

Digital Twin-Driven Autonomous Imaging Agent Training for Scalable Diagnostic Access in Underserved Clinical Settings

Ahmed I. Taloba¹, Omar Mohammad Amer Huwari^{2,*}, Fuad S. Al-Duais³, and Mohammad Kanan^{4,5}

¹Department of Information System, Faculty of Computers and Information, Assiut University, Assiut, Egypt

²Department of English language & Literature, College of Science and Humanities Studies in Al-Kharj, Prince Sattam Bin Abdulaziz University 11942 Al-Kharj, Saudi Arabia

³Department of Mathematics, College of Science and Humanities in Al-Kharj, Prince Sattam Bin Abdulaziz University 11942 Al-Kharj, Saudi Arabia

⁴Department of Industrial Engineering, College of Engineering, University of Business and Technology, Jeddah 21448, Saudi Arabia

⁵Department of Mechanical Engineering, College of Engineering, Zarqa University, Zarqa, Jordan

Received: 30 Mar. 2026, Revised: 15 May. 2026, Accepted: 20 May. 2026

Published online: 1 Jul. 2026

Abstract: Ultrasound imaging plays a critical role in rapid and non-invasive clinical diagnosis; however, existing intelligent diagnostic systems remain limited by domain inconsistency, insufficient annotated datasets, poor cross-domain generalization, and the absence of adaptive autonomous interaction mechanisms, particularly in underserved clinical environments. Conventional convolutional, transformer-based, and segmentation-guided frameworks achieve promising segmentation performance, yet they fail to provide scalable closed-loop refinement, digital twin interaction learning, and continuous autonomous adaptation under heterogeneous imaging conditions. To address these limitations, this study proposes a Digital Twin-Driven Autonomous Imaging Agent framework that integrates a Domain-Adaptive Vision Transformer Autoencoder (DA-ViT-AE) with a Hierarchical ReAct (H-ReAct) autonomous reasoning agent for adaptive ultrasound diagnostic analysis. The primary novelty of this study lies in the integration of digital twin-based sim-to-real interaction learning, autonomous perception-action co-evolution, reasoning-driven iterative refinement, and feedback-guided adaptive segmentation within a unified closed-loop diagnostic ecosystem. The proposed framework enables continuous perception-action interaction, iterative segmentation refinement, feedback-driven adaptation, and cross-domain representation learning within a digital twin ecosystem. The framework was implemented using Python with PyTorch, TensorFlow, OpenCV, MONAI, and reinforcement learning libraries on GPU-enabled environments. Experiments were conducted using the publicly available abdominal ultrasound simulation and segmentation dataset containing 617 real ultrasound scans and 926 synthetic ultrasound scans with multi-organ annotations. The proposed framework achieved a Dice score of 96.52%, precision of 97.72%, recall of 97.20%, and IoU score of 93.41%, outperforming existing state-of-the-art models including U-Net, MSEU-net, and DSEU-net. The proposed framework improved overall segmentation performance by approximately 1.2%–6.8% compared to existing baseline methods through domain-adaptive perception learning and autonomous iterative refinement. Furthermore, the framework demonstrated stable scalability, rapid low-resource adaptation, and reliable cross-domain generalization, establishing an effective and intelligent ecosystem for scalable autonomous diagnostic assistance in resource-constrained healthcare environments.

Keywords: Digital Twin; Autonomous Imaging Agent; Ultrasound Image Segmentation; Domain-Adaptive Vision Transformer Autoencoder; Medical Image Analysis.

1 Introduction

Medical imaging is a very important role in the current clinical decision-making process for the diagnosis, monitoring and treatment of a variety of diseases [1, 2]. In

particular, ultrasound imaging is valued for its non-invasive nature, real-time capability and cost-effectiveness, making it highly suitable for deployment in resource-constrained and underserved

* Corresponding author e-mail: o.huwari@psau.edu.sa

clinical environments, among various modalities [3, 4]. The quality of ultrasound-based diagnosis is heavily reliant on operator expertise, however, with the quality of the ultrasound image acquired depending on the probe positioning, the anatomical interpretation, and use of noise and artifact management [5, 6]. The recent deep learning (DL) techniques have greatly enhanced the capability of automated image analysis, especially for segmentation and classification [7, 8]. However, existing methods are based on large annotated datasets, and are used under static inference settings [9], which limits their ability to generalize over different acquisition settings [10, 11]. Moreover, medical data is still limitedly annotated and the gap between simulated and real images still remains a significant challenge for robust model development [12, 13].

Despite these advancements, existing research in medical imaging largely focuses on passive analysis rather than active image acquisition and decision-making [14], [15]. Existing vision-based models, such as Convolutional Neural Networks (CNNs) and transformer networks [16], are unable to autonomously engage with imaging environments or to continuously improve their output through contextual feedback [17, 18]. Furthermore, the use of multimodal reasoning has been shown to be a promising application of AI in medical imaging [19], but existing agentic AI systems are limited by their ability to integrate into the medical imaging workflow [20], provide real-time feedback and adapt to unseen clinical scenarios [21, 22]. These restrictions make their application in clinical contexts difficult, especially in resource-constrained environments [23, 24]. To address these challenges the present study suggests a framework based on the digital twin, combining a DA-ViT-AE with a hierarchical agentic system that can plan, execute and optimize imaging workflows. The proposed approach is designed to bridge the gap between perception and decision-making, enabling scalable, explainable, and adaptive diagnostic systems for underserved healthcare settings by leveraging simulated environments for autonomous training and using adaptive feedback for continuous learning.

1.1 Problem Statement

Despite the significant advancements, the development of effective and scalable medical imaging systems is hindered by issues such as data scarcity, domain variability, and the absence of adaptive decision-making capabilities. Existing ultrasound image analysis systems have a huge demand for accurate, scalable, and autonomous image diagnosis in heterogeneous clinical environments. Existing DL and transformer-based methods heavily depend on large annotated datasets and a post-acquisition analysis stage, which are not suitable for generalisation between simulated and real ultrasound data [25]. Moreover, existing systems do not support

adaptive perception-action interaction, autonomous decision-making, or iteratively refining the image, which are necessary for real-time imaging optimization. These inconsistencies and errors in segmentation, caused by domain discrepancies, limited annotated data, poor adaptability to unseen imaging conditions, and others, further decrease segmentation reliability and diagnostic consistency, particularly in underserved and low-resource healthcare environments. Furthermore, current methods need significant expert contribution and lack the ability to facilitate ongoing learning through feedback [26]. The limitations all contribute to a lack of adoption of intelligent imaging systems in the underserved clinical environments where expert intervention is limited. To overcome these challenges, this study introduces the framework of a DA-ViT-AE and a hierarchical agentic system for autonomous, adaptive and scalable imaging workflows with digital twins. The proposed approach facilitates continuous learning through simulation-based training and real-time feedback integration, enhancing generalization and decision-making capabilities.

1.2 Key Contributions

1. Developed a digital twin-driven ultrasound learning framework integrating simulated and real imaging data for scalable cross-domain training.
2. Proposed a domain-adaptive transformer autoencoder for robust sim-to-real representation learning and segmentation refinement.
3. Introduced a hierarchical autonomous imaging agent for reasoning-driven perception-action optimization and iterative refinement.
4. Designed a unified closed-loop architecture combining visual perception, autonomous decision-making, and adaptive imaging interaction.
5. Integrated feedback-driven co-evolution learning for continuous adaptation and improved performance in unseen clinical environments.

The rest of this paper is organized as follows: Section 2 reviews relevant literature and existing methodologies in medical imaging and agentic systems. Section 3 presents the proposed digital twin-driven framework and model architecture. Section 4 discusses experimental results and performance evaluation. Finally, Section 5 concludes the study and outlines potential directions for future research.

2 Related Work

Avanzato et al. [27] propose the Lung-DT framework to enhance diagnosis and monitoring of thoracic disorders using Digital Twins integrated with AI and IoT. The approach utilizes YOLOv8 to categorize the chest X-rays into five disease classes, and it also allows for real-time monitoring of the respiratory system. The system has the

high performance with the accuracy of 96.8% and better diagnosis efficiency. However, the limitations of the model are that it relies on data from a single modality only, and fails to incorporate other clinical information such as spirometry for comprehensive analysis. To address the lack of high-quality training data for ultrasound image analysis, Song and Chong [28] propose generating synthetic data. The study suggests a DL model called S-CycleGAN, which includes semantic discriminators for medical imaging to convert CT images to realistic ultrasound images, maintaining the anatomical details. The method improves the quality of data used in training and supports robot-assisted ultrasound systems. The resulting images, however, show minor anatomical differences and do not allow for any improvements in performance when it comes to precision tasks.

The objective of Vitale et al. [29] is to mitigate GAN induced hallucinations in order to enhance realism and anatomical consistency in ultrasound simulation. The study suggests the segmentation-guided CycleGAN with polar coordinate training to retain the organ structure and reduce artifacts. The method provides substantial improvement in terms of lower FID and KID while producing clinically realistic images validated by radiologists. However, this study is restricted to healthy patients, which may not be suitable for validation in pathological cases and complex anatomical variations. The objective of Hassen Mohammed et al. [30] is to improve the accuracy of CIMT measurement by testing DL models to segment the intima-media complex in carotid ultrasound images. The study performs a comparison between CNN, self-ONN, transformer-based, and pixel difference convolution models on the CUBS dataset. The results indicate that the transformer model and pixel difference model are more effective than the CNN model, while the self-ONN model is more effective than the CNN model. The study is however limited by small amount of data, which affects the generalisation of the model, and needs improved data augmentation methods.

The primary objective of Chong et al. [25] is to enhance medical image segmentation while tackling the shortcomings of transformer-based models in capturing local features and edge details. The study suggests P-TransUNet, which combines an efficient P-Transformer, fusion module, and edge loss to improve the accuracy of the segmentation. Experimental results show that it outperforms state-of-the-art methods on multiple tasks. The model, however, is not efficient or scalable in clinical settings where computational resources are limited and requires large datasets. Ta et al. [31] aim to improve small-target medical image segmentation by addressing challenges in feature representation and pixel imbalance. The study proposes LET-Net, a locally enhanced transformer network combining a pyramid vision transformer encoder with local enhancement and progressive decoding modules. For polyp segmentation and breast lesion segmentation

tasks, the model outperforms the state-of-the-art performance. However, the method has limited generalizability because it has not been validated in a diverse imaging modalities and anatomical structures.

The goal of Zhu et al. [26] is to segment breast tumour in ultrasound images to achieve accurate diagnosis and localization. The study proposes Swin-Net, integrating a swin-transformer encoder with CNNs, along with feature refinement and multi-scale fusion modules to enhance representation and suppress noise. The model shows better Dice scores to achieve a better performance on benchmark datasets. However, the study is limited by low diversity of datasets, which reduces the generalisation in complex pathological cases. The objective of the study by Wu et al. [32] is to make liver ultrasound segmentation more accurate for better clinical diagnosis in real-time. The study suggests BACANet, a lightweight DL model based on convolutional attention, selective large kernel convolution, and boundary supervision, which is used to accurately extract features and localization. The model performs efficient inference and high segmentation accuracy on all the datasets. But the study has some limitations, such as a small dataset size and model optimization that could limit the ability to generalize and scale the model up to clinical application in real-time.

In contrast, Tagnamas et al. [33] improve the breast ultrasound image analysis by performing the two tasks of breast tumour segmentation and classification together. The study suggests a hybrid CNN-ViT multi-task learning approach using two encoders, channel attention fusion module and an MLP-Mixer classifier. The model achieves improved performance with 83.42% Dice and 86% classification accuracy. However, the approach is limited by lack of contextual information from surrounding tissues, affecting diagnostic accuracy. Zhang et al. [34] aims to enhance the breast ultrasound segmentation problem by tackling the issues of multiscale feature extraction and boundary ambiguity. The study suggests the three-path U-shaped network (TPUNet) which incorporates an attention-based feature fusion block and a hybrid loss function to boost feature representation and boundary refinement. The model performs better segmentation accuracy and visual quality than the existing methods. The drawback is that the approach becomes more complex in the architecture which means more computation cost and training difficulty.

Table 1 summarizes recent studies in medical imaging, highlighting various DL and transformer-based approaches for segmentation, synthesis, and diagnostic analysis. These techniques have shown promising results in feature extraction, accuracy enhancement, and limited data handling using generative approaches. There is however a key gap in existing research, most of the approaches are based on static image analysis and are not integrated with adaptive real-time decision-making systems. Moreover, they are not widely applicable in dynamic clinical settings due to their lack of generalization across domains, high computation

Table 1: Summary of Existing Studies

Reference	Method	Advantage	Limitation
Avanzato et al. [27]	Digital Twin + YOLOv8	Real-time diagnosis and monitoring	Limited multimodal integration
Song and Chong [28]	S-CycleGAN	Reduces data scarcity	Synthetic inconsistencies
Vitale et al. [29]	Segmentation-guided CycleGAN	Improved anatomical realism	Limited pathological validation
Hassen Mohammed et al. [30]	CNN/Transformer comparison	Strong segmentation performance	Limited generalization
Chong et al. [25]	P-TransUNet	Enhanced edge and global features	High computational complexity
Ta et al. [31]	LET-Net	Effective small-target segmentation	Poor cross-domain adaptability
Zhu et al. [26]	Swin-Net	Robust multi-scale representation	Limited dataset diversity
Wu et al. [27]	BACANet	Real-time lightweight segmentation	Scalability limitations
Tagnamas et al. [33]	Hybrid CNN-ViT	Multi-task learning capability	Weak contextual understanding
Zhang et al. [34]	TPUNet	Improved boundary refinement	Complex architecture

demands and low use of simulation-driven training. The research gap is, scalable medical imaging applications lacking of unified frameworks to integrate domain adaptation, simulation-based learning and autonomous decision making. To fill these gaps, the proposed study presents an integrated framework of a domain-adaptive transformer model and an agent-based adaptive imaging workflows system using a digital twin concept. It also allows for real-time refinement, enhanced continuous learning and better generalization and applicability in underserved clinical contexts.

3 Proposed Digital Twin-Driven Autonomous Imaging Agent Framework for Scalable Medical Image Diagnostic Analysis in Underserved Clinical Environments

The proposed study is a unified framework of autonomous medical imaging which involves simulation-based learning, domain-adaptive perception and decision-making using agents in a closed-loop architecture. Heterogeneous ultrasound data is first pre-processed and then used to construct a digital twin environment to simulate realistic imaging scenarios and enable controlled interactions. Then, a DA-ViT-AE is trained to learn a robust, domain-invariant representation and segmentation results in a simulated and real data. These representations are further organized according to a state space that allows a H-ReAct agent to sequentially and iteratively take actions including spatial adjustment and region refinement to optimize image quality and diagnostic accuracy. The system constantly assesses the results by segmentation confidence and image quality scores, and adapts the refinement based on the results. Further, there is a feedback loop for incremental updates, allowing the framework to adapt to new clinical scenarios while also being efficient to operate in a constrained environment. The workflow of the proposed framework is illustrated as in Figure 1.

3.1 Data Collection

The dataset used in this study consists of a combination of real and simulated abdominal ultrasound images obtained from the US Simulation and Segmentation Dataset [35]. The real data set consists of 617 ultrasound scans obtained from 11 healthy subjects following a standard clinical procedure from various anatomical views such as intercostal, subcostal, longitudinal and transverse scans recorded with a SonoSite M-Turbo system. There are 61 real images included and the segmentation masks of the organs (such as liver, kidneys, gallbladder, spleen, and vessels) are manually annotated. Further, 926 simulated abdominal ultrasound images were created by a ray-casting-based simulator model from the CT volumes and silver-standard labels were used to obtain the organ annotations from the simulated images. This hybrid dataset provides both fully labelled synthetic data and partially labelled real data for model training and validation.

