Is denoising necessary for ultrasound image segmentation deep learning: review and benchmark

Fei Liu, Zhixia Dong, Pei Qin, Binjie Qin, *Member, IEEE*, Sijie Xu, Xiangyun Zhao, Xinjian Wan, Xu Chen, Lu Chen

***Abstract*—Ultrasound image segmentation deep learning still has performance bottleneck due to an inherent speckle noise having complex non-Gaussian statistics in the images. Denoised input data, multi-task segmentation & denoising, and holistically robust feature learning are three solutions to the speckle challenge in deep ultrasound image segmentation. To assess whether denoising (or despeckling) is necessary for ultrasound image segmentation deep learning in addressing speckle challenge and improving performance, we review deep learning ultrasound image segmentation and denoising as well as establish an ultrasound image denoising-segmentation cross benchmarking considering the abovementioned solutions, with the following core components. *Datasets*: 4 public ultrasound datasets and 2 self-collected datasets. *Despeckling methods*: 7 typical despeckling methods, such as non-local means and diffusion methods. *Basic models*: U-Net [1], SK-U-Net [2], and CE-Net [3] for segmenting breast ultrasound images, and U-Net and DAEFF-Net [4] for echocardiography. *Multi-task model*: SFS block [5] for segmentation and despeckling feature fusion. *Heuristically speckle-robust models*: residual feedback & refinement network RF-Net [6] and transformer-assisted CNN network CDM [7]. We eliminate the nondeterminism effect [8], [9] in the deep learning model training via deterministic training or averaging 30 repeated training runs. We conduct comprehensive experimental evaluations in both intra- and cross-dataset testings in terms of segmentation evaluation metrics and statistical analysis with the Friedman test and two paired tests. We demonstrate that the performance improvement from denoising pre-processing is more unstable and slighter (if exists) 6nsti7583ce challenge**

