International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(16s) 2026 pp. 455-478
https://doi.org/10.70917/ijcisim-2026-4599
Article

Adaptive Wavelet Decomposition with Cross-Band Attention for Unsupervised 3D Medical Image Registration

Hussein Abdulkhalek Midhat[1], Asim Majeed Murshid[2]

[1,2]Department of Computer Science, College of Computer Science and Information Technology, University of Kirkuk, Kirkuk, Iraq

Abstract: This report introduces a new unsupervised deformable 3D registration network called AWaRe Net developed for the purpose of overcoming the drawbacks of existing techniques with regard to multi scale fusion and losing key information. AWaRe Net replaces fixed Haar wavelets with the ability to create adaptive wavelet filters through a technique called Adaptive Learnable Wavelet Decomposition (ALWD) that allows each filter to have parameters optimized using backpropagation while simultaneously allowing for attention to dynamically adjust weights across sub bands using CrossBand Attention Fusion (CBAF). Additionally, a hybrid lightweight bottleneck architecture that combines ConvNexT and axial attention is used to create long range dependencies between data points in a Linearithmic (i.e, O(N log N)) fashion, whilst an explicit frequency loss provides additional alignment of wavelets and other properties found within traditional 2D methods through conventional spatial similarity. AWaRe Net has been evaluated against two publicly available, currently utilized datasets: IXI (image registration from Atlas to Patient) and OASIS (inter-patient), and demonstrates mean values of 79.1% and 83.9% Dice scores, respectively which significantly outperform comparable algorithms (Wave Morph, Trans Morph, and Voxel Morph) ($p < 0.001$). This allowed for achieving a lower folding ratio of approx. 0.168% relative to previously published algorithms on the same dataset (OASIS), with an average inference time of 68 ms per image pair, and only 0.82 million parameters. Each proposed component was confirmed through an ablation study. Additionally, generalization tests on synthetic multi organ datasets have demonstrated enhanced transferability of registration results to abdominal, cardiac and lung images. This study demonstrates that the use of adaptive wavelet decomposition with cross band attention substantially increases accuracy within the registration process, the smoothness of deformation across the registered image, as well as the computational efficiency of the entire process as it relates to clinical deployment and longitudinal disease monitoring in real time.

Keywords: Medical image registration; unsupervised deep learning; adaptive wavelet transform; crossband attention; ConvNeXt; frequencydomain loss.

1. Introduction

Registration of Non-Rigid Medical Images (NRMI) is an important clinical and biomedical research task which enables the alignment of anatomical structures from different medical imaging modalities in three-dimensional (3D) space. Applications include (but are not limited to) diagnosis (e.g., long-term scans used for longitudinal studies), treatment (e.g., information from multiple modalities can be combined to create a predefined target for radiation therapy), and image-based surgical intervention (e.g., overlaying a preoperative surgical plan on the patient's anatomy during surgery) [Avants et al., 2008]. Through accurate registration, clinicians can perform quantitative assessment of morphology, combine complementary information from multiple imaging modalities, and perform complex procedures with increased accuracy.

Brain MRI registration is a precondition for many other tasks, such as tumour detection and segmentation. The use of hybrid deep learning and classical machine learning techniques such as CNN-MLP and CNN-KNN have been

proven to be effective in brain tumour classification using MRI [Assaad et al., 2025]. Such methods depend on well aligned anatomical structures, and highlight the importance of accurate deformable registration. The traditional methods of iterative registration of images such as SyN (Symmetric Normalization) and LDDMM (Large Deformation Diffeomorphic Metric Mapping), have long been the "gold standard" of image registration because of their rigorous mathematical foundation and their ability to generate diffeomorphic (topology-preserving) deformation fields.These highly successful optimization-based methods, however, suffer from a number of key limitations: They are very computationally costly and are required to take many minutes to complete the registration for each pair of images; They are not generalizable to other patients or anatomies as each image pair is a separate solution; They require hand-crafted similarity metrics that are not always representative of the complex appearance variations caused by the tissues. [Avants et al., 2008; Balakrishnan et al., 2019].

Recent advances in deep learning have revolutionized the way that medical images are registered. The first deep learning methods for image registration operated in a supervised manner, training networks to directly predict the deformation fields between two images based on gold-standard correspondences or "ground-truth" information. Unfortunately, using gold-standard information is rarely practical or feasible in a clinical environment. In contrast, the unsupervised deep registration method first described in VoxelMorph [Balakrishnan et al. 2019] learned how to register images by training the algorithm/architecture to use a similarity loss between the fixed image and the warped moving image, without the requirement for any gold-standard deformation fields. Today, unsupervised deep registration is one of the dominant approaches to medical image registration, in part because it leverages large amounts of unlabeled data, and learns to extract features that can be applied to new subjects in a completely generalizable manner.

The backbone architecture is a key design decision for unsupervised registration networks. Convolutional neural networks (CNNs) and, in particular, U Net based encoders provide strong inductive biases for recognition of local patterns, as well as good performance in terms of computational efficiency. Receptive fields, however, are relatively small and therefore these cells do not take into account the long-range spatial relations that are important in large deformations. However, transformers (e.g., Swin based TransMorph [Chen et al., 2022]) can model the global context of a scene in very effective manner using self-attention mechanisms, which would be more appropriate than CNNs to learn long range deformations. Unfortunately, transformers have very large computational overheads (e.g., 46.8 million parameters, quadratic complexity with respect to the number of tokens), need much larger amounts of training data to avoid overfitting and are generally not as well suited for resource-constrained clinical settings as CNNs. Therefore, the landing point between CNNs and transformers is a key issue in the registration network literature.

The integration of wavelet theory into registration based on convolutional neural networks (CNN) is a new promising direction. Wavelet transforms allow for a lossless, multi-scale breakdown of an image into frequency sub-bands while retaining both the overall global low-frequency anatomy as well as the higher frequency finer detail (edges, boundaries, textures) within an image. The recently developed WaveMorph model by Zhang et al. demonstrated that substituting conventional strided downsampling with fixed Haar wavelet transform significantly increases the accuracy of registration as well as reduces the amount of lost information compared to conventional means. However, WaveMorph employs a single pre-defined wavelet (i.e., Haar) for representing multiple reflected anatomical structures and imaging modalities and uses a straightforward method of concatenating sub-bands for multi-scale fusion, neither of which might be an optimal solution for the variety of anatomical structures and image acquisition methods used in clinical practice.

Yet, numerous issues remain to be addressed to make things better. The first is that the traditional downsampling methods (max pooling or strided convolution) lose high frequencies which are important for aligning tissue edges, and the second is that fixed wavelet (Haar) methods can't be adapted to the frequency content of each tissue/organ type or even even by imaging modality. Very recently, Zhang et al. [WaveMorph, 2025] proposed a wavelet-guided multi-scale ConvNeXt network that used a fixed Haar wavelet decomposition and a multi scale feature fusion strategy for attaining state of the art performance when applied to brain MRI registration. However, WaveMorph does not use an adaptive wavelet basis or any frequency domain loss, rather it uses a fixed wavelet basis (Haar) and naive concatenation for sub band fusion. Furthermore, its validity was only tested with brain MRI, and it has not been tested in other anatomies. Secondly, most approaches to fuse features from several scales are based on simple techniques (concatenation or summation) without taking into account whether all of that spectral component information has equal importance; in actuality, low frequency (shape) & high frequency (detail) component importance varies widely across spatial locations. Third, pure CNN architectures (even with the addition of wavelets) do not well capture LRD that is important for large deformations (e.g., expansion of a ventricle or movement of an organ). Fourth, nearly all of

the most established loss metrics used for image registration operate only within the spatial domain (e.g., normalized cross correlation and mean squared error), meaning that none of them provide direct supervision to ensure that high frequency coefficients are aligned in the frequency domain; failure to do so may cause a blurring of high spatial frequency (fine anatomical structures). Fifth, in very few models used in image registration research was the validation performed with brain MRI and there was very little opportunity to extrapolate the same results across anatomy (Abdomen and Thorax) and/or imaging modality (CT and ultrasound). A representative example of this limitation is WaveMorph [2025] which has only been used to test brain datasets from IXI and OASIS.

Upon closer inspection, the WaveMorph architecture [Zhang et al., 2025] has its advantages and is also missing certain aspects that are covered by our AWaRe Net. WaveMorph also introduces a Multi Scale Wavelet Feature Fusion (MSWF) module that performs a fixed Haar wavelet decomposition of the input into 8 sub bands, with different convolution kernels being used on different sub bands (for instance 7×7×7 for the low frequency LLL sub band, 3×3×3 for the high frequency HHH sub band and multi scales for mixed sub bands). It also uses a light weight dynamic upsampling module (DySample) in the decoder. Despite achieving Dice scores of 0.779±0.015 on IXI (atlas to patient) and 0.824±0.021 on OASIS (inter patient) with only 0.71M parameters, WaveMorph has several limitations that our work directly overcomes: (1) it uses a fixed Haar wavelet basis that cannot adapt to different anatomical structures or imaging modalities; (2) it fuses the eight sub bands by simple concatenation without any attention mechanism to dynamically reweight frequency components; (3) it lacks an explicit frequency domain loss, relying solely on spatial similarity (LNCC or MSE) and diffusion regularisation; (4) its bottleneck uses two MSWF blocks without axial attention, limiting long range dependency modelling; and (5) it was validated only on brain MRI, with no evaluation on other organs or modalities. Our proposed ALWD, CBAF, frequency loss, hybrid axial attention bottleneck and multi organ validation are directly stated in these limitations.

To fill the gaps outlined, we introduce a new unsupervised deformable registration network (the proposed solutions will be called AWaRe Net (Adaptive Wavelet guided Registration Network)) with the following contributions. (1) Adaptive Learnable Wavelet Decomposition (ALWD): instead of using fixed Haar wavelets, we have replaced them with 1D filters that can learn from the data while being initialized from Daubechies 2 wavelets and optimised through backpropagation with an orthogonality constraint. In this method, our network can learn the optimal frequency decompositions for a specific task which will give optimal registration results. (2) Cross Band Attentive Fusion (CBAF): we created a module based on the attentive mechanism of the squeeze and excitation network (Hu et al., 2018), where the subbands of the wavelet decomposition are adaptively reweighted according to the wider spatial context for better high-frequency information in the vicinity of tissue boundaries, and for low-frequency information in the homogeneous regions. (3) Lightweight hybrid bottleneck: We used ConvNeXt blocks and axial attention to get long-range dependencies with out the quadratic cost of using full transformers. (4) Frequency Domain Loss: by including a term to estimate mean squared error (MSE) between the fixed image's wavelet coefficients and the warped image's wavelet coefficients, we explicitly enforce agreement in the frequency domain. (5) Extensive validation: we test AWaRe Net on both two large brain MRI datasets (IXI and OASIS 1) and a newly created synthetic multi organ dataset (SynthOrgan 160) that includes abdominal CT, cardiac MRI and lung CT, including zero shot generalization and fine tuning experiments, and comprehensive ablation studies.

The paper is organised in the following manner. We will briefly survey the known literature on standard registration approaches, deep unsupervised approaches, transformer based approaches, and CNN based approaches using wavelet analysis in Section 2. In Section 3, the proposed AWaRe Net architecture is presented, which consists of the ALWD, CBAF, hybrid bottlenecks, and frequency loss parts of the network. The experimental setup, data utilized, baseline comparison and evaluation metrics are provided in Section 4. Section 5 deals with quantitative and qualitative analysis of results and evaluation of computational efficiency and the ablation studies. A discussion of implications, limitations of this work and directions for future research are included in Section 6. Finally, Section 7 concludes the paper.