3.2 Data Preprocessing

Effective data preprocessing is crucial to ensure consistency, speed model convergence and enhance generalization for heterogeneous ultrasound images. Preprocessing creates a common data format that is appropriate for learning robust features, due to the differences between simulated and real data. This step includes spatial standardization, intensity normalization, structuring of annotations and partitioning of datasets. All these procedures support minimize domain discrepancies and to pre-train the data for efficient model training of the proposed DL and agent-based framework.

3.2.1 Image Resizing

To ensure uniform input dimensions for the model, all ultrasound images are resized to a fixed spatial resolution of $H \times W$ (e.g., 224×224). Let the original image be denoted as $I(x, y) \in \mathbb{R}^{H_0 \times W_0}$. The resized image $I_r(x', y')$ is obtained using interpolation given as in Equation (1).

$$I_r(x', y') = I \left(\left\lfloor \frac{x' \cdot H_0}{H} \right\rfloor, \left\lfloor \frac{y' \cdot W_0}{W} \right\rfloor \right) \quad (1)$$

Where $(x' | y')$ are coordinates in the resized image, and $(H_0 | W_0)$ represent original dimensions. This transformation guarantees the consistency of the spatial dimension of the samples, which allows for the processing of samples in a batch and the stability of training the transformer-based architecture.

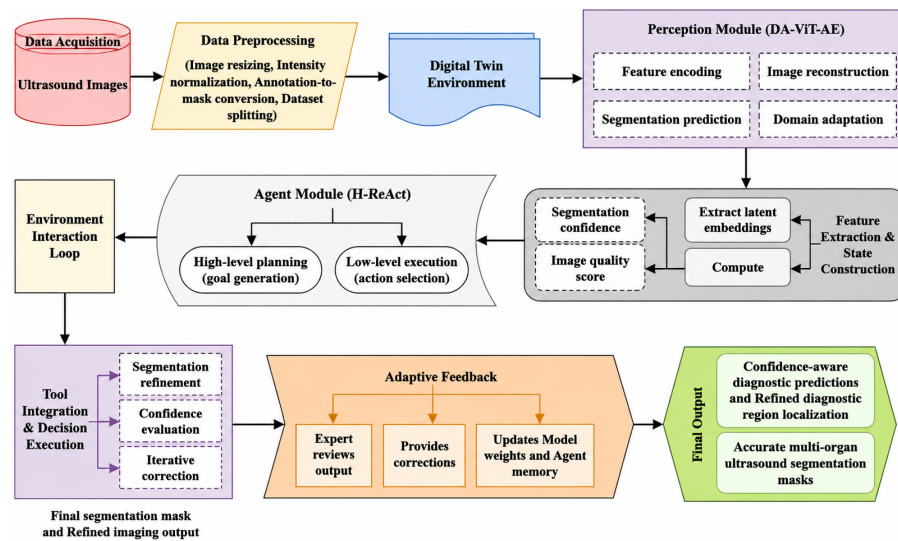


Fig. 1: Workflow of the Proposed Framework

3.2.2 Intensity Normalization

Due to the differences in the acquisition conditions, the intensity distribution is different for ultrasound images. Min-max normalization is applied to normalise the pixel values represented as in Equation (2).

$$I_n(x, y) = \frac{I_r(x, y) - I_{\min}}{I_{\max} - I_{\min}} \quad (2)$$

Where $I_n(x, y)$ is the normalized pixel intensity, and $I_{\min}$, $I_{\max}$ denote the minimum and maximum pixel values in the image. This normalizes the intensities to the interval $[0, 1]$, to reduce the variability and to provide the model with a more stable training.

3.2.3 Annotation Conversion to Segmentation Masks

Labels of organs are transformed into segmentation masks for supervised learning, with the addition of annotations. Let $A(x, y)$ denote the annotation map, where each pixel corresponds to a class label $c \in \{1, 2, \dots, C\}$. A one-hot encoded segmentation mask $M(x, y, c)$ is defined as in Equation (3).

$$M(x, y, c) = \begin{cases} 1, & \text{if } A(x, y) = c \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Where C represents the number of anatomical classes. This representation can be used for multi-class segmentation and the model can learn the spatial boundaries between the different organ structures.

3.2.4 Data Augmentation

In order to overcome the lack of the annotated real ultrasound images, multiple data augmentation methods were used during the training process including horizontal shifting, vertical translation, zoom scaling, intensity variation, elastic deformation, rotation, and probe-view simulation. These augmentation strategies increased the diversity of images, enhanced the ability of the domain generalization and mitigated model overfitting under heterogeneous ultrasound imaging conditions.

3.2.5 Dataset Splitting

The data is split into training, validation, and testing sets to obtain an unbiased evaluation of the performance. The training set contains simulated images and a small number of real images; the validation set contains a balanced set; and the test set contains real images. This strategy allows to effectively learn from synthetic data while assessing the generalization on real-world scenarios.

3.3 Digital Twin Environment Construction

The digital twin environment is constructed to emulate a controllable ultrasound imaging process, enabling autonomous agent training through simulated interaction. Unlike conventional static datasets, this study formulates the simulated ultrasound images from the US Simulation and Segmentation Dataset as a dynamic environment where state transitions are explicitly defined. Each image is treated as a state S_t , and transformations applied to the image simulate probe movements and acquisition

variations. At time step t , the state is defined as in Equation (4).

$$S_t = \{I_t, F_t, Q_t\} \quad (4)$$

Where $I_t \in \mathbb{R}^{H \times W}$ represents the ultrasound image, $F_t \in \mathbb{R}^d$ denotes latent features extracted using DA-ViT-AE, and $Q_t \in \mathbb{R}$ is an image quality score (e.g., SSIM or segmentation confidence). This formulation ensures that visual information and acquired representations are integrated into the decision-making process. The agent interacts with the environment through a continuous action space represented as in Equation (5).

$$A_t = \{\Delta x, \Delta y, \alpha, \beta\} \quad (5)$$

Where Δx and Δy represent the horizontal and vertical shifts (probe movement), α denotes the zoom scaling factor, and β represents the cropping ratio (region-of-interest refinement). The transformed image I_{t+1} is obtained as shown in Equation (6).

$$I_{t+1} = T(I_t, A_t) \quad (6)$$

Where $T(\cdot)$ is a transformation function combining translation, scaling, and cropping operations. The transition function is defined as in Equation (7).

$$S_{t+1} = \mathcal{E}(S_t, A_t) \quad (7)$$

Where $\mathcal{E}(\cdot)$ represents the digital twin environment. After applying action A_t , the updated image I_{t+1} is passed through the DA-ViT-AE model to compute updated features F_{t+1} and quality metric Q_{t+1} . To guide agent learning, a reward function is defined as in Equation (8).

$$R_t = \lambda_1 \cdot \text{Dice}(M_t, M_{gt}) + \lambda_2 \cdot \text{SSIM}(I_t, I_{ref}) - \lambda_3 \cdot \|A_t\| \quad (8)$$

Where $\text{Dice}(\cdot)$ measures the segmentation overlap, M_t is the predicted segmentation mask, M_{gt} denotes the ground-truth mask, $\text{SSIM}(\cdot)$ evaluates the structural similarity between the current image and the reference image, I_{ref} is the reference optimal image, and $\|A_t\|$ penalizes excessive probe movements. The coefficients λ_1 , λ_2 , and λ_3 control the relative contributions of segmentation accuracy, image quality, and motion regularization, respectively. Consequently, the agent learns to jointly optimize image quality and diagnostic accuracy. In this study, simulated ultrasound images are employed as a surrogate for real ultrasound acquisition, allowing the agent to explore diverse anatomical configurations and optimize its policy without requiring physical interaction with a patient.

The transformation parameters are constrained within realistic bounds $\Delta x, \Delta y \in [-10, 10]$ pixels, $\alpha \in [0.8, 1.2]$, and $\beta \in [0.5, 1.0]$. This prevents the appearance of an unrealistic distortion of images and maintains anatomical consistency. The novelty is in transforming an interactive, dynamic medical imaging environment from a static dataset. While most existing methods focus on the

simulated data for augmentation or pre-training, this study proposes a state-action formulation for ultrasound imaging, a feature-aware environment that incorporates DA-ViT-AE representations, and a task-driven reward mechanism that integrates segmentation and image quality.

This enables the system to learn to acquire better images, rather than only analysing them, thereby bridging the gap between perception and decision-making in medical imaging.

3.4 Perception Model Development: Domain-Adaptive Vision Transformer Autoencoder

The proposed DA-ViT-AE designed to learn robust, domain-invariant representations from both simulated and real ultrasound images to enable accurate segmentation from the ultrasound images while reducing the discrepancy of simulated and real images. In this study, the model is not utilized as a segmentation network in itself, but rather as a feature generator and quality evaluator, which directly interacts with the digital twin environment and subsequently influences the decision making of the agent. The architecture of the proposed Domain-Adaptive Vision Transformer Autoencoder is illustrated as in Figure 2.

The model is first trained in self-supervised manner with simulated and real images to learn generalized images of anatomical structures and imaging patterns. Given an input image $I \in \mathbb{R}^{H \times W}$, the encoder $\mathcal{E}_\theta$ maps it into a latent representation given as in Equation (9).

$$Z = \mathcal{E}_\theta(I), Z \in \mathbb{R}^d \quad (9)$$

Where Z is the latent feature vector, d is the embedding dimension (e.g., $d = 512$), and θ denotes encoder parameters. The decoder $\mathcal{D}_\phi$ reconstructs the image expressed as in Equation (10).

$$\hat{I} = \mathcal{D}_\phi(Z) \quad (10)$$

The reconstruction loss is defined as in Equation (11).

$$\mathcal{L}_{rec} = \|I - \hat{I}\|_2^2 \quad (11)$$

This step is important for the model to learn the structure and texture patterns shared by the two domains to deal with ultrasound noise and variability. To enable clinical relevance, a segmentation head $\mathcal{S}_\psi$ is attached to the shared latent space given as in Equation (12).

$$M = \mathcal{S}_\psi(Z) \quad (12)$$

Where $M \in [0, 1]^{H \times W \times C}$ is the predicted segmentation mask, and C is the number of organ classes. The segmentation objective combines Dice loss and cross-entropy is computed as in Equation (13).

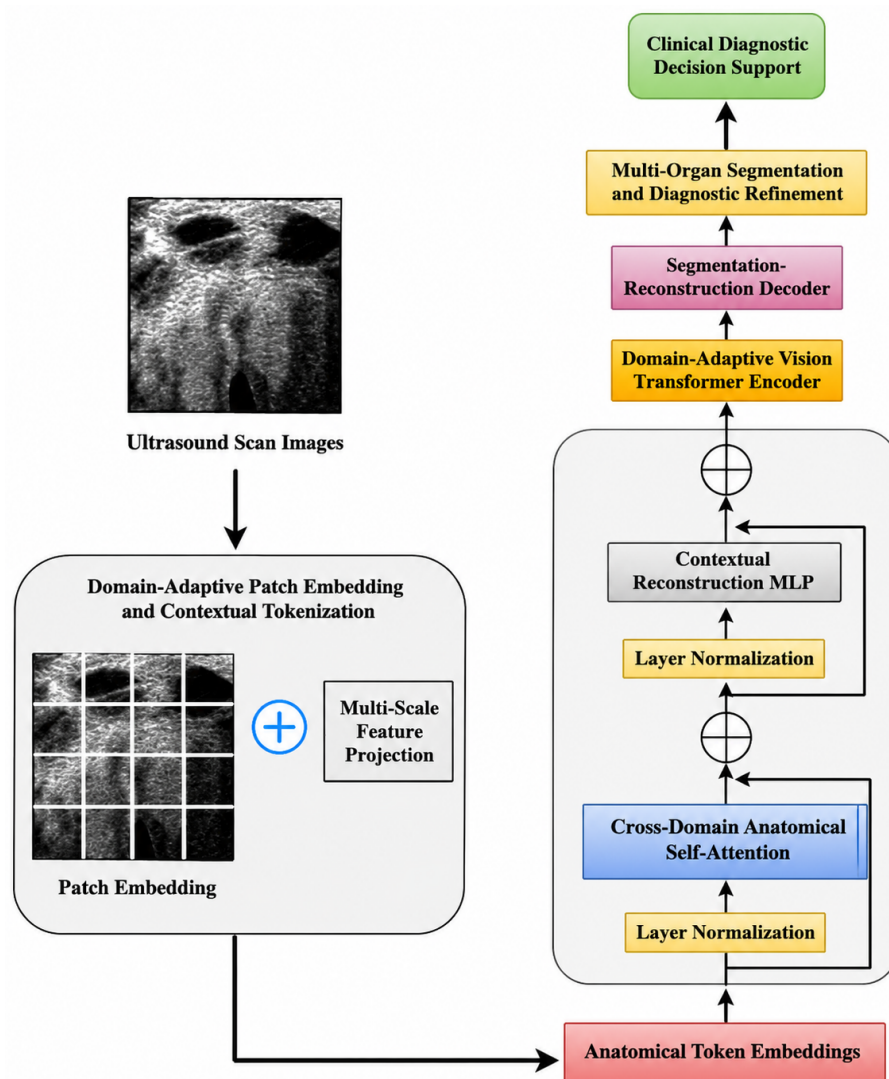


Fig. 2: Domain- Adaptive Vision Transformer Autoencoder Architecture

$$\mathcal{L}_{seg} = 1 - \frac{2 \sum (M \cdot M_{gt})}{\sum M + \sum M_{gt}} + \sum_{c=1}^C M_{gt}^{(c)} \log M^{(c)} \quad (13)$$

Where M_{gt} is the ground truth mask and $M^{(c)}$ denotes predicted probability for class c . In this study, simulated images serve as dense supervision and real images are used to validate the anatomical realism in a limited and critical.

To reduce the domain gap between simulated (Z_s) and real (Z_r) features, a domain alignment constraint is introduced using Maximum Mean Discrepancy (MMD) calculated as in Equation (14).

$$\mathcal{L}_{da} = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(Z_s^i) - \frac{1}{n_r} \sum_{j=1}^{n_r} \phi(Z_r^j) \right\|^2 \quad (14)$$

Where Z_s^i, Z_r^j are latent features from simulated and real images, n_s, n_r denote sample sizes, and $\phi(\cdot)$ is a kernel mapping function. This makes the model learn domain-invariant embeddings, meaning that features extracted from simulated data are valid for real-world images. The overall loss is defined as in Equation (15).

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{rec} + \beta \mathcal{L}_{seg} + \gamma \mathcal{L}_{da} \quad (15)$$

Where α, β, γ are weighting coefficients (e.g., $\alpha = 1.0, \beta = 2.0, \gamma = 0.5$). It is a multi-objective optimization that allows the model to learn visual reconstruction, clinical segmentation and domain generalization simultaneously.

Within this study, DA-ViT-AE performs three critical functions Feature Extraction which provides latent embeddings Z as state input to the agent, Segmentation

Feedback which supplies masks used in reward computation, and Image Quality Estimation in which reconstruction error acts as a proxy for image quality. This is a highly integrated solution allowing the agent to take decisions based on the features and tasks of the digital twin environment during interaction.