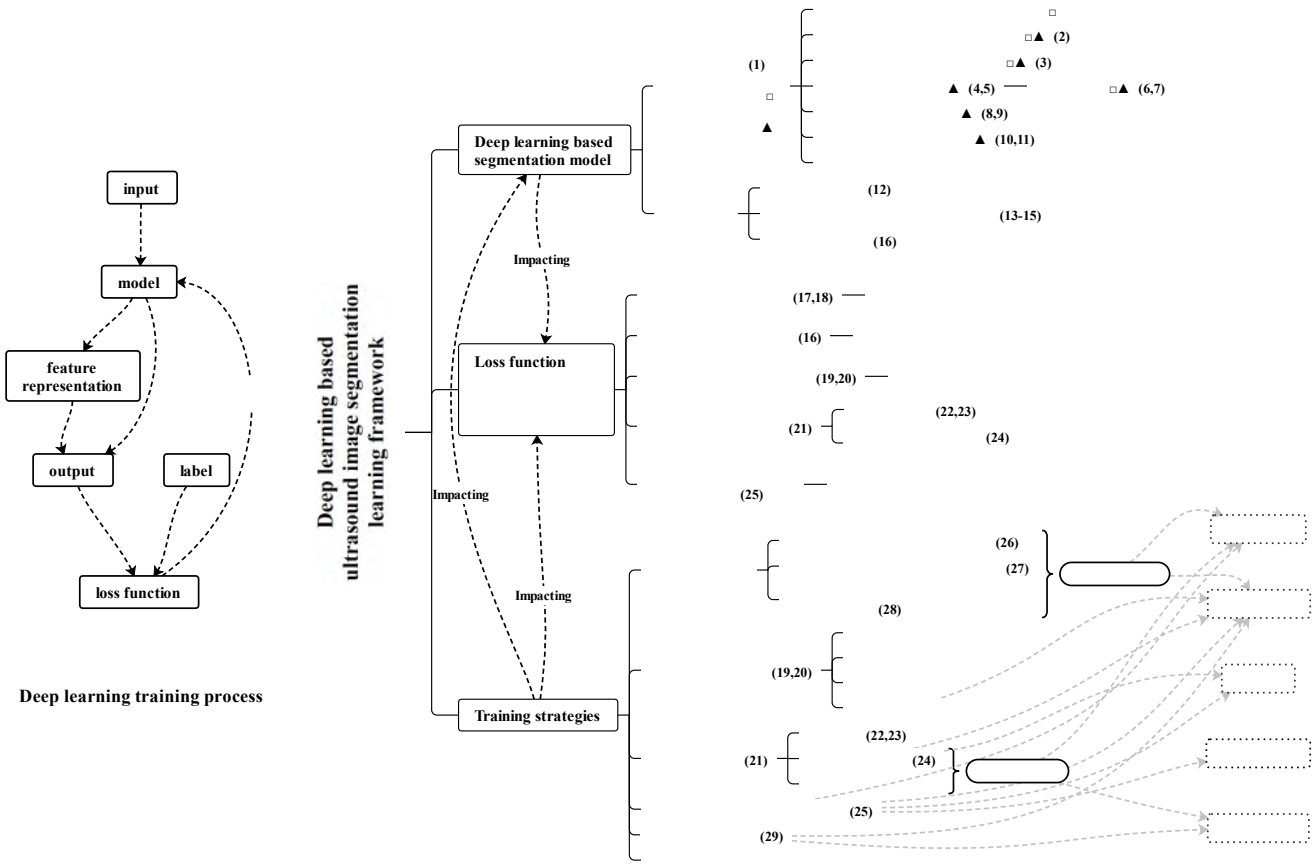

Fig. 1: Research directions for deep learning based ultrasound image segmentation. For the sake of brevity, we denote the references for corresponding topics in the form of numbers in the bracket. **Deep learning based segmentation model**: segmentation building blocks (1) [19], [20], RNNs (2) [21], GNNs (3) [22], attention (4,5) [23], [24], Transformer (6,7) [25], [26], multi-scale (8,9) [27], [28], boundary correction block (10,11) [6], [29]; architecture, Encoder-Decoder (12) [1], detection based segmentation architecture (13-15) [19], [27], [30]–[35], generative models (16) [36]. **Loss function**: segmentation task oriented (17,18) [37], [38]; generative models oriented (16) [36]; supervision strategies oriented (19,20) [39], [40]; transfer learning oriented (21-24) [41]–[44]; disentangled representation oriented (25) [45]. **Training strategies**: data augmentation (26-28) [46]–[48]; supervision strategies (19,20) [39], [40]; transfer learning (21-24) [41]–[44]; disentangled representation learning (25) [45]; curriculum learning (29) [49].

learning scenarios with different anatomical sites, imaging settings, supervision strategies, and dataset sizes.

To the best of our knowledge, no previous research has applied multi-task denoising-segmentation deep learning for segmentation purpose in ultrasound image analysis. Xie et al [58] applied the main task of denoising to preserve retinal structural information, where auxiliary segmentation task provided retina-related region information. Huang et al. [5] applied multi-task denoising-segmentation for segmentation purpose, where the scan noise — generated from moving 2D scanning, 3D formation and anatomical plane projection — is significantly different from the non-Gaussian statistics of speckle noise.

Regarding the holistic deep learning ultrasound segmentation that inherently reduces speckle noise, some state-of-the-art methods have designed speckle noise resistant framework based on holistic and heuristic rules about deep learning model, loss function, and training strategy. For example, Wu et al. [7] assumed that within- and cross-image long-range dependency modeling can extract consistent feature to alleviate noise disturbance. Observing that speckle noise and heart motion in echocardiography video make the inter-frame correspondence problem worse for the video segmentation, Wu et al. [18] designed context-aware U-Net encoders to extract feature map from 3 consecutive frames, while also designing spatiotemporal semantic calibration and bi-directional fusion modules to align the feature maps of consecutive frames for speckle-mitigating correspondence calculation. However, the overall robustness of segmentation performance cannot be explicitly credited to any denoising solutions, and it remains uncertain whether the segmentation performance can find potential rooms for further improvement if applying denoising with the other two solutions due to the lack of denoising-segmentation cross benchmarking, that is exactly the purpose of this work. We therefore explore the differences of segmentation performance among denoising pre-processing,

multi-task denoising-segmentation, and holistic deep learning segmentation framework to demonstrate whether denoising is necessary in segmentation performance improvement. This work has a fourfold contribution:

(1) We review deep learning ultrasound image segmentation and denoising methods in a holistic view, covering all possible combinations between the denoising-segmentation deep learning frameworks for the challenge of complex speckle noise.

(2) To the best of our knowledge, this is the first comprehensive denoising-segmentation cross benchmarking for addressing speckle noise in deep ultrasound segmentation. This benchmarking has concluded that denoising pre-processing brings unstable and slight (if exists) performance improvement to the downstream task of deep learning ultrasound segmentation, and we recommend to regard it as a kind of deep learning hyper-parameter, which should be checked in clinical application to judge its real effect on segmentation. While multi-task denoising-segmentation [5] actually results in a segmentation performance degradation, which may be the limitation of the cross-task gap that is possibly generated from the different context reasoning and input-output workflows in generalization or domain transfer.

(3) The holistic deep leaning ultrasound segmentation framework, including elaborate segmentation building blocks such as attention mechanism, transformer, and multi-scale mechanism, have been proved to effectively explore the contextual information and simultaneously reduce speckle noise in deep segmentation. Furthermore, the effectiveness of an intrinsic boundary correction with contextual perception in a holistic design, has been verified by experiments in our denoising-segmentation cross benchmarking. The context-aware holistic segmentation design with self-correction is much more evident and stable in superior performance improvements than the denoising pre-processing and multi-task denoising & segmentation.

(4) Acknowledging the performance improvement brought by an inherent denoising in semi-/weakly-/un-supervised deep learning, we recommend the proposed denoising-segmentation cross benchmarking to select denoising strategies in these cases. It should be noted that semi-/weakly-/un-supervised training might be far from clinical application, potential investigation should allow the recurrent emergence of context-aware generalization for a variety of heterogeneity in holistic deep learning.

II. Ultrasound Image Segmentation Deep Learning Framework

To tackle the challenges of multiplicative speckle noise, low-contrast features, ambiguous boundaries, and structural variations (see the challenge column in Table I), deep learning based ultrasound image segmentation is designed to learn **robust feature representations** via deep learning model and loss function as well as training strategy (see Figure 1) for enhancing segmentation performance. The whole procedure of deep learning training has been drawn on the left of Figure 1. Specifically, given a label-guided optimization [48], [62] loss function, the deep learning model can be updated to learn **feature representations** from the input data for an optimized output segmentation performance [40]. The training strategies are designed for better performance, training efficiency, generalization, stability, or interpretability. We additionally list out representative breast ultrasound and echocardiography segmentation algorithms in Table I according to ultrasound segmentation challenges and our review taxonomy.

A. Deep Learning Model

From the perspective of deep learning training, deep learning model is expected to extract and combine global semantic features and local detailed features appropriately for contextual understanding to get fine segmentation results [19], [20], [27], [28], [77]–[82]. We decompose the deep learning segmentation model into segmentation building blocks and overall segmentation architecture. The segmentation building blocks extract features, while the overall segmentation architecture coordinates the extracted features to produce the final output.

1) Segmentation Building Blocks: Segmentation building blocks, including convolutional neural networks (CNNs), recurrent neural networks (RNNs) [21], graph neural networks (GNNs) [22], attention mechanism [4], [17], [63]–[67], transformer [25], [26], and multi-scale mechanism [19], [83], are all designed with certain inductive bias [84]. CNNs are the most basic segmentation building blocks that apply local convolution in sliding windows, resulting in computational efficiency, local-space-invariance, grid spatial relationship modeling and training data efficiency. Compared with CNNs, other blocks introduce better segmentation oriented features from long-range dependencies, graph relationship, attention-selection and multi-scale combination.

RNNs capture sequential data dependencies by utilizing memory and hidden states in feedback connections [21]. Three typical ways of applying RNNs to image segmentation are: 1) organizing image information in a sequential order, such as associating intermediate feature grid points [85], encoding neighbouring patch relationships [86], or utilizing natural slice connections in 3D data [87]; 2) progressively optimizing RNN-based deep latent feature representation in multiple segmentation rounds [49]; and 3) constructing interaction/aggregation of multi-scale feature maps with ConvLSTM/GRU [88], [89]. The broad meaning of “recurrent” can be extented to feedback segmentation refinement without RNNs [6].

