low–low (LL), low–high (LH), high–low (HL), and high–high (HH) subbands. The low-frequency branch can be further decomposed hierarchically to obtain multi-level spatial-frequency representations, as illustrated in Figure 2.

$$\begin{cases} x_L[n] = \sum_k x[2n-k] \cdot g_\theta[k], \\ x_H[n] = \sum_k x[2n-k] \cdot h_\theta[k], \end{cases} \tag{4}$$

where $x_L[n]$ and $x_H[n]$ provide a compact notation for the low- and high-frequency responses generated by the wavelet filtering operation. For a two-dimensional feature map, the filtering operation is applied separably along the two spatial dimensions.

Specifically, for an input feature tensor $\mathbf{X}^{(\ell-1)} \in \mathbb{R}^{C \times H_{\ell-1} \times W_{\ell-1}}$, the four subbands at decomposition level ℓ are obtained as

$$\begin{aligned} X_{LL}^{(\ell)}(c,p,q) &= \sum_m \sum_n X^{(\ell-1)}(c, 2p-m, 2q-n) g_\theta[m] g_\theta[n], \\ X_{LH}^{(\ell)}(c,p,q) &= \sum_m \sum_n X^{(\ell-1)}(c, 2p-m, 2q-n) g_\theta[m] h_\theta[n], \\ X_{HL}^{(\ell)}(c,p,q) &= \sum_m \sum_n X^{(\ell-1)}(c, 2p-m, 2q-n) h_\theta[m] g_\theta[n], \\ X_{HH}^{(\ell)}(c,p,q) &= \sum_m \sum_n X^{(\ell-1)}(c, 2p-m, 2q-n) h_\theta[m] h_\theta[n]. \end{aligned} \tag{5}$$

Here, g_θ and h_θ denote the learnable low-pass and high-pass filters, respectively, and the factor of two in both spatial directions represents downsampling. For an input tensor of size $C \times H \times W$ with even spatial dimensions, the first-level LL, LH, HL, and HH subbands therefore have size $C \times H/2 \times W/2$. EWTConv employs a two-level decomposition, in which the first-level LL subband is further decomposed to produce second-level LL, LH, HL, and HH subbands of size $C \times H/4 \times W/4$.

The learnable filters g_θ and h_θ are initialized using Daubechies-4 wavelet coefficients and are optimized together with the remaining network parameters during training.

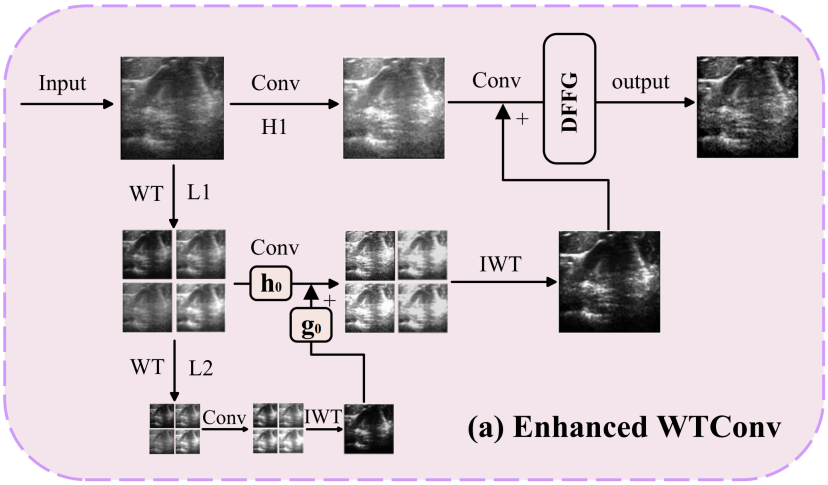

Figure 2. Detailed structure of the Enhanced WTConv module corresponding to module (a) in Figure 1.

To adaptively integrate the decomposed frequency features, a Dynamic Frequency Fusion Gating (DFFG) mechanism is employed. At each decomposition level, the LL subband is treated as the low-frequency branch, while the LH, HL, and HH subbands represent high-frequency detail information. The three high-frequency subbands are first concatenated along the channel dimension and then projected through a 1×1 convolution to obtain a unified high-frequency representation x_H with the same number of channels as the low-frequency representation x_L. Because the four subbands generated at the same decomposition level have the same spatial resolution, no additional spatial resampling is required before frequency fusion.

$$\begin{cases} z = \sigma\big(W_\phi([x_L; x_H])\big), \\ x_{\text{fused}} = z \odot x_L + (1 - z) \odot x_H, \end{cases} \tag{6}$$

where W_ϕ denotes a 1×1 convolution and $\sigma(\cdot)$ is the sigmoid function. The 1×1 convolution outputs the same number of channels as the low- and high-frequency branches. Therefore, for feature maps with spatial size $H_\ell \times W_\ell$ and C channels, the gating response satisfies $z \in \mathbb{R}^{C \times H_\ell \times W_\ell}$. Thus, z is a spatial-channel gating map rather than a scalar or a purely channel-wise coefficient. This design enables DFFG to adaptively regulate the relative contributions of low- and high-frequency information at different spatial locations and feature channels. A larger gating response assigns greater weight to the low-frequency representation, whereas a smaller response increases the contribution of the high-frequency representation.

For features from different decomposition levels, DFFG does not directly fuse tensors with mismatched spatial resolutions. In the two-level decomposition, the second-level representation is first restored to the first-level spatial resolution through inverse wavelet transformation (IWT). The reconstructed second-level representation is then incorporated into the first-level low-frequency branch before the first-level IWT is performed. Conse-

quently, cross-level feature alignment is achieved hierarchically through IWT rather than through explicit interpolation or direct fusion between tensors of different resolutions.