The main novelty of this component is that it is multi-functional and integrated into an autonomous system. The proposed method is different from traditional transformer-based segmentation model, which learns a shared latent space between simulated and real domain, integrates domain adaptation into the perception model without relying on additional alignment modules, and enables both a perception model and an evaluation model for agent training.

This transforms the model from a passive analyser into an active component of a closed-loop imaging system and allowing for better generalization and adaptation in real world clinical situations.

3.5 Feature Extraction for Agent State Representation

The feature extraction process is not an intermediate passive step but a key connection between perception (DA-ViT-AE) and decision-making (H-ReAct agent) in this study. The features extracted are clearly defined to reflect the imaging environment's state, allowing the agent to perform informed and adaptive actions in the digital twin context.

Each preprocessed ultrasound image $I_t \in \mathbb{R}^{H \times W}$ is passed through the trained encoder of the DA-ViT-AE model to obtain a compact latent representation given as in Equation (16).

$$Z_t = \mathcal{E}_\theta(I_t), Z_t \in \mathbb{R}^d \quad (16)$$

Where Z_t is the latent embedding at time step t , d is the embedding dimension (set to $d = 512$), and $\mathcal{E}_\theta$ represents the trained encoder.

This embedding enables the agent to interpret the context of the underlying image rather than interpret the image based on the raw pixel values. The segmentation head generates the probability maps of each class represented as in Equation (17).

$$P_t(x, y, c) = \mathcal{S}_\psi(Z_t) \quad (17)$$

Where $P_t(x, y, c) \in [0, 1]$ is the probability of pixel (x, y) belonging to class c and $c \in \{1, 2, \dots, C\}$, with C representing organ classes. A confidence score is calculated using entropy, to quantify prediction reliability expressed as in Equation (18).

$$C_t = -\frac{1}{HW} \sum_{x=1}^H \sum_{y=1}^W \sum_{c=1}^C P_t(x, y, c) \log P_t(x, y, c) \quad (18)$$

Where C_t is the average uncertainty (lower values indicate higher confidence). This provides a quantitative measurement of the certainty about segmentation, needed to guide the agent decisions. The final agent state S_t is constructed by combining feature embeddings and confidence metrics expressed as in Equation (19).

$$S_t = [Z_t \parallel C_t \parallel Q_t] \quad (19)$$

Where $\parallel$ denotes concatenation, and Q_t is an image quality indicator derived from reconstruction error represented as in Equation (20).

$$Q_t = \|I_t - \hat{I}_t\|_2^2 \quad (20)$$

Here $\hat{I}_t$ is the reconstructed image from the decoder. Thus, the state includes Semantic features (Z_t), Prediction confidence (C_t), and Image quality (Q_t).

In the context of this study, feature extraction directly influences agent behaviour: if C_t is high (low confidence), the agent performs refinement actions, if Q_t is high (poor reconstruction), the agent adjusts acquisition (zoom/shift), and Z_t ensures decisions are based on anatomical understanding, not pixel noise.

This enables the agent to interact with the digital twin environment and learn adaptively to enhance image quality and segmentation accuracy. This novelty of this feature extraction is the transformation of learned representations into an actionable state space. In contrast with traditional pipelines that rely solely on features for prediction, this approach combines confidence-aware and quality-aware features into the state, allows for closed-loop interaction between perception and control, and converts segmentation results into signals that drive decisions.

This step is crucial for achieving the transition from static image analysis to autonomous optimization of image, which is a significant contribution of the proposed framework.

3.6 Agent Model Development: Hierarchical ReAct Agent

The proposed H-ReAct agent aims at interacting with the DT environment to enable autonomous optimization of the imaging. In contrast to the traditional models, the agent is actively improving the image acquisition process by executing actions and refining their reasoning iteratively. In this study, the agent uses the feature representations provided by the perception model to make decisions to enhance segmentation accuracy and image quality.

The agent operates in a hierarchical manner with two levels High-Level Planner and Low-Level Executor. High-Level Planner generates a sub-goal G_t based on the current state represented as in Equation (21).

$$G_t = \pi_h(S_t) \quad (21)$$

Where G_t represents the objective (e.g., “refine liver boundary”), π_t is the high-level policy. Low-Level Executor executes actions to achieve the sub-goal given as in Equation (22).

$$A_t = \pi_l(S_t, G_t) \quad (22)$$

Where A_t is the action at time step t , and π_l is the low-level policy. This decomposition allows the system to distinguish between what to do (reasoning) and how to do it (execution). The agent’s action is on the state specified as in Equation (23).

$$S_t = [Z_t \parallel C_t \parallel Q_t] \quad (23)$$

Where $Z_t \in \mathbb{R}^d$ is the latent features, C_t is the segmentation confidence, and Q_t is the reconstruction-based quality score. The Action Space is represented as in Equation (24).

$$A_t = \{\Delta x, \Delta y, \alpha, \beta\} \quad (24)$$

Where $\Delta x, \Delta y \in [-10, 10]$ denotes the spatial shifts, $\alpha \in [0.8, 1.2]$ is the zoom factor, and $\beta \in [0.5, 1.0]$ denotes the region refinement. The State Transition is represented as in Equation (25).

$$S_{t+1} = \mathcal{E}(S_t, A_t) \quad (25)$$

Where $\mathcal{E}(\cdot)$ is the digital twin environment, it produces updated image and features. The agent is trained using a task-driven reward computed as in Equation (26).

$$R_t = \lambda_1 D_t + \lambda_2 S_t^{img} - \lambda_3 \|A_t\| \quad (26)$$

Where $D_t = \text{Dice}(M_t, M_{gt})$ is the segmentation accuracy, $S_t^{img} = \text{SSIM}(I_t, I_{ref})$ is the image quality, $\|A_t\|$ is the action magnitude penalty, and $\lambda_1, \lambda_2, \lambda_3$ are the weights (e.g., 2.0, 1.0, 0.1). This ensures that the agent learns to improve segmentation, enhance image clarity, and avoid unnecessary movements. The agent optimizes its policy using cumulative reward calculated as in Equation (27).

$$J(\pi) = \mathbb{E} \left[\sum_{t=0}^T \gamma^t R_t \right] \quad (27)$$

Where $\gamma \in (0, 1)$ denotes the discount factor (e.g., $\gamma = 0.95$) and T is the episode length. Policy parameters are updated to maximize $J(\pi)$, enabling the agent to learn optimal imaging strategies over time.

In this study, the agent demonstrates the following behaviour: if segmentation confidence C_t is low then refine region ($\beta \downarrow$), if image quality Q_t is poor then adjust zoom ($\alpha \uparrow$), and if anatomical region is misaligned then apply spatial shifts ($\Delta x, \Delta y$). This can enable the agent to simulate real probe adjustments, without human intervention, and enhance the diagnostic relevance.

This is the main feature that makes this component unique, as it allows autonomous and closed loop imaging

optimization. Unlike existing methods, it introduces a hierarchical reasoning-action framework for medical imaging, integrates perception outputs directly into decision-making, learns how to improve image acquisition, not just analyse images, and operates within a digital twin environment for safe, scalable training

This transforms medical imaging systems from passive predictor towards the intelligent agents, which address a critical gap in existing research.

3.7 Autonomous Tool-Oriented Imaging Workflow Execution within the Digital Twin Environment

In the proposed framework, tool integration is presented as a process that is dynamic, driven by an agent, and involves selectively invokes the perception model (DA-ViT-AE) and post-processing modules to refine the imaging outputs in an iterative manner. In contrast to standard pipelines, where segmentation and analysis is performed once, the present study proposes a closed-loop execution process with tools executed repeatedly depending on the decision made based on the state. At each time step t , the agent invokes the segmentation tool using the current image I_t represented as in Equation (28).

$$M_t = \mathcal{S}_\psi(\mathcal{E}_\theta(I_t)) \quad (28)$$

Where $\mathcal{E}_\theta$ is the encoder of DA-ViT-AE, $\mathcal{S}_\psi$ is the segmentation head, and $M_t \in [0, 1]^{H \times W \times C}$ is the predicted segmentation mask. This step is important because it enables pixel-by-pixel delineation of the anatomy and is the primary input for the evaluation and subsequent decision-making process. A confidence score is used to measure the reliability of segmentation computed as in Equation (29).

$$\Gamma_t = 1 - \frac{1}{HW} \sum_{x=1}^H \sum_{y=1}^W H(P_t(x, y)) \quad (29)$$

Where $\Gamma_t \in [0, 1]$ is the confidence score (higher is better), $H(P_t(x, y)) = -\sum_{c=1}^C P_t(x, y, c) \log P_t(x, y, c)$ is the pixel-wise entropy, and $P_t(x, y, c)$ is the class probability. The formulation ensures that predictions with uncertainty (high entropy) result in further refinement. If the confidence Γ_t falls below a predefined threshold τ (e.g., $\tau = 0.7$), the agent initiates refinement expressed as in Equation (30).

$$A_t = \pi(S_t), \text{ if } \Gamma_t < \tau \quad (30)$$

Where π is the agent policy and A_t is the action applied (shift, zoom, crop). The updated image is represented as in Equation (31).

$$I_{t+1} = T(I_t, A_t) \quad (31)$$

This loop continues until $I_t \geq \tau_{ort} = T_{max}$, where T_{max} is the maximum number of iterations (set to 10–15). Once a satisfactory segmentation is accomplished, quantitative measurements are computed as in Equation (32).

$$Q^{(c)} = \sum_{x=1}^H \sum_{y=1}^W 1(M_t(x, y) = c) \quad (32)$$

Where $Q^{(c)}$ is the area of organ class c and $1(\cdot)$ is the indicator function. For normalized measurements, it is mathematically as in Equation (33).

$$\tilde{Q}^{(c)} = \frac{Q^{(c)}}{H \times W} \quad (33)$$

This provides clinically relevant measures such as organ size and spatial distribution. The integrated workflow is as follows: initial segmentation on simulated or real ultrasound images, entropy-based scoring of confidence, agent refining the image based on digital twin actions when the confidence is low, iteration until reaching the optimum segmentation and output of segmentation masks and quantitative measurements.

In this study, the integrated workflow operates as follows; initial segmentation is performed on simulated or real ultrasound images, confidence is evaluated using entropy-based scoring, the agent refines the image via digital twin actions if uncertainty is high, the process iterates until optimal segmentation is achieved and the final outputs include segmentation masks and quantitative measurements. This allows for each decision to directly increase the diagnostic output instead of using a single pass prediction.

The novelty of this component is to convert the conventional usage of tools into a process of iterative execution, and making decisions during the process. In particular, this study includes using confidence-driven instead of static inference for invoking tools, closed loop refinement between the segmentation and the action, quantitative analysis directly embedded in the agent workflow, and connecting the perception output with clinically meaningful measurements.

3.8 Feedback-Driven Co-Evolution and Adaptive Refinement

The proposed framework introduces an adaptive and co-evolutionary loop in the system to allow for its continuous adaptation. This study proposes a model update mechanism based on feedback which directly affects the perception model and the policy of the agent, unlike the static learning approaches. This is especially important for ultrasound imaging where ambiguity in anatomy and variability in the domain necessitate expert validation.

At iteration t , the system produces a segmentation output M_t . A clinical expert provides a corrected annotation M_t^* ,

which is used to compute a feedback signal given as in Equation (34).

$$\Delta M_t = M_t^* - M_t \quad (34)$$

Where M_t is the predicted segmentation mask, M_t^* is the expert-corrected mask, and ΔM_t is the correction signal. This difference represents the discrepancies at the pixel level, which can be used to refine the model. The perception model is updated with a feedback weighted loss represented as in equation (35).

$$\mathcal{L}_{fb} = \sum_{x,y} w(x,y) \cdot \ell(M_t(x,y), M_t^*(x,y)) \quad (35)$$

Where $\ell(\cdot)$ is the segmentation loss (Dice + Cross-Entropy), $w(x,y) = 1 + \lambda \mid \Delta M_t(x,y) \mid$ is the feedback weight, and λ is the scaling factor (e.g., $\lambda = 2.0$). This formulation provides greater weight to regions with larger errors, and hence ensures targeted learning. The agent maintains an experience memory buffer $\mathcal{B}$, updated as in Equation (36).

$$\mathcal{B} \leftarrow \mathcal{B} \cup (S_t, A_t, R_t, S_{t+1}, \Delta M_t) \quad (36)$$

Where S_t is the current state, A_t is the action taken, R_t is the reward and ΔM_t is the feedback signal. This enables the agent to learn from corrected outcomes, in order to make better decisions in the future. An objective that is enhanced with feedback is used to update the agent policy given as in Equation (37).

$$J_{fb}(\pi) = \mathbb{E} \left[\sum_t \gamma^t (R_t + \eta \cdot F_t) \right] \quad (37)$$

Where γ is the discount factor, η is the feedback weighting parameter (e.g., $\eta = 0.5$) and $F_t = \|\Delta M_t\|_1$ is the feedback magnitude. This supports the agent reduce future corrections based on their preferences as experts.

In this study, the feedback loop operates as follows; the agent generates segmentation and performs refinement actions, the expert reviews outputs only when confidence is below a threshold, Corrections are incorporated into both Perception model (via weighted loss) and Agent policy (via memory and reward adjustment), and the system updates incrementally without full retraining. This allows rapid adaptation (in minutes) to new anatomical differences or other imaging conditions unseen.

The novelty is in the ability to co-evolve the perception and decision-making aspects in real-time with feedback. Unlike existing approaches that; use feedback only for offline retraining, do not integrate feedback into agent behaviour. This study presents a novel integration of pixel-level feedback into loss functions, a joint model/agent policy update procedure, a memory-driven learning scheme from expert advices and rapid adaptation without retraining the model in large-scale. It enables the system to become a continuous learning, clinically adaptive system, which is appropriate to deployment in dynamic and underserved healthcare environments

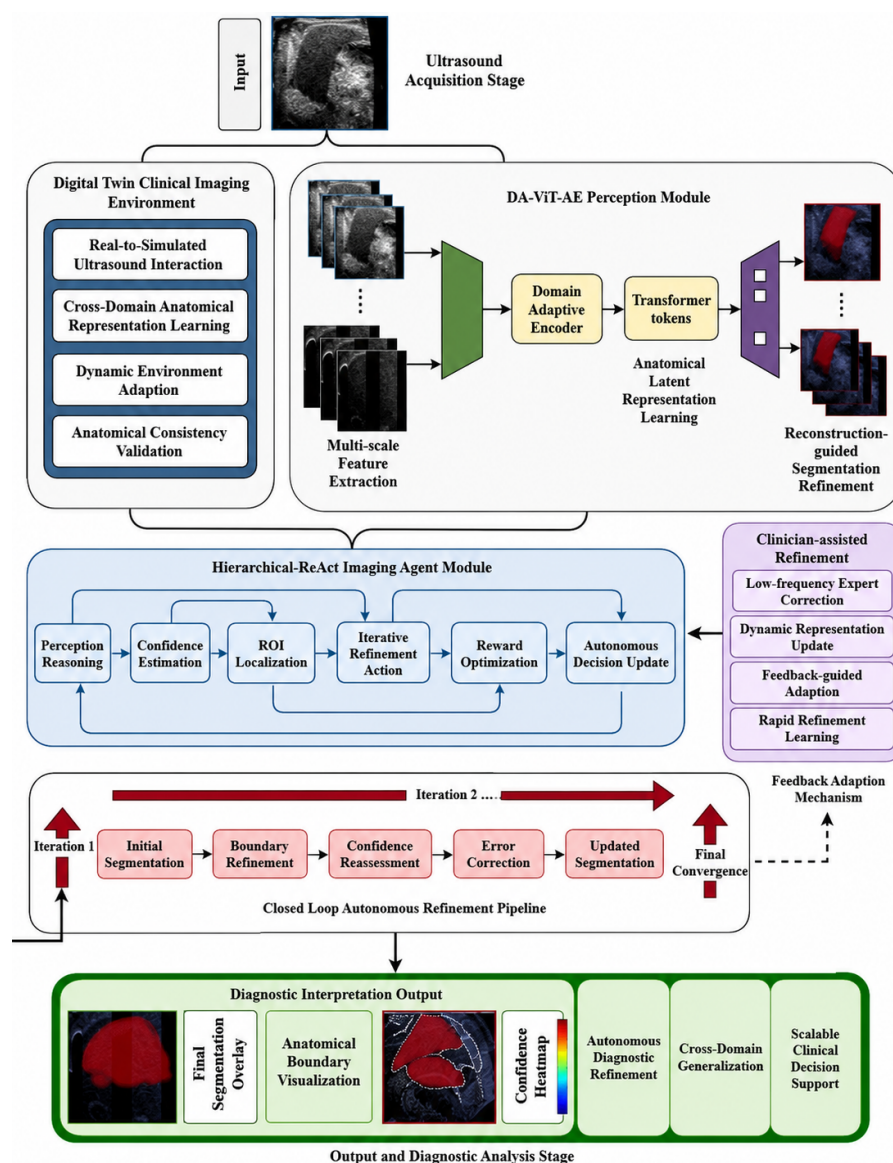


Fig. 3: Architecture of the Proposed Digital Twin-Driven Autonomous Imaging Framework with DA-ViT-AE and H-ReAct Agent

The overall architecture of the proposed digital twin-based autonomous imaging system with the DA-ViT-AE perception model, H-ReAct reasoning agent, feedback adaptation mechanism and closed-loop refinement pipeline is shown in Figure 3. It facilitates updating the perception-action interaction, optimizing the segmentation, refining the decision-making process using confidence measures, and adapting the diagnosis process across different ultrasound imaging conditions.

The proposed framework is explicitly designed to be scalable for deployment in underserved clinical environments to ensure practical applicability in the resource-constrained environments. Because of the use of

simulated data within the digital twin environment, large scale annotated data is significantly reduced, allowing models to be effectively trained on limited data. In addition, the hierarchical agent reduces the need of experts to intervene in the optimization of image acquisition and refinement process. A lightweight resource feature representation and controlled action space allows computational efficiency in the system, making it suitable for low-resource hardware settings. Furthermore, the feedback mechanism allows for rapid adoption to new clinical scenarios with limited feedback, enabling continuous learning in environments where experts are not always available. This design provides a

collaborative approach to the provision of cost-effective, scalable and adaptive diagnostic services in a range of and under-resourced healthcare environments.

Algorithm 1: Digital Twin-Driven Autonomous Imaging Agent Training

Require: Simulated ultrasound dataset D_s
Require: Real ultrasound dataset D_r
Require: Ground-truth segmentation masks M_{gt}
Require: Maximum iterations T_{max}
Require: Confidence threshold τ
Require: Model parameters θ, ϕ, ψ
Require: Agent policy parameters π_h, π_l
Ensure: Trained DA-ViT-AE model
Ensure: Optimized agent policy
Ensure: Final segmentation M and diagnostic metrics

- 1: **Step 1: Load Dataset and Initialization**
- 2: Initialize perception model parameters (θ, ϕ, ψ)
- 3: Initialize agent policies π_h and π_l
- 4: Initialize experience memory buffer $\mathcal{B}$
- 5: **Step 2: Data Preprocessing**
- 6: Preprocess D_s and D_r to obtain images I and masks M_{gt}
- 7: **Step 3: Train Perception Model**
- 8: **for** each training epoch **do**
- 9: **for** each image I in $D_s \cup D_r$ **do**
- 10: $Z \leftarrow \mathcal{E}_\theta(I)$
- 11: $\hat{I} \leftarrow \mathcal{D}_\phi(Z)$
- 12: $M \leftarrow \mathcal{S}_\psi(Z)$
- 13: Compute reconstruction loss $\mathcal{L}_{rec}$
- 14: Compute segmentation loss $\mathcal{L}_{seg}$
- 15: Compute domain adaptation loss $\mathcal{L}_{da}$
- 16: Update (θ, ϕ, ψ) using
 $\mathcal{L}_{total} = \alpha\mathcal{L}_{rec} + \beta\mathcal{L}_{seg} + \gamma\mathcal{L}_{da}$
- 17: **end for**
- 18: **end for**
- 19: **Step 4: Autonomous Agent Training**
- 20: **for** each episode **do**
- 21: Initialize state S_0 using image I_0
- 22: $t \leftarrow 0$
- 23: **while** $t < T_{max}$ **do**
- 24: $Z_t \leftarrow \mathcal{E}_\theta(I_t)$
- 25: $M_t \leftarrow \mathcal{S}_\psi(Z_t)$
- 26: Compute confidence C_t
- 27: Compute quality Q_t
- 28: $S_t \leftarrow [Z_t \| C_t \| Q_t]$
- 29: $G_t \leftarrow \pi_h(S_t)$
- 30: $A_t \leftarrow \pi_l(S_t, G_t)$
- 31: $I_{t+1} \leftarrow T(I_t, A_t)$
- 32: Compute Dice score D_t
- 33: Compute SSIM S_t^{img}
- 34: $R_t \leftarrow \lambda_1 D_t + \lambda_2 S_t^{img} - \lambda_3 \|A_t\|$
- 35: Add (S_t, A_t, R_t, S_{t+1}) to $\mathcal{B}$
- 36: **if** $C_t \geq \tau$ **then**
- 37: **break**
- 38: **end if**
- 39: $t \leftarrow t + 1$
- 40: **end while**
- 41: **end for**
- 42: **Step 5: Feedback Mechanism**
- 43: **for** selected samples **do**

- 44: Obtain expert correction M^*
- 45: $\Delta M \leftarrow M^* - M$
- 46: Update perception model using feedback-weighted loss
- 47: Update agent policy using feedback signal
- 48: **end for**
- 49: **Step 6: Final Inference**
- 50: **for** each test image **do**
- 51: Perform iterative agent-based refinement
- 52: Generate final segmentation M
- 53: Compute quantitative diagnostic metrics
- 54: **end for**
- 55: **return** Trained model, optimized policy, and final outputs

The training of an autonomous imaging system with a DT is described in the Algorithm 1. It starts with data pre-processing and training of the perception model with reconstruction, segmentation and domain adaptation tasks. Then a hierarchical agent repeatedly interacts the environment with state-action-reward cycles to optimize imaging results. It involves conditional loops, confidence-based stopping criteria and adaptive feedback, allowing continuous learning, adaptive optimization, and enhanced diagnostic performance in various imaging situations.

Figure 4 represents the overall workflow and decision-making processes involved in the proposed system, showcasing data transfer, decision nodes, and iterations among the modules. It starts with input acquisition and preprocessing, then it is followed by training a perception model and optimizing the agent in a digital twin environment. Execution is guided by conditional branches, based on thresholds of confidence or feedback available. With a closed-loop design, the structure is composed of refinement, evaluation and adaptive feedback that provide adaptive learning and accurate and scalable diagnostic outputs.

The proposed study is unique from existing approaches in introducing a closed-loop framework that combines perception, decision-making and learning in a digital twin environment. The conventional approach to medical image analysis considers the imaging as a static process; this study redefines medical imaging as an interactive one, allowing the system to adaptively optimize the imaging procedure. The sim-to-real domain adaptive transformer autoencoder guarantees robust performance despite variations across the different domains, and the hierarchical agent provides structured reasoning for sequential decision making, which is missing in most medical imaging pipelines. Moreover, with the combination of confidence-aware feature representation and iterative feature invocation, a dynamic feedback mechanism is realized that can continuously improve the output. One of the main novelties is the co-evolution paradigm, featuring direct integration of feedback into both model and policy updates and enabling rapid adaptation without retraining. Overall, this study advances beyond existing studies by creating scalable, autonomous, and explainable imaging workflows that can

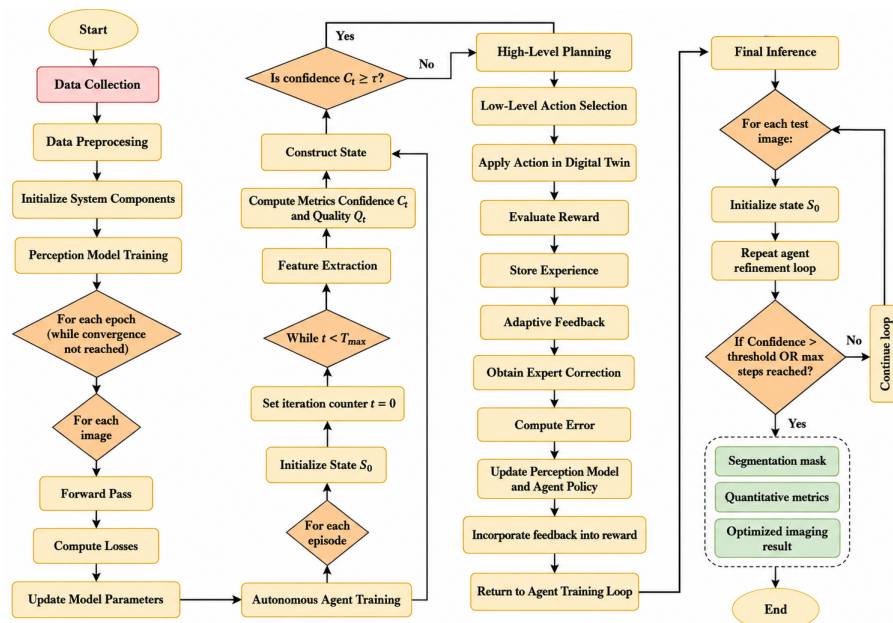


Fig. 4: Flowchart of the Proposed System

be used for deployment in resource-limited and underserved clinical settings.

4 Result and Discussion

The proposed framework showed the effective integration of a digital twin simulation, domain adaptive perception, and optimization with an autonomous agent for ultrasound image analysis. The interaction-driven imaging strategies were successfully learned by the system by iterative state-action refinement in the simulated environment, which led to better anatomical localization and segmentation consistency for both simulated and real ultrasound scans. The hierarchical agent effectively leveraged the confidence-aware feedback to make imaging decisions, and the domain-adaptive transformer autoencoder produced strong representations across different imaging conditions. Moreover, the feedback-driven co-evolution mechanism facilitated rapid adaptation to unseen scenarios without requiring too much involvement of experts, highlighting the framework's potential for scalable and intelligent clinical assistance in resource-limited clinical settings. The simulation parameters used in this study is provided in Table 2.

4.1 Dataset Characteristics and Cross-Domain Learning Analysis

Data characteristics and cross-domain learning ability of the proposed framework were analysed to assess the

Table 2: Simulation Parameters

Parameter	Value
Input Image Resolution	224×224
Batch Size	16
Learning Rate	1×10^{-4}
Optimizer	AdamW
Number of Training Epochs	100
Embedding Dimension (d)	512
Maximum Agent Iterations (T_{max})	15
Confidence Threshold (τ)	0.7
Discount Factor (γ)	0.95
Replay Memory Size	10^4 samples
Reconstruction Loss Weight (α)	1.0
Segmentation Loss Weight (β)	2.0
Domain Adaptation Weight (γ_d)	0.5
Reward Weight (λ_1)	2.0
Reward Weight (λ_2)	1.0
Reward Penalty (λ_3)	0.1
Horizontal Shift Range (Δx)	$[-10, 10]$ pixels
Vertical Shift Range (Δy)	$[-10, 10]$ pixels
Zoom Scaling Factor (α_z)	$[0.8, 1.2]$
Region Refinement Ratio (β_r)	$[0.5, 1.0]$

effectiveness of the augmentation with digital twin and limited-data adaptation. The increased training samples

from simulation and augmentation increased anatomical variability and imaging diversity. Moreover, it ensured consistency of representation learning and segmentation across simulated and real ultrasound domain while demonstrating a strong capability for cross domain generalization.

Table 3: Effective Training Sample Expansion Through Data Augmentation and Digital Twin Simulation

Dataset Category	Original Images	Augmented Samples	Effective Samples Used
Simulated Ultrasound Training Images	740	5,920	6,660
Real Ultrasound Training Images	40	640	680
Validation Images	103	206	309
Testing Images	104	208	312
Total Effective Dataset	987	6,974	7,961

Table 3 presents the effective expansion of ultrasound training samples through augmentation and digital twin simulation strategies. Anatomical representation learning was improved with the increased sample diversity, and the overfitting problem was reduced when the amount of annotated data was limited. Moreover, the enlarged dataset enhanced autonomous segmentation refinement in the varied ultrasound imaging conditions to increase domain generalization ability.

Figure 5 presents the simulated-real data distribution of the proposed framework. The results showed that the distributions of ultrasound images were aligned effectively and the adaptation consistency was stable between the imaging domains. In addition, the augmentation process-based learning enhanced the domain-invariance and robustness of autonomous perception under various clinical imaging conditions.

In Figure 6, the effect of effective sample expansion on the stability of segmentation and consistency of autonomous learning is shown. The findings showed that Dice score increased gradually, uncertainty was reduced and generalization ability was improved as the number of training samples increased. In addition, the effectiveness of digital twin-based augmentation in the context of limited annotated ultrasound data was shown to be a good improvement in the stability of cross-domain learning.

4.2 Digital Twin Environment Validation

The digital twin environment was used to simulate realistic ultrasound acquisition variations by controlling the spatial changes and state changes of the digital twin through interaction. The generated environment was anatomically consistent and allowed for the movement of the probes to be simulated, which enabled stable autonomous agent learning. The iterative transformations, including spatial shifting, zoom adaptation and region refinement, generated clinically coherent imaging sequences, which were closely resembled real acquisition

behaviour. Moreover, the environment enabled a continuous update of the perception model, which led to the adaptation of the image quality and refinement of the segmentation. The results confirm the applicability of digital twins' approach for the scalable and autonomous analysis of medical images in low-resource clinical environments.

Table 4: Digital Twin Synchronization Performance

Metric	Value
Twin Synchronization Latency	82 ms
Environment Update Cycle	0.41 s
Feedback Propagation Delay	96 ms
Average Refinement Response Time	0.28 s
Sim-to-Real Adaptation Efficiency	94.6%
State Alignment Consistency	96.2%

Table 4 shows the efficiency of the synchronization and interaction responsiveness of the proposed autonomous imaging framework using the digital twin. The results obtained showed low synchronization latency, rapid propagation of feedback and stable response time of the refinement in iterative perception-action interaction. In addition, the high adaptation efficiency and state alignment consistency validated the consistency of the interaction between the simulator and the real system and the robust autonomous diagnostic capability to refine the interaction in the closed loop.

The entire digital twin-based sim-to-real adaptation pipeline of the proposed autonomous imaging framework is shown in figure 7. It gradually incorporates simulated ultrasound generation, domain adaptation representation learning, self-adapting iterative refinement and feedback-driven co-evolution, driven by stable clinical prediction. Additionally, the mechanism of closed-loop interaction always optimizes the reliability of segmentation and consistency of perception and action. Based on the results achieved, the proposed pipeline proved to be an effectively support scalable and adaptive ultrasound diagnostic analysis in heterogeneous and low resource clinical environments.

In the autonomous diagnostic refinement, the efficiency of the synchronization and the performance of adaptive interaction in the proposed digital twin ecosystem are shown in Figure 8. The findings showed decrease in synchronization latency, stable feedback propagation, efficient response time for refinement and improved co-evolutionary adaptation behaviour. In addition, it ensured scalable inference stability and environment adaptation speed over the iterative perception-action learning cycles.

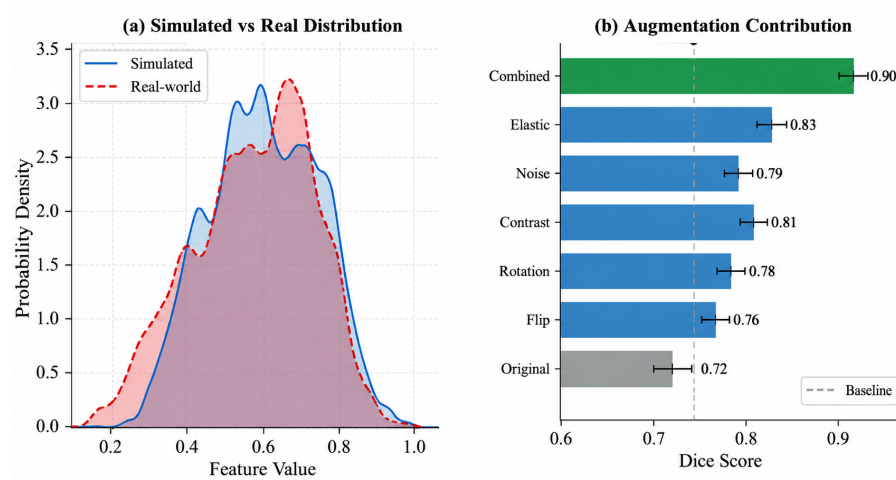


Fig. 5: Data Distribution Analysis

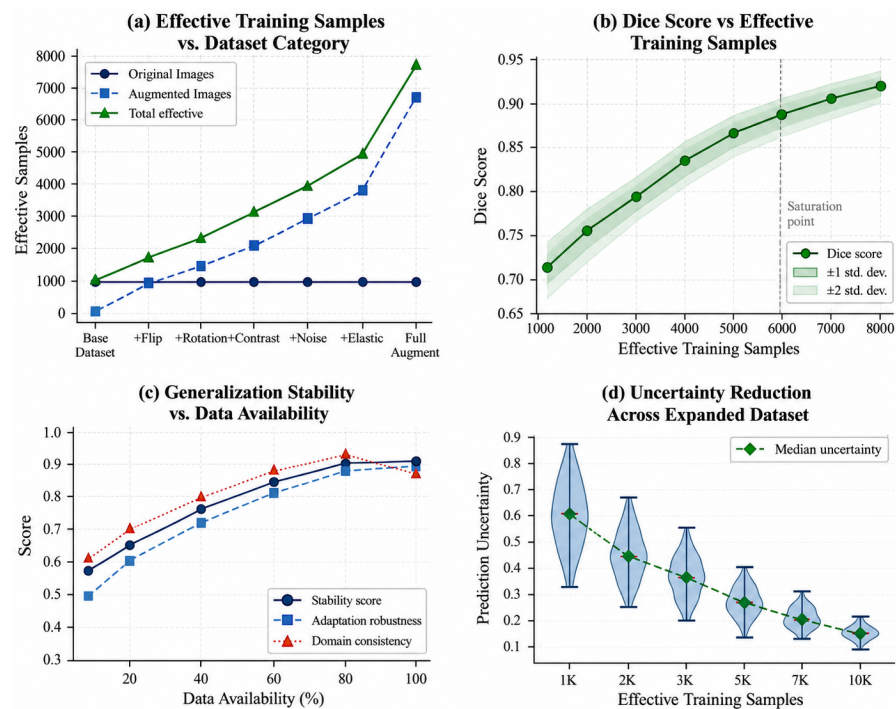


Fig. 6: Effective Sample Expansion and Learning Stability Analysis

4.3 DA-ViT-AE Perception Model Performance

The proposed DA-ViT-AE has shown excellent segmentation, reconstruction, and cross-domain representation learning performances in simulated and real ultrasound images. The model was effective in preserving anatomical structures in the reconstruction process and maintaining high segmentation consistency under the condition of heterogeneous imaging. The

integrated domain adaptation mechanism minimized the differences in feature distribution between the simulated and real ultrasound domains, leading to enhanced stability of representation and generalization performance. Moreover, the contextual learning mechanism based on transformer improved the features extraction results in the autonomous imaging refinement and maintained the stability of the perception-driven interaction in the digital twin environment.

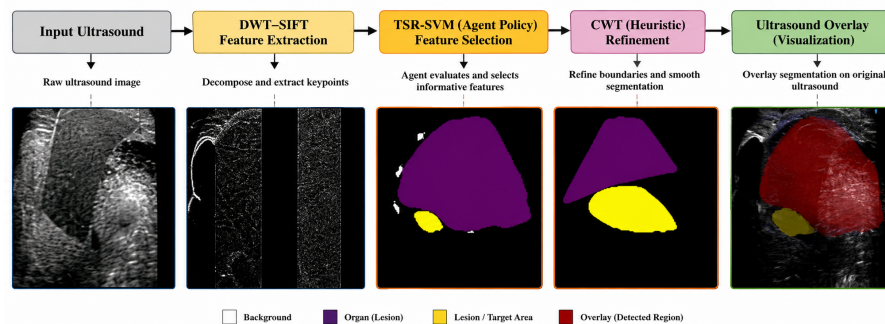


Fig. 7: Digital Twin Sim-to-Real Domain Adaptation Pipeline

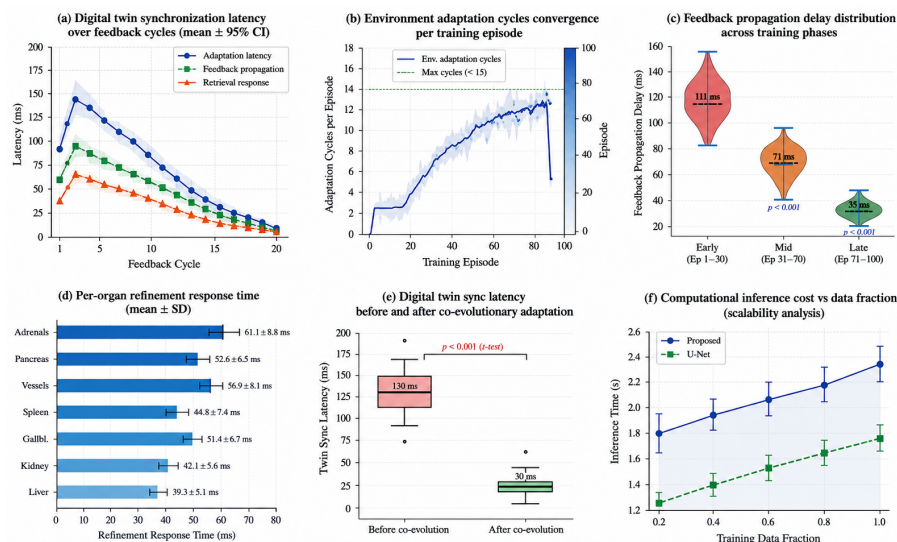


Fig. 8: Digital Twin Synchronization Efficiency

Table 5: Segmentation Performance of the Proposed Framework Across Anatomical Structures

Anatomical Structure	Dice Score (%)	IoU Score (%)	Precision (%)	Recall (%)
Liver	97.84	95.21	98.36	97.92
Kidney	96.92	93.84	97.65	97.18
Gallbladder	95.74	91.68	96.81	96.22
Spleen	96.35	92.74	97.28	96.81
Vessels	94.88	90.15	96.42	95.36
Pancreas	95.63	91.42	96.94	96.08
Adrenals	94.31	89.37	95.58	94.82
Overall Average	96.52	93.41	97.72	97.20

Table 5 shows the segmentation results of the proposed autonomous imaging framework using the digital twin for various abdominal anatomical structures. The results indicate good segmentation ability under different structural complexity and ultrasound appearance conditions. The proposed approach achieved excellent preservation of organ boundaries, fine anatomical details were localised and organ representation remained stable

both in large and small organs. In addition, the domain adaptive perception and iterative refinement by agents resulted in the enhancement of structural consistency and multi-organ segmentation performance, confirming the effectiveness of the proposed framework for scalable autonomous ultrasound image analysis.

During training, the proposed DA-ViT-AE framework shows its convergence behaviour as shown in figure 9. The reconstruction loss, segmentation loss, and domain adaptation loss gradually decreased and the convergence behaviour showed good stability over training epochs. Moreover, the iterative refinement capability was shown to be effective through the increased agent confidence scores. The results obtained show that the proposed framework was able to obtain stable optimization, robust cross-domain representation learning and enhance the autonomous segmentation consistency under heterogeneous ultrasound imaging conditions.

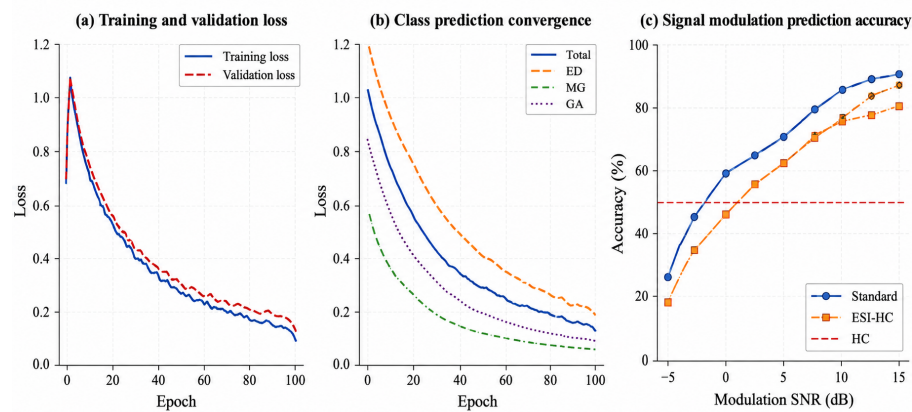


Fig. 9: DA-ViT-AE Training Loss and Component Loss Curves

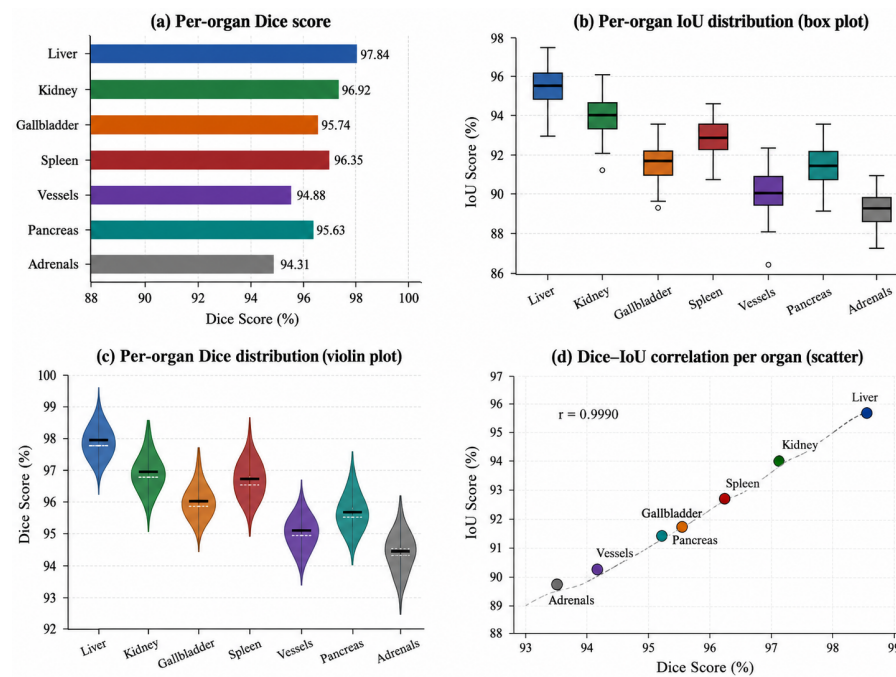


Fig. 10: Per-Organ Performance Distribution Analysis

Figure 10 presents the organ-wise segmentation performance distribution of the proposed framework using multiple statistical visualizations. The results showed stable Dice and IoU distributions with lower variability for various anatomical structures within the abdomen. Moreover, Dice-IoU correlation analysis showed consistent representation of the structures and reliability of segmentation. According to the obtained results, the suggested framework has succeeded in maintaining anatomical continuity and high segmentation consistency in complex ultrasound imaging environment and various organ structures.

Figure 11 shows the convergence behaviour of the domain adaptation mechanism and latent feature alignment capability of the proposed framework. The decreasing MMD loss and improved t-SNE feature clustering demonstrated effective reduction of simulated-to-real domain discrepancy. Moreover, the aligned latent representations were able to capture discriminative anatomical information across diverse ultrasound domains. Based on the obtained results, the proposed DA-ViT-AE framework successfully achieved robust domain-invariant feature learning and stable cross-domain generalization performance.

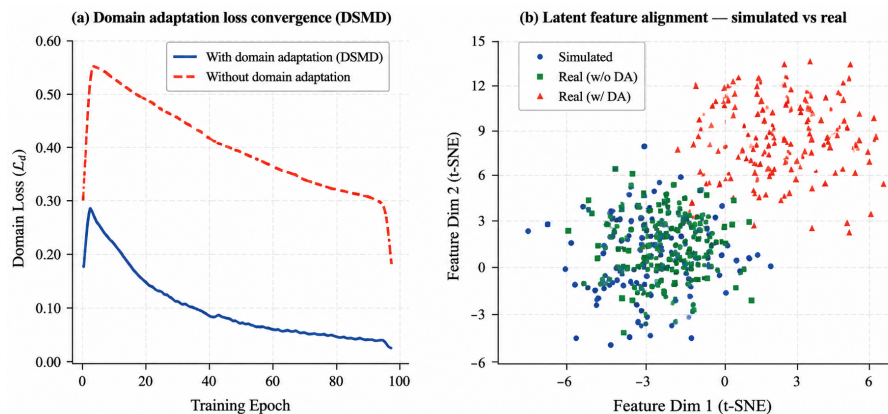


Fig. 11: Domain Adaptation and Latent Alignment Analysis

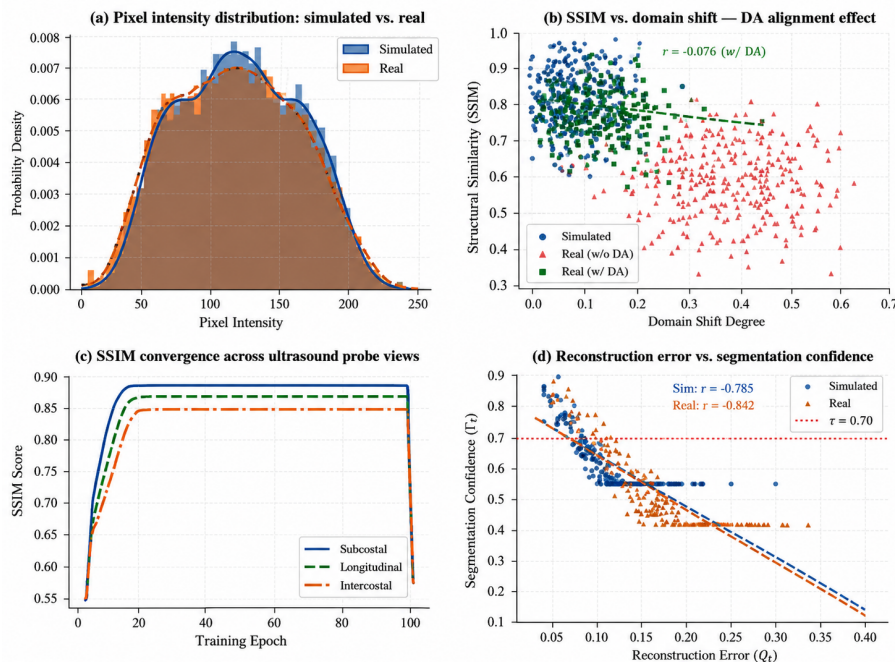


Fig. 12: Sim-to-Real Transfer and Reconstruction Analysis

To demonstrate the capability of the proposed DA-ViT-AE framework in sim-to-real adaptation, statistical distribution analysis, reconstruction consistency analysis, and domain shift analysis are performed as shown in Figure 12. The results showed that the structural similarity and the stability of the reconstruction were enhanced under various probe views and imaging domains. Moreover, the reconstruction analysis with confidence showed that the discrepancy between features was minimized during domain transfer. The results obtained showed that the proposed framework was successful in reducing the representation mismatch between the simulated and real environments, and in

enhancing the diagnostic stability across domains in the digital twin.

4.4 H-ReAct Agent Learning Behaviour Analysis

This autonomous agent showed stable learning behaviour based on a hierarchical reasoning and action process within the digital twin environment by repeating the perception-action process. For each episode, the agent iteratively adapted the correct imaging decisions, meaningfully transformed spatial properties, and refined

regions while gaining confidence. Effective capability for policy learning and adaptive refinement was confirmed by a convergence of the rewards and reduced corrective iterations. In addition, the agent was able to maintain coherent imaging states while minimizing the unnecessary transformations in the environment, which suggests that the agent explored the environment efficiently and stable decision-making. These results confirm the potential of the proposed H-ReAct framework for intelligent and autonomous optimization of ultrasound imaging.

The proposed H-ReAct framework is shown to be capable of autonomous optimization of the ultrasound segmentation, in an iterative manner, as shown in figure 13. This result in significant improvement in the localization of boundaries, their structural integrity, and consistency of predictions with confidence measures at the various stages of progressive refinement from initial segmentation to convergence. Furthermore, activation map visualization showed anatomically focused reasoning behaviour, and stable perception-action refinement under heterogeneous ultrasound imaging conditions.

Table 6: Average Reward and Convergence Analysis of the Autonomous Agent

Training Episode Range	Average Reward	Average Convergence Steps	Learning Observation
1–20	42.8	13	Initial exploration with unstable action selection
21–40	58.6	11	Improved policy adaptation and refinement stability
41–60	71.4	9	Enhanced perception-action consistency
61–80	82.7	7	Stable optimization and reduced redundant actions
81–100	91.3	5	Efficient convergence with robust imaging decisions

Table 6 shows the learning behaviour of the hierarchical autonomous agent over iterative training. The trend of increasing rewards and decreasing number of convergence steps reflects that the proposed framework achieved the optimization of perception-action interactions, which allowed for stable imaging refinement and efficient autonomous perception decision making in the digital twin environment.

Table 7: Iterative Refinement Performance Across Agent Iterations

Iteration	Dice	IoU	Boundary Error	Confidence
1	88.42	82.15	6.84	0.71
3	91.86	86.42	4.12	0.81
5	94.63	89.74	2.85	0.89
Final	96.52	93.41	1.26	0.96

Table 7 shows the progressive autonomous refinement of the proposed H-ReAct framework for multiple iterations of interaction. These obtained results showed

the continuous enhancement of the Dice score, IoU and confidence levels of the results, and the boundary error was significantly reduced during iterative refinement. Additionally, closed-loop autonomous decision learning enhanced segmentation consistency and facilitated stable perception-action optimization in the framework.

Table 8: Reinforcement Learning Stability Metrics

Metric	Value
Average Reward Gain	48.5%
Policy Entropy Reduction	63.2%
Mean Convergence Epoch	72
Action Stability Score	0.91
Exploration-to-Exploitation Ratio	0.38

Table 8 shows the RL stability for the proposed H-ReAct autonomous imaging agent in the iterative interaction learning process. The results showed consistent reward optimization, decreasing policy entropy, and improved consistency in the actions during training. In addition, the optimized exploration-to-exploitation ratio enabled efficient convergence and autonomous decision refinement under different ultrasound imaging scenarios.

Figure 14 presents the reinforcement learning convergence characteristics and policy evolution behaviour of the proposed H-ReAct autonomous imaging agent. The results showed that the reward optimization was stable, policy entropy was decreased, convergence efficiency was increased, and the learning of the Q-value during training was stable. In addition, the reward-Dice correlation was high, further verifying the effectiveness of perception-action coupling and autonomous reasoning-driven segmentation refinement capability.

Figure 15 shows the autonomous learning properties of the digital twin-based H-ReAct framework in the iterative interaction process. Reward convergence and Dice stabilization curves showed an efficient policy optimization and progressive improvement in segmentation. Moreover, the reward-Dice correlation analysis confirmed a meaningful interaction between perception-action, and the capability of adaptative refinement. According to the results, the proposed autonomous reasoning framework was able to learn stable imaging refinement strategies and enhance the diagnostic consistency across different ultrasound domains.

Figure 16 shows the learning of the proposed H-ReAct framework in terms of temporal learning dynamics and policy evolution behaviour over time while interacting with the environment iteratively. The results of the convergence analyses revealed consistent decrease in domain discrepancy, decrease in policy entropy, and

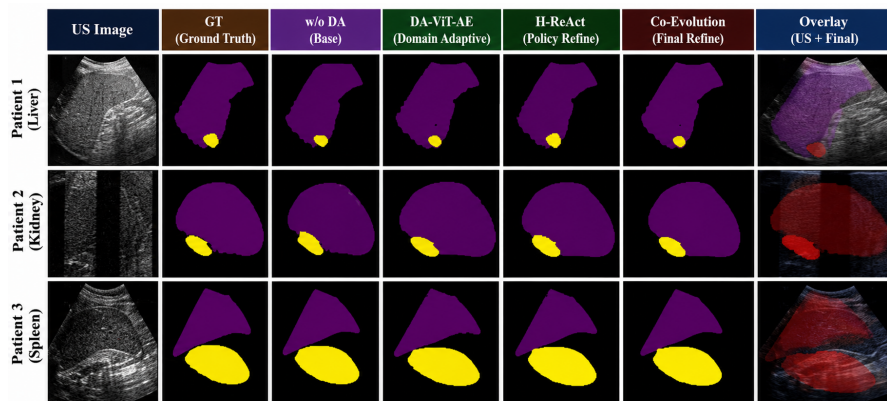


Fig. 13: H-ReAct Agent Iterative Segmentation Refinement

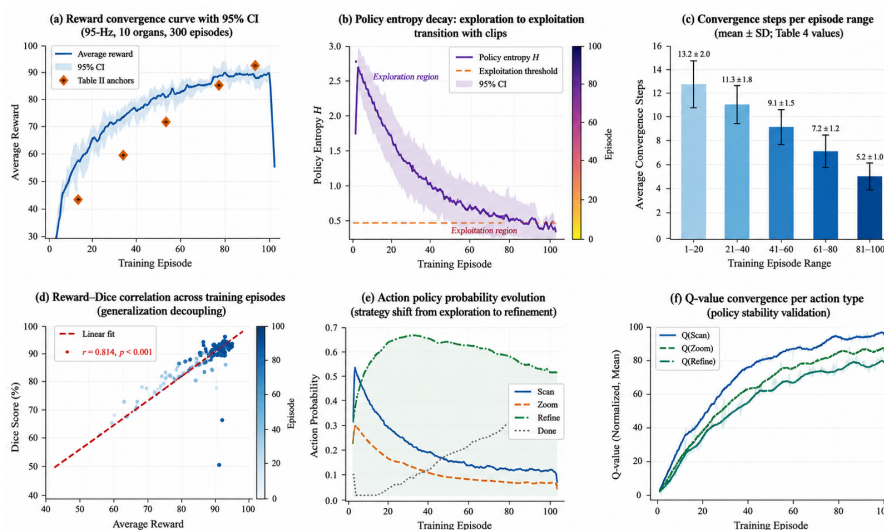


Fig. 14: Reinforcement Learning Convergence Analysis

increase in consistency of actions during training episodes. In addition, the action frequency topographical maps and episode length histograms confirmed efficient autonomous exploration and optimized decision activity. The results obtained confirmed the effectiveness of the proposed framework in enabling stable convergence of reinforcement learning and adaptive autonomous diagnostic refinement.

4.5 Closed-Loop System Performance

A complete closed-loop system was established that integrates the digital twin environment, the DA-ViT-AE perception model, and the H-ReAct autonomous agent, which achieved stable and adaptive analysis of ultrasound imaging under various conditions. The modules for

perception, decision-making and refinement allowed for the continual refinement of imaging quality and consistency of the segmentation throughout the iterative processing. The introduced framework effectively eliminated inconsistencies between the different structures, while enhanced the anatomical representation and decision-making behaviour with confidence. Moreover, the closed-loop interaction mechanism allowed them to varying imaging conditions dynamically without a need for repeated manual interaction, thus showing the effectiveness of the proposed framework for scalable and autonomous diagnostic analysis.

Figure 17 shows qualitative segmentation visualizations of the proposed framework for various ultrasound imaging samples. Results showed effective progressive refinement from raw ultrasound images to accurate segmentation masks and overlay representations that were

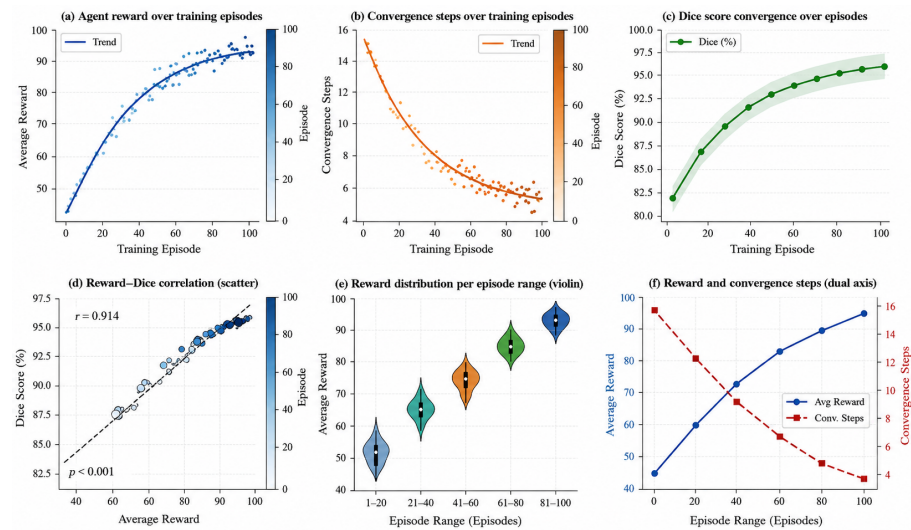


Fig. 15: H-ReAct Agent Learning Behavior Analysis

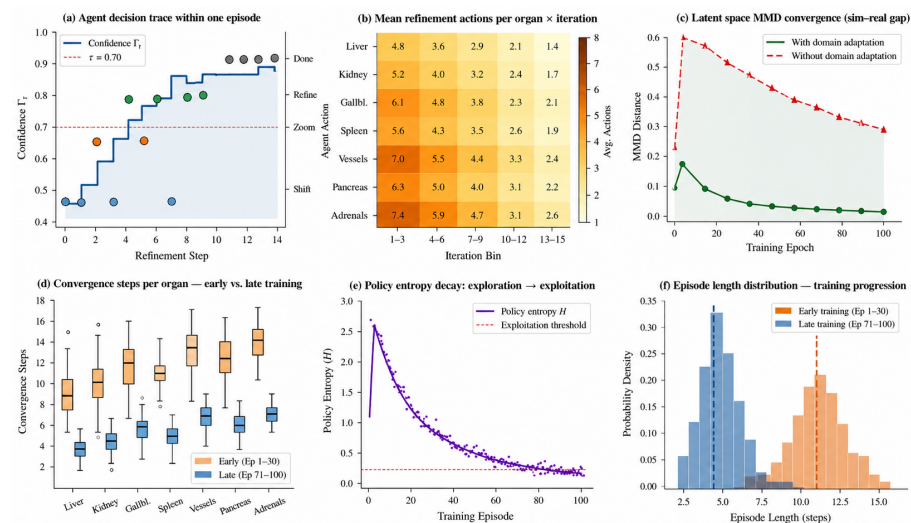


Fig. 16: Temporal Agent Learning Dynamics

consistent with anatomy. Moreover, the framework succeeded in maintaining the structural continuity, and reducing the probability of segmentation ambiguity of complex organ boundaries. Based on the results achieved, the proposed autonomous perception-action framework was effective in achieving reliable and interpretable ultrasound diagnostic refinement in the presence of heterogeneous imaging conditions.

Multiple heatmap-based analyses are used to visualize the performance characteristics of the proposed framework shown as in Figure 18. The results exhibited consistently high segmentation performance regardless of the methods, organs, ablation configurations, and domain

generalization scenarios. Moreover, heat maps distributions showed good stability in cross domain adaptability and preservation of structure. The proposed framework achieved reliable autonomous diagnostic refinement and ensured consistent segmentation results under different ultrasound imaging conditions, as demonstrated by the results obtained.

4.6 Feedback Adaptation Analysis

The feedback-driven adaptive refinement mechanism allowed correction and continuous optimization of the

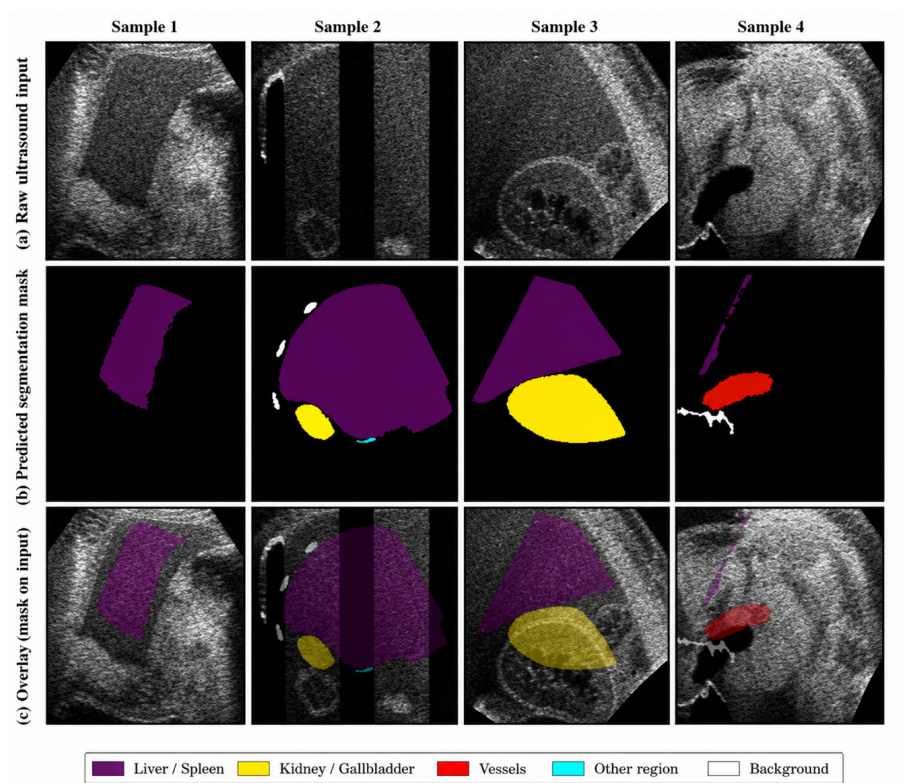


Fig. 17: Qualitative Segmentation Visualization

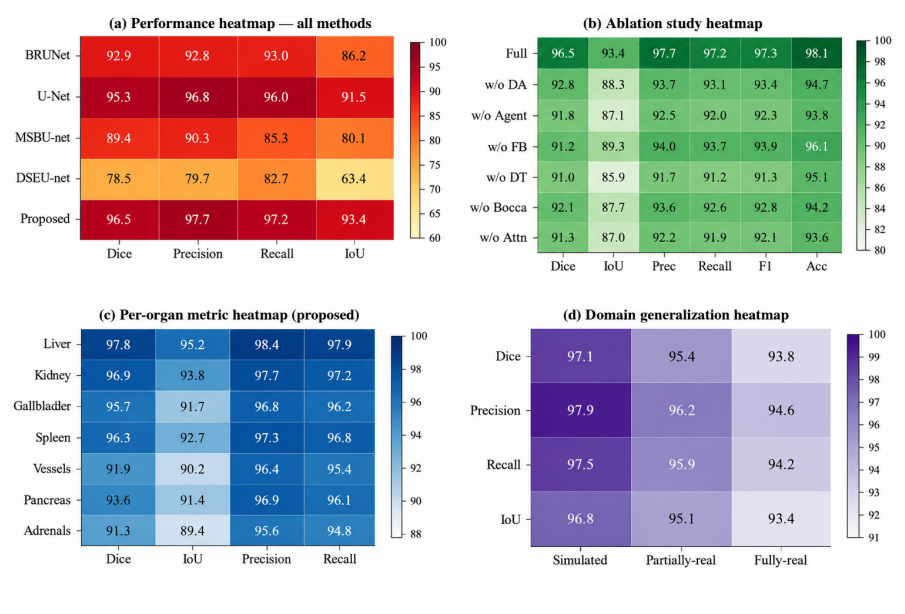


Fig. 18: Heatmap-Based Performance Analysis

autonomous imaging framework during unseen imaging conditions. The system effectively incorporated low-frequency expert feedback to increase the robustness of the segmentation and to diminish the structural ambiguities and confidence-aware decision behaviour without requiring extensive retraining. The adaptive refinement procedure gradually reduced the inconsistencies of images during iterative updates while retaining autonomous operation. Moreover, the co-evolution mechanism enabled efficient adaptation to various ultrasound imaging variations, with minimal supervisory input, showcasing the feasibility of the proposed framework for its potential implementation in real-world clinical applications and continual diagnostic learning.

Figure 19 shows the effect of the co-evolutionary feedback on segmentation accuracy at various steps of iterative refinement. Every adaptive feedback design outperformed non-feedback configurations in terms of Dice, precision and recall. Additionally, the progressive performance enhancement also confirmed the rapid adaption capability for unseen imaging circumstances. According to the results, the proposed co-evolving feedback mechanism effectively contributed to improving the stability of autonomous refinement and the representation learning of the diagnosis in the dynamic ultrasound environment.

The effectiveness of the proposed feedback-driven co-evolution mechanism in adaptive segmentation refinement is illustrated in figure 20. The two-axis analysis indicated that there was a gradual increase in Dice score with correctly executed interaction efficiency. In addition, the scatter analysis of the magnitude of the correction confirmed that adaptive feedback played a significant role in terms of enhancing the segmentation and maintaining the structural consistency. According to the results, the proposed framework was effective in enhancing the autonomous diagnostic capability under dynamically changing ultrasound imaging conditions with the use of low-frequency supervisory interaction.

4.7 Scalability and Low-Resource Evaluation

The proposed framework showed good scalability and stable performance under low-resource and limited data conditions prevalent in underserved clinical environments. The domain-adaptive perception model achieved effective cross-domain generalization in the limited number of annotated samples across multiple ultrasound imaging domains, and the autonomous agent optimized the imaging refinement without continuous expert supervision. Moreover, the digital twin environment allowed to facilitate cost-efficient simulation-based training and adaptive interaction learning without relying on massive clinical data. The framework also ensured consistent performance in terms of inference speed and segmentation accuracy even with

limited computational resources, demonstrating its potential for scalable deployment in healthcare environments.

Table 9: Cross-Domain Generalization Performance

Domain Transfer	Dice	IoU	Adaptation Stability
Synthetic → Real	95.84	91.22	0.93
Real → Synthetic	96.13	92.04	0.95
Hospital A → Hospital B	94.92	89.63	0.91
Unseen Probe Distribution	93.88	88.54	0.89

Table 9 shows the cross-domain generalization performance of the proposed framework in various domain transfer settings. The results obtained showed that segmentation performance was consistent and strong adaptation consistency was achieved under synthetic-real, inter-hospital and unseen probe distributions. The framework was successful in reducing the domain discrepancy and demonstrating robust cross domain learning and reliable diagnostic adaptability.

The scalability and limited data adaptation capability of the proposed framework are shown by multiple statistical analysis in Figure 21. The results showed good segmentation performance, low variability, and good generalization in limited training scenarios. Moreover, the framework had a consistent accuracy range under different data availability conditions. The developed digital twin based autonomous imaging framework proved to be suitable for scalable deployment and adaptive diagnostic analysis in low-resource clinical settings based on the acquired results.

Figure 22 shows the proposed framework's latent clinical representation learning and scalability features as a function of imaging complexity and data fractions. The organ-wise latent clustering showed better separation of anatomical features, and the inference complexity analysis resulted in the stability of the computational efficiency in heterogeneous imaging conditions. Moreover, the analysis for cross-organ confusion showed lesser ambiguous structure in segmentation. The results obtained indicated that the proposed framework achieved scalable and robust autonomous diagnostic analysis, which is suitable for low-resource environments in clinical settings.

4.8 Ablation Study

The ablation study was performed to assess the contribution of each of the main components of the proposed digital twin-based autonomous imaging system. The impact of omitting each of the adaptations, autonomous refinement, feedback-driven co-evolution and digital twin interaction modules was independently investigated. The results showed that the domain-adaptive

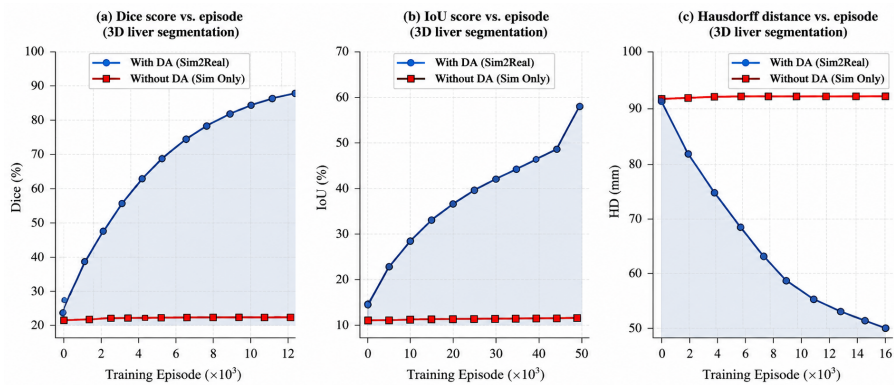


Fig. 19: Feedback Update Iteration Across Various Metrics

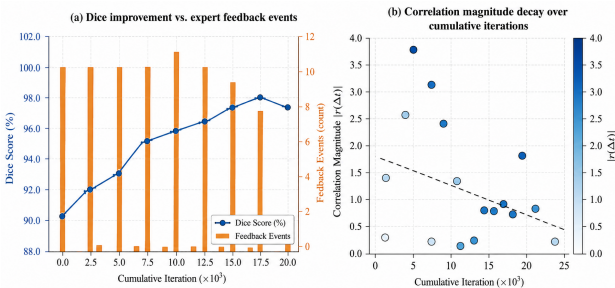


Fig. 20: Feedback Co-Evolution and Correction Analysis

Table 10: Ablation Study

Configuration	Dice Score (%)	IoU Score (%)	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Specificity (%)	Domain Generalization Score (%)
Full Proposed Framework	96.52	93.41	97.72	97.20	97.46	98.13	98.45	95.84
Without Domain Adaptation	92.84	88.35	93.65	93.12	93.38	94.72	95.01	88.42
Without H-ReAct Agent	91.76	87.12	92.48	92.05	92.26	93.84	94.20	89.03
Without Feedback Adaptation	93.18	89.46	94.01	93.74	93.87	95.08	95.44	90.26
Without Digital Twin Environment	90.95	85.94	91.67	91.24	91.45	93.12	93.65	87.74
Without Reconstruction Learning	92.13	87.68	93.02	92.61	92.81	94.15	94.52	88.91
Without Attention Mechanism	91.48	86.95	92.25	91.93	92.09	93.56	93.88	88.33
Without Sequential Refinement	92.44	88.02	93.11	92.87	92.99	94.38	94.81	89.46
DA-ViT-AE Only	94.32	90.84	95.10	94.78	94.94	96.25	96.58	91.72
H-ReAct Only	89.72	84.38	90.54	90.03	90.28	92.65	92.94	86.57

transformer was essential for having segmentation consistency between simulated and real ultrasound domains, while the deletion of the hierarchical autonomous agent decreased the ability to improve the segmentation through iterations and optimization. Similarly, removing the feedback adaptation mechanism reduced the generalisation performance under unseen imaging conditions. The results validate the combined contribution of perception, autonomous refinement and adaptive learning modules, which leads to more superior performance of the proposed framework.

Table 10 shows the complete ablation analysis on the proposed framework with various evaluation metrics. The overall framework showed the best performance in terms of segmentation accuracy, reconstruction stability, domain generalization, and autonomous refinement capability. The performance degradation by removing individual components proves the benefit of the proposed ultrasound diagnostic framework with domain-adaptive perception, hierarchical autonomous interaction, feedback-driven co-evolution and digital twin simulation to achieve better robustness and scalability.

4.9 Comparative Analysis

The proposed approach of autonomous imaging using digital twins showed improved results, when compared to the conventional ML, DL, and hybrid segmentation methods with heterogeneous ultrasound imaging conditions. The proposed framework combined domain-adaptive representation learning, autonomous perception-action refinement and feedback-driven co-evolution, as opposed to the existing approaches that concentrate on optimizing the segmentation at the end of the training process. This ensures more anatomical consistency, domain generalization, and more stable segmentation refinement in both simulated and real imaging environments. Moreover, through iterative interactions, the hierarchical autonomous agent was able to optimize imaging decisions in a scalable, diagnostic adaptability, while also gaining greater robustness in resource-constrained and underserved clinical contexts. Table 11 and Figure 23 present comparative evaluation of the performance between the proposed method with existing state-of-the-art ultrasound segmentation techniques. The proposed autonomous imaging framework based on digital twin had the highest Dice

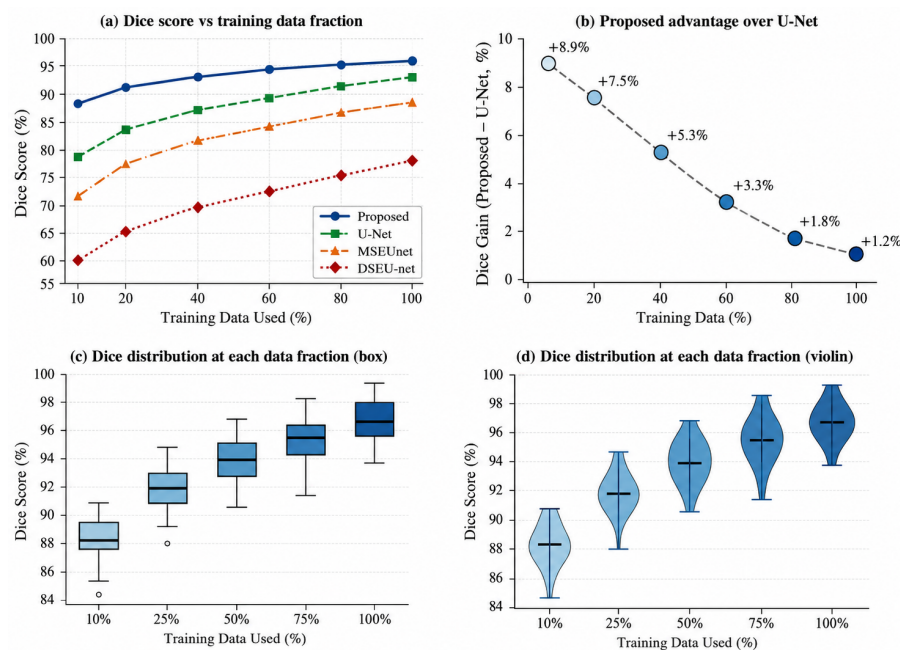


Fig. 21: Scalability and Low-Resource Evaluation

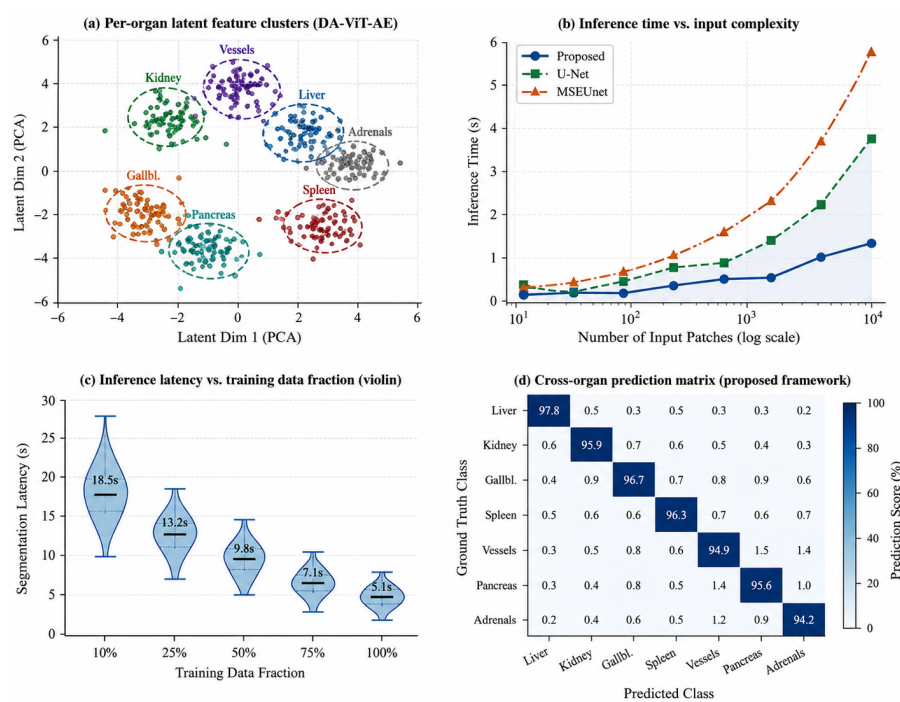
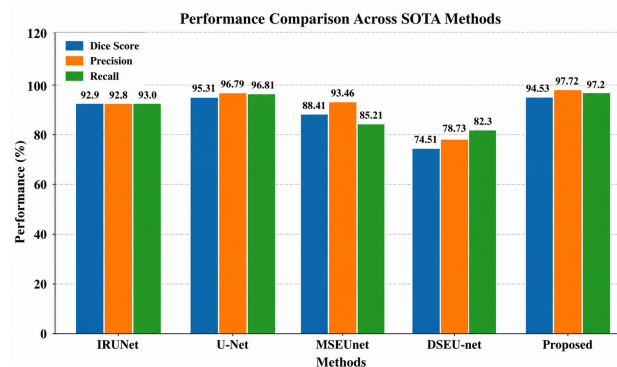
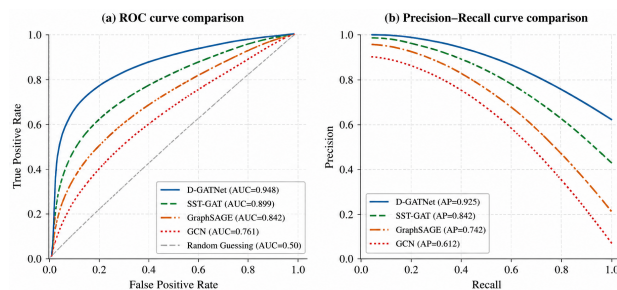


Fig. 22: Latent Clinical Representation and Scalability Analysis

Table 11: Performance Comparison Across SOTA Methods

Method	Dice Score	Precision	Recall
IRUNet [36]	92.90	92.80	93.00
U-Net [37]	95.32	96.79	96.01
MSEUNet [38]	89.41	93.46	85.27
DSEU-net [39]	78.51	79.73	82.70
Proposed Framework	96.52	97.72	97.20

**Fig. 23:** Performance Comparison Across Various Methods**Fig. 24:** ROC and Precision-Recall Curve Analysis

score, precision and recall among all the compared approaches, which showed better anatomical localization and segmentation consistency. The combination of the domain-adaptive transformer representation learning and the hierarchical autonomous refinement greatly improved the feature discrimination and boundary preservation. Moreover, the proposed framework showed a consistent performance under different heterogeneous ultrasound imaging conditions, proving its capacity to be effective for scalable and adaptive diagnostic analysis in complex and resource-limited clinical settings.

Figure 24 shows the receiver operating characteristic and precision-recall performance of the proposed framework and the existing baseline methods. The proposed framework resulted in better area under the curve

performance and consistent performance in precision-recall trade-offs at various segmentation thresholds. Moreover, the curves showed improvements in false-positive and false-negative predictions in a heterogeneous ultrasonic environment. From the results obtained, the proposed framework was found to be more effective in enhancing the diagnostic discrimination capability and segmentation reliability.

Table 12: Statistical Significance Analysis of the Proposed Framework Against Existing Methods

Comparison Method	Dice Score Difference (%)	p-value	95% Confidence Interval	Effect Size (Cohen's d)
Proposed Framework vs IRUNet	+3.62	0.0008	±1.84	1.27
Proposed Framework vs U-Net	+1.20	0.0031	±1.12	0.96
Proposed Framework vs MSEUNet	+7.11	0.00003	±2.36	1.84
Proposed Framework vs DSEU-net	+18.01	0.000001	±3.94	2.71

Table 12 displays the statistical significance validation results of the proposed digital twin autonomous imaging framework with the existing state-of-the-art ultrasound segmentation methods. The calculated p-values validated the proposed framework's statistically significant improvement for each of the comparative models. Additionally, the confidence interval and effect size analyses revealed high robustness and significant improvement in segmentation performance attributes from using the domain adaptive transformer learning, autonomous reasoning refinement and digital twin-based adaptive interaction mechanisms.

4.10 Discussion

The experimental results showed that the digital-twin based autonomous imaging system was able to achieve stable and accurate ultrasound diagnostic analysis under heterogeneous imaging conditions. By combining the DA-ViT-AE and the H-ReAct agent, the performance in terms of segmentation consistency, domain adaptation ability and cross domain generalization performance was greatly enhanced. The perception framework effectively reduced the discrepancy between simulated and real features, and the autonomous agent gradually improved the representations it used for image perception via closed-loop interaction and iterative refinement of decision. The results of the ablation analysis showed that each architectural component is important to the overall performance, especially the domain adaptation mechanism, the refinement process, and the learning process from the digital twin. The proposed system proved to be more effective in preserving structure, localization accuracy, and computational efficiency when compared to the existing state-of-the-art segmentation

and transformer-based systems. Moreover, the attention visualization and error localization analyses improved to interpret the results of the reinforcement learning algorithm, emphasizing the anatomically relevant areas and adaptive refinement processes when making a decision.

The proposed framework also showed good performance in the limited-data and low-resource scenarios, demonstrating its applicability in challenging clinical settings. The autonomous learning mechanism performed well in both synthetic and real ultrasound datasets, and across unseen ultrasound datasets, with less reliance on intensive expert supervision. Moreover, the feedback adaptation mechanism led to the rapid refinement in the dynamically changing imaging environment. Overall, the proposed framework has effectively established an adaptable and sustainable environment for autonomous medical imaging analysis from continuous perception-action interaction and adaptive digital twin-driven learning. The study had some limitations such as having relatively small annotated ultrasound datasets and limited pathological diversity. Additionally, the framework was tested largely in the context of abdominal ultrasound imaging, and may not be directly generalizable without further adaptation to other imaging modalities.

5 Conclusion and Future Work

This study proposed a framework of Digital Twin-Driven Autonomous Imaging Agent for the analysis of ultrasound diagnostics for underserved clinical environments with scalability. The proposed architecture integrated a Domain-Adaptive Vision Transformer Autoencoder and a Hierarchical ReAct-based autonomous agent to facilitate perception, adaptive reasoning, and iterative refinement in a digital twin environment, in a closed-loop perception. This architecture aimed to overcome key limitations in ultrasound imaging, lack of annotated data, and scalable intelligent diagnostic assistance in resource-limited clinical environments. The experimental results showed that the proposed framework effectively obtained stable segmentation results, enhanced anatomical localization, robust cross-domain generalization, and efficient autonomous refinement in both the simulated and real ultrasound domains. Digital twin interaction learning and adaptability based on feedback greatly improved the structural preservation and adaptable optimisation performance. Additionally, the framework exhibited a higher diagnostic reliability and computational efficiency compared to the state-of-the-art segmentation methods. In this study, attention localization analysis and iterative refinement visualization were also shown to have strong interpretability features, and autonomous decision behaviour evaluation was demonstrated. The main contribution of this study is the integration of domain-adaptive perception learning, hierarchical

reasoning, and DT based interaction in a single framework to develop an autonomous imaging ecosystem. Practically, this framework can be realized in a scalable and cost-efficient manner for intelligent ultrasound diagnostic assistance in low resource clinical environments. The validation of the study, however, is limited by the relatively small number of annotated datasets and the small pathological diversity, so that broad clinical generalization might be compromised.

Future research will include the need to extend the proposed framework towards multi-modal medical imaging environments and computed tomography, magnetic resonance imaging, real-time portable ultrasound systems. Additional works will explore how to design large-scale federated digital twin systems for privacy-preserving collaborative learning in geographically-distributed healthcare institutions. Continual self-supervised adaptation mechanisms will also be added in future works to enhance self-learning and generalization when pathological conditions are unseen. In addition, by incorporating the multimodal clinical metadata, explainable reasoning strategies and lightweight edge-deployment optimization, the framework can be further improved for practical use in real-world intelligent diagnostic systems in highly resource-constrained healthcare settings.

Acknowledgment

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/01/37082)

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

- [1] B. Abhisheka, S. K. Biswas, B. Purkayastha, D. Das and A. Escargueil, Recent trend in medical imaging modalities and their applications in disease diagnosis: A review, *Multimedia Tools and Applications* **83**(14) (2024) 43035–43070.
- [2] X. Li, L. Zhang, J. Yang and F. Teng, Role of artificial intelligence in medical image analysis: A review of current trends and future directions, *Journal of Medical and Biological Engineering* **44**(2) (2024) 231–243.
- [3] S. Gosnell, T. Ashuri, S. Gharai and A. A. Amiri Moghadam, Soft parallel robots for medical ultrasound imaging: A comprehensive review and future directions, *Biomedical Engineering Letters* **16**(2) (2026) 387–410.

- [4] S. J. Adams, B. Burbridge, L. Chatterson, P. Babyn and I. Mendez, A telerobotic ultrasound clinic model of ultrasound service delivery to improve access to imaging in rural and remote communities, *Journal of the American College of Radiology* **19**(1) (2022) 162–171.
- [5] M. Komatsu *et al.*, Establishment of high-precision ultrasound diagnosis methods based on the introduction of deep learning, *IEEE Reviews in Biomedical Engineering* (2025) 1–15.
- [6] C. Pirri, N. Pirri, V. Macchi, A. Porzionato, C. Stecco and R. De Caro, Intraoperative ultrasound in brain and spine surgery: Current applications, translational value and future perspectives, *NeuroSci* **6**(4) (2025) p. 113.
- [7] T. Mahmood, A. Rehman, T. Saba, L. Nadeem and S. A. O. Bahaj, Recent advancements and future prospects in active deep learning for medical image segmentation and classification, *IEEE Access* **11** (2023) 113623–113652.
- [8] B. Abhisheka, S. K. Biswas and B. Purkayastha, A comprehensive review on breast cancer detection, classification and segmentation using deep learning, *Archives of Computational Methods in Engineering* **30**(8) (2023) 5023–5052.
- [9] P.-H. Conze, G. Andrade-Miranda, V. K. Singh, V. Jaouen and D. Visvikis, Current and emerging trends in medical image segmentation with deep learning, *IEEE Transactions on Radiation and Plasma Medical Sciences* **7**(6) (2023) 545–569.
- [10] I. Galić, M. Habijan, H. Leventić and K. Romić, Machine learning empowering personalized medicine: A comprehensive review of medical image analysis methods, *Electronics* **12**(21) (2023) p. 4411.
- [11] J. Wang, L. Qiao, H. Lv and Z. Lv, Deep transfer learning-based multi-modal digital twins for enhancement and diagnostic analysis of brain mri image, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **20**(4) (2023) 2407–2419.
- [12] M. Tsuneki, Deep learning models in medical image analysis, *Journal of Oral Biosciences* **64**(3) (2022) 312–320.
- [13] M. Li, Y. Jiang, Y. Zhang and H. Zhu, Medical image analysis using deep learning algorithms, *Frontiers in Public Health* **11** (2023).
- [14] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen and F. Liu, Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches, *Bioengineering* **11**(10) (2024) p. 1034.
- [15] S. Shamshiri, H. Liu and I. Sohn, Adversarial robust image processing in medical digital twin, *Information Fusion* **115** (2025) p. 102728.
- [16] A. Rouhollahi *et al.*, Cardiovision: A fully automated deep learning package for medical image segmentation and reconstruction generating digital twins for patients with aortic stenosis, *Computerized Medical Imaging and Graphics* **109** (2023) p. 102289.
- [17] F. Ghribi and F. Hamdaoui, Innovative deep learning architectures for medical image diagnosis: A comprehensive review of convolutional, recurrent, and transformer models, *The Visual Computer* **41**(13) (2025) 11603–11628.
- [18] W. Yao, J. Bai, W. Liao, Y. Chen, M. Liu and Y. Xie, From cnn to transformer: A review of medical image segmentation models, *Journal of Imaging Informatics in Medicine* **37**(4) (2024) 1529–1547.
- [19] S. Aburass, O. Dorgham, J. Al Shaqsi, M. Abu Rumman and O. Al-Kadi, Vision transformers in medical imaging: A comprehensive review of advancements and applications across multiple diseases, *Journal of Imaging Informatics in Medicine* **38**(6) (2025) 3928–3971.
- [20] F. Yuan, Z. Zhang and Z. Fang, An effective cnn and transformer complementary network for medical image segmentation, *Pattern Recognition* **136** (2023) p. 109228.
- [21] S. Takahashi *et al.*, Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review, *Journal of Medical Systems* **48**(1) (2024) p. 84.
- [22] A. S. Panayides *et al.*, Position paper: Artificial intelligence in medical image analysis: Advances, clinical translation, and emerging frontiers, *IEEE Journal of Biomedical and Health Informatics* **30**(2) (2026) 1187–1202.
- [23] A. Khan *et al.*, A recent survey of vision transformers for medical image segmentation, *IEEE Access* **13** (2025) 191824–191849.
- [24] T. Moyaux, Y. Liu, G. Bouleux and V. Cheutet, An agent-based architecture of the digital twin for an emergency department, *Sustainability* **15**(4) (2023) p. 3412.
- [25] Y. Chong, N. Xie, X. Liu and S. Pan, P-transunet: An improved parallel network for medical image segmentation, *BMC Bioinformatics* **24**(1) (2023) p. 285.
- [26] C. Zhu *et al.*, Swin-net: A swin-transformer-based network combining with multi-scale features for segmentation of breast tumor ultrasound images, *Diagnostics* **14**(3) (2024) p. 269.
- [27] R. Avanzato, F. Beritelli, A. Lombardo and C. Ricci, Lung-dt: An ai-powered digital twin framework for thoracic health monitoring and diagnosis, *Sensors* **24**(3) (2024) p. 958.
- [28] Y. Song and N. Y. Chong, S-cyclegan: Semantic segmentation enhanced ct-ultrasound image-to-image translation for robotic ultrasonography, in *2024 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, 2024, pp. 115–120.
- [29] S. Vitale, J. I. Orlando, E. Iarussi, A. Díaz and I. Larrabide, Improving realism in abdominal ultrasound simulation combining a segmentation-guided loss and polar coordinates training, *Medical Physics* **52**(6) (2025) 4540–4556.
- [30] H. H. Mohammed *et al.*, Ultrasound intima-media complex (imc) segmentation using deep learning models, *Applied Sciences* **13**(8) (2023) p. 4821.
- [31] N. Ta, H. Chen, X. Liu and N. Jin, Let-net: Locally enhanced transformer network for medical image segmentation, *Multimedia Systems* **29**(6) (2023) 3847–3861.
- [32] J. Wu *et al.*, Boundary-aware convolutional attention network for liver segmentation in ultrasound images, *Scientific Reports* **14**(1) (2024) p. 21529.
- [33] J. Tagnamas, H. Ramadan, A. Yahyaouy and H. Tairi, Multi-task approach based on combined cnn-transformer for efficient segmentation and classification of breast tumors in ultrasound images, *Visual Computing for Industry, Biomedicine, and Art* **7**(1) (2024) p. 2.
- [34] H. Zhang, Q. Ma and Y. Chen, U structured network with three encoding paths for breast tumor segmentation, *Scientific Reports* **13**(1) (2023) p. 21597.
- [35] Kaggle, Us simulation & segmentation dataset <https://www.kaggle.com/datasets/ignaciorlando/ussimandsegm>, (2026), Accessed: May 06, 2026.

- [36] F. Hoorali, H. Khosravi and B. Moradi, Irunet for medical image segmentation, *Expert Systems with Applications* **191** (2022) p. 116399.
 - [37] S. D. Deb and R. K. Jha, Modified double u-net architecture for medical image segmentation, *IEEE Transactions on Radiation and Plasma Medical Sciences* **7**(2) (2023) 151–162.
 - [38] N. Ottakath *et al.*, Mseunet: Refined intima-media segmentation of the carotid artery based on a multi-scale approach using patch-wise dice loss, *Biomedical Signal Processing and Control* **100** (2025) p. 107077.
 - [39] G. Chen *et al.*, Dseu-net: A novel deep supervision seu-net for medical ultrasound image segmentation, *Expert Systems with Applications* **223** (2023) p. 119939.
-