GNNs [22] process graph data that is composed of nodes and edges as well as initial node features in medical image segmentation by representing anatomical structure associations [71], [90], [91]. The node comes from initial mask [71], intermediate convolutional features [90], [92] or VAE latent distribution [91]. The edge is related to spatial distance [93] or self-attention relationship [94]. The role of GNN building block varies in the architecture, for example, completely utilizing graph and GNNs in decoder for feature representation and output representation [91], adding a GNN module for supplementary bottleneck feature [92], or adding a graph convolutional network based boundary rendering to further improve the segmentation accuracy by vertex adjustment [71].

As an effective approach for brain-like contextual understanding, attention mechanism [4], [17], [63]–[67] dynamically

TABLE I: Summary of deep learning based ultrasound segmentation challenges and methods. Segmentation challenges consist of the presence of multiplicative speckle noise, low-contrast features, ambiguous boundaries, structural variations, limited annotation, domain-gap, and etc. Attention mechanism [4], [17], [63]–[67], transformer [7], [16], multi-scale mechanism [17], [68]–[70] are hypothesized to reduce speckle noise effect. Boundary/residual correction and refinement [6], [17], [71] is popular in breast ultrasound segmentation, while motion-enhanced representation [72], [73] and multi-analysis task-aware learning are popular in echocardiography segmentation [66], [67], [72], [73].

Reference	Dimension	Anatomy	Challenge	Hypothesis/Idea	Deep learning model	Loss function	Training Strategy
Wang et al. [6]	2D	breast	(1) missing/ambiguous boundaries, speckle noise (2) large lesion variety (3) significant individual differences	a novel residual feedback network by learning residual representation of hardly-predicted pixels	a novel residual feedback network residual representation module residual feedback transmission strategy	seg: iou loss of initial segmentation and residual-guided segmentation, bce loss of residual representation	Not mentioned
Wu et al. [7]	2D	breast	(1) lesion variations (2) ambiguous boundaries (3) speckle noise and artifacts	within- and cross-image long-range dependency modeling	two parallel encoders: CNN and Transformer bottleneck: a cross-image dependency modeling module	seg: bce+dice model: cross-image dependency Loss	transforming based data augmentation transfer learning: pretrained CNN & transformer encoder backbone
Xue et al. [17]	2D	breast, prostate	(1) speckle artifacts (2) blurry boundaries (3) inhomogeneous intensity distributions	long-range non-local dependencies and boundary detection	multi-scale ASPP bottleneck global guidance block, spatial and channel attention boundary detection module on encoder CNNs	seg: bce+dice model&seg: mse loss of boudary map	transforming based data augmentation transfer learning: pre-trained ResNext backbone; multi-task of segmentation and boudary detection
Chen et al. [63]	2D	breast	(1) similar intensity distributions (2) variable morphologies (3) blurred boundaries (4) irregular shapes	adaptive attention	channel&spatial adaptive self-attention module for all convolutional blocks	seg: bce	not mentioned
Chen et al. [74]	2D	pneumonia, COVID-19, breast tumour	ImageNet pre-training has domain gap with medical training.	Self-supervised medical models are highly transferable.	feature extraction encoder (ResNet, ShuffleNet-v2) meta-weighting network Mask R-CNN for finetune: classification/segmentation heads	self-supervision: contrastive loss, weighted InfoNCE loss meta-learning loss	contrastive learning well-designed transforming based data augmentation, in geometric, clolor and mixup
Huang et al. [71]	2D	breast	(1) blurry or occluded edges (2) irregular nodule shapes	Boundary is important for automated BUS nodule segmentation.	Multi-scale ASPP bottleneck boundary selection module graph convolutional-based boundary rendering module	seg: cross entropy, point Cross-Entropy, L2 point matching loss	data augmentation: intensity jittering and flipping multi-task: region segmentation and boundary selection
Ning et al. [15]	2D	breast	(1) pattern complexity (2) Similar foreground-background intensity (3) low-contrast features and blurry boundary (4) lesion shape and position variations	background-salient representations for assisting foreground segmentation	foreground\background saliency maps U-Net shaped foreground path U-Net shaped background path straight middle path, background-assisted fusion unit, shape-aware unit, edge-aware unit and position-aware unit	seg: bce+dice model/multi-task: shape-related morphological information loss in shape-aware unit	not highlighted
Zhou et al. [64]	3D	breast	(1) tumour shape and size variations (2) uncertain tumour locations (3) blurry boundary, low signal-to-noise ratio (4) speckle noise and artifacts	tumour location information is essential	3D Mask R-CNN head for tumour location V-Net Cross-model attention mechanism in skip-connection layers aggregate Mask R-CNN location to V-Net feature level		

adjusts feature weights to highlight salient features and ignore irrelevant ones [23], [24]. According to the weighted targets, attention mechanism in image segmentation can be categorized into spatial attention that selects attentive regions, channel attention that selects useful feature channel [95], and hybrid attention. SK-U-Net [2] added channel attention alike SENet [95] after convolutional blocks in the encoder for better breast tumour segmentation with fewer model parameters. DAEFF-Net [4] used additional path of ECA module [96] based channel attention in feature extraction module and spatial attention for selective high- and low-level feature fusion for paediatric echocardiographic segmentation. Considering the low signal-to-noise of echocardiography and stronger feature coherence of local pixels than global pixels, PLANet [65] proposed pyramidal multi-head local attention module to enhance neighbouring feature while accommodating size variability. DSCG-Net [97] included a scale-based spatial attention to fuse multi-level features extracted by the encoder, and connected a centerline heatmap reconstruction side-branch network to the end of the encoder for increasing the network generalization in segmenting the common and internal carotid arteries.

Self-attention mechanism [98] characterizes the pairwise embedding correlations of all positions in a sequence (such as text, image patch, audio) to calculate weighted better embedding features. It is formulated as a weighted query mapping from key-value pairs to an output. It can be applied as a kind of spatial attention in image segmentation tasks, for example, Wu et al. [99] designed a non-local block for long-range dependency modeling in computer vision applications. Originated from self-attention, transformer has become a popular deep learning building block [25], [26]. Compared with CNNs, transformer can capture better long-range dependencies. On the other hand, however, it needs more training data to learn potential rules due to the weak inductive bias. Researchers in medical image segmentation have turned the research directions from pure transformer to the combination of CNNs and transformer as well as from transformer-based fully supervised training to transformer based pre-training [25], [26]. For example, Wu et al. [7] utilized parallel CNN&transformer encoder and a cross-image dependency modeling module for within- and cross-image long-range dependency modeling for breast ultrasound segmentation. Zhao et al. [16] constructed interactive fusion and learning between local convolution features and global transformer context information for key point locating in pediatric echocardiographic.

Multi-scale blocks collect long-range multi-scale dependencies and propagate local geometric contextual information through parallel pooling and atrous convolution of different scales in different deep intermediate layers [17], [19], [27], [28], [68], [70] as well as parallel encoder [69]. Specifically, atrous spatial pyramid pooling (ASPP) embedded the global contextual information in deep convolutional neural networks for semantic image segmentation [83]. CE-net [3] added an inception structure with multi-branches of astrous convolutions and a pooling block with multi-kernel of different sizes in the bottleneck, enlarging the receptive field and encoding global context information. An unsupervised multi-scale shape-aware strategy [100] captured long-range relationships in the high-order statistics that measure the joint distribution of classes at relative positions corresponding to different orientation and distances in cross-domain image segmentation. MDF-Net [70] employed a two-stage architecture with a multiscale feature selection sub-network and a structurally optimized refinement sub-network, mitigating speckle noise and inter-subject variation via better feature exploration and fusion. Multi-scale semantic features in the different intermediate layers can be further refined by attention [17], graph representation [71] or joint alignment of cross-domain invariant information [100], and then be received at decoder to progressively recover geometric details from the interaction of rich but noisy contexts for the fine segmentation.

Recently, uncertainty-based boundary correction or boundary edge refinement [29], [101]–[103] with context perception is proposed to address the high ambiguity of object boundary representation and the high variability of poor ultrasound image quality. By exploiting dynamic boundary preservation block to predict a key boundary point (KBP) map [102] for enhancing the semantic features from images, SCCNet [29] proposed an iterative training strategy to update the importance value of the KBP map for U-Net training and use a weighted cross-entropy loss to give more attention to the KBP. A context module also incorporated a class-level context using the predicted segmentation map to construct a dynamic multi-scale filter with adaptive kernel weights for more contextual perception in discriminating similar objects. RF-Net [6] incorporated a novel residual representation module to grasp the residual characteristics of the ambiguous boundaries and perplexing regions. This incorporation facilitates the network in directing increased attention towards the pixels that are challenging to predict. SABR-Net [103] addressed the missing and ambiguous boundