AI Medicine
https://www.sciltp.com/journals/aim

Article

Deep Learning-Based Segmentation of the Brachial Plexus in Ultrasound Images: A Cross-Device Generalization Assessment

Xinlong Zhao [*] and Dingcheng Tian

College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110024, China
[*] Correspondence: 2301378@stu.neu.edu.cn

How To Cite: Zhao, X.; Tian, D. Deep Learning-Based Segmentation of the Brachial Plexus in Ultrasound Images: A Cross-Device Generalization Assessment. *AI Medicine* **2025**, *2*(2), 7. https://doi.org/10.53941/aim.2025.100007

Received: 4 August 2025
Revised: 11 November 2025
Accepted: 25 November 2025
Published: 2 December 2025

Abstract: Ultrasound-guided brachial plexus regional anesthesia is a commonly used clinical technique for upper limb surgery analgesia. It provides excellent efficacy and high safety. However, due to the presence of substantial noise in ultrasound image pixels and the complex anatomical structures, accurate identification of nerves heavily depends on the operator's experience. Precise nerve identification is crucial for patient recovery. Deep learning-based image segmentation can automatically identify the location of the brachial plexus in ultrasound images, thereby assisting clinicians in performing brachial plexus nerve blocks.In this study, we systematically compared the performance of three neural network architectures for brachial plexus segmentation in ultrasound images, including CNN-based, Transformer-based, and Mamba-based models. The experimental data come from ultrasound images acquired by three different devices (eSaote, Sonosite, and Butterfly). All models were trained on data from the eSaote device and tested on images from both eSaote and the Sonosite and Butterfly devices.The results indicate that on the eSaote device dataset, ConvUNeXt and UNet achieved the highest mean Intersection over Union (mIoU), with scores of 0.9027 and 0.9043, respectively. However, in cross-device testing, TransUNet and VMUNet exhibited better generalization ability. On the low-quality Butterfly test set, TransUNet maintained strong segmentation performance.In addition, the models showed some limitations in data dependency and cross-domain adaptability, and possible directions for improvement are suggested. This study can serve as a reference for selecting and optimizing ultrasound nerve segmentation models.

Keywords: ultrasound-guided regional anesthesia; brachial plexus segmentation; model generalization; cross-device robustness

1. Introduction

In recent years, Ultrasound-Guided Regional Anesthesia (UGRA) has been increasingly used in clinical practice as an alternative to General Anesthesia (GA). It offers good analgesic effects and shortens recovery time. Among the various UGRA techniques, the Brachial Plexus Block (BPB) is one of the most frequently applied methods for upper limb surgeries. In particular, the Supraclavicular Brachial Plexus Block (SCBP) is often used because it can reduce opioid use, shorten hospital stays, and promote patient recovery [1–3].

The safety and success of SCBP depend on the clinician's ability to identify nerve structures in ultrasound images. This is particularly difficult for less experienced anesthesiologists, as misidentification may increase the risk of nerve injury [4]. To improve UGRA procedures, researchers have begun using artificial intelligence (AI), especially deep learning (DL), for nerve recognition and segmentation. Convolutional neural network (CNN) models can extract neural features from ultrasound images, enabling real-time automatic nerve segmentation and more consistent SCBP performance [5].

Multiple neural network architectures have been applied to medical image segmentation tasks, including U-Net [6], VGG16-UNet [7], and Attention-UNet based on attention mechanisms [8]. In recent years, studies have also introduced Transformer-based architectures (e.g., Swin-UNet [9]) and state-space modeling networks such as VMUNet [10] and H-vmunet [11] to enhance modeling capabilities for long-range dependencies. Additionally, CNN variants optimized for feature extraction and spatial attention mechanisms (e.g., ConvUNeXt [12], DCSAU-Net [13]) have shown excellent segmentation performance. Other research has proposed neural networks that combine region-aware global modeling strategies to improve the accuracy and real-time performance of nerve segmentation in ultrasound images. This approach builds Region-aware Pyramid Aggregation (RPA) and Adaptive Pyramid Fusion (APF) modules to effectively reduce speckle noise interference and enable efficient context modeling, achieving outstanding performance on public nerve segmentation datasets [14]. Further studies have incorporated Recurrent Neural Networks (RNNs) into the U-Net structure to enhance contextual modeling of brachial plexus nerves in ultrasound images. By using auxiliary loss to address sample imbalance, these models have achieved significantly better segmentation performance than the original U-Net on public datasets [15].

In this study, we conduct a comprehensive evaluation of multiple advanced deep learning models on the task of nerve segmentation in SCBP ultrasound images. We propose a systematic video frame sampling strategy to improve the representativeness and diversity of training data. We also evaluate various network architectures including Transformer-based models (Swin-UNet, TransUnet), Mamba-based models (VMUNet, H-VMunet), and CNN-based optimized architectures across multiple ultrasound datasets. Evaluation metrics include standard segmentation performance indicators such as Dice coefficient, IoU, accuracy, recall, and specificity.The main works and contributions of this study are as follows:

1. We conduct a comprehensive comparison of both classical and state-of-the-art deep learning models (including VMUNet, H-VMunet, Swin-UNet, ConvUNeXt, and TransUnet) in the SCBP segmentation task, providing benchmarks for future research.
2. We analyze the generalization ability of these models under cross-device data distribution shifts, exploring their robustness in different clinical environments.

2. Materials and Methods

2.1. Dataset

2.1.1. Data Description

The dataset used in this study was collected from clinical ultrasound-guided brachial plexus block (UGRA) [16], All videos were initially annotated by a senior anesthesiologist with extensive experience in ultrasound-guided nerve blocks. When encountering frames with ambiguous anatomical structures or uncertain boundaries, the annotations were reviewed by a second expert of equal seniority. Only frames where both experts reached a consensus were retained as the final ground truth masks. Frames with disagreement or unresolved anatomical uncertainty were discarded to avoid introducing potential noisy labels.

The Esaote MyLab One™ dataset consists of clinical ultrasound videos from 196 patients, with a total of 196 independent video samples, specifically including 157 videos from the supraclavicular brachial plexus (SCBP) region and 33 videos from the interscalene (ISC) region. By extracting frames from these videos, a total of 34,926 ultrasound images were obtained. Among them, 26,966 frames were manually annotated by experienced anesthesiologists for nerve structures. All images were uniformly resized to a resolution of 542 × 562 pixels and underwent standardized preprocessing to ensure compatibility with various deep learning model input requirements.

The Sonosite M-Turbo™ dataset comprises 15 clinical videos, containing a total of 2186 frames, each with a resolution of 512 × 360 pixels.

The Butterfly iQ™ dataset includes 15 ultrasound videos, with a total of 3088 frames, each resized to a resolution of 566 × 534 pixels.

To enhance model generalizability across imaging devices, this study leveraged a pre-existing multi-vendor ultrasound dataset acquired from:

1. Esaote MyLab One™: A high-end cart-based ultrasound system featuring high image resolution and tissue contrast, commonly used in standardized clinical procedures.
2. Sonosite M-Turbo™: A portable device widely used for bedside operations and emergency settings. It offers relatively consistent image quality under diverse operating conditions.
3. Butterfly iQ™: A handheld, smartphone-connected ultrasound device known for its high accessibility and mobility, albeit with higher image noise and relatively lower quality.

Example images from this dataset can be found in Figure 1.

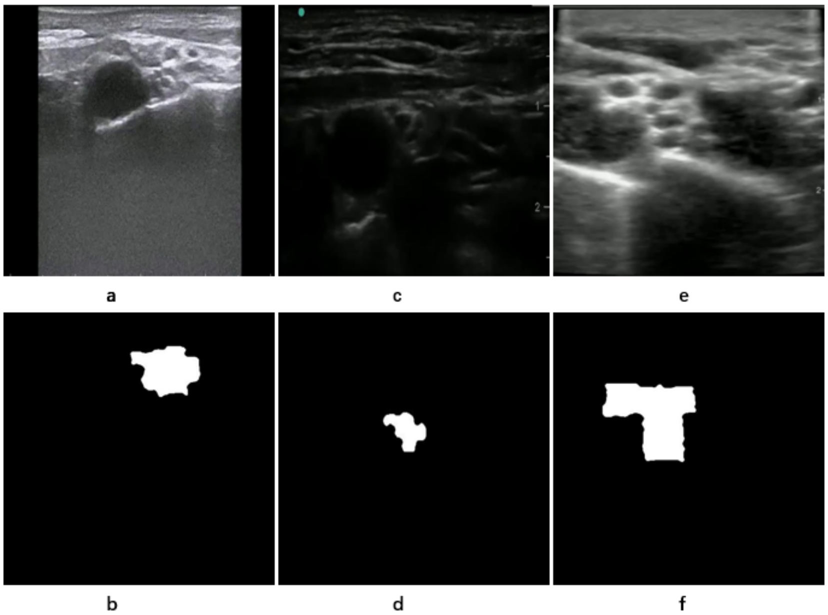

Figure 1. Examples from the ultrasound brachial plexus dataset. Subfigures (**a**,**c**,**e**) show original ultrasound images acquired by eSaote, Sonosite, and Butterfly devices, respectively. Subfigures (**b**,**d**,**f**) show the corresponding binary nerve annotation masks (ground truth).

Figure 2 shows the grayscale distribution differences across all images acquired by the three ultrasound devices before preprocessing, clearly reflecting variations in brightness characteristics and contrast. After applying Contrast-Limited Adaptive Histogram Equalization (CLAHE) for image preprocessing, as shown in Figure 3, the grayscale distributions become more distinguishable. Specifically, Esaote images exhibit a unimodal distribution, while Sonosite images show a slight peak in the low-gray region but still predominantly occupy the mid-gray range. In contrast, Butterfly images are noticeably darker, with a higher proportion of low-gray pixels and prominent high-frequency noise, resulting in lower overall contrast and increased difficulty for nerve segmentation. Consequently, the grayscale distributions of Esaote and Sonosite images are similar, mainly concentrated in the mid-gray range with relatively wide dynamic ranges, which facilitates the distinction between nerve regions and surrounding tissues. In comparison, Esaote and Butterfly images are more dissimilar in their grayscale characteristics.

This dataset, built on a diversified data acquisition strategy, serves as a rich resource for scenario simulation and model evaluation. It facilitates a thorough analysis of model performance and generalization across varying imaging environments

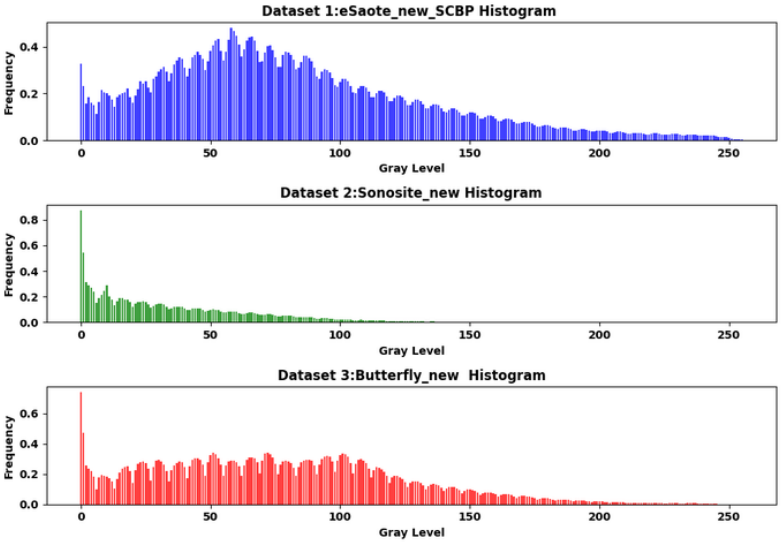

Figure 2. Image histograms before preprocessing, showing grayscale distribution differences among the three ultrasound devices: (1) Esaote MyLab One, (2) Sonosite M-Turbo, and (3) Butterfly iQ.

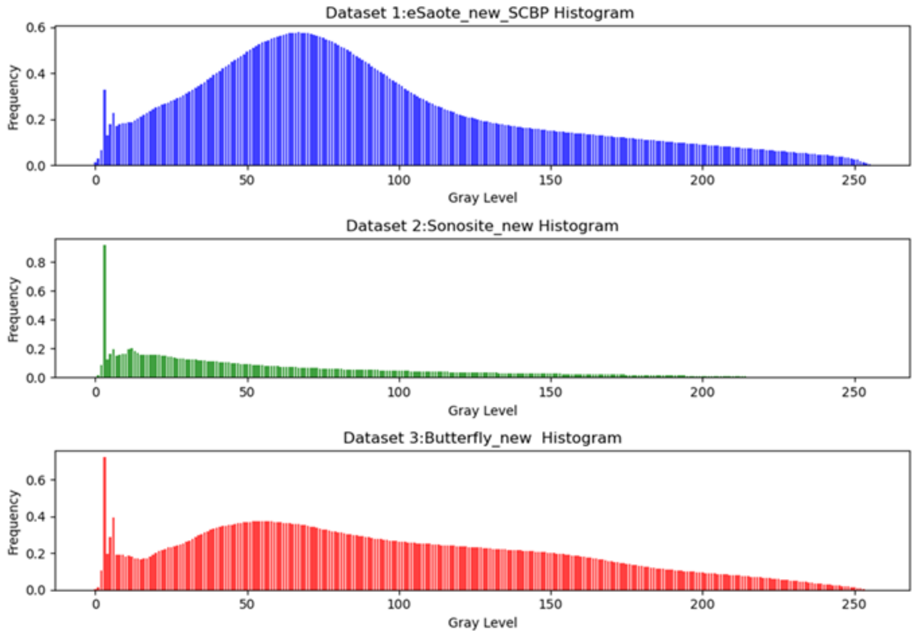

Figure 3. Image histograms after applying Contrast-Limited Adaptive Histogram Equalization (CLAHE), showing grayscale distribution differences among the three ultrasound devices. (1) Esaote MyLab One, (2) Sonosite M-Turbo, and (3) Butterfly iQ.

2.1.2. Data Processing

In the data preprocessing stage, we first filtered the ISC images to remove low-quality or invalid frames, thereby improving data consistency and reliability. Subsequently, to fully leverage the temporal characteristics of ultrasound videos and effectively reduce image redundancy, we adopted a frame sampling strategy based on time series. Compared with other medical imaging modalities such as MRI and CT, ultrasound images exhibit stronger temporal continuity and more complex noise structures. All frames were extracted from the original ultrasound-guided brachial plexus block (UGRA) videos, and uniform sampling was performed by selecting every 5th frame. This approach preserves key temporal variations, enhances the model's ability to learn temporal features, and avoids the adverse effects of redundant frames on training efficiency.

To ensure stable model training and reproducibility of results, the dataset was divided into training, validation, and test sets with a ratio of 60%, 20%, and 20%, respectively. The number of frames for each ultrasound device is shown in Table 1. This partitioning strategy provides a sufficient number of training samples while enabling effective evaluation and comparison across different stages of model development.

Table 1. Dataset Information for Each Ultrasound Device.

Device	Training Set	Validation Set	Test Set	Test Set Positive Frames
Esaote MyLab One™	2791	930	932	-
Sonosite M-Turbo™	-	-	-	477
Butterfly iQ™	-	-	-	2302

Prior to model training, all images underwent standardized preprocessing. First, to enhance the visibility of nerve structures and suppress speckle noise, we applied Contrast-Limited Adaptive Histogram Equalization (CLAHE) [17] for local contrast enhancement. Compared with conventional histogram equalization, CLAHE prevents over-amplification of background noise and significantly improves the discernibility of nerve boundaries. In our implementation, the clip limit parameter was set to 2.0, which controls the degree of histogram clipping and prevents excessive contrast amplification in homogeneous regions. The tile grid size was set to (8, 8), meaning the image was divided into 8 × 8 contextual regions, each of which was adaptively equalized to preserve local contrast. These parameter values were selected empirically to balance the enhancement of nerve visibility and the suppression of noise artifacts. Additionally, to mitigate variations in dynamic range caused by different ultrasound devices, we applied Min–Max normalization to scale all grayscale values to the [0, 1] range, thereby improving input consistency and training stability. An example of an ultrasound image after CLAHE preprocessing is shown in Figure 4.

To further enhance the model's generalization capability and simulate diverse clinical imaging conditions, we incorporated various data augmentation techniques during training. These included random horizontal flipping (probability = 0.3), random brightness and contrast adjustments (probability = 0.2), translation ($\pm 10\%$), scaling ($\pm 20\%$), rotation ($\pm 45°$), and gamma correction ($\gamma \in [0.8, 1.2]$). These augmentations effectively simulate the variability in ultrasound images caused by differences in probe orientation, imaging angles, and device settings, thus improving the model's adaptability to real-world scenarios. All images were ultimately resized to 224 $\times$ 224 pixels.

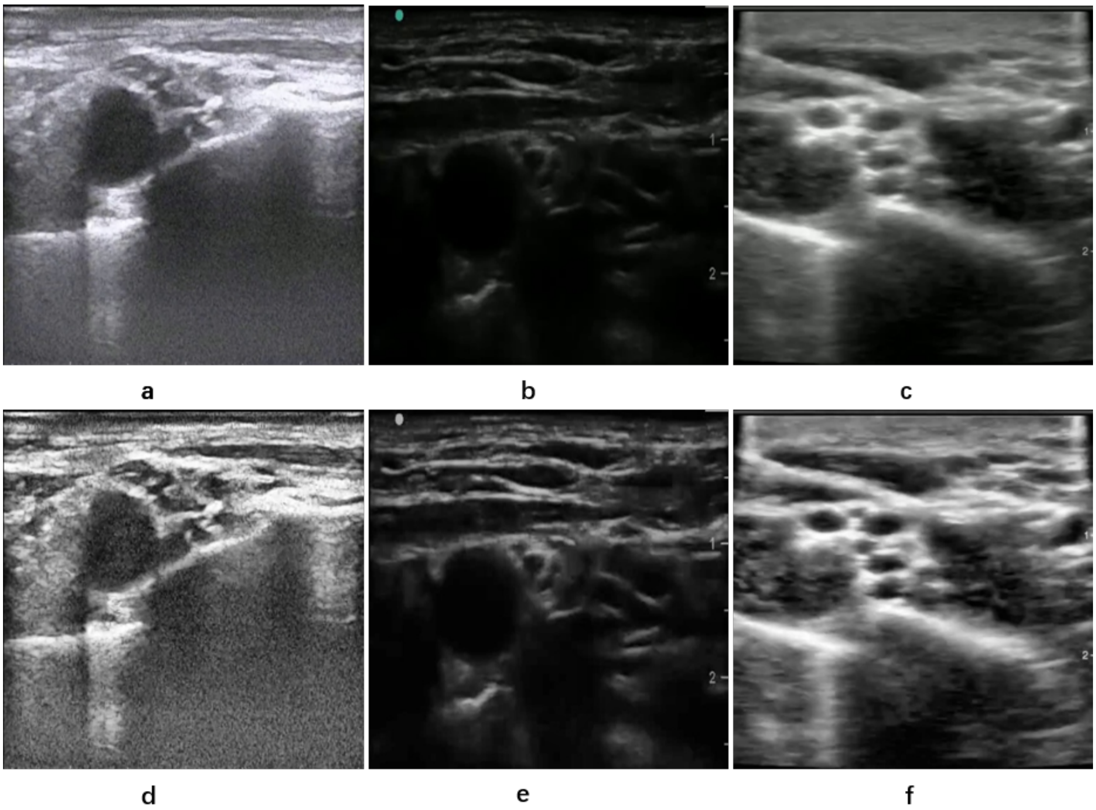

Figure 4. An example of an ultrasound image after CLAHE preprocessing. Panels (**a**–**c**) are the original images, while panels (**d**–**f**) are the images after CLAHE processing. The Contrast-Limited Adaptive Histogram Equalization (CLAHE) method enhances local contrast and suppresses speckle noise, thereby improving the visibility of nerve structures.

2.2. Deep Learning Models

In this study, six representative deep learning models were selected to perform automatic segmentation of the brachial plexus in ultrasound images. These models span convolutional neural networks (CNNs), state space models (Mamba), and Transformer-based architectures. Each selected model features distinct structural designs, parameter scales, and computational efficiencies, providing a comprehensive evaluation of the adaptability and generalization capabilities of different network types in ultrasound segmentation tasks.

The selected models include the classical UNet, the lightweight ConvUNeXt, the state space modeling-based SwinUNet and H-VMUNet, as well as the Transformer-based architectures SwinUNet and TransUNet. The number of parameters among these models ranges from several megabytes to hundreds of megabytes, are summarized in Table 2.

Table 2. Summary of model characteristics, including parameter size, model class, and key architectural remarks. All models are evaluated under a unified dataset and metrics to explore their performance in ultrasound brachial plexus segmentation.

Model	Ref.	Param Size	Class	Remarks
ConvUNeXt	[12]	3.25 MB	CNN	a. Incorporates depthwise separable convolution and Layer Normalization. b. Improves efficiency and scalability, ideal for lightweight deployments. c. Uses Kaiming Uniform initialization for convolution and linear layers, and sets BatchNorm gamma = 1, beta = 0.

Table 2. *Cont.*

Model	Ref.	Param Size	Class	Remarks
UNet	[6]	17.26 MB	CNN	a. U-shaped encoder-decoder with skip connections. b. Preserves spatial resolution and extracts semantic features, widely used in medical segmentation. c. Uses Kaiming Uniform initialization for convolution and linear layers, and sets BatchNorm gamma = 1, beta = 0.
VMUNet	[10]	27.42 MB	Mamba	a. Uses Visual State Space (VSS) modules and patch-level extraction. b. Outperforms CNNs/ViTs in long-sequence modeling with a lightweight design. c. Uses pretraining (ImageNet) to initialize weights
H-VMUNet	[11]	8.97 MB	Mamba	a. Vision Mamba backbone with global-local enhancement modules. b. Balances global structural understanding and fine detail localization. c. Convolutional layers are initialized with a normal distribution; linear layers and special parameters use truncated normal (std = 0.02); LayerNorm/GroupNorm with weight = 1 and bias = 0.
SwinUNet	[9]	27.16 MB	Transformer	a. Swin Transformer blocks in UNet with windowed attention and skip connections. b. Combines global context and local features for better segmentation accuracy. c. Uses pretraining (ImageNet) to initialize weights
TransUNet	[18]	105.27 MB	Transformer	a. CNN encoder for local, ViT for global modeling, and decoder for output recovery. b. Strong generalization and robustness for diverse medical imaging tasks. c. Uses pretraining (ImageNet) to initialize weights

3. Experiments and Analysis

3.1. ExperimentalConfiguration

All model training and testing procedures in this study were implemented using the PyTorch 2.3.0 deep learning framework and conducted on a single NVIDIA RTX 4090D GPU. All input images were uniformly resized to $224 \times 224 \times 3$ to match the input requirements of various models and improve training efficiency.

To ensure reproducibility, we set the random seed to 42 and conducted all experiments in a non-distributed environment. The batch size was set to 8, and the models were trained for a total of 200 epochs. The optimizer used was AdamW, with an initial learning rate of 0.001 and a weight decay of 0.01. A cosine annealing learning rate scheduler (CosineAnnealingLR) was applied to dynamically adjust the learning rate, with a maximum cycle $T_{\text{max}} = 50$ and a minimum learning rate $\eta_{\text{min}} = 1 \times 10^{-5}$, to facilitate stable convergence in the later stages of training.

In this study, the BCE-Dice Loss was adopted as the primary training objective. This composite loss function integrates the Binary Cross-Entropy (BCE) loss and the Dice loss, with weighting to jointly optimize pixel-wise classification accuracy and overall segmentation consistency. During the training process, model parameters are updated using the training set, while at the end of each epoch, the model is evaluated on the validation set to compute its BCE-Dice Loss. The model exhibiting the best performance on the validation set (i.e., the lowest loss) is saved as the final model, ensuring optimal generalization capability. Formally, the loss function is defined as:

$$\mathcal{L}_{\text{total}} = w_b \cdot \mathcal{L}_{\text{BCE}} + w_d \cdot \mathcal{L}_{\text{Dice}} \tag{1}$$

where $w_b = 1$ and $w_d = 1$ are the weighting coefficients for the BCE and Dice components, respectively. During training, the BCE term accelerates the convergence of pixel-wise predictions, while the Dice term is critical for addressing foreground-background class imbalance and improving boundary continuity. This loss function has been widely used in medical image segmentation tasks and is particularly well-suited for the binary segmentation scenario of this study (i.e., distinguishing nerve regions from non-nerve regions).The detailed hyperparameter settings used in this study are summarized in Table 3.

Table 3. Training hyperparameter settings used in this study.

Parameter Category	Configuration/Option
Framework	PyTorch 2.3.0
Hardware	NVIDIA RTX 4090D (24 GB)
Input Size	$224 \times 224 \times 3$

Table 3. *Cont.*

Parameter Category	Configuration/Option
Random Seed	42
Batch Size	8
Optimizer	AdamW (lr = 0.001, wd = 0.01)
Learning Rate Scheduler	CosineAnnealingLR ($T_{\text{max}} = 50$, $\eta_{\text{min}} = 1 \times 10^{-5}$)
Loss Function	BCE-Dice Loss ($w_b = 1$, $w_d = 1$)

3.2. Performance Metrics

To comprehensively evaluate model performance on the nerve segmentation task, we adopt several commonly used semantic segmentation metrics: mean Intersection over Union (mIoU), F1 Score, Accuracy, Specificity, Recall, and Mean Absolute Error (MAE). These metrics collectively assess model performance from multiple perspectives, including region overlap, pixel-wise classification precision, foreground/background discrimination capability, and overall segmentation error.

The mIoU measures the average intersection-over-union between the predicted and ground truth masks. The F1 Score is the harmonic mean of precision and recall, particularly effective in class-imbalanced scenarios. Accuracy reflects the proportion of correctly classified pixels among all pixels. Specificity evaluates the model's ability to correctly identify background pixels, while Recall emphasizes sensitivity to the foreground (i.e., nerve regions). MAE quantifies prediction bias as pixel-wise absolute error.

3.3. Results on the eSaote Dataset (Train: eSaote, Test: eSaote)

To assess model performance on in-domain data, we conducted detailed evaluations on the eSaote test set. Since the test set shares a distribution consistent with the training set, it effectively reflects each model's fitting capacity and feature learning effectiveness.

As shown in Table 4, both UNet and ConvUNeXt achieved outstanding results across multiple evaluation metrics, with mIoU scores of 0.9043 and 0.9027, respectively. These results highlight the robustness and reliability of classic convolutional architectures under high-quality ultrasound imaging conditions.TransUNet and H-VMUNet follow closely, benefiting from their hybrid structures that incorporate Transformer or state space mechanisms, which better capture both global and local contextual information.In contrast, SwinUNet performed less favorably, with a noticeably lower mIoU compared to other models. Qualitative results also showed issues such as edge blurring and structural discontinuities, possibly due to the limitations of its local window attention mechanism in modeling fine-grained texture details.

Figure 5 illustrates the nerve segmentation visualization results of different models on various test samples. To demonstrate segmentation performance across multiple test cases, the figure highlights differences among models in preserving spatial structures. TransUNet, ConvUNeXt, and UNet generally maintain relatively complete nerve region morphology, with smooth segmentation boundaries and continuous structures, indicating their strong capability in modeling the morphology of nerve tissues. In contrast, the prediction masks of SwinUNet often exhibit structural loss and discontinuity, revealing its instability in perceiving local image structures. Although VMUNet perform well overall, they occasionally show slight prediction shifts or discontinuous edges in some samples, suggesting that there is still room for improvement in local detail depiction.

Table 4. Comparison of segmentation performance across different models on the eSaote test set. Metrics include mIoU, Loss, F1, Accuracy, Specificity, Recall, and MAE.

Model	mIoU	Loss	F1	Acc	Spec	Recall	MAE
ConvUNeXt	0.9027	0.1488	0.8987	0.9899	0.9943	0.9054	0.0100
UNet	0.9043	0.1505	0.9004	0.9901	0.9947	0.9027	0.0098
VMUNet	0.8698	0.2002	0.8601	0.9857	0.9906	0.8907	0.0142
H-VMUNet	0.8961	0.1605	0.8911	0.9891	0.9936	0.9029	0.0108
SwinUNet	0.7718	0.3997	0.7266	0.9738	0.9877	0.7062	0.0261
TransUNet	0.9002	0.1535	0.8958	0.9896	0.9940	0.9051	0.0103

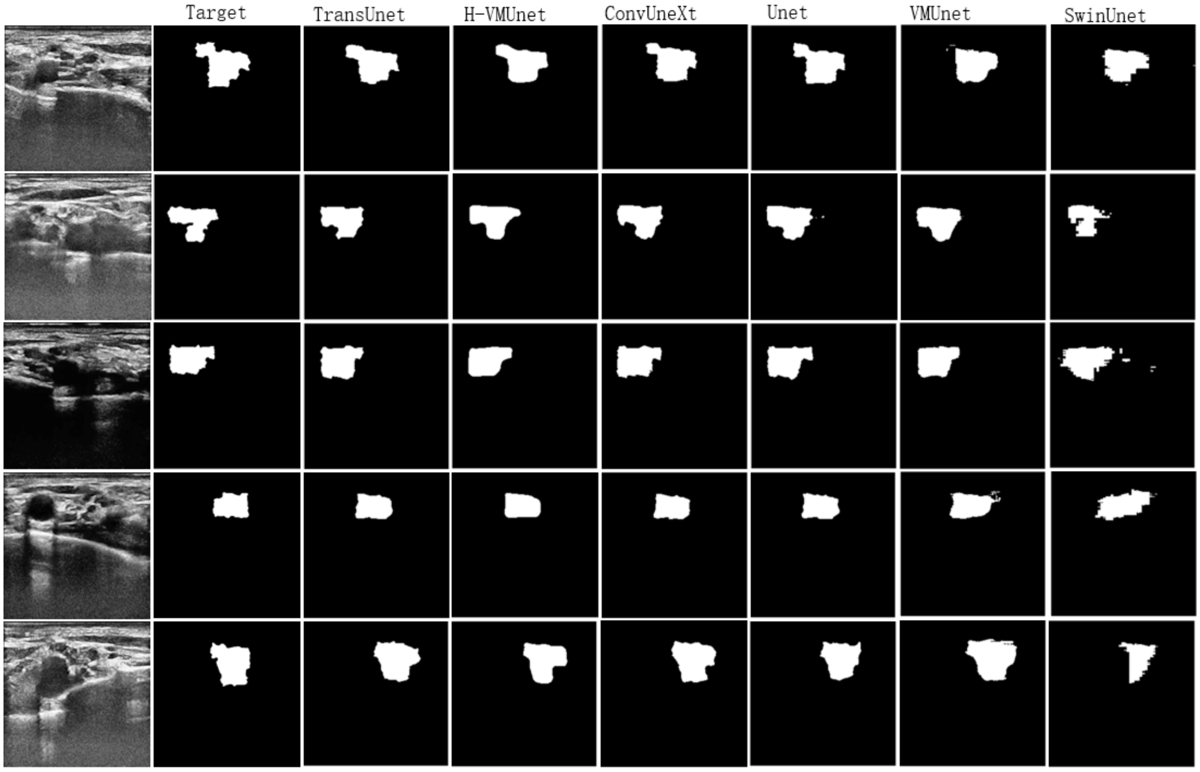

Figure 5. Visualization of nerve segmentation results of each model on the eSaote test images. The left column shows the original images and ground truth masks, while the remaining columns present the predicted results of different models.

It should be noted that the binary segmentation threshold for all models is set to 0.5. Specifically, for the probability maps predicted by the models, pixels with values greater than 0.5 are classified as foreground (nerve regions), while pixels with values less than or equal to 0.5 are classified as background (non-nerve regions). This threshold selection ensures a balance between foreground and background and simplifies the comparison of segmentation results across different models or devices.

Figure 6 shows the loss variation trends of different models during the training and validation processes, revealing their differences in fitting capability and generalization performance. In the training phase, all models generally exhibit fast convergence, especially within the first 10 epochs, where the loss drops rapidly, indicating that the models can effectively capture feature information from the training data. As training progresses, the loss gradually stabilizes, with most models converging to values below 0.2, suggesting no obvious underfitting. In the validation phase, the loss curves of different models show varying degrees of fluctuation. Among them, ConvUNeXt, UNet, and TransUNet have relatively stable validation losses, demonstrating strong predictive ability on unseen samples and good generalization performance. In contrast, the validation loss curves of H-VMUNet and SwinUNet fluctuate more significantly, indicating potential overfitting during training or weaker robustness to data perturbations.

3.4. Results on the Sonosite Dataset (Train: eSaote, Test: Sonosite)

To evaluate the generalization ability of the models on data from heterogeneous devices, we conducted cross-device testing using ultrasound images acquired with the Sonosite device. Since the imaging style of this device differs from that of the eSaote series used for training, this experiment effectively reveals the models' adaptability to distribution shifts.

As shown in Table 5, the overall performance of all models on the Sonosite test set drops significantly compared to the eSaote test set, reflecting the greater challenges faced under cross-device conditions.Among all models, SwinUNet performs the best, ranking first in several key metrics such as mIoU, F1 score, and Recall, with an mIoU of 0.7937, demonstrating the excellent cross-domain adaptability of its state-space-based architecture. H-VMUNet and TransUNet follow closely, achieving mIoUs of 0.7814 and 0.7710, respectively, suggesting that architectures incorporating global modeling mechanisms exhibit stronger robustness in transfer scenarios.In contrast, although ConvUNeXt performs well on the training device (eSaote), its mIoU drops to 0.7364 on Sonosite, indicating higher sensitivity to distribution shifts. A similar trend is observed in UNet, whose mIoU decreases to 0.7276, showing limited generalization ability. SwinUNet once again ranks lowest in performance, with an mIoU of 0.6682 and an F1 score of only 0.5612.

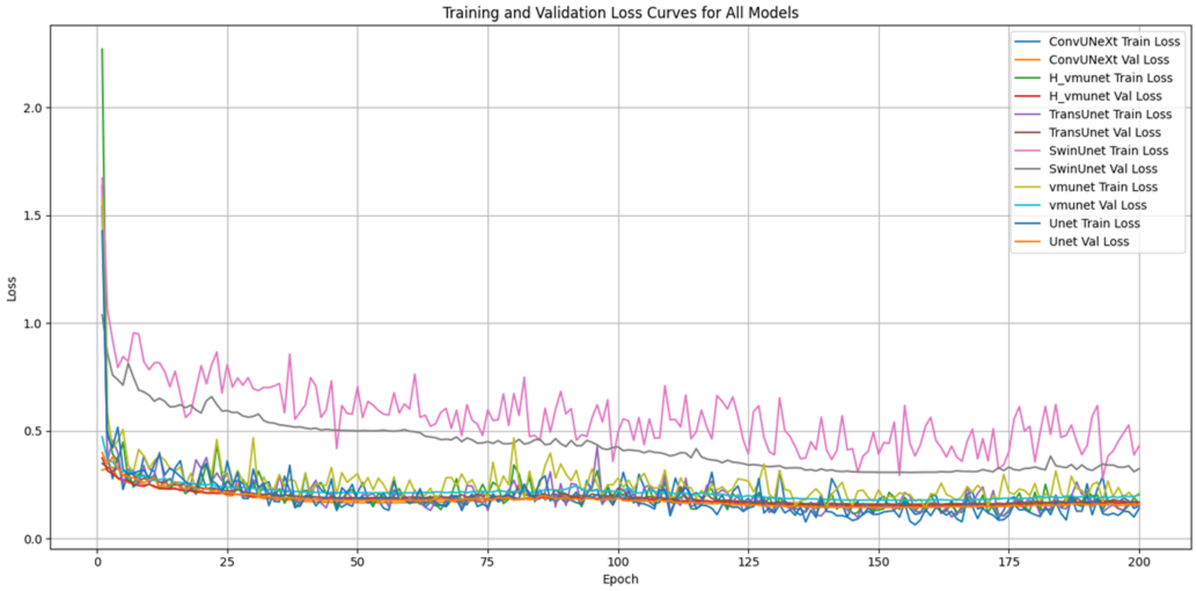

Figure 6. Loss curve trends of different models on the training and validation sets.

Table 5. Segmentation performance comparison of different models on the Sonosite test set. Metrics include mIoU, Loss, F1, Accuracy, Specificity, Recall, and MAE. Bold indicates the best performance, and bold italic indicates the second-best

Model	mIoU	Loss	F1	Acc	Spec	Recall	MAE
ConvUNeXt	0.7364	0.4877	0.6731	0.9666	0.9856	0.6351	0.0333
UNet	0.7276	0.5072	0.6590	0.9650	0.9844	0.6250	0.0349
VMUNet	**0.7937**	**0.3436**	**0.7621**	**0.9729**	***0.9827***	**0.8016**	**0.0270**
H-VMUNet	***0.7814***	***0.3853***	***0.7432***	***0.9726***	**0.9862**	0.7336	***0.0273***
SwinUNet	0.6682	0.6215	0.5612	0.9482	0.9674	0.6121	0.0517
TransUNet	0.7710	0.4004	0.7291	0.9695	0.9816	***0.7581***	0.0304

Figure 7 presents the visualization results of different models on the Sonosite dataset. It can be observed that SwinUNet and TransUNet show clear advantages in maintaining the shape and boundary continuity of the nerve regions, with segmentation contours that are relatively sharp and closely aligned with the ground truth masks. Although H-VMUNet also performs well, some samples show slightly blurred edge structures.In contrast, the predictions of ConvUNeXt and UNet exhibit varying degrees of displacement and deformation, indicating their limited adaptability to changes in image feature distributions. The prediction maps of SwinUNet contain more fragmented regions and false positives, further confirming its inadequate modeling capacity under low-resolution and high-noise conditions.

These experimental results suggest that architectural design plays a critical role in the cross-device robustness of segmentation models. Architectures with stronger long-range dependency modeling capabilities, such as SwinUNet and TransUNet, can maintain stable performance when facing imaging style shifts, whereas models relying only on local convolutions or local attention mechanisms are more prone to performance degradation.

3.5. Results on the Butterfly Dataset (Train: eSaote, Test: Butterfly)

To further evaluate the generalization ability of the models under conditions with poorer imaging quality and more challenging clinical scenarios, we conducted experiments on ultrasound images collected by the Butterfly device. This dataset features images with obvious blurriness, low contrast, and noise interference, which are significantly different from the eSaote data used for training, thus effectively testing the models' adaptability to severe image distribution shifts.

As shown in Table 6, the overall performance of all models drops significantly compared to the previous two test groups, with some models almost completely failing.In this experiment, TransUNet still maintains relatively leading performance, achieving an mIoU of 0.6725 and an F1 score of 0.5682. It is the only model that maintains moderate segmentation accuracy (IoU > 0.5) under such extreme conditions. In contrast, the other models perform poorly on the Butterfly test set. Models such as UNet, SwinUNet, and SwinUNet have F1 scores close to zero, and mIoUs do not exceed 0.47, indicating severe degradation of segmentation ability under strong noise and high blurriness.It is noteworthy that although ConvUNeXt scores relatively high on Accuracy and Specificity, its F1 score is only 0.0895, indicating that most of its predictions are negative, achieving high scores mainly on background classification but failing to cover the actual target regions. H-VMUNet attempts to generate shape structures, but its mIoU is only 0.4581, indicating deviation of its predicted mask coverage from the ground truth.

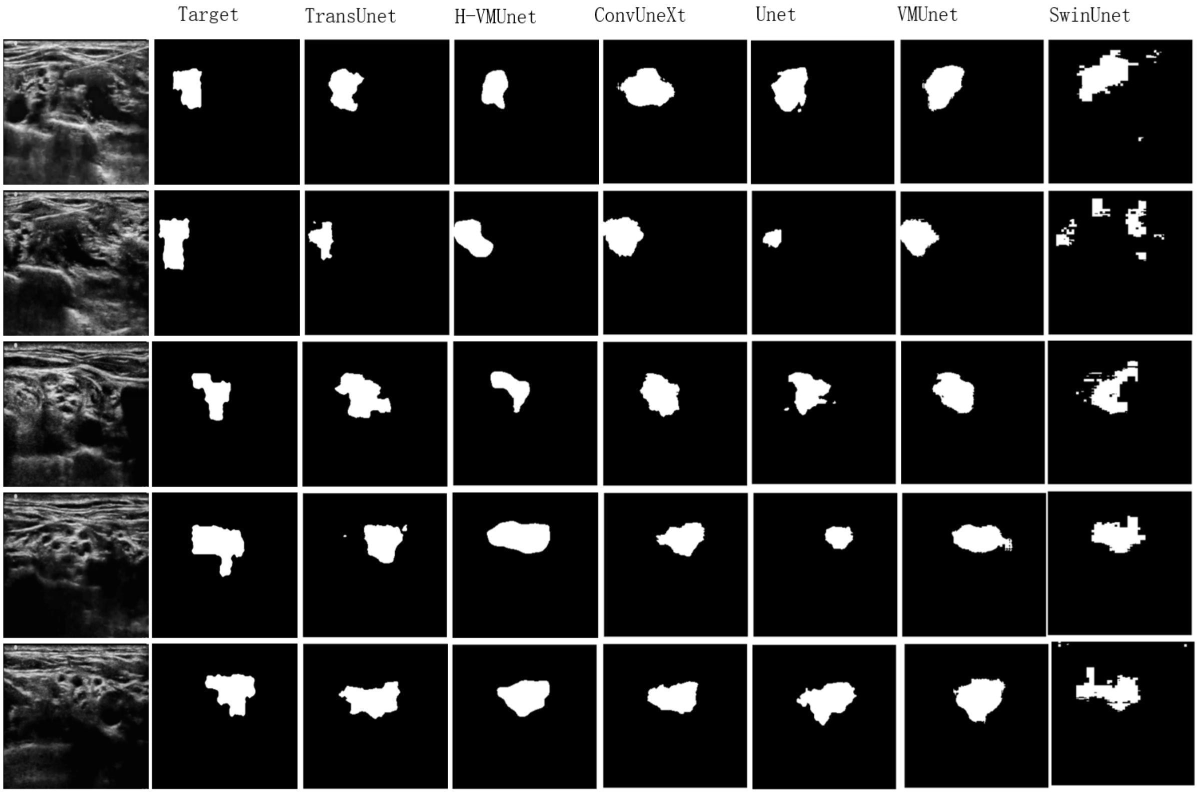

Figure 7. Visualization of nerve segmentation results of different models on the Sonosite test images. The left column shows the original images and ground truth masks, while the remaining columns display the predicted results of the models.

Figure 8 shows the visualization results of different models on the Butterfly dataset, further confirming the conclusions of the quantitative evaluation above. TransUNet is the only model that can accurately recover the nerve structure boundaries in multiple samples, with predicted region shapes closely resembling the ground truth masks. Other models generally fail to produce effective segmentation, showing blank outputs, random noise blocks, or severe misalignments. UNet, SwinUNet, ConvUNeXt, and VMUNet all failed to achieve satisfactory performance in this brachial plexus segmentation task.

Table 6. Segmentation performance comparison of different models on the Butterfly test set. Metrics include mIoU, Loss, F1, Accuracy, Specificity, Recall, and MAE. Bold indicates the best performance, and bold italic indicates the second-best.

Model	mIoU	Loss	F1	Acc	Spec	Recall	MAE
ConvUNeXt	***0.4697***	6.6466	***0.0895***	0.8930	0.9566	***0.0731***	0.1069
UNet	0.4640	8.1890	0.0	0.9280	**1.0000**	0.0	**0.0719**
VMUNet	0.4592	1.9420	0.0	0.9184	***0.9896***	0.0	0.0815
H-VMUNet	0.4581	***1.7568***	0.0079	0.9124	0.9827	0.0048	0.0875
SwinUNet	0.4640	1.8063	0.0	***0.9281***	**1.0000**	0.0	0.0718
TransUNet	**0.6725**	**0.7225**	**0.5682**	0.9500	0.9881	**0.4574**	**0.0499**

Overall, under extreme low-quality imaging conditions, most mainstream models face significant performance challenges, and only some hybrid architectures (such as TransUNet) exhibit a certain degree of generalization ability. This result suggests that future research should focus more on designing model robustness to real clinical characteristics such as imaging noise and boundary blurring.

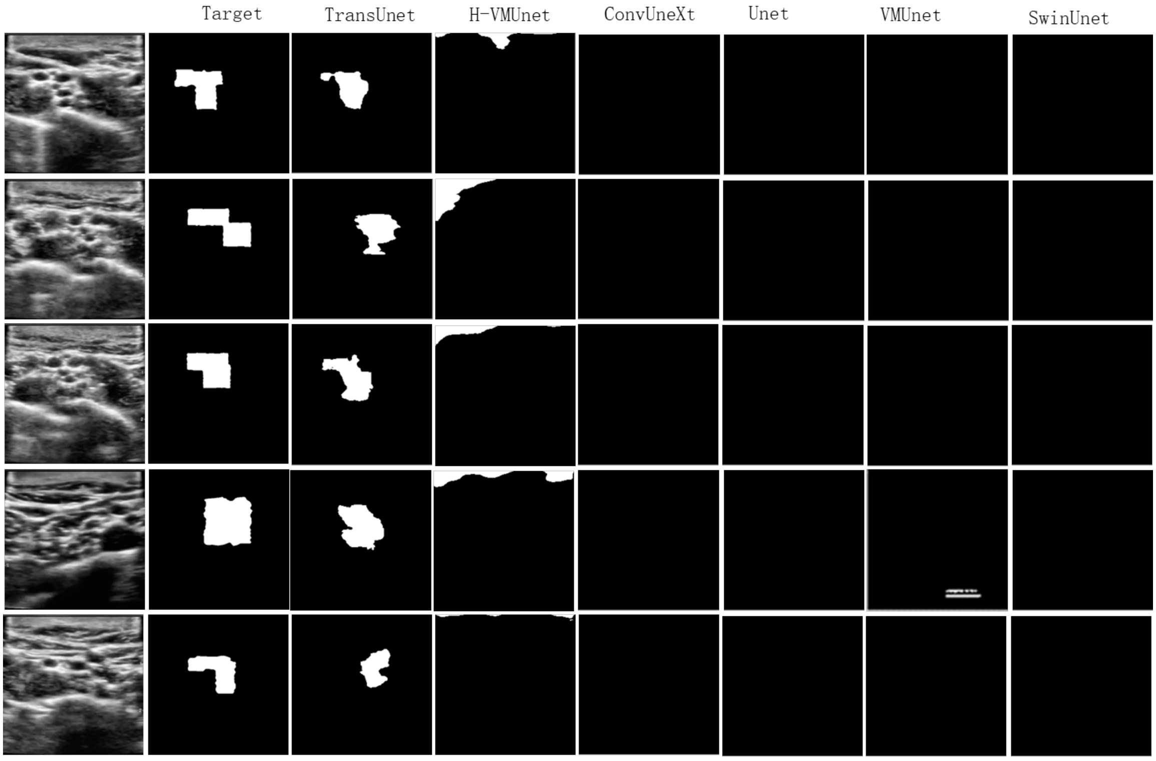

Figure 8. Visualization of nerve segmentation results of different models on the Butterfly test images. The left column shows the original images and ground truth masks, while the remaining columns display the predicted results of the models.

4. Discussions

To deeply analyze the generalization ability of different models in cross-device semantic segmentation tasks, we computed and compared the mean Intersection over Union (mIoU) metric on three test datasets eSaote, Sonosite, and Butterfly), as shown in Figure 9. All models were trained on the eSaote dataset and tested with fixed weights to evaluate their performance differences on both homogeneous and heterogeneous data.

In the within-training-distribution test, all models achieved high segmentation accuracy, with mIoU scores above 0.86, demonstrating strong feature learning capabilities and good fitting performance. Especially, ConvUNeXt, UNet, TransUNet, and H-VMUNet all achieved mIoU scores exceeding 0.89 under the training distribution.

However, cross-device test results showed significant performance degradation. Particularly on the Sonosite test set, all models had mIoU scores lower than those on the training set, indicating that distribution shifts substantially affect segmentation performance. Nevertheless, SwinUNet (mIoU ≈ 0.7937) and H-VMUNet demonstrated stronger cross-domain adaptability, suggesting that their architectures are better at capturing generalized features.

On the more challenging Butterfly dataset, except for TransUNet which still maintained a relatively high mIoU of 0.6725, the segmentation performances of other models generally dropped below 0.5, indicating insufficient robustness of current architectures when facing low resolution, strong noise, and drastic imaging style changes.

From a methodological perspective, the fully supervised training strategy adopted in this study works well in same-distribution scenarios but still has limitations in domain-shift conditions. On one hand, training data highly depends on manual annotations, with limited data volume, and image quality and imaging standards vary greatly across devices, limiting further improvement in model generalization. On the other hand, although existing model architectures have strong representational power, they lack explicit domain adaptation mechanisms.

To address this issue, future work may consider incorporating model distillation mechanisms, i.e., constructing high-performance large teacher models and performing thorough pretraining and fine-tuning on multi-source data to guide lightweight student models to learn more robust representations. This approach is expected to enhance the generalization performance of small models under unknown devices or low-quality images, better meeting real clinical deployment needs.Future research can also explore semi-supervised learning, domain generalization strategies, or multimodal auxiliary inputs (such as image-guided annotations or manual region markings) to further optimize model usability in resource-constrained and heterogeneous data environments. Semi-supervised learning can leverage unlabeled data to mitigate annotation scarcity, while domain generalization aims to learn device-agnostic features to improve cross-device robustness. These approaches have demonstrated effectiveness in other medical imaging tasks [19], suggesting their potential for enhancing ultrasound nerve segmentation models.

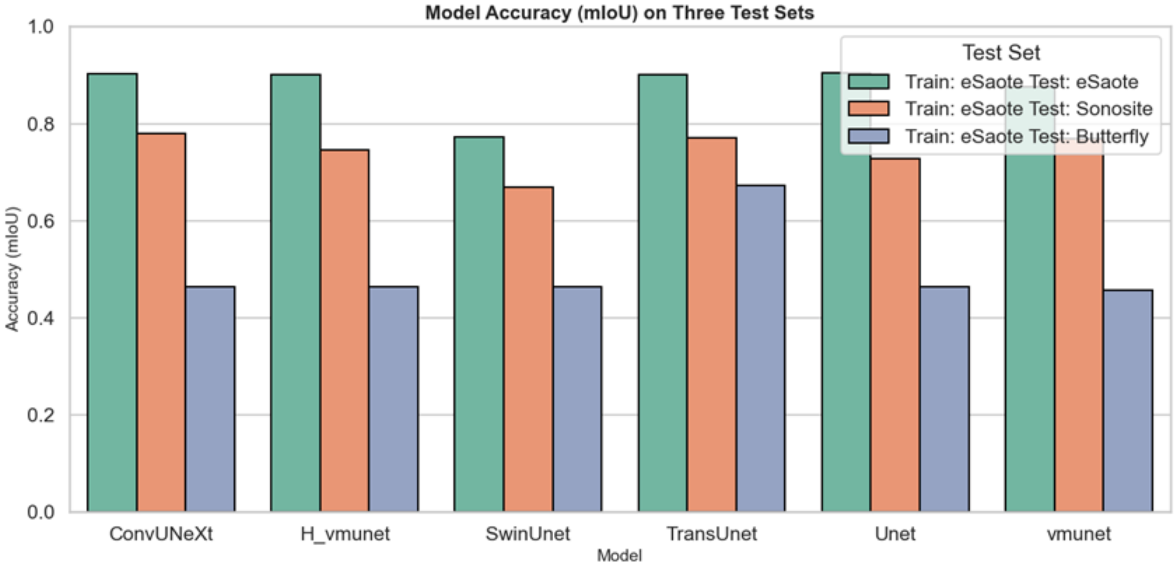

Figure 9. Comparison of mIoU performance of different models on three test datasets (Train: eSaote).

The dataset used in this study has certain limitations, as it lacks some surgical data from clinicians, making more detailed experiments infeasible. Future work will incorporate additional clinical data and operative records to conduct a more thorough investigation of the model's practicality and safety.

5. Conclusions

This study systematically compared the performance of six representative deep learning models in ultrasound-guided brachial plexus nerve segmentation tasks, covering convolutional neural networks (UNet, ConvUNeXt), Transformer architectures (TransUNet, SwinUNet), and state-space modeling networks (VMUNet, H-VMUNet). All models performed well on the homogeneous eSaote dataset, with UNet and ConvUNeXt achieving the highest segmentation accuracy. On the Sonosite and Butterfly datasets with device domain differences, overall performance declined. TransUNet and VMUNet demonstrated stronger robustness and generalization capabilities. Particularly on the low-quality Butterfly dataset, TransUNet was the only model maintaining moderate segmentation performance.

Furthermore, visualization analyses revealed differences and similarities among the models in boundary preservation, structural continuity, and erroneous region identification, providing intuitive support for the quantitative metrics. Future work will focus on developing lighter, more efficient, and cross-domain adaptive neural network architectures to provide more reliable automated assistance tools for ultrasound image navigation in nerve block anesthesia.

Author Contributions

X.Z. and D.T.: conceptualization, methodology, software, data curation, writing—original draft preparation, visualization, investigation, supervision, validation. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Ethical review and approval were waived for this study, due to REASON (This study was conducted using a publicly available dataset).

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used ChatGPT to polish the language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Van Boxtel, J.; Vousten, V.; Pluim, J.; et al. Hybrid deep neural network for brachial plexus nerve segmentation in ultrasound images. In Proceedings of the 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021; pp. 1246–1250.
2. Pincus, E. Regional anesthesia: An overview. *AORN J.* **2019**, *110*, 263–272.
3. Banerjee, S.; Acharya, R.; Sriramka, B. Ultrasound-guided inter-scalene brachial plexus block with superficial cervical plexus block compared with general anesthesia in patients undergoing clavicular surgery: A comparative analysis. *Anesth. Essays Res.* **2019**, *13*, 149–154.
4. Chui, J.; Murkin, J.M.; Posner, K.L.; et al. Perioperative peripheral nerve injury after general anesthesia: A qualitative systematic review. *Anesth. Analg.* **2018**, *127*, 134–143.
5. Mateo, J.L.; Fernández-Caballero, A. Finding out general tendencies in speckle noise reduction in ultrasound images. *Expert Syst. Appl.* **2009**, *36*, 7786–7797.
6. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
7. Li, X.; Fang, J.; Zhao, Y. A Multi-Target Identification and Positioning System Method for Tomato Plants Based on VGG16-UNet Model. *Appl. Sci.* **2024**, *14*, 2804.
8. Trebing, K.; Stanczyk, T.; Mehrkanoon, S. SmaAt-UNet: Precipitation nowcasting using a small attention-UNet architecture. *Pattern Recognit. Lett.* **2021**, *145*, 178–186.
9. Cao, H.; Wang, Y.; Chen, J.; et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 205–218.
10. Ruan, J.; Li, J.; Xiang, S. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv* **2024**, arXiv:2402.02491.
11. Wu, R.; Liu, Y.; Liang, P.; et al. H-vmunet: High-order vision mamba unet for medical image segmentation. *Neurocomputing* **2025**, *624*, 129447.
12. Han, Z.; Jian, M.; Wang, G.G. ConvUNeXt: An efficient convolution neural network for medical image segmentation. *Knowl.-Based Syst.* **2022**, *253*, 109512.
13. Xu, Q.; Ma, Z.; Duan, W.; et al. DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation. *Comput. Biol. Med.* **2023**, *154*, 106626.
14. Wu, H.; Liu, J.; Wang, W.; et al. Region-aware global context modeling for automatic nerve segmentation from ultrasound images. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 2907–2915.
15. Wang, Y.; Geng, J.; Zhou, C.; et al. Segmentation of ultrasound brachial plexus based on U-Net. In Proceedings of the 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), Beijing, China, 14–16 May 2021; pp. 482–485.
16. Tyagi, A.; Tyagi, A.; Kaur, M.; et al. Nerve Block Target Localization and Needle Guidance for Autonomous Robotic Ultrasound Guided Regional Anesthesia. In Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Abu Dhabi, United Arab Emirates, 14–18 October 2024; pp. 5867–5872.
17. Tian, D.; Zhu, B.; Wang, J.; et al. Brachial plexus nerve trunk recognition from ultrasound images: A comparative study of deep learning models. *IEEE Access* **2022**, *10*, 82003–82014.
18. Chen, J.; Lu, Y.; Yu, Q.; et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
19. Hu, M.; Li, Y.; Yang, X. Skinsam: Empowering skin cancer segmentation with segment anything model. *arXiv* **2023**, arXiv:2304.13973.