In the proposed EWT-HA-Net, EWTConv is incorporated into the modified MBConv block after the 1×1 expansion convolution. The EWTConv operation uses a 5×5 receptive field for spatial-frequency feature extraction, followed by batch normalization and Swish activation. The learnable low-pass and high-pass wavelet filters are initialized using Daubechies-4 coefficients and are jointly optimized with the remaining network parameters. The resulting fused representation is subsequently passed to the hybrid attention module and the 1×1 pointwise projection layer of the MBConv block. Low-frequency bands provide structural context (e.g., tissue layout), while high-frequency bands enhance detail (e.g., boundary textures). This fusion mechanism suppresses irrelevant or noisy components (e.g., speckle noise in homogeneous regions) and enhances informative features for transformation learning.

The EWTConv module, equipped with learnable frequency decomposition and content-aware fusion, facilitates robust spatiotemporal feature extraction. It effectively captures both intra-frame anatomical structures and inter-frame motion cues, enabling precise modeling of the transformation matrices for 3D ultrasound reconstruction. Moreover, the lightweight design of EWTConv reduces computational complexity, which may facilitate its future deployment in resource-constrained reconstruction scenarios.

2.3. Hybrid CoordAtt-SE Attention Module

The attention mechanism in deep learning models enables the network to focus on specific regions of an image that contain crucial information for the task at hand. This mechanism has led to significant improvements in various computer vision tasks, such as object detection and image segmentation. In our sensorless freehand 3D ultrasound reconstruction task, regions with strong speckle patterns are important because they provide structural cues for estimating inter-frame transformations. To address this, we introduce a hybrid attention module, which combines Coordinate Attention (CoordAtt) [16] and Squeeze-and-Excitation (SE) modules [17], as shown in Figure 3.

In the network architecture, CoordAtt is employed in the first MBConv block of each stage, while the conventional SE mechanism is retained in the second MBConv block. CoordAtt separately aggregates feature information along the horizontal and vertical directions to preserve position-sensitive spatial dependencies, whereas SE performs channel-wise recalibration through global average pooling and channel weighting. This stage-wise complementary configuration enables the network to successively emphasize spatially informative regions and discriminative feature channels without simultaneously introducing both attention mechanisms into the same MBConv block.

Specifically, the CoordAtt module captures global information along both horizontal and vertical directions using adaptive pooling operations. By processing and combining features from these two directions, CoordAtt effectively captures spatial dependencies across different locations, generating a spatial attention map. This allows the network to focus on regions that exhibit strong local structures or texture information in the speckle pattern, which are critical for the reconstruction task.

The SE module further enhances the feature representation by re-calibrating the important channels in the feature maps. It leverages global context information obtained through global average pooling and employs two fully connected layers to weight each channel, generating a scaling factor that is multiplied with the input features. This mechanism helps the model focus on the most informative channels, thus improving the feature representation.

By combining CoordAtt with the SE module, our hybrid attention block effectively utilizes both spatial and channel-wise information. This allows the model to better identify and enhance

the key regions in ultrasound images that are affected by speckle noise. By combining CoordAtt with SE, the hybrid attention design jointly exploits position-sensitive spatial information and channel-wise dependencies, thereby enhancing the representation of motion-informative features in ultrasound sequences while maintaining a lightweight architecture.

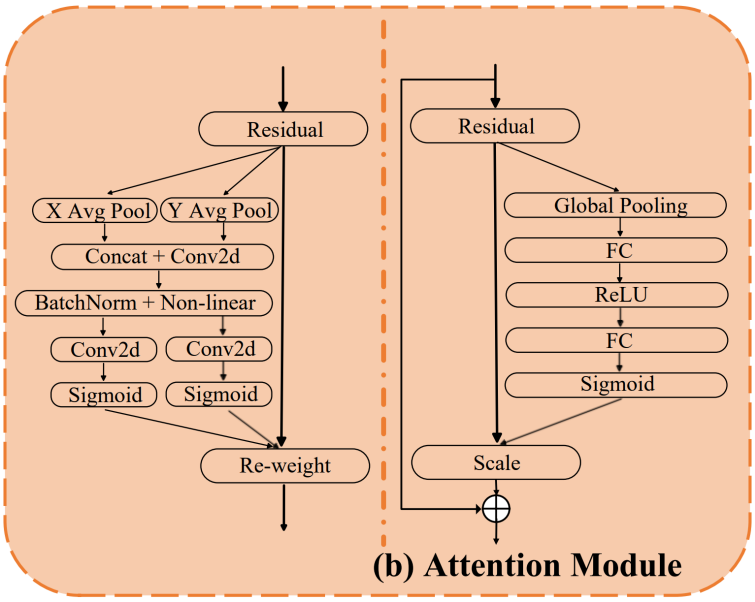

Figure 3. Detailed structure of the Hybrid Attention Module corresponding to module (b) in Figure 1.

2.4. Loss Functions

The training objective consists of two complementary components: a Root Mean Squared Distance (RMSD) loss and a Pearson-correlation-based loss. The RMSD loss directly penalizes the point-wise spatial discrepancy between the predicted and ground-truth representations, thereby providing an explicit constraint on local geometric alignment. Given the predicted representation $\hat{\mathbf{y}}_i$ and the corresponding ground truth $\mathbf{y}_i$, the RMSD loss is defined as

$$\mathcal{L}_{\text{RMSD}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2}, \tag{7}$$

where $\hat{\mathbf{y}}_i, \mathbf{y}_i \in \mathbb{R}^d$ denote the predicted and ground-truth inter-frame transformation representations for the i-th supervised frame pair, respectively; d denotes the dimensionality of the transformation representation; and N denotes the number of supervised frame pairs involved in the loss calculation. The loss is applied directly to the inter-frame transformation representations predicted by the network rather than to the final reconstructed trajectory or the frame-corner coordinates used for evaluation.

Although the RMSD loss effectively constrains point-wise spatial errors, it does not explicitly enforce consistency in the overall variation pattern between the predicted and ground-truth representations. Therefore, a Pearson-correlation-based loss is further introduced to provide a complementary global structural constraint.

Let $\hat{\mathbf{y}}$ and $\mathbf{y}$ denote the vectorized prediction and ground truth, respectively. Their Pearson correlation coefficient is calculated as

$$\rho(\hat{\mathbf{y}}, \mathbf{y}) = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{8}$$

where $\bar{\hat{y}}$ and $\bar{y}$ denote the mean values of the predicted and ground-truth representations, respectively. Based on the Pearson correlation coefficient, the correlation loss is defined as

$$\mathcal{L}_{\text{corr}} = 1 - \rho(\hat{\mathbf{y}}, \mathbf{y}). \tag{9}$$

The RMSD loss and the correlation loss are equally weighted in our implementation, i.e., $\lambda_{\text{RMSD}} = \lambda_{\text{corr}} = 1$. Accordingly, the final training objective is defined as

$$\mathcal{L} = \lambda_{\text{RMSD}}\mathcal{L}_{\text{RMSD}} + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}} = \mathcal{L}_{\text{RMSD}} + \mathcal{L}_{\text{corr}}. \tag{10}$$

In all experiments, the two terms are equally weighted, with $\lambda_{\text{RMSD}} = \lambda_{\text{corr}} = 1$. This equal-weight setting was used consistently without additional weight tuning. The RMSD term constrains the direct discrepancy between the predicted and ground-truth inter-frame transformations, whereas the correlation term promotes consistency in their overall variation pattern. The predicted inter-frame transformations are subsequently accumulated to recover the probe trajectory and perform 3D reconstruction. Their combination enables the network to jointly optimize point-wise spatial alignment and overall transformation consistency.

3. Experiments and Results

3.1. Datasets

We utilize an open-source dataset from [18,19] to validate our model. The ultrasound image dataset used in this study was acquired using an Ultrasonix machine (BK, Europe (BK Medical, Herlev, Denmark)) with a curvilinear probe (4DC7-3/40 model). The position information for each frame of ultrasound images was recorded using an optical tracker (NDI Polaris Vicra, Northern Digital Inc., Waterloo, ON, Canada). The data was collected at a frame rate of 20 fps with an image size of 480 × 640, and no speckle reduction was applied. The ultrasound system was set to a frequency of 6 MHz, with a dynamic range of 83 dB, a gain of 48%, and a depth of 9 cm.

The scanning subjects in the dataset were volunteers' left and right forearms. For each forearm, the ultrasound probe was moved along three different trajectories, that are linear line("L" shape), "C" shape, and "S" shape, scanning in both distal-to-proximal and proximal-to-distal directions, with the ultrasound plane kept perpendicular or parallel to the scanning direction. The complete dataset consists of 1200 scans from 50 subjects, with each subject contributing 24 scans. The data are organized into 50 subject-specific folders, and each scan contains the ultrasound image sequence and the corresponding transformation matrices stored in an .h5 file. To prevent subject-level data leakage, the dataset was partitioned strictly at the subject level rather than at the scan level. Specifically, the 50 subjects were divided into training, validation, and test sets at a ratio of 3:1:1, corresponding to 30 subjects (720 scans), 10 subjects (240 scans), and 10 subjects (240 scans), respectively. All scans associated with a given subject were assigned exclusively to one subset, and no subject was shared among the training, validation, and test sets. Therefore, the reported test performance was evaluated on subjects that were completely unseen during model training and validation.

The original dataset contained 400 linear (L-shaped), 400 C-shaped, and 400 S-shaped scans. Compared with linear scanning, the C-shaped and S-shaped trajectories involve more pronounced changes in probe translation and orientation and therefore represent more challenging cases for trajectory estimation. To increase the representation of nonlinear scanning trajectories, 40 linear scans were removed and replaced by 20 C-shaped scans and 20 S-shaped scans selected from the same source dataset. The replacement scans were acquired under the same data-collection protocol and were selected solely according to

the predefined trajectory category, independently of model predictions or reconstruction performance. As a result, the final dataset contained 360 linear, 420 C-shaped, and 420 S-shaped scans, while the total number of scans remained unchanged at 1200.

The 50 subjects were divided into training, validation, and test sets using a 3:1:1 subject-level split. All scans from the same subject were assigned exclusively to one subset, and no subject was shared across the training, validation, and test sets. The subject assignment was kept fixed during the trajectory-composition adjustment, and the same subject-independent partition and final trajectory composition were used for EWT-HA-Net and all comparison methods.

3.2. Implementation Details

To ensure a fair comparison, the proposed model and the comparison methods were evaluated under a unified experimental protocol. All experiments were conducted on a machine equipped with an NVIDIA RTX 4090 GPU. The model was implemented using the PyTorch deep learning framework (version 2.1.0). The training and evaluation processes were performed on a Linux operating system (Ubuntu 20.04). The relevant libraries and their versions used for the implementation included h5py 3.8.0, matplotlib 3.8.2, numpy 1.26.2, torch 2.1.0, torchvision 0.16.0, and tensorboard 2.15.1.

The proposed EWT-HA-Net was trained using the Adam optimizer with an initial learning rate of 1×10^{-4}, which was decayed by a factor of 0.9 every 5 epochs. The batch size was set to 16. Each network input consisted of two ultrasound frames with a spatial resolution of 480×640 pixels. Validation was performed after every training epoch, and the checkpoint achieving the lowest geometric distance error on the validation set was selected as the final model for test-set evaluation.

For all comparative experiments, the same subject-independent training, validation, and test subsets described in Section 3.1 were used. Specifically, all methods were trained and evaluated using the same fixed subject partition, trajectory composition, image preprocessing pipeline, and evaluation metrics. The validation set was used for model selection and hyperparameter adjustment, whereas the test set was kept completely held out during model development and was used only for the final performance evaluation.

For task-specific comparison methods with publicly available implementations, including DCL-Net, the multi-task model, and the TUS-REC baseline, we adopted the official publicly released code. ResNet101, 2D ResNeXt, and 3D ResNeXt were reimplemented according to their original publications and adapted to the same inter-frame transformation regression task. All comparison models were retrained and evaluated using the same subject-independent data partition as EWT-HA-Net.

To ensure a fair comparison while respecting the architectural characteristics of different methods, method-specific optimization settings and training durations followed those recommended in the corresponding official implementations or original publications whenever applicable, rather than forcibly imposing identical training hyperparameters on all architectures. When additional adjustment was required, the hyperparameters were determined exclusively according to performance on the common validation set. The test set remained completely held out and was not used for hyperparameter tuning, checkpoint selection, or model development.

Preprocessing included normalization of the ultrasound images and data augmentation using random rotations, cropping, and flipping to improve model generalization. The same preprocessing and data-augmentation procedures were consistently applied to the training data of all comparison methods.

3.3. Evaluation Metrics

We evaluated the model's performance using several metrics to assess accuracy, robustness, and efficiency. The Mean Corner Point Distance Error (MDE) measures the average distance between corresponding frame corners, reflecting variations in speed and orientation across frames. The Final Drift (FD) quantifies the displacement error between the last frame and the ground truth, indicating long-term tracking accuracy. The Drift Rate (DR) normalizes FD by the sequence length, providing a measure of error relative to sequence duration. These three metrics together assess the model's accuracy and error accumulation. For efficiency, we used Params (total parameters indicating model memory requirements) and FLOPs (total floating-point operations), with lower values indicating a more efficient model. Additionally, we introduced the Average Percentage Difference (APD) to compare our method with baseline models, quantifying improvements in key performance metrics like MDE and FD. These evaluations offer a comprehensive view of the model's accuracy, robustness, and computational efficiency in 3D ultrasound image sequence reconstruction.

3.4. Comparison to Baseline Methods

Table 1 presents a comparative analysis of EWT-HA-Net and the evaluated comparison methods in terms of reconstruction accuracy and computational complexity. EWT-HA-Net achieves the lowest MDE, FD, and DR, with values of 15.74 mm, 26.87 mm, and 11.74%, respectively. Meanwhile, the proposed method requires only 8.81 M parameters and 692.41 M FLOPs, indicating a favorable balance between reconstruction accuracy and computational complexity.

Table 1. Performance of different methods using test set which includes 240 scans.

Model	MDE (mm)		FD (mm)		DR (%)		Params (M)		FLOPs (M)	
	Value	APD (%)	Value	APD (%)	Value	APD (%)	Value	APD (%)	Value	APD (%)
ResNet101 [20]	25.46	38.18	46.78	42.56	13.82	15.05	44.55	80.22	7866.44	91.20
2D ResNeXt [21]	22.35	29.57	40.17	33.11	11.98	2.00	25.03	64.80	4288.18	83.85
3D ResNeXt [22]	19.82	20.78	31.54	14.81	12.03	2.41	28.79	69.40	8825.51	92.15
DCL-Net [23]	17.21	8.72	30.05	10.58	11.94	1.68	30.01	70.64	8404.60	91.76
Multi-task model [13]	20.03	21.62	34.81	22.81	12.67	7.34	9.20	4.24	1024.30	32.40
Baseline [19]	19.10	17.59	33.00	18.58	12.54	6.38	12.23	27.96	1019.63	32.09
Our Method	**15.74**	0	**26.87**	0	**11.74**	0	**8.81**	0	**692.41**	0

Note: Bold values indicate the best result for each metric among the compared methods.

To further assess the statistical reliability of the reconstruction accuracy, we performed a subject-level analysis using MDE as the primary accuracy metric. MDE was calculated separately for each of the 10 independent test subjects, and the results are reported as mean $\pm$ standard deviation together with 95% bootstrap confidence intervals obtained by resampling subjects. Because all methods were evaluated on the same test subjects, paired Wilcoxon signed-rank tests were used to compare EWT-HA-Net with each comparison method. Holm correction was applied to account for multiple pairwise comparisons, with an adjusted p-value below 0.05 considered statistically significant. The subject was treated as the independent statistical unit because multiple scans acquired from the same subject are not statistically independent.

As shown in Table 2, EWT-HA-Net achieves the lowest subject-level MDE of 15.74 ± 1.77 mm. The paired Wilcoxon signed-rank tests show that the differences between EWT-HA-Net and all evaluated comparison methods remain statistically significant after Holm correction ($p_{adj} < 0.05$). In particular, the comparison with DCL-Net, which achieves the second-lowest MDE, yields an adjusted p-value of 0.04883. The present statisti-

cal analysis characterizes inter-subject variability, whereas variability arising from repeated training with different random seeds remains to be investigated in future work.

Table 2. Subject-level statistical analysis of MDE on the independent test set.

Model	MDE (mm), Mean ± SD	95% CI (mm)	p_{adj}
ResNet101	25.46 ± 3.27	23.56–27.41	0.01172
2D ResNeXt	22.35 ± 4.35	20.10–25.16	0.01172
3D ResNeXt	19.82 ± 2.07	18.67–21.08	0.01172
DCL-Net	17.21 ± 2.19	15.97–18.53	0.04883
Multi-task model	20.03 ± 2.73	18.34–21.53	0.01172
Baseline	19.10 ± 2.17	17.88–20.42	0.01172
EWT-HA-Net	15.74 ± 1.77	14.66–16.74	–

Figure 4 provides a complementary visualization of the reconstruction-error distributions across the evaluated methods. Because MDE and FD are measured in millimeters whereas DR is expressed as a percentage, DR is multiplied by 2.5 only for visualization so that the three metrics can be displayed within a comparable plotting range. This scaling is used solely for graphical presentation and does not affect any quantitative results or comparisons. The original, unscaled values of MDE, FD, and DR are reported in Table 1.

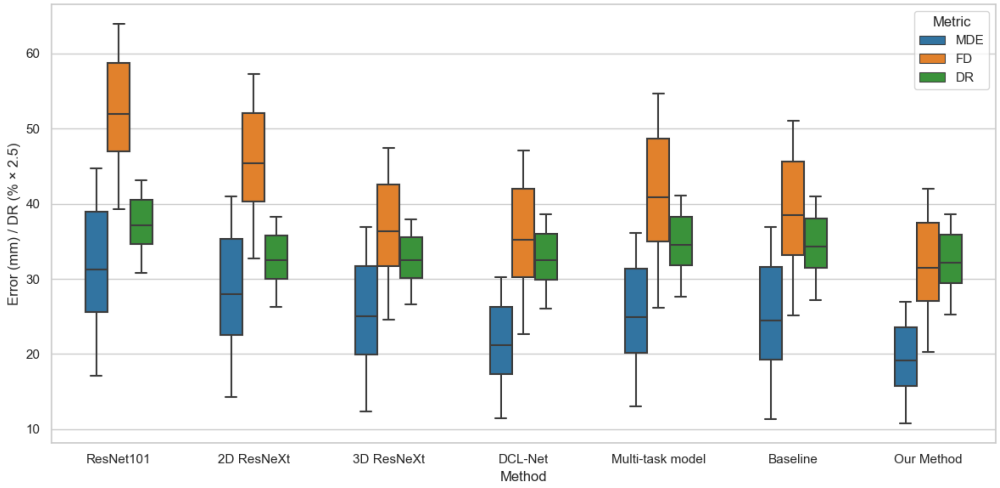

Figure 4. Comparison of the reconstruction error range between our method and the others.

We further evaluated the methods on the subsets containing only C-shaped and S-shaped scans, as reported in Tables 3 and 4. Under these more complex nonlinear scanning trajectories, EWT-HA-Net achieves lower MDE, FD, and DR than the evaluated comparison methods. Compared with the results on the complete test set, the performance differences become more pronounced for several competing methods, whereas EWT-HA-Net maintains relatively lower reconstruction errors. These results suggest that the proposed method is less sensitive to increasing trajectory complexity and maintains comparatively favorable reconstruction performance under the evaluated complex scanning patterns.

Figure 5 presents a frame-wise MDE comparison between the baseline and EWT-HA-Net under different scanning trajectories, including (a) L-shaped, (b) C-shaped, and (c) S-shaped scans. The solid curves represent the mean MDE at each frame index, while the shaded regions indicate the corresponding minimum–maximum error ranges. The dashed boundaries denote the minimum and maximum observed errors.

Table 3. Performance of different methods using test set which includes only "C" scans.

Model	MDE (mm)		FD (mm)		DR (%)	
	Value	APD (%)	Value	APD (%)	Value	APD (%)
ResNet101	42.96	45.46	46.89	32.48	16.34	16.46
2D ResNeXt	40.91	17.48	47.86	16.20	18.08	24.50
3D ResNeXt	36.10	35.10	43.68	27.52	16.01	14.74
DCL-Net	27.36	14.36	35.79	11.54	14.06	2.91
Multi-task model	37.28	37.15	44.97	29.60	15.50	11.94
Baseline	35.13	33.30	40.38	21.59	14.66	6.89
Our Method	**23.43**	0	**31.66**	0	**13.65**	0

Note: Bold values indicate the best result for each metric among the compared methods.

Table 4. Performance of different methods using test set which includes only "S" scans.

Model	MDE (mm)		FD (mm)		DR (%)	
	Value	APD (%)	Value	APD (%)	Value	APD (%)
ResNet101	53.69	48.05	50.12	33.26	19.87	29.09
2D ResNeXt	47.31	41.05	50.93	17.48	19.36	27.22
3D ResNeXt	40.34	30.86	45.26	26.09	16.18	12.92
DCL-Net	31.69	11.99	37.10	9.84	15.28	7.79
Multi-task model	44.63	37.51	46.98	28.80	17.85	21.06
Baseline	39.64	29.64	41.08	18.57	16.55	14.86
Our Method	**27.89**	0	**33.45**	0	**14.09**	0

Note: Bold values indicate the best result for each metric among the compared methods.

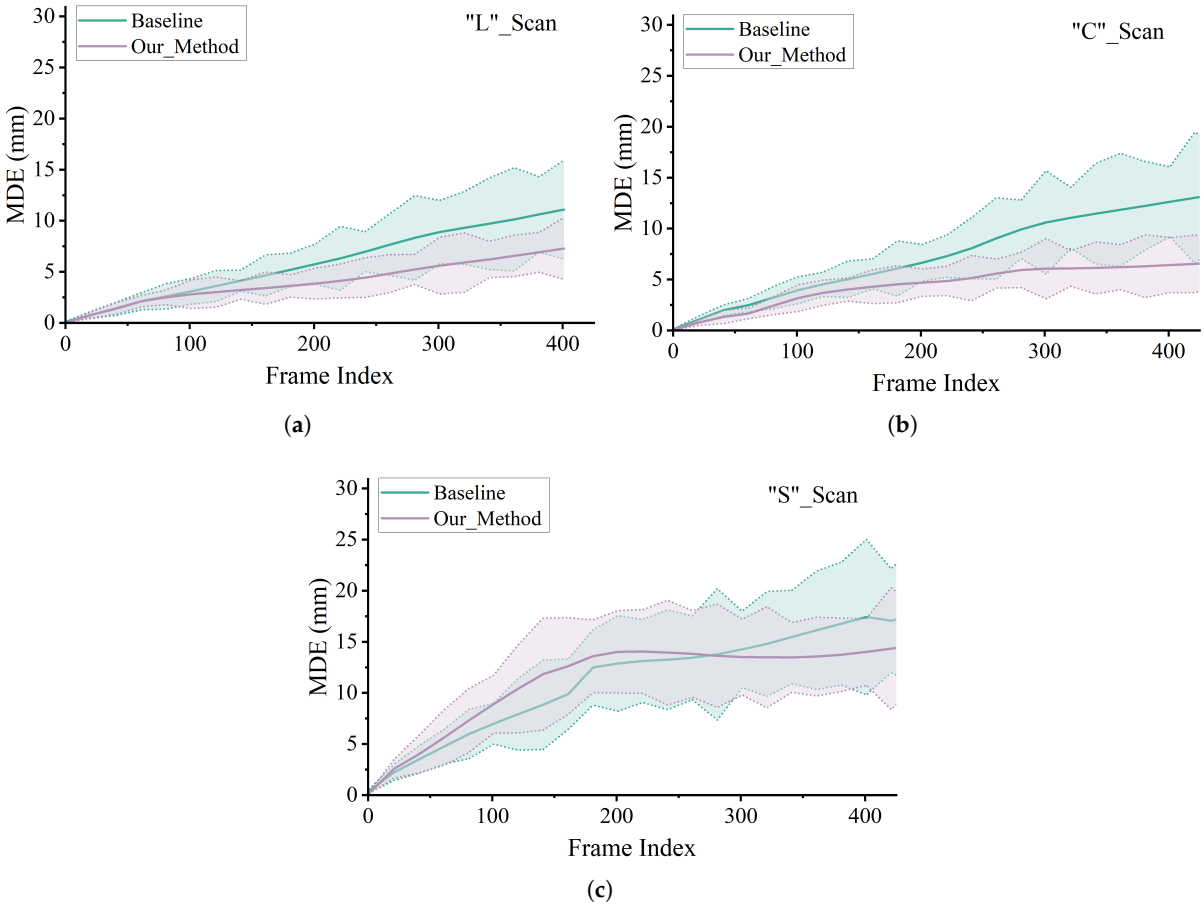

Figure 5. Comparison of performance between baseline and proposed method in different scanning strategies at varying frame index: (**a**) L-shaped scan; (**b**) C-shaped scan; and (**c**) S-shaped scan.

Figure 6 presents representative qualitative comparisons of reconstructed scanning trajectories for linear, C-shaped, and S-shaped scans. The ground-truth trajectory (GT), the prediction of EWT-HA-Net (Pred), and the baseline reconstruction are shown for visual comparison. The purpose of this figure is to illustrate the overall agreement between the reconstructed and ground-truth scanning paths rather than to provide an independent quantitative assessment of spatial-feature preservation. Quantitative reconstruction accuracy is evaluated using MDE, FD, and DR and is reported in Tables 1, 3 and 4. The larger errors observed for the more complex nonlinear trajectories, particularly the S-shaped scans, also indicate that such scanning patterns remain challenging for the current method.

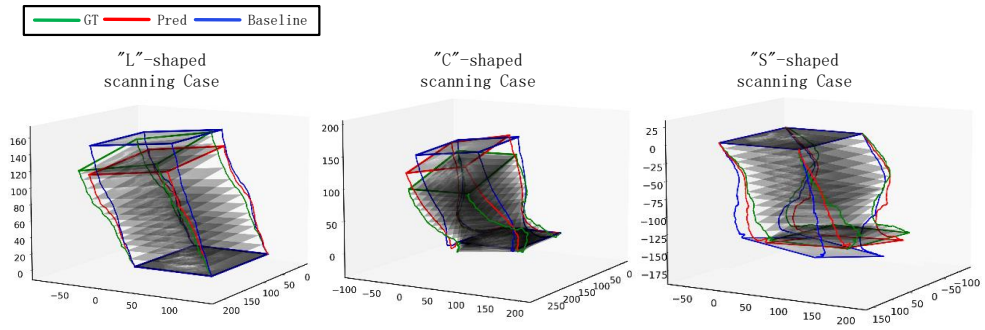

Figure 6. Comparison of the reconstruction results between the baseline and our method.

Although EWT-HA-Net demonstrates improved reconstruction performance across the evaluated scanning trajectories, the results also reveal several limitations. In particular, reconstruction errors increase for more complex nonlinear trajectories, especially S-shaped scans, indicating that large and rapidly varying probe motions remain challenging. In addition, the present experiments are limited to forearm ultrasound data acquired using a specific scanner and probe configuration. Therefore, the generalizability of the proposed method to other anatomical regions, ultrasound systems, probes, operators, and acquisition conditions requires further validation.

3.5. Ablation Study

Table 5 presents a controlled set of module-level configurations designed to clarify the respective roles of the loss formulation, attention mechanisms, and EWTConv in EWT-HA-Net. The same data partition, training settings, and evaluation protocol were used across all configurations. The first two configurations examine the effect of the loss formulation, the intermediate configurations compare different attention designs, and the final comparison evaluates the contribution of EWTConv within the complete attention-based configuration.

Table 5. Ablation study of the major module-level configurations of EWT-HA-Net. The configurations are designed to evaluate the effects of the loss formulation, attention mechanisms, and EWTConv.

Configuration	EWTConv	CoordAtt	SE	RMSD+Corr.	MDE (mm)	FD (mm)	DR (%)	Params (M)	FLOPs (M)
C1: SE + RMSD			✓		20.07	33.92	14.79	9.20	735.61
C2: SE + RMSD + Corr.			✓	✓	19.10	33.00	13.54	9.20	735.61
C3: CoordAtt + RMSD + Corr.		✓		✓	20.74	36.26	13.24	9.38	759.55
C4: CoordAtt + SE + RMSD + Corr.		✓	✓	✓	**15.51**	27.04	12.10	9.22	741.04
C5: Full EWT-HA-Net	✓	✓	✓	✓	15.74	**26.87**	**11.74**	**8.81**	**692.41**

Note: Bold values indicate the best result for each metric among the compared methods. A checkmark indicates that the corresponding component is included in the configuration.

A controlled set of module-level configurations was evaluated to clarify the respective roles of the loss formulation, CoordAtt, SE, and EWTConv. The same data partition, training settings, and evaluation protocol were used across these configurations. The first two configurations differ only in the loss function, the intermediate configurations examine the attention design, and the final comparison evaluates the effect of introducing EWTConv into the complete attention-based configuration.

First, the effect of the loss function can be observed by comparing the configurations using RMSD alone and the combined RMSD-correlation loss. With only the RMSD loss, the model achieves an MDE of 20.07 mm, an FD of 33.92 mm, and a DR of 14.79%. After introducing the correlation term, MDE decreases to 19.10 mm, FD to 33.00 mm, and DR to 13.54%, while the parameter count and FLOPs remain unchanged at 9.20 M and 735.61 M, respectively. These results indicate that the correlation term complements point-wise alignment by encouraging consistency in the overall spatial relationships between the predicted and ground-truth transformations.

Second, the attention-module configurations further illustrate the effect of different attention designs. With CoordAtt alone, the model obtains an MDE of 20.74 mm, whereas the combined CoordAtt-SE configuration reduces MDE to 15.51 mm. These results suggest that combining position-sensitive spatial modeling with channel-wise recalibration is beneficial for inter-frame transformation estimation under the evaluated configuration.

It should be noted that the present ablation study evaluates EWTConv as an integrated module. The learnable wavelet filters and DFFG were not independently isolated, and therefore their individual contributions should not be inferred from the current ablation results.

Finally, the contribution of EWTConv can be evaluated by comparing the complete attention-based configuration before and after its introduction, with the other components kept unchanged. After EWTConv is introduced, the parameter count decreases from 9.22 M to 8.81 M and FLOPs decrease from 741.04 M to 692.41 M, corresponding to reductions of approximately 4.4% and 6.6%, respectively. Meanwhile, FD decreases from 27.04 mm to 26.87 mm and DR decreases from 12.10% to 11.74%. Although MDE slightly increases from 15.51 mm to 15.74 mm, the overall results indicate that EWTConv provides a favorable balance between computational efficiency and long-sequence drift robustness. Its contribution is therefore mainly reflected in efficient multi-frequency feature extraction and reduced error accumulation rather than a uniform improvement across all accuracy metrics.

Overall, the ablation study reveals distinct but complementary roles of the proposed components. The combined RMSD-correlation loss improves the optimization of spatial relationships, the CoordAtt-SE hybrid attention mechanism provides the major improvement in reconstruction accuracy, and EWTConv further reduces computational complexity while maintaining competitive MDE and improving FD and DR. Their integration enables EWT-HA-Net to achieve a favorable balance among reconstruction accuracy, drift robustness, and model efficiency.

4. Discussion

Several limitations of the present study should be acknowledged. First, the current experiments were conducted on forearm ultrasound data acquired using a single ultrasound system and probe configuration. Therefore, the generalizability of EWT-HA-Net to other anatomical regions, ultrasound scanners, probe types, operators, and acquisition conditions has not yet been fully established. Second, although the proposed model substantially reduces the number of parameters and FLOPs, computational efficiency alone does not directly demonstrate real-time performance, and a dedicated latency analysis on different hardware platforms remains necessary. Third, the present study focuses on reconstruction accuracy and computational efficiency rather than prospective clinical validation. Future

work will therefore investigate cross-device and cross-anatomy generalization, inference latency on different computational platforms, and prospective evaluation under more diverse clinical acquisition conditions. The present statistical analysis characterizes inter-subject variability, whereas variability arising from repeated training with different random seeds remains to be investigated in future work.

5. Conclusions

This paper proposes EWT-HA-Net, an efficient deep learning framework for sensorless freehand 3D ultrasound reconstruction. The proposed method estimates inter-frame spatial transformations by exploiting the selected frame pair together with sequence contextual information and recovers the probe trajectory through sequential transformation accumulation. EWTConv enhances multi-scale spatial-frequency feature extraction through learnable wavelet decomposition and dynamic frequency fusion, while the stage-wise CoordAtt-SE attention strategy strengthens position-sensitive and channel-wise feature representation. Experimental results demonstrate that EWT-HA-Net achieves improved reconstruction performance with low computational complexity and maintains comparatively stable performance under more complex scanning trajectories. Overall, the proposed framework provides a favorable balance between reconstruction accuracy, drift robustness, and model efficiency. Future work will focus on generalization across different anatomical regions and ultrasound systems, inference latency evaluation, and prospective clinical validation.

Author Contributions: Conceptualization Y.Y.; methodology, Y.Y.; software, Y.Z.; validation, Y.Y. and Y.Z.; formal analysis, Y.Y.; resources, Z.B.; writing—original draft preparation, Y.Y.; writing—review and editing, Y.Y.; visualization, Y.Z.; supervision, Z.B.; project administration, Z.B. and Q.N.; funding acquisition, Y.Y. and Q.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Jiangsu Province [BK20231059] and the Key Research and Development Projects in Xuzhou [XWKYHT20240092].

Institutional Review Board Statement: The ethical approval and informed-consent procedures associated with the original data acquisition are reported in the original TUS-REC2024 dataset documentation and source publication.

Informed Consent Statement: Informed consent for participation was obtained by the original data collectors from all subjects involved in the primary data collection for the TUS-REC 2024 challenge. The present study is a secondary analysis of this publicly available, anonymized dataset, and therefore, no additional informed consent was required.

Data Availability Statement: The data presented in this study are openly available in the TUS-REC 2024 Challenge dataset at https://github-pages.ucl.ac.uk/tus-rec-challenge/TUS-REC2024/data.html (accessed on 6 October 2024). These data were originally published and described by Qi Li, et al. [19] in the challenge overview paper "TUS-REC2024: A Challenge to Reconstruct 3D Freehand Ultrasound Without External Tracker" (https://doi.org/10.48550/arXiv.2506.21765, accessed on 13 November 2025). No new data were created in this study.

Acknowledgments: The authors would like to thank the anonymous referees and the editors for their helpful comments and suggestions.

Conflicts of Interest: All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EWT-HA-Net	Efficient Wavelet-convolution-enhanced Hybrid Attention Network
EWTConv	Enhanced Wavelet Transform Convolution Module

References

1. Tenajas, R.; Miraut, D.; Illana, C.I.; Alonso-Gonzalez, R.; Arias-Valcayo, F.; Herraiz, J.L. Recent advances in artificial intelligence-assisted ultrasound scanning. *Appl. Sci.* **2023**, *13*, 3693. [CrossRef]
2. Adriaans, C.A.; Wijkhuizen, M.; van Karnenbeek, L.M.; Geldof, F.; Dashtbozorg, B. Trackerless 3D freehand ultrasound reconstruction: A review. *Appl. Sci.* **2024**, *14*, 7991. [CrossRef]
3. Prevost, R.; Salehi, M.; Jagoda, S.; Kumar, N.; Sprung, J.; Ladikos, A.; Bauer, R.; Zettinig, O.; Wein, W. 3D freehand ultrasound without external tracking using deep learning. *Med. Image Anal.* **2018**, *48*, 187–202. [CrossRef]
4. Luo, M.; Yang, X.; Wang, H.; Du, L.; Ni, D. Deep motion network for freehand 3D ultrasound reconstruction. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 290–299.
5. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
6. Maturana, D.; Scherer, S. Voxnet: A 3D convolutional neural network for real-time object recognition. In *Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015*; IEEE: Piscataway, NJ, USA, 2015; pp. 922–928.
7. Li, Q.; Shen, Z.; Li, Q.; Barratt, D.C.; Dowrick, T.; Clarkson, M.J.; Vercauteren, T.; Hu, Y. Trackerless freehand ultrasound with sequence modelling and auxiliary transformation over past and future frames. In *Proceedings of the 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Cartagena de Indias, Colombia, 18–21 April 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
8. Yan, Z.; Yang, X.; Luo, M.; Chen, J.; Chen, R.; Liu, L.; Ni, D. Fine-grained context and multi-modal alignment for freehand 3D ultrasound reconstruction. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Saint-Malo, France, 26–29 September 2004*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 340–349.
9. Dou, Y.; Mu, F.; Li, Y.; Varghese, T. Sensorless end-to-end freehand 3-D ultrasound reconstruction with physics-guided deep learning. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control.* **2024**, *71*, 1514–1525. [CrossRef]
10. van der Pol, H.G.; van Karnenbeek, L.M.; Wijkhuizen, M.; Geldof, F.; Dashtbozorg, B. Deep learning for point-of-care ultrasound image quality enhancement: a review. *Appl. Sci.* **2024**, *14*, 7132. [CrossRef]
11. Finder, S.E.; Amoyal, R.; Treister, E.; Freifeld, O. Wavelet convolutions for large receptive fields. In *Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 363–380.
12. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019*; PMLR: Cambridge, MA, USA, 2019; pp. 6105–6114.
13. Li, Q.; Shen, Z.; Li, Q.; Barratt, D.C.; Dowrick, T.; Clarkson, M.J.; Vercauteren, T.; Hu, Y. Long-term Dependency for 3D Reconstruction of Freehand Ultrasound Without External Tracker. *IEEE Trans. Biomed. Eng.* **2023**, *71*, 1033–1042. [CrossRef]
14. Zhang, J.; Walter, G.G.; Miao, Y.; Lee, W.N.W. Wavelet neural networks for function learning. *IEEE Trans. Signal Process.* **1995**, *43*, 1485–1497. [CrossRef]
15. Sun, X.; Yu, Y.; Cheng, Q. Adaptive multimodal feature fusion with frequency domain gate for remote sensing object detection. *Remote Sens. Lett.* **2024**, *15*, 133–144. [CrossRef]
16. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Conference, 19–25 June 2021; pp. 13713–13722.
17. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
18. Li, Q.; Saeed, S.U.; Barratt, D.C.; Clarkson, M.J.; Vercauteren, T.; Hu, Y. Trackerless 3D Freehand Ultrasound Reconstruction Challenge. *Zenodo* **2024**. [CrossRef]
19. Li, Q.; Saeed, S.U.; Huang, Y.; Luo, M.; Yan, Z.; Chen, J.; Yang, X.; Ni, D.; Winter, N.; Nguyen, P.; et al. TUS-REC2024: A Challenge to Reconstruct 3D Freehand Ultrasound Without External Tracker. *arXiv* **2025**, arXiv:2506.21765.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
21. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.

22. Wu, P.; Cui, Z.; Gan, Z.; Liu, F. Three-Dimensional ResNeXt Network Using Feature Fusion and Label Smoothing for Hyperspectral Image Classification. *Sensors* **2020**, *20*, 1652. [CrossRef]
23. Guo, H.; Xu, S.; Wood, B.; Yan, P. Sensorless freehand 3D ultrasound reconstruction via deep contextual learning. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Virtual, 4–8 October 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 463–472.