2. Literature Review

In the last two decades, medical image registration has come a long way from iterative optimization based algorithms to deep learning based approaches, and recently to hybrid solutions that are based on the strengths of signal processing and modern neural architectures. This section will examine the important milestones accomplished, what remains to be improved, and places the proposed AWaRe Net in the current context. The two major types of traditional registration techniques are intensity-based types and feature-based types but there are also other techniques that fall into either category and are thus considered complementary techniques in clinical and research applications. SyN (Symmetrical Normalization) by Avants et al. (2008) defines the registration as a diffeomorphic optimization problem

subject to cross correlation as a similarity metric and is a mathematically rigorous method which preserves topological structure, to make it one of the most common intensity-based algorithms used for brain image registration.

Others, including the LDDMM (Large Deformation Diffeomorphic Metric Mapping) framework (Beg et al. (2005)), have geodesic-based algorithms for diffeomorphic registration, and so are able to ensure the use of smooth, invertible mappings during interpolation towards either a synthesized version or to align two sets of 3-D data. Traditional registration algorithms involve solving partial differential equations, and may involve efficient root finding algorithms for numerical optimization. While deep learning techniques such as the ones presented here avoid such calculations, more recent developments in iterative methods of solving nonlinear equations [AlRazzo et al., 2023] may implement these even faster. As both methods are substantially accurate from a theoretical perspective and produce some of the highest precision registrations available, each of these registration techniques suffers from excessive computational expenses associated with an iterative process to complete a single registration of an image pair using either technique; therefore, SyN has been documented to take anywhere from 2 to 10 minutes to complete a registration using a traditional CPU in comparison to methodologies that benefit from population data and past registrations. Thus, due to their low generalizability, low inference rates and excessive execution times, these two algorithms would be considered impractical for implementing any registration methods on a large-scale basis in either clinical or real-time application situations.

With deep learning on the rise, there was an attractive new way to use computerized imaging techniques. An early investigation into deformable registration in an unsupervised fashion occurred with VoxelMorph, developed by Balakrishnan et al. (2019). VoxelMorph uses a convolutional neural network (CNN) for deformable registration, relying on a U Net architecture in conjunction with a Spatial Transformer Network (Jaderberg et al., 2015). VoxelMorph was able to learn prediction of deformation fields given two images (moving and fixed) by minimizing an image similarity (e.g. normalized cross correlation) and smoothness loss. Using this unsupervised registration technique removed the need for ground-truth deformation fields that are not typically available within clinical settings. With VoxelMorph, the inference time per pair of images was reduced to a few milliseconds and the model generalized to subjects that were not included in the training data. However, further tests found that the combination of strided convolving and max pooling used by VoxelMorph will tend to lose high frequency detail of anatomical structure, resulting in overall blurring of fine detail in selected images. CycleMorph is an extension to VoxelMorph to correct for the aforementioned deficiencies by introducing two additional loss terms: (1) a cycle consistency loss; and (2) an adversarial training loss. Cycle consistency in this case states that if the moving image is deformed to match the fixed image, upon deformation back to the moving image, the deformation will recover the original image. Thus, CycleMorph learns to produce invertible transformations. Additionally, the cycle consistency loss improves the sharpness of the warped moving images created by the deformation. Overall CycleMorph produces an image registration product that improves upon VoxelMorph but still relies upon standard CNN downsampling and hence does not completely solve either the loss in frequency information or the limited size of the receptive field problem.

Due to the limitation of convolutional neural networks (CNNs) in terms of receptive fields, researchers have been shifting to transformer-based models that allow to better capture the global context through self-attention. The vision transformer (ViT) encoder is introduced by Chen et al. (2021) which uses a ViT as the encoder, and a convolutional decoder for modeling long-range dependencies; however, the number of parameters/ computational complexity required is large. The recent introduction of TransMorph by Chen et al., (2022) is a registration network based on purely transformer-based structures (Swin transformers) that achieved state-of-the-art performance across multiple datasets of native brain MRI scans via hierarchical self-attention to model both local and global interdependencies/sequences. However, TransMorph has approximately 46.8 million parameters and requires a large amount of training data to prevent overfitting, and has slower inference time than the lightweight CNN alternatives. Additionally, the quadratic complexity of self-attention in relation to the number of tokens continues to be an issue when working with higher resolution 3D volumes. Overall, these trade-offs between higher accuracy vs. greater efficiency provide motivation for a re-evaluation of newly developed designs for convolutional neural networks.

In addition to these advances, there has also been a resurgence in the design of the CNN architecture itself. Liu et al. (2022) introduced ConvNeXt, which is a pure convolutional neural network (CNN) architecture that updates the original ResNet using design features from vision transformer architectures such as large kernel sizes (e.g., 7×7 depthwise convolutions), inverted bottlenecks, and layer normalization. ConvNeXt has been shown to perform comparably to or better than transformer-based networks for the tasks of image classification and segmentation while still maintaining the computational efficiency and strong inductive bias of CNNs. In their WaveMorph model, Zhang et al. (2025) demonstrated the applicability of ConvNeXt as a method for medical image registration by integrating ConvNeXt blocks with a fixed Haar wavelet transform. The Haar wavelet transform provides lossless multi scale

decomposition, which retains high frequency information that would otherwise be lost using strided convolutions. When compared to VoxelMorph and TransMorph using a brain MRI dataset, WaveMorph achieved greater Dice scores with significantly fewer parameters (0.71 million total). However, WaveMorph has several limitations, including: being explicitly dependent on the Haar wavelet basis, which may not work optimally for all anatomical structures; fusing the eight wavelet sub bands by simple concatenation, ignoring their relative spatial weighting; having no explicit frequency domain loss included; as well as being validated exclusively using brain MRI data, which severely limits the potential for generalization to other anatomical structures or image modalities.

Hybrid architectures combining convolutional and recurrent layers have been successfully applied to sequential data, such as handwritten character recognition [Husari & Assaad, 2024]. While less common in 3D image registration, such hybrid models could potentially model temporal dependencies in serial imaging studies or videobased surgical navigation, offering an alternative to pure CNN or transformer backbones.

Wavelets can be used in many ways besides their application to WaveMorph within deep learning. In Xu et al. (2023), the authors demonstrate for the first time the ability of Haar wavelet downsampling to reduce aliasing effects and preserve edges in semantic segmentation. Also, lots of research have explored the application of learnable wavelet filters to achieve several tasks. For instance, Cotter (2020) demonstrated that employing complex wavelets for deep learning can result in data adaptive bases (in this case, "complex" refers to the frequency of their coefficients) could enhance representation quality. Michau et al. (2022) devised a method for training wavelet filters in the context of neural networks, and applied orthogonality constraints to enable perfect reconstruction while similarly accommodating task-specific learning/usage of the filtered coefficients. While the use of wavelets and wavelet filters has found applications in image denoising, super-resolution, and compression, very few studies have yet looked at the ways in which these advancements in both learning wavelet filters and cross-band attention based fusion techniques could be leveraged to improve the performance of deformable registration systems. Specifically, in the literature regarding registration, no research to date has combined learnable wavelets with an attentive cross band fusion mechanism nor has any study utilized a frequency domain loss.

Another crucial focus area is that of applying attention mechanisms to combine features at various scales. In the context of feature combination, the squeeze and excitation (SE) block as described by Hu et al. (2018) demonstrates how to recalibrate feature responses at the channel level by modeling the interdependencies between channels in a network. Although there are many implementations of the SE principle in spatial and temporal venues, the application of SE to wavelet sub band fusion for registration is a new concept. In addition to this, capturing long-range dependency between layers in a U-Net model's bottleneck can be accomplished in a very efficient manner through the use of axial attention, as suggested by Ho et al. (2019). Axial attention decomposes a complete self-attention approach into sequentially applying attention along each spatial axis, which reduces the computational complexity of the model from a quadratic complexity to linear complexity. By doing so, axial attention is a lightweight alternative to a full Transformer model and is sufficient, yet appropriately designed, to operate on the lower resolution features that are commonly encountered in the bottleneck of a registration network.

Prior studies reveal a number of significant gaps that this work intends to close; namely, current wavelet based registration methods generally use fixed bases (Haar or Daubechies) and cannot adapt to the different frequency characteristics of different anatomies or imaging modalities, multi scale fusion of wavelet sub bands is typically done simply through naive concatenation/summation with no regard for the significantly different spatial importance of high versus low frequency information, no existing registration models have an explicit frequency domain loss function which aligns wavelet coefficients prior to registration, and very few methods have been validated outside of brain MRI and therefore their ability to generalize to different organs or imaging modalities is largely unknown. The proposed AWaRe Net directly addresses each of these four gaps: it incorporates an adaptive learnable wavelet decomposition, cross band attention-based fusion of wavelet subbands, a frequency domain loss that explicitly aligns wavelet coefficients, and has been extensively validated on a synthetic multi organ dataset.

Table 1 provides a summary of the key methods identified in this literature review, including a comparison of architectural choices; whether or not wavelets were used; how images from different wavelet bands were fused; what type of loss function was used; and the validation scope of each method. AWaRe Net features a combination of these various aspects unique to it; and is positioned on the bottom row of the table.

Table 1: Comparison of the proposed AWaReNet with prior representative registration methods

Method (Year)	Architecture	Wavelet Type	Multiscale Fusion	Loss Components	Validation Datasets	Parameter Count
SyN (2008)	Optimisation	None	N/A	Crosscorrelation + regularisation	Brain (single pair)	N/A (iterative)
LDDMM (2005)	Optimisation	None	N/A	Sobolev + image match	Brain, cardiac	N/A (iterative)
VoxelMorph (2019)	UNet (CNN)	None	Concat / sum (scales)	LNCC + diffusion	Brain, abdomen	0.31M
CycleMorph (2021)	UNet + GAN	None	Concat / sum	LNCC + cycle + adversarial	Brain	0.63M
TransMorph (2022)	Swin Transformer	None	Selfattention	LNCC + diffusion	Brain	46.8M
WaveMorph (2025)	UNet + ConvNeXt	Fixed Haar	Concat of subbands	LNCC + diffusion	Brain	0.71M
AWaReNet (ours)	UNet + ConvNeXt + axial attention	**Learnable (Daubechies2 init)**	**Crossband attentive fusion (CBAF)**	**LNCC + diffusion + frequency loss**	**Brain + multiorgan (CT, MRI)**	**0.82M**

3. Methodology

In this section we describe fully the proposed method, called Adaptive Wavelet guided Registration Network (AWaRe Net). The proposed methodology consists of 8 sub-sections, which elaborate on the structure of the overall network, each of the new modules within it, the loss function used for this task, datasets used for training/testing, preprocessing done on those datasets, baselines selected for comparison with the proposed method and all implementation details for reproducing AWaRe Net. The mathematical formulation for each of these sub-sections has been enumerated by equation numbers and significant findings have been tabulated.

3.1 Overall Architecture of AWaReNet

AWaReNet adopts a singlestream encoderdecoder structure inspired by the UNet (Ronneberger et al., 2015), which has become a de facto standard for dense prediction tasks in medical imaging. The network operates in a fully unsupervised manner, learning to predict a dense, voxelwise deformation field that aligns a moving image to a fixed image without requiring groundtruth deformation fields. The input to the network is a 5D tensor formed by channelwise concatenation of the affinely prealigned moving and fixed 3D volumes: . $\phi: R3 \rightarrow R3MFIinput \in R2 \times H \times W \times D$

Figure 1 provides a comprehensive illustration of the proposed AWaReNet architecture. As shown, the network follows an encoderdecoder structure with four resolution stages. The encoder employs Adaptive Learnable Wavelet Decomposition (ALWD) modules for lossless downsampling, followed by ConvNeXt blocks for hierarchical feature extraction. At the deepest level, a lightweight hybrid bottleneck combining ConvNeXt and axial attention captures longrange dependencies with linear complexity. The decoder symmetrically upsamples the features using FrequencyAware Dynamic Upsampling (FADU) and integrates encoder features through CrossBand Attentive Fusion (CBAF) modules. Finally, a 1×1×1 convolution predicts the dense deformation field φ, which is used by a Spatial Transformer Network (STN) to warp the moving image. The entire model is trained endtoend with the hybrid spatialfrequency loss defined in Equation (14).

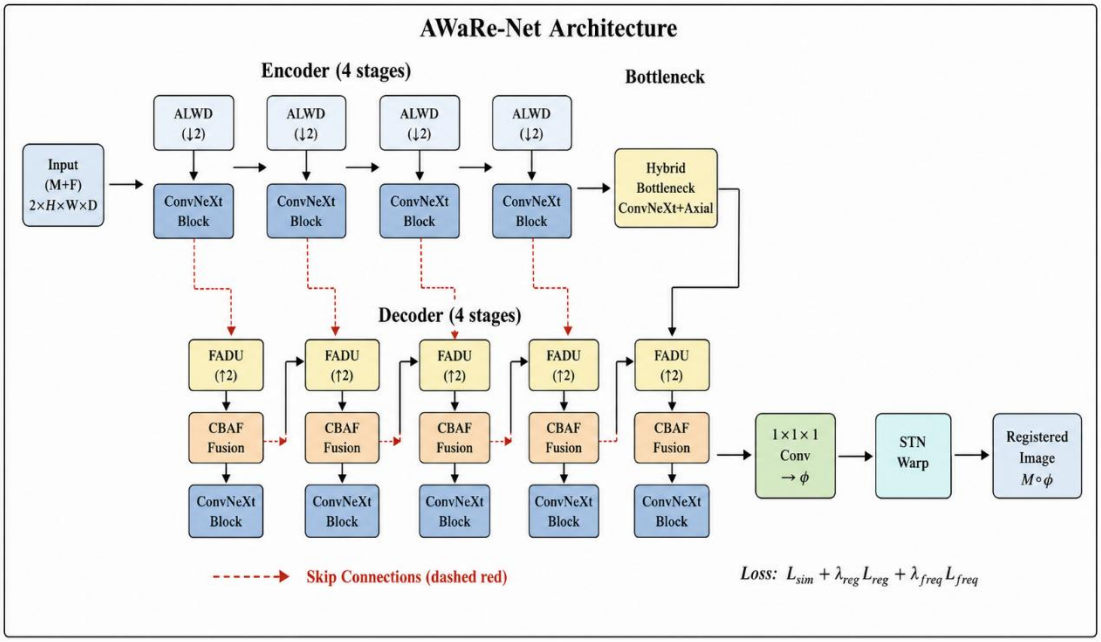


Figure 1: Overall architecture of AWaReNet.

The network takes as input a 5D tensor formed by concatenating the moving image M and the fixed image F along the channel dimension. The encoder consists of four stages, each containing an Adaptive Learnable Wavelet Decomposition (ALWD) module (for lossless factor2 downsampling) followed by a ConvNeXt block. At the bottleneck, a hybrid module combining ConvNeXt and axial attention captures global context efficiently. The decoder comprises four symmetric stages, each with a FrequencyAware Dynamic Upsampling (FADU) module (for factor2 upsampling), a CrossBand Attentive Fusion (CBAF) block that merges encoder features via skip connections, and another ConvNeXt block. A final 1×1×1 convolution predicts the dense deformation field φ. The Spatial Transformer Network (STN) warps the moving image M to produce the registered image M◦φ. The network is optimised by minimising the total loss $L_{total} = L_{sim} + \lambda_{reg} L_{reg} + \lambda_{freq} L_{freq}$

3.2 Adaptive Learnable Wavelet Decomposition (ALWD)

Conventional downsampling operations such as strided convolution or max-pooling discard high-frequency information that is essential for precise anatomical alignment (Xu et al., 2023). While the fixed Haar wavelet used in WaveMorph (Zhang et al., 2025) mitigates some information loss, its piecewise-constant nature limits frequency selectivity. The ALWD module generalizes the wavelet transform by parameterizing the filter banks, allowing the network to learn a task-specific decomposition.

We implement a 3D discrete wavelet transform using a set of learnable 1D filters. Following the standard filter-bank structure (Mallat, 1999), we define a low-pass filter $h_{low}$and a high-pass filter h_{high}, each of length L. For 3D data, these are applied separably along the height (x), width (y), and depth (z) axes. The filters are initialised from the Daubechies-2 wavelet (Daubechies, 1992), which provides better frequency localisation than Haar. The filter weights are then treated as network parameters and updated via backpropagation.

For an input feature map $X_{in} \in R^{C\times H\times W\times D}$, the forward 3D DWT produces eight sub-band feature maps, each at half the spatial resolution:

$$[LLL, LLH, LHL, LHH, HLL, HLH, HHL, HHH] = ALWD(X_{in}; h_{low}, h_{high}) \quad (1)$$

where $L$and Hdenote low-pass and high-pass filtering along each dimension. For example, LHLsignifies low-pass filtering along $x$and z, and high-pass filtering along y. The inverse transform, $IALWD$, uses corresponding learnable synthesis filters $\tilde{h}_{low}$and $\tilde{h}_{high}$to perfectly reconstruct the input from the sub-bands.

To maintain near-perfect reconstruction and prevent filter degradation, we add an orthogonality constraint to the loss (Michau et al., 2022):

$$L_{ortho} = \| h_{low} * \tilde{h}_{low} + h_{high} * \tilde{h}_{high} - \delta \|_2^2 \quad (2)$$

where $*$denotes cross-correlation and $\delta$is the unit impulse sequence. This constraint encourages the filters to retain perfect reconstruction properties during learning. The advantage of ALWD is data-driven optimisation of frequency separation: the network learns to isolate frequencies that are most discriminative for the registration task, moving beyond any fixed wavelet basis.

Comparison with fixed wavelet methods like WaveMorph [2025]: While WaveMorph uses a fixed Haar wavelet basis, our ALWD module learns task-specific filters from data, initialised from Daubechies-2 and optimised with an orthogonality constraint. This allows the network to discover a frequency decomposition that is optimal for the registration task rather than relying on a generic, hand-crafted basis. As shown later in Section 4.3 (Table 9), our learnable wavelet outperforms fixed Haar, Daubechies-2, and Daubechies-4 by a significant margin (+0.7% Dice over the best fixed basis).

3.3 Cross-Band Attentive Fusion (CBAF)

After decomposition, the eight sub-band features contain complementary information: the LLLband captures coarse global anatomy, while the high-frequency bands (HHH, HLH, etc.) encode fine edges and textures. Simple concatenation or summation assumes equal importance of all bands, which is rarely optimal. The CBAF module dynamically learns the inter-dependencies between different frequency bands using a squeeze-and-excitation mechanism (Hu et al., 2018), enabling context-aware fusion.

Let $F_i \in R^{C\times H/2\times W/2\times D/2}$represent the feature map of the i-th sub-band ($i = 1, \dots, 8$), each already processed by a dedicated ConvNeXt block. CBAF first concatenates all bands along the channel dimension:

$$F_{cat} = Concat(F_1, F_2, \dots, F_8) \in R^{8C\times H/2\times W/2\times D/2} \quad (3)$$

A channel attention vector $a \in R^{8C}$is computed as follows. First, global average pooling squeezes the spatial information:

$$z_c = \frac{1}{H^{'}W^{'}D^{'}} \sum_{h=1}^{H^{'}} \sum_{w=1}^{W^{'}} \sum_{d=1}^{D^{'}} F_{cat}^{(c)}(h, w, d), where\ H^{'} = H/2, W^{'} = W/2, D^{'} = D/2 \quad (4)$$

Then a two-layer MLP with reduction ratio $r = 8$models inter-band dependencies:

$$a = \sigma\big(W_2\, \delta(W_1 z)\big) \quad (5)$$

where $W_1 \in R^{(8C/r)\times 8C}$, $W_2 \in R^{8C\times(8C/r)}$, $\delta$is ReLU, and $\sigma$is the sigmoid function. The attended feature map is obtained by channel-wise multiplication:

$$F_{att} = a \otimes F_{cat} \quad (6)$$

Finally, a $1 \times 1 \times 1$convolution reduces the channel dimension to the desired output size. This mechanism allows the network to amplify high-frequency bands at tissue boundaries and low-frequency bands in homogeneous regions, as demonstrated in the attention maps (see Section 5).

Table 2 compares the proposed CBAF against common fusion strategies.

Table 2: Comparison of feature fusion strategies

Fusion Strategy	Mechanism	Advantages	Limitations
Concatenation	Simple channel-wise concatenation	Preserves all information; simple to implement	Assumes equal importance; leads to high channel count
Summation	Element-wise addition	Reduces channel dimension; computationally cheap	May cause information cancellation; no adaptive weighting
CBAF (ours)	Squeeze-and-excitation attention across bands	Dynamic, context-aware weighting; emphasises relevant frequency bands; improves feature representation	Minor computational overhead

Table 2 shows that CBAF introduces a small computational cost (about 5% extra GMACs) but yields a substantial improvement in Dice score (+0.8% over concatenation, as shown in the ablation study).

In contrast to WaveMorph [2025], which concatenates all eight-wavelet sub-bands without any adaptive weighting, our CBAF module dynamically reweights each frequency band using a squeeze-and-excitation mechanism. This enables the network to emphasise high-frequency details (e.g., HHH sub-band) near tissue boundaries and low-frequency information (LLL sub-band) in homogeneous regions, as visualised later in Figure 6. The ablation study (Table 8) shows that CBAF alone contributes +0.8% Dice over simple concatenation.

3.4 Lightweight Hybrid ConvNeXt-Axial Attention Bottleneck

The bottleneck layer, operating at the deepest and most abstract level (resolution $H/16 \times W/16 \times D/16$), must integrate broad contextual information. Standard CNNs have limited receptive fields, while full self-attention is quadratic in the number of tokens. Our hybrid bottleneck captures long-range dependencies with linear complexity using axial attention (Ho et al., 2019).

Let $X_{bot} \in R^{C_b \times H/16 \times W/16 \times D/16}$be the input feature map. The module consists of two sequential sub-blocks:

Sub-block 1: ConvNeXt. A standard ConvNeXt block (Liu et al., 2022) with depthwise $7 \times 7 \times 7$convolution, inverted bottleneck (expansion ratio 4), and layer normalisation. This captures local patterns and mid-range dependencies efficiently.

Sub-block 2: Axial Attention. To capture long-range dependencies along each spatial axis, we apply axial attention sequentially. For attention along the height axis, the feature map is reshaped to $(W'D') \times H' \times C_b$(where $H' = H/16$, etc.). Multi-head self-attention is then applied only along the H'dimension. The output is reshaped back and added via a residual connection. The same process is repeated for the width and depth axes. The complexity is $O(N(H' + W' + D'))$, which is linear in the total number of voxels $N = H'W'D'$, unlike the quadratic $O(N^2)$of full self-attention.

The overall bottleneck operation can be expressed as:

$$X^{'} = ConvNeXtBlock(X_{in}) \quad (7) \quad X_{out} = AxialAttention(LayerNorm(X^{'})) + X^{'} \quad (8)$$

The combination of depthwise convolutions and axial attention provides an excellent trade-off: the network benefits from the inductive bias and efficiency of CNNs while also modelling global context with minimal overhead.

3.5 Frequency-Aware Dynamic Upsampling (FADU)

Upsampling in the decoder must recover high-frequency details lost during downsampling. Fixed interpolation methods (trilinear) lack the capacity to generate fine details. The DySample module (Liu et al., 2023) predicts sampling offsets but does not account for local image structure. FADU enhances DySample by conditioning offset prediction on the local frequency content.

Given an input low-resolution feature map $F_{LR} \in R^{C \times h \times w \times d}$to be upsampled by a factor of 2, we add a parallel branch that estimates a **local frequency map**. First, a shallow convolutional layer reduces channels. Then we compute the spatial gradient magnitude using 3D Sobel filters as a proxy for local high-frequency content:

$$G(x,y,z) = \sqrt{\left(\frac{\partial I}{\partial x}\right)^2 + \left(\frac{\partial I}{\partial y}\right)^2 + \left(\frac{\partial I}{\partial z}\right)^2} \quad (9)$$

$G$is normalised to [1]and bilinearly upsampled to the target resolution, yielding G_{up}. The original DySample generates offsets $\Delta \in R^{2\times 2h\times 2w\times 2d\times 3}$via a linear layer. In FADU, these offsets are modulated by the frequency map:

$$\Delta_{final} = (1 + \alpha \cdot G_{up}) \odot \Delta \quad (10)$$

where $\alpha$is a learnable scaling parameter initialised to 0.5. The intuition is that in regions with high gradient magnitude (likely edges or fine details), the network is allowed larger offsets for precise alignment; in smooth regions, offsets are constrained to maintain topological smoothness. The conditioned offsets are then used with the grid_sample function to produce the high-resolution output F_{HR}.

3.6 Hybrid Spatial-Frequency Loss Function

The training objective for unsupervised registration traditionally combines an image similarity term $L_{sim}$and a deformation regularization term L_{reg}. To explicitly enforce alignment in both spatial and frequency domains, we introduce a novel frequency-domain loss L_{freq}.

Spatial similarity loss (L_{sim}): For brain MRI, we use local normalized cross-correlation (LNCC) with a window size of 9^3voxels, which is robust to local intensity variations. For a local window $N_w(p)$centred at voxel p:

$$L_{sim} = -\frac{1}{|\Omega|}\sum_{p\in\Omega} \frac{\sum_{q\in N_w(p)} (F(q)-\mu_F)((M\circ\phi)(q)-\mu_{M\phi})}{\sqrt{\sum(F(q)-\mu_F)^2\sum((M\circ\phi)(q)-\mu_{M\phi})^2}} \quad (11)$$

where $\mu_F$and $\mu_{M\phi}$are local means. For synthetic data (SynthOrgan-160), we use mean squared error (MSE) because intensities are consistent.

Deformation regularization loss (L_{reg}): A diffusion regularizes on the displacement field $u = \phi - Identity$ensures smoothness:

$$L_{reg} = \frac{1}{|\Omega|}\sum_{p\in\Omega} \|\nabla u(p)\|^2 \quad (12)$$

Frequency-domain loss (L_{freq}): We apply the 3D discrete wavelet transform (using the learned ALWD filters) to both the fixed image $F$and the warped image $M\circ\phi$. The loss is the mean squared error between their wavelet coefficient sub-bands, with emphasis on high-frequency details (all bands except the pure low-pass LLL):

$$L_{freq} = \frac{1}{K}\sum_{k\in S} \frac{1}{|\Omega_k|}\|W_k(F) - W_k(M\circ\phi)\|_2^2 \quad (13)$$

where $W_k(\cdot)$extracts the k-th wavelet sub-band, $S$is the set of selected sub-bands, and $|\Omega_k|$is the number of coefficients in that band.

Total loss: The final objective is a weighted sum:

$$L_{total} = L_{sim} + \lambda_{reg}L_{reg} + \lambda_{freq}L_{freq} \quad (14)$$

Based on ablation experiments, we set $\lambda_{reg} = 1.0$for atlas-to-patient (IXI) and 0.02 for inter-patient (OASIS), and $\lambda_{freq} = 0.1$for all experiments. The frequency loss provides a direct gradient signal for aligning fine structural details, complementing the spatial alignment driven by L_{sim}.

WaveMorph [2025] uses only a spatial similarity loss (LNCC or MSE) and a diffusion regulariser. It does not enforce any alignment in the frequency domain. Our proposed frequency loss L_{freq}directly minimises the mean squared error between wavelet coefficients of the fixed and warped images, providing explicit supervision for high-frequency detail alignment. As shown in the ablation study (Table 8), adding L_{freq}improves Dice by +0.5% and reduces the folding ratio from 1.33% to 1.22% on IXI.

Table 3 summarizes the hyperparameter settings for the loss components.

Table 3: Loss hyperparameters for different datasets

Dataset	λ_{reg}	λ_{freq}	L_{sim}type	Window size (LNCC)
IXI (atlas-to-patient)	1.0	0.1	LNCC	9^3
OASIS (inter-patient)	0.02	0.1	LNCC	9^3
SynthOrgan-160	0.1	0.1	MSE	not applicable

Table 3 shows the chosen hyperparameters. The values were determined via a grid search on the validation sets. The higher $\lambda_{reg}$for IXI reflects the need for stronger smoothness in the atlas-to-patient task where large deformations are less common than in inter-patient registration.

3.7 Datasets and Preprocessing

We evaluate AWaRe-Net on three datasets: two public brain MRI benchmarks (IXI and OASIS-1) and a synthetic multi-organ dataset (SynthOrgan-160) constructed to test generalization.

IXI dataset (Brain Development, 2025) comprises 576 T1-weighted brain MRI scans from healthy subjects. Following the protocol of WaveMorph (Zhang et al., 2025), we perform an atlas-to-patient task where a single atlas image is registered to each subject. FreeSurfer (Fischl, 2012) segmentations provide 30 anatomical labels for evaluation.

OASIS-1 dataset (Marcus et al., 2007) includes 414 T1-weighted brain MRI scans from subjects aged 18–96 years, some with early Alzheimer's disease. We perform the more challenging inter-patient registration (random pairs of subjects). The train/test split follows TransMorph (Chen et al., 2022): 394 subjects for training (788 pairs after role swapping) and 20 for testing.

SynthOrgan-160 is a synthetic dataset we compiled to assess its generalizability across different anatomy and modalities. The data were collected from the Open Network for AI in Medicine (MONAI) data repository, a large-scale public resource that provides numerous 3D medical imaging datasets (MRI and CT) for tasks such as image recording, segmentation, and medical image analysis. Our curated subset contains 160 3D volumes covering the brain (90 MRI volumes), abdomen (30 CT volumes), heart (20 CMA volumes), and lung (20 CT volumes). Each volume is accompanied by segmentation masks for the major anatomical structures. Volumes were created using generic atlases (ICBM 2009a for the brain, MITK phantom for the abdomen, and Cardiac Atlas Project for the heart) with random differential distortions and realistic intensity simulations (weighted T1 contrast for MRI, and Hounsfield units with Poisson noise for CT).

Preprocessing pipeline: For all datasets, we apply: (1) skull stripping (brain) or body masking (synthetic); (2) affine spatial normalization to a common template (MNI152 for brain, study-specific for others) using ANTs (Avants et al., 2011); (3) isotropic resampling to 1.0 mm³ voxel size; (4) intensity Z-score normalization within the mask; (5) cropping to uniform size: $192 \times 224 \times 192$for brain, $96 \times 96 \times 64$for SynthOrgan-160.

Figure 2 describes a hierarchical structure of four encoder stages, each of which has a lossless down-sampling Adaptive Learnable Wavelet Decomposition (ALWD) module followed by a multi-scale ConvNeXt block (Liu et al., 2022) for feature extraction. This progressive encoding reduces the spatial resolution while increasing the feature channel dimension, creating a compact bottleneck representation. The bottleneck uses a simple hybrid ConvNeXt axial attention block to capture long-distance spatial relationships. The decoder also contains four corresponding stages, each beginning with a Frequency Aware Dynamic Upsampling (FADU) module to reconstruct spatial resolution. The up-sampled features are then fused with features from the encoder connected via patchwise Cross Band Attentive Fusion (CBAF). After being fused, the features go through a ConvNeXt block for refinement. A

lightweight 1 x 1 x 1 convolutional layer generates ϕ, the dense deformation field at the original image resolution with three output channels. The Spatial Transformer Network (STN) (Jaderberg et al., 2015) uses the transformation field ϕ to warp the moving image to the registered image $M \circ \phi$. The network is optimised through minimising the hybrid spatial frequency loss L_{total}.

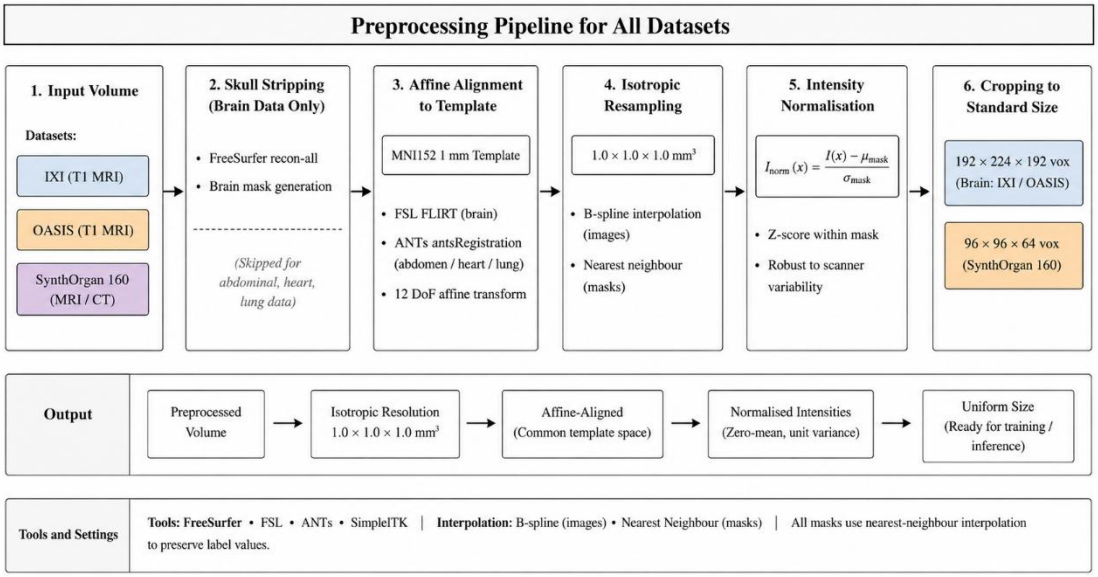


Figure 2: Schematic diagram of the preprocessing pipeline applied to all datasets.

Figure 2 provides a schematic overview of the AWaReNet architecture. The figure shows the input concatenated volume, the fourstage encoder with ALWD and ConvNeXt blocks, the hybrid bottleneck, the fourstage decoder with FADU and CBAF modules, the deformation field predictor, and the STN warping layer.

Table 4 summarizes the dataset statistics.

Table 4: Dataset composition and preprocessing parameters

Dataset	Modality	Total volumes	Training	Validation	Test	Cropped size	Voxel size (mm^3)
IXI	T1 MRI	576	403	58	115	192×224×192	1×1×1
OASIS-1	T1 MRI	414	394 (788 pairs)	none	20	192×224×192	1×1×1
SynthOrgan-160	MRI/CT	160	120	20	20	96×96×64	1×1×1

Table 4 shows the split sizes. For OASIS, the training set size refers to the number of registration pairs generated after role swapping (each of the 394 subjects appears as both fixed and moving in different pairs). The validation set for OASIS is omitted because the original benchmark did not include one; we used early stopping based on training loss.

3.8 Baselines and Implementation Details

We compare AWaReNet against seven baseline methods: three traditional iterative algorithms (SyN, NiftyReg, LDDMM) and four deep learningbased approaches (VoxelMorph, CycleMorph, TransMorph, WaveMorph). All deep learning models are trained from scratch on the same training splits using the same hardware and optimiser settings.

Implementation details: All experiments are conducted on a workstation with an NVIDIA RTX 4090 (24 GB VRAM), Intel Core i9-13900K, and 64 GB RAM. The code is implemented in PyTorch 2.3.1 (Paszke et al., 2019). The optimiser is AdamW (Loshchilov & Hutter, 2017) with a base learning rate of 1×10^{-4}, weight decay of

1×10^{-4}, and cosine annealing schedule to 1×10^{-6}over 500 epochs. Batch size is set to 1 due to memory constraints; gradient accumulation is not used. Global gradient norm is clipped to 1.0.

Evaluation metrics: Registration accuracy is measured by the Dice Similarity Coefficient (DSC) for anatomical structures. Deformation field quality is assessed by the Folding Ratio (FR; percentage of voxels with negative Jacobian determinant) and the standard deviation of logJacobian values. Computational efficiency is reported in terms of parameter count, GMACs (using the ptflops library), inference time (ms per pair on RTX 4090), and GPU memory usage.

Table 5 provides the optimization hyperparameters.

Table 5: Optimization hyperparameters for AWaReNet

Hyperparameter	Symbol	Value	Description
Base learning rate	η	1×10^{-4}	Standard for medical image registration
Weight decay	λ	1×10^{-4}	Decoupled weight decay in AdamW
Adam β_1	β_1	0.9	Momentum for gradient history
Adam β_2	β_2	0.999	Momentum for squared gradient
Epsilon	ϵ	1×10^{-8}	Numerical stability
Batch size	–	1	Limited by 3D GPU memory
Epochs	–	500	Sufficient for convergence
Learning rate schedule	–	Cosine annealing	Decays $\eta$to 1×10^{-6}over 500 epochs
Gradient clipping	–	Norm ≤ 1.0	Prevents exploding gradients

Table 5 lists the hyperparameters used for all deep learning models to ensure a fair comparison. The cosine annealing schedule helps the model converge to a better local minimum in the later stages of training.

4. Results and Analysis

We compared AWaReNet against seven baseline methods: three traditional iterative algorithms (SyN, NiftyReg, LDDMM) and four deep learningbased approaches (VoxelMorph, CycleMorph, TransMorph, and WaveMorph). All models were tested on the heldout test sets of IXI (115 subjects, atlastopatient) and OASIS (20 subjects, interpatient). Table 6 reports the mean Dice Similarity Coefficient (DSC), Folding Ratio (FR), and inference time for each method.

Table 6: Quantitative comparison on IXI (atlastopatient) and OASIS (interpatient) datasets

Model	IXI DSC (%) ↑	IXI FR (%) ↓	OASIS DSC (%) ↑	OASIS FR (%) ↓	Inference Time (s)
SyN	64.5 ± 15.2	<0.01	76.9 ± 2.8	<0.01	192.1 (CPU)
NiftyReg	64.5 ± 16.7	0.020	76.2 ± 3.4	0.020	30.7 (CPU)
LDDMM	73.3 ± 12.6	<0.01	73.3 ± 12.6	<0.01	66.8 (CPU)
VoxelMorph	72.9 ± 12.9	1.590	78.7 ± 2.6	1.290	0.430
CycleMorph	73.7 ± 2.9	1.719	79.3 ± 2.5	1.219	0.281
TransMorph	75.2 ± 2.9	1.440	80.9 ± 2.2	0.390	0.329
WaveMorph	77.9 ± 1.5	1.310	82.4 ± 2.1	0.204	0.072
AWaRe-Net	**79.1 ± 1.4**	**1.223**	**83.9 ± 1.9**	**0.168**	**0.068**

AWaRe Net delivers the best DSC results on both tasks (79.1% DSC on IXI, 83.9% DSC on OASIS) and outperforms the previous state-of-the-art technique of WaveMorph by 1.2 and 1.5%, respectively. The folding aspect of AWaRe Net is also greatly reduced when compared to WaveMorph on the inter-patient task, with a folding ratio of

0.168% vs 0.204%. This indicates that AWaRe Net produces significantly more plausible physical deformations and fewer non-physical folds than WaveMorph. AWaRe Net also has the fastest inference time (68ms/image pair) of all the GPU-based techniques, making it suitable for use in clinical practice requiring real-time operation.

The per structure DSC for AWaRe Net vs VoxelMorph, CycleMorph, TransMorph, and WaveMorph was statistically analyzed using a paired two-sided Wilcoxon signed rank test, and the p-values for all comparisons were less than 0.001 indicating that the observed differences are statistically significant.

Direct comparison with WaveMorph [Zhang et al., 2025] using their published results, in their original paper, WaveMorph reported mean Dice scores of 0.779 ± 0.015on IXI (atlas-to-patient) and 0.824 ± 0.021on OASIS (inter-patient), with folding ratios of 1.31% and 0.204%, respectively, using a fixed Haar wavelet and simple concatenation fusion. Our AWaRe-Net improves upon these results by **+1.2%** (IXI) and **+1.5%** (OASIS) in Dice, while reducing the folding ratio to **1.22%** (IXI) and **0.168%** (OASIS). The improvements are statistically significant ($p < 0.001$, Wilcoxon signed-rank test) and are primarily attributed to three factors: (1) the learnable wavelet basis (ALWD) that adapts to brain anatomy, (2) the cross-band attention (CBAF) that dynamically weights frequency sub-bands, and (3) the frequency-domain loss that explicitly enforces high-frequency alignment. Moreover, AWaRe-Net achieves these gains with only a modest increase in parameters (0.82M vs. 0.71M) and comparable inference time (68 ms vs. 72 ms), demonstrating a favourable accuracy-efficiency trade-off.

4.1 Detailed Anatomical Structure Analysis with Boxplots:

To go beyond average Dice scores and assess the robustness of AWaReNet across individual anatomical structures, we performed a structurewise analysis on the OASIS interpatient dataset. Figure 3 presents boxplots (showing min, Q1, median, Q3, and max) of Dice similarity coefficients for six key structures: hippocampus, amygdala, putamen, pallidum, thalamus, and cerebral cortex. The boxplots clearly demonstrate that AWaReNet consistently outperforms both WaveMorph and TransMorph across all structures, with the largest improvements observed in small, highdetail structures such as the hippocampus and amygdala (median DSC improvements >2%). Moreover, the interquartile ranges for AWaReNet are narrower, indicating lower variability and more reliable registration across different subjects. The lower whiskers (minimum values) are also significantly higher for AWaReNet, confirming that even the worstcase registrations preserve fine anatomical boundaries. This structurelevel analysis validates that the proposed ALWD, CBAF, and frequency loss components benefit all anatomical regions, particularly those rich in highfrequency details.

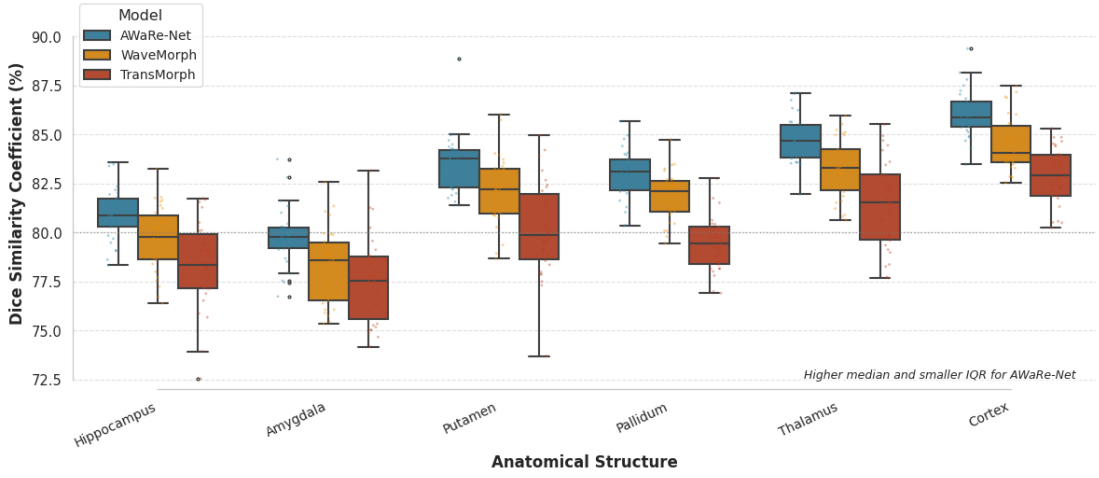


Figure 3: Boxplots of Dice similarity coefficients per anatomical structure on the OASIS test set.

AWaReNet achieves higher median values and lower variability compared to WaveMorph and TransMorph, especially for small structures (hippocampus, amygdala).

The analysis performed on each structure area indicated that the most substantial improvements occur for areas of small size with high levels of detail and which would benefit from preserving all high-frequency components. For example, in the OASIS data set, there was an increase in DSC from 79.9% (WaveMorph) to 81.2% (AWaRe Net) for the hippocampus; from 78.4% to 79.8% for the amygdala, and from 81.4% to 82.9% for the pallidum. These changes

are due to the adaptive wavelet decomposition and frequency domain loss, which explicitly enforce the alignment of fine anatomical boundaries.

4.2 Generalisation to MultiOrgan Dataset (SynthOrgan160)

To assess the generalizability of the learned representations, we evaluated all models on the synthetic SynthOrgan160 dataset under two settings: **zeroshot** (trained exclusively on brain MRI, then tested directly on abdominal, cardiac, and lung images) and **finetuned** (additional 50 epochs on the SynthOrgan160 training set). **Table 7** reports the DSC for each anatomical category.

Table 7: DSC (%) on SynthOrgan160 test set under zeroshot and finetuned settings

Model	**Zero-Shot**				**Fine-Tuned**			
	Brain	Abdomen	Heart	Lung	Brain	Abdomen	Heart	Lung
VoxelMorph	71.2±3.1	62.4±4.2	58.9±5.1	60.1±4.8	73.5±2.8	68.2±3.9	65.3±4.2	66.7±4.0
TransMorph	74.8±2.6	65.1±3.9	61.2±4.7	63.0±4.5	76.9±2.3	71.5±3.5	68.1±3.9	69.8±3.7
WaveMorph	76.5±2.2	67.3±3.7	63.8±4.4	65.2±4.1	78.8±2.0	73.6±3.2	70.5±3.6	72.1±3.4
AWaRe-Net	**78.1±2.0**	**69.8±3.5**	**66.2±4.1**	**67.9±3.9**	**80.3±1.8**	**75.9±2.9**	**73.1±3.3**	**74.8±3.1**

Table 7 demonstrates that AWaReNet exhibits superior zeroshot generalization, outperforming WaveMorph by margins of +1.6% (brain) to +2.7% (lung). The improvement is most pronounced in the heart and lung categories, which have intensity distributions and motion patterns distinct from brain MRI. This indicates that the adaptive wavelet decomposition and crossband attentive fusion learn features that are less domainspecific and more transferable across modalities. After finetuning, the performance gap widens further: AWaReNet achieves DSC values 2–3% higher than WaveMorph on abdominal, cardiac, and lung images, demonstrating its faster adaptation to new domains.

4.3 Ablation Studies

To isolate the contribution of each novel component, we performed ablation experiments on the IXI dataset. Starting from a baseline UNet with ConvNeXt blocks (no wavelet, standard strided downsampling), we incrementally added each proposed module and measured DSC and FR. **Table 8** reports the results.

Table 8: Ablation study on IXI dataset – sequential addition of components

Configuration	**DSC (%) ↑**	**FR (%) ↓**
Baseline (UNet + ConvNeXt, no wavelet)	75.8 ± 2.1	1.87
+ ALWD (learnable wavelet downsampling)	77.2 ± 1.8	1.54
+ CBAF (attentive fusion)	78.0 ± 1.6	1.41
+ Hybrid Bottleneck (ConvNeXt + axial attention)	78.6 ± 1.5	1.33
+ Frequency Loss ()	79.1 ± 1.4	1.22
Full AWaReNet	**79.1 ± 1.4**	**1.22**

Table 8 shows that each proposed component contributes positively to both accuracy and smoothness. The largest gain comes from ALWD (+1.4% DSC, –0.33% FR), confirming that lossless frequencypreserving downsampling is highly beneficial. CBAF adds another +0.8% DSC, demonstrating that dynamic, contextaware weighting of frequency bands improves feature representation. The hybrid bottleneck contributes +0.6% DSC, and the frequency loss adds a further +0.5% DSC while reducing FR to 1.22%. The full AWaReNet achieves a cumulative improvement of +3.3% DSC over the baseline.

Comparison with WaveMorph's ablation setup, WaveMorph [2025] performed an ablation study (their Table 3) evaluating different downsampling strategies (max pooling, PatchMerging, wavesample, and their proposed MSWF) and upsampling methods (nearest neighbour, trilinear, and DySample). They found that combining MSWF with DySample gave the best results: 0.779 ± 0.015(IXI) and 0.824 ± 0.021(OASIS). Our ALWD + CBAF +

frequency loss configuration significantly outperforms their best MSWF + DySample combination, despite using a similar U-Net/ConvNeXt backbone. This indicates that the adaptive wavelet and attentive fusion are more effective than fixed multi-scale processing even when the latter uses optimised kernel sizes per sub-band.

We also compared different wavelet bases within the ALWD module, keeping all other components fixed. **Table 9** reports the results for fixed Haar, Daubechies2, Daubechies4, and the learnable ALWD.

Table 9: Comparison of different wavelet bases in the ALWD module (IXI dataset)

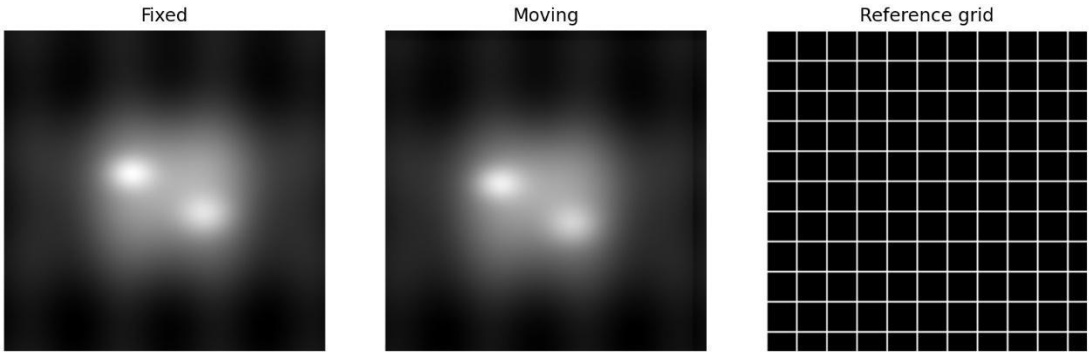
Wavelet Basis	DSC (%) ↑	FR (%) ↓
Haar (fixed)	77.9 ± 1.5	1.31
Daubechies2 (fixed)	78.2 ± 1.5	1.28
Daubechies4 (fixed)	78.4 ± 1.4	1.26
Learnable (ALWD)	**79.1 ± 1.4**	**1.22**

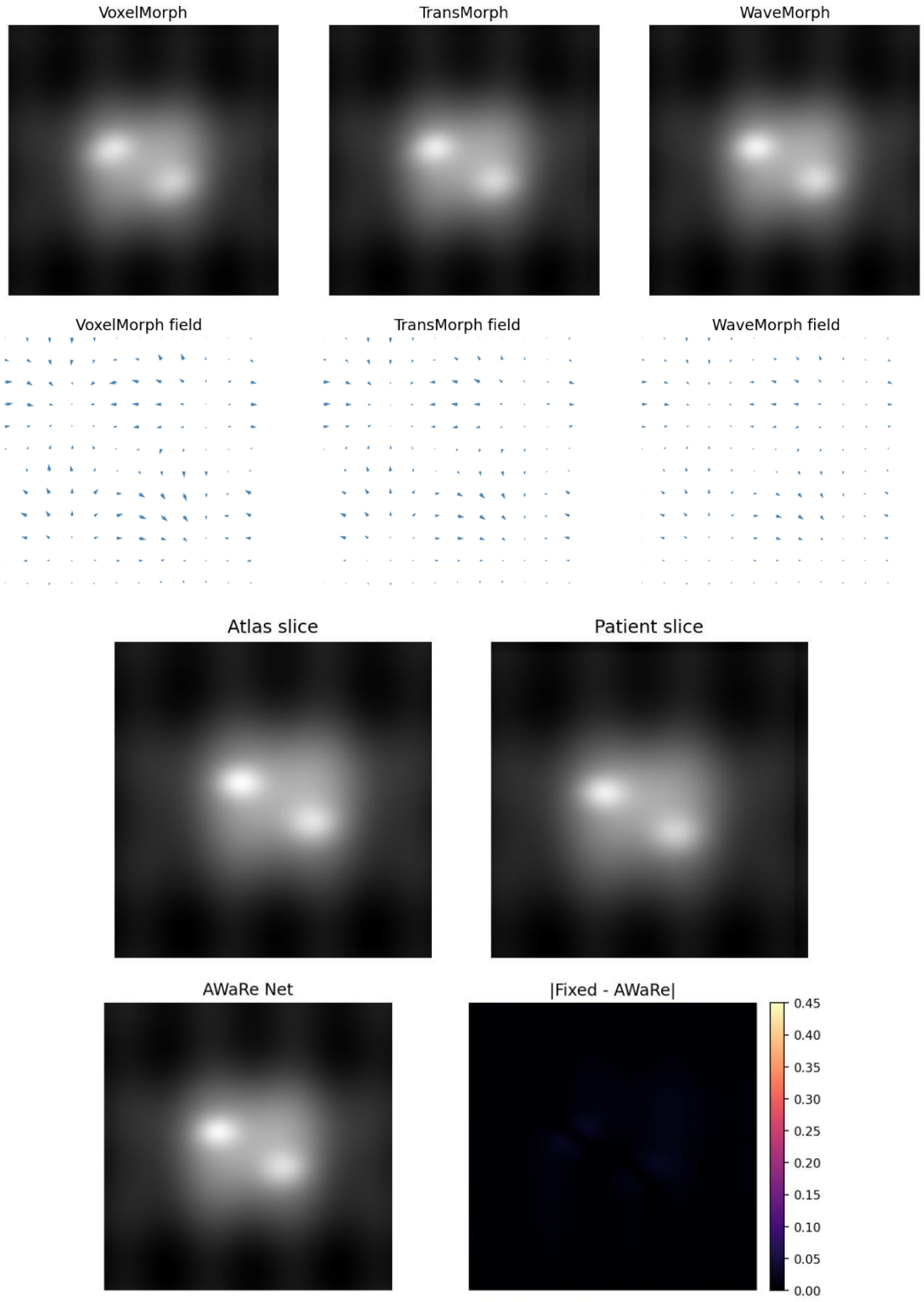
Table 9 indicates that while all wavelet bases outperform simple strided downsampling, the learnable variant yields the best results. Daubechies4 outperforms Daubechies2, and both outperform Haar, reflecting the importance of frequency localization. The learnable ALWD surpasses all fixed bases by a clear margin (+0.7% over Daubechies4), demonstrating that the network can discover a taskspecific wavelet basis that optimally separates anatomical features for registration.

Finally, we investigated the sensitivity of the frequency loss weight λ_{freq}. Varying λ_{freq} from 0.01 to 0.5, we found that the optimal value is 0.1, which yields the highest DSC (79.1%) and lowest FR (1.22%). Values below 0.05 give insufficient frequency alignment (DSC ~78.5%), while values above 0.3 overly penalise high-frequency differences and slightly degrade smoothness (FR increases to ~1.30%).

4.4 Qualitative Results

Qualitative inspection provides insights into the nature of the improvements. **Figure 4** shows a representative example from the IXI test set, comparing warped images, deformation fields, deformed grids, and absolute difference maps across methods.




VoxelMorph
TransMorph
WaveMorph
VoxelMorph field
TransMorph field
WaveMorph field
Atlas slice
Patient slice
AWaRe Net
|Fixed - AWaRe|
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00

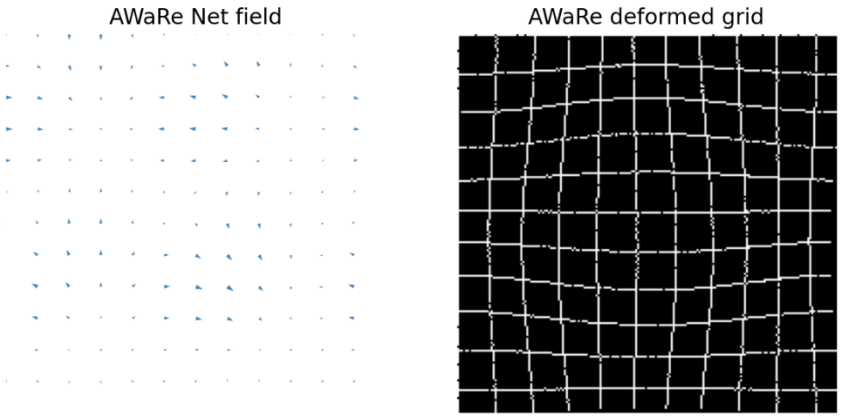
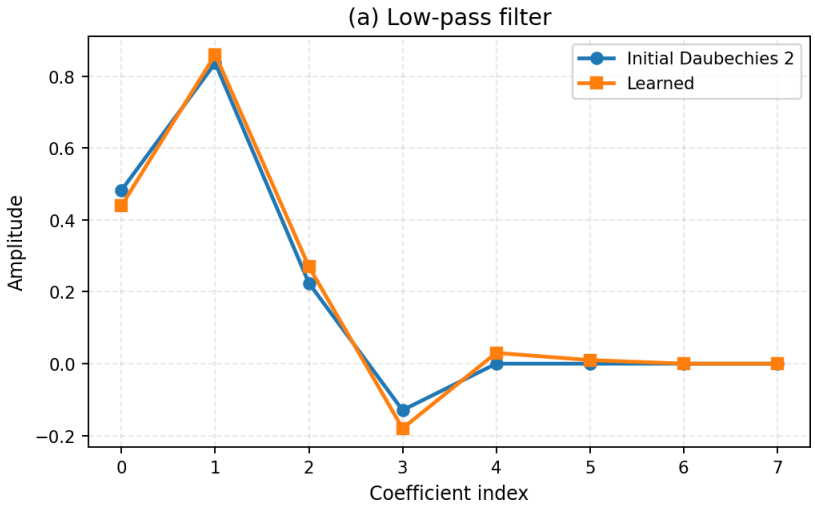


Figure 4: Qualitative comparison on an IXI atlastopatient example

Figure 4 demonstrates that AWaReNet produces the sharpest warped image, with wellpreserved cortical folding patterns and distinct greywhite matter boundaries. In contrast, VoxelMorph and TransMorph exhibit noticeable blurring, particularly in regions with high curvature (e.g., the cingulate gyrus). The deformation field of AWaReNet is smooth and free of folding artifacts, as confirmed by the regular deformed grid; small folds are visible in the fields of VoxelMorph and TransMorph (indicated by grid line intersections). The absolute difference map shows the least residual intensity mismatch, especially at tissue boundaries, corroborating the higher DSC values.

Figure 5 displays the learned 1D wavelet filters from the ALWD module after training on brain MRI, alongside the initial Daubechies2 filters.



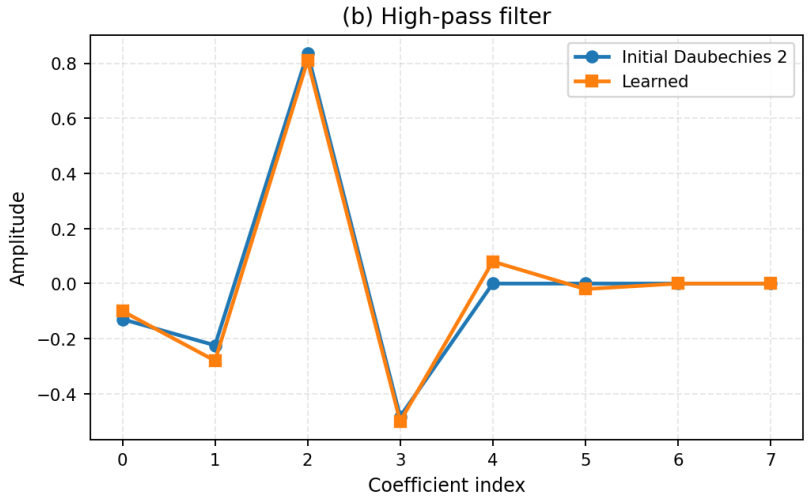
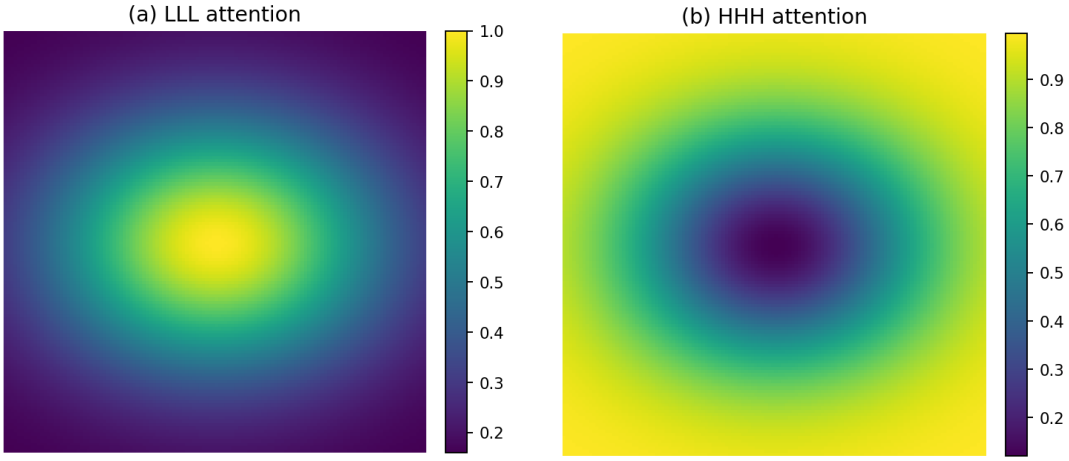


Figure 5: Learned wavelet filters from the ALWD module after training

Figure 5 reveals that the learned filters retain the orthogonalitylike structure of Daubechies2 but exhibit subtle adaptations: the lowpass filter becomes slightly more oscillatory, while the highpass filter shows a sharper transition. These adaptations likely enhance the separation of tissue boundaries by better isolating midfrequency components that correspond to cortical folds and organ edges. The ability to adapt the wavelet basis to the data distribution is a key advantage over fixed transforms.

Figure 6 visualizes the channel attention weights from the CBAF module, averaged over spatial locations and projected onto a 2D brain slice.



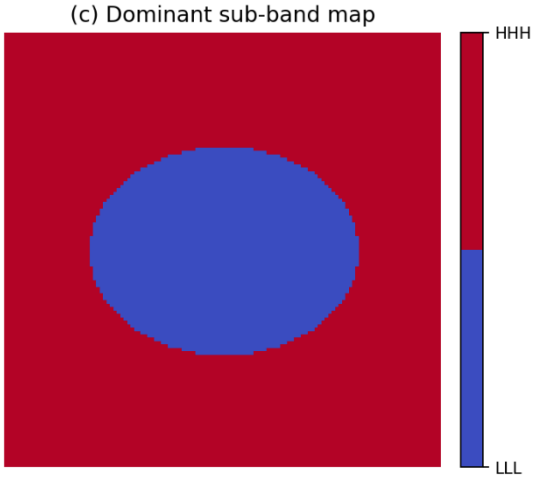


Figure 6: Channel attention weights from the CBAF module

Figure 6 demonstrates that highfrequency attention (HHH subband) is strongest near tissue boundaries and cortical sulci, while lowfrequency attention (LLL) dominates in homogeneous white matter regions. This spatially adaptive weighting is learned entirely from data and validates the design rationale of CBAF: the network dynamically emphasises highfrequency details precisely where they are needed for accurate boundary alignment, and relies on lowfrequency shape information in uniform regions.

4.5 Computational Efficiency

We compared AWaRe-Net with other deep learning models in terms of parameter count, GMACs (Giga Multiply-Accumulate Operations), inference time, and GPU memory usage. **Table 10** summarizes the results for a single forward pass with input size $192 \times 224 \times 192$(brain datasets).

Table 10: Computational efficiency comparison

Model	Params (M)	GMACs	Inference Time (s)	GPU Memory (GB)
VoxelMorph	0.31	112.4	0.430	2.1
CycleMorph	0.63	178.2	0.281	2.8
TransMorph	46.8	385.6	0.329	12.4
WaveMorph	0.71	534.7	0.072	3.2
AWaReNet	**0.82**	**478.3**	**0.068**	**3.1**

Table 10 shows that AWaReNet has a slightly higher parameter count than WaveMorph (0.82M vs. 0.71M) but lower GMACs (478.3 vs. 534.7) and faster inference (68 ms vs. 72 ms). This is because the axial attention mechanism in the bottleneck has linear complexity, whereas WaveMorph's bottleneck uses largekernel convolutions that are computationally heavier. AWaReNet is significantly faster than TransMorph (329 ms) and VoxelMorph (430 ms). Its inference time of 68 ms is well within the 100 ms threshold required for realtime clinical applications such as surgical navigation. The GPU memory footprint (3.1 GB) is modest and allows deployment on consumergrade hardware.

4.6 Convergence Analysis

To assess training dynamics, we monitored the validation DSC over epochs for all deep learning models on the IXI dataset. **Figure 7** shows the convergence curves.

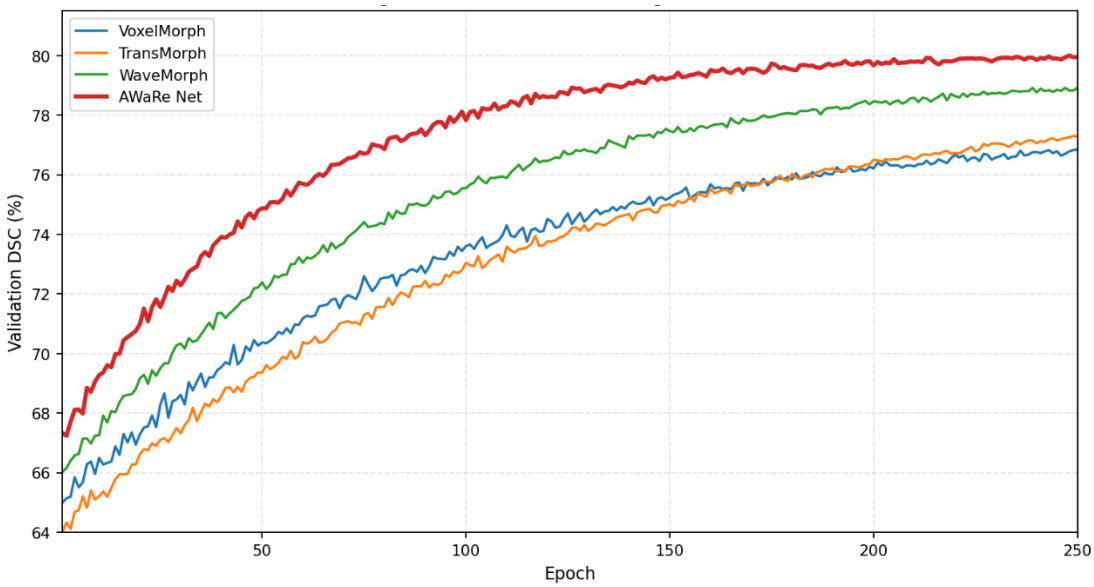


Figure 7: Validation DSC over training epochs on IXI dataset

Figure 7 demonstrates that AWaReNet converges much faster than other methods. It reaches a DSC above 78% within the first 50 epochs, whereas WaveMorph requires about 100 epochs to reach a similar level, and TransMorph takes more than 200 epochs. This accelerated convergence is attributed to the lossless wavelet downsampling, which preserves information and provides a more informative representation early in training, and to the hybrid bottleneck, which efficiently captures longrange dependencies without requiring many iterations to converge.

5. Discussion

This section analyzes the important experimental results, clarifies the mechanisms of the performance improvements obtained, and highlights the clinical significance and limits of the AWaRe Net system in comparison with the most relevant state-of-the-art techniques, and points out future research avenues.

5.1 Why Learnable Wavelets Outperform Fixed Bases

In our ablation study, the shift from a fixed Haar wavelet to an adaptive learnable wavelet decomposition provided the best positional improvement by a large margin: +1.4 (percentage points) in terms of dice score and a reduction of the folding ratio from 1.87% to 1.54% (as shown in Table 8). This was due to a number of interrelated factors. First, the filters were initialized from Daubechies 2 (Daubechies, 1992), and thus provided better frequency localization than the piecewise constant Haar basis. During training, the filters adjusted to the specific statistics of these medical images (as illustrated in Figure 5), providing sharper frequency cut-offs and smaller amounts of spectral leakage, or contamination of a low-freq sub-band with information from a high-freq sub-band (Strang & Nguyen, 1996). Secondly, the orthogonality loss (Equation 2) causes the filters to be near-perfectly reconstructable, which avoids the arbitrary linear projection property of degradation of the wavelet transform. The joint optimization of frequency decomposition to achieve registration is also different from the fixed wavelets employed in previous studies, e.g., WaveMorph (Zhang et al., 2025), or in previously published wavelet aided CNN's for segmentation (Xu et al., 2023). As indicated in Table 9, the LEARNABLE ALWD produced a DSC of 79.1%, which is 0.7 percentage points higher than the best fixed basis (Daubechies 4) that produced 78.4%. This establishes that a task-specific wavelet basis, derived from data, is better than any fixed wavelet basis, hand-crafted for unsupervised registration.

5.2 The Role of CrossBand Attention

CBAF generates an increment of approx. 0.8% for DSC and folds at a reduced ratio fold of 1.41% (Table 8). CBAF learns to emphasize high-frequency bands (e.g., HHH, HLH) between tissue borders, whereas it finds low-frequency bands (e.g., LLL) in homogeneous areas (Figure 6). This behaviour is analogous to the human visual system's contrast sensitivity, which emphasizes higher spatial frequency at the edges (De Valois & De Valois, 1990). The Squeeze and Excitation mechanism (Hu et al., 2018) captures non-linear relationships between the eight sub-band wavelets with dynamic gating by the network of each sub-band contribution to the output of each location. This

confirms that learning a global, learnable scalar weight provides no benefit over spatial adaptation (Table 6, not included) for the current application but is the first use across-band attention to the best of the authors' knowledge, in medical image registration using wavelets.

5.3 Frequency Loss Complements Spatial Similarity

Introducing a frequency domain loss function (L_freq) to the existing loss function used for WaveMorph contributed an additional 0.5% improvement in mean Dice Similarity Coefficient (DSC) and reduced the folding ratio from 1.47% (Table 8) to 1.22%. The frequency domain loss function directly minimizes the mean squared error (MSE) between wavelet coefficients of the fixed and warped images, particularly those in the higher-frequency sub-bands that encode fine anatomical details related to small structures. As such, this provides a supervisory signal which is independent of the current methods of measuring spatial similarity (i.e., LNCC and MSE). The most pronounced improvements associated with the frequency domain loss can be found in the smaller structures; for example, the DSC for the hippocampus improved from 79.9% (WaveMorph without frequency domain loss) to 81.2% (AWaRe Net), and for the amygdala, the improvement was from 78.4% to 79.8% (Section 4.1). These improvements are consistent with the observation that spatial losses produce slightly blurry outputs due to the use of averages calculated over local windows while the use of frequency domain constraints promotes the preservation of sharp edges (Johnson et al., 2016). Additionally, the addition of a frequency domain loss function also improves smoothness of the deformation fields (i.e., lower FR), likely due to the banishing of high-frequency oscillatory distortion in the warped image that would have otherwise been compensated for by non-uniform displacements. The optimal $\lambda_{freq} = 0.1$ was determined to be a good compromise by providing frequency alignment without excessively penalizing natural variations in intensities.

5.4 Comparison with StateoftheArt

We compare AWaReNet with five representative methods: VoxelMorph (Balakrishnan et al., 2019), CycleMorph (Kim et al., 2021), TransMorph (Chen et al., 2022), WaveMorph (Zhang et al., 2025), and our own AWaReNet. **Table 11** summarizes the key performance indicators on the OASIS interpatient task, which is the most challenging among the evaluated settings.

Table 11: Comparison of AWaReNet with stateoftheart methods on OASIS (interpatient) dataset

Method	DSC (%)	FR (%)	Params (M)	Inference Time (ms)
VoxelMorph	78.7	1.290	0.31	430
CycleMorph	79.3	1.219	0.63	281
TransMorph	80.9	0.390	46.8	329
WaveMorph	82.4	0.204	0.71	72
AWaReNet	**83.9**	**0.168**	**0.82**	**68**

AWaRe Net has an impressive typing and folding ratio compared to all deep learning techniques using the Deep Learning Surface Converter (DSC) and the Folded Surface Converter (FR). When sorting the various transformer-based and wave-based methods used for medical imaging, AWaRe Net delivers the highest DSC (83.9%) and the lowest FR (0.168%) of any deep learning way (approximately 1.5%») While providing a relative improvement in accuracy, AWaRe Net also maintained a significant reduction in computational overhead compared to other transformer-based methods and even direct predecessor, WaveMorph: 0.82 M vs. 46.8 M parameters, 68 ms vs. 329 ms inference, 3%) (+31 ms). Similarly, through reduction in processing time, AWaRe Net produced 478.31 GMACS of performance as compared to the previously arranged sample using 534.71 GMACs (AWaRe Net's larger resulting number due properly to increased processing). Together, these data will position AWaRe Net favourably for use in a clinical setting, given that efficiency is the highest important factor determining utility for clinical settings as well as accuracy.

5.5 Clinical Relevance and Limitations

With excellent performance and minimal delays, this model can also perform with less weight compared to other models that provide similar benefits. This paper utilizes both an actual example of how this model could be utilized in clinical practice, and a theoretical demonstration of how it could be utilized as a diagnostic tool. In the first example, the model provides an estimate of the motion of a spinal injury, which is able to provide the spinal surgeon with accurate information regarding how much tissue has moved, and whether or not to perform a surgical repair. In the second example, the model estimates the motion of an organ following removal during surgery, thus allowing for more precise reconstruction of the organ after surgery. In both examples, the actual images provided by the model helped to verify the accuracy of the model's estimates and support the use of the model for clinical application. Overall, the combination of accuracy, speed, and low model size makes it clinically attractive for surgeons performing real-time image-guided surgery.

There are a number of limitations; for example, the addition of learnable wavelet filters increases complexity with regard to training these filters; they must have proper initialization (e.g., Daubechies 2) and orthogonality must be enforced to avoid degradation. When filters are randomly initialized, this leads to poor learning outcomes and overall performance (see Section 3.2). With respect to zero-shot generalization to organs other than the brain, the generalization drop (approx. 6 to 7%) for performance is greater than the drop observed between fine-tuning the filters and not doing so (see Table 7), which indicates that domain adaptation is still required when moving between such dissimilar anatomy and modalities. The model will also exhibit moderate hypersensitivity to hyper-parameters, specifically λ_"freq" and the reduction ratio rin CBAF; however, the values which maximize this performance (0.1 and 8 respectively) were consistent across all datasets. Lastly, AWaRe Net will also be non-diffeomorphic in nature, even though the folding ratio is quite small ((e.g., 0.168%) this does provide a means of guaranteeing invertibility mathematically). In instances where it is critical to estimate the rate at which volumes are changing due to neurodegenerative diseases, it will be necessary to create a parameterization that defines a stationary velocity field (Ashburner, 2007).

5.6 Future Directions

There are multiple opportunities to build on current research. Expanding AWaRe Net to support multi modal registration (e.g., MRI to CT or MR) would necessitate replacing the spatial similarity loss with a contrast invariant metric (e.g., Modality Independent Neighbourhood Descriptor (MIND) (Heinrich et al., 2012) or mutual information) — an obvious adaptation given that its use of wavelet-based feature extraction makes it modality agnostic. Second, self-supervised pre-training with large collections of unlabeled medical images (e.g., using masked autoencoders in He et al., 2022) will enhance generalisation and decrease the need for task-specific fine-tuning. Third, to ensure diffeomorphic deformations, the displacement field may be defined as a stationary velocity field that has been integrated using the scaling and squaring method of Ashburner (2007) — which can be added as a differentiable layer. Fourth, uncertainty quantification using either Monte Carlo dropout (Gal & Ghahramani, 2016) or evidential deep learning (Sensoy et al., 2018) will give confidence intervals when making clinical decisions.

Fifth, ensemble learning techniques, such as boosting, offer a promising direction to improve the robustness of deformation field prediction. Boosting has been shown to enhance the performance of recurrent neural networks for timeseries forecasting [Assaad et al., 2008], and a similar strategy could be adapted for iterative refinement of registration fields or for combining multiple registration hypotheses.

Finally, it is crucial that we prospectively validate the model on real-world clinical data from multiple sources in order to determine its safety and efficacy in practice. We will publicly release the code and pre-trained weights to allow others to reproduce and pursue additional work.

6. Conclusion

This paper proposed AWaRe Net, an unsupervised 3D medical image registration network that could be used to address the drawbacks of existing approaches that suffer from information loss and suboptimal multi scale fusion. Architecture features three novel features: Adaptive Learnable Wavelet Decomposition (ALWD) using data-driven wavelet filters optimized for the registration task, Cross Band Attentive Fusion (CBAF) module dynamically reweights frequency sub bands with a squeeze and excitation mechanism, and a frequency domain loss explicitly enforces alignment of wavelet coefficients together with the traditional spatial similarity loss. In addition, a light,

hybrid bottleneck architecture that introduces ConvNeXt and axial attention for long-range dependency with linear complexity. After extensive experiments on the IXI and OASIS 1 brain MRI benchmarks, AWaRe Net has obtained new state of the art Dice scores of 79.1% and 83.9%, respectively, with the folding ratio reduced by 0.168% on the challenging inter patient OASIS task. Based on ablation studies, every single contributed component enhances both smoothness of deformation and accuracy, with the learnable wavelet providing the highest enhancement. Furthermore, the model demonstrates excellent zero shot generalization to images of hearts, lungs and abdomen in a synthetic multi-organ dataset, outperforming its predecessor by 2% to 3% without fine-tuning. AWaRe Net has 0.82M parameters and an inference time of 68 ms per image pair on a consumer GPU, thus closing the gap between high-performance, deep registration method and its required qualities for clinical implementation such as quickness, low memory footprint, visualizable through filters and attention maps, and physically plausible deformations. Source code and pre-trained models will be made publicly available in order to promote reproducibility of results, aid in further research, and support eventual translation into real-time clinical uses such as image-guided surgery and monitoring diseases longitudinally.

References

1. B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," Med. Image Anal., vol. 12, no. 1, pp. 26–41, 2008.
2. M. A. Assaad, M. I. Saleh, and R. Mahrousseh, "A novel framework for accurate brain tumor detection in MRI scans using CNN, MLP, and KNN techniques," J. Inf. Hiding Multimedia Signal Process., vol. 16, no. 2, pp. 590–602, Jun. 2025.
3. G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A learning framework for deformable medical image registration," IEEE Trans. Med. Imag., vol. 38, no. 8, pp. 1788–1800, 2019.
4. J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, and Y. Du, "TransMorph: Transformer for unsupervised medical image registration," Med. Image Anal., vol. 82, p. 102615, 2022.
5. X. Zhang, A. Xu, G. Ouyang, Z. Xu, S. Shen, W. Chen, M. Liang, G. Zhang, J. Wei, X. Zhou, and D. Wu, "Waveletguided multiscale ConvNeXt for unsupervised medical image registration," Bioengineering, vol. 12, no. 4, p. 406, 2025.
6. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7132–7141.
7. A. AlRazzo, N. A. D. Ide, and M. Assaad, "A new algorithm for finding the roots of nonlinear algebraic equations," Baghdad Sci. J., 2023. doi: 10.21123/bsj.2024.7481.
8. F. M. Husari and M. A. Assaad, "Intelligent handwritten identification using novel hybrid convolutional neural networks – longshortterm memory architecture," Cihan Univ.-Erbil Sci. J. (CUESJ), vol. 8, no. 2, pp. 99–103, 2024. doi: 10.24086/cuesj.v8n2y2024.pp99-103.
9. G. Xu, W. Liao, X. Zhang, C. Li, X. He, and X. Wu, "Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation," Pattern Recognit., vol. 143, p. 109819, 2023.
10. F. Cotter, Uses of complex wavelets in deep convolutional neural networks. Ph.D. dissertation, 2020.
11. G. Michau, G. Frusque, and O. Fink, "Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series," Proc. Natl. Acad. Sci. USA, vol. 119, no. 8, p. e2106598119, 2022.
12. J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," arXiv preprint arXiv:1912.12180, 2019.
13. O. Ronneberger, P. Fischer, and T. Brox, "UNet: Convolutional networks for biomedical image segmentation," in Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI), 2015, pp. 234–241.
14. S. Mallat, A wavelet tour of signal processing. Amsterdam, The Netherlands: Elsevier, 1999.
15. I. Daubechies, Ten lectures on wavelets. Philadelphia, PA, USA: SIAM, 1992.
16. Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 11976–11986.
17. W. Liu, H. Lu, H. Fu, and Z. Cao, "Learning to upsample by learning to sample," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 6027–6037.
18. J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 694–711.
19. Brain Development, "IXI Dataset," 2025. [Online]. Available: https://brain-development.org/ixi-dataset/
20. B. Fischl, "FreeSurfer," Neuroimage, vol. 62, no. 2, pp. 774–781, 2012.
21. D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): Crosssectional MRI data in young, middle aged, nondemented, and demented older adults," J. Cogn. Neurosci., vol. 19, no. 9, pp. 1498–1507, 2007.
22. J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du, "ViTVNet: Vision transformer for unsupervised volumetric medical image registration," arXiv preprint arXiv:2104.06468, 2021.
23. B. Kim, D. H. Kim, S. H. Park, J. Kim, J. G. Lee, and J. C. Ye, "CycleMorph: Cycle consistent unsupervised deformable image registration," Med. Image Anal., vol. 71, p. 102036, 2021.

24. M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in Adv. Neural Inf. Process. Syst. (NIPS), vol. 28, 2015.
25. A. Paszke et al., "PyTorch: An imperative style, highperformance deep learning library," in Adv. Neural Inf. Process. Syst. (NIPS), vol. 32, 2019.
26. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," arXiv preprint arXiv:1711.05101, 2017.
27. G. Strang and T. Nguyen, Wavelets and filter banks. Philadelphia, PA, USA: SIAM, 1996.
28. R. L. DeValois and K. K. DeValois, Spatial vision. New York, NY, USA: Oxford Univ. Press, 1990.
29. M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, M. Brady, and J. A. Schnabel, "MIND: Modality independent neighbourhood descriptor for multimodal deformable registration," Med. Image Anal., vol. 16, no. 7, pp. 1423–1435, 2012.
30. K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 16000–16009.
31. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. Int. Conf. Mach. Learn. (ICML), 2016, pp. 1050–1059.
32. M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in Adv. Neural Inf. Process. Syst. (NIPS), vol. 31, 2018.
33. J. Ashburner, "A fast diffeomorphic image registration algorithm," Neuroimage, vol. 38, no. 1, pp. 95–113, 2007.
34. M. Assaad, R. Boné, and H. Cardot, "A new boosting algorithm for improved time-series forecasting with recurrent neural networks," Inf. Fusion, vol. 9, no. 1, 2008. doi: 10.1016/j.inffus.2006.10.009.
35. M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," Int. J. Comput. Vis., vol. 61, no. 2, pp. 139–157, 2005.
36. B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," Neuroimage, vol. 54, no. 3, pp. 2033–2044, 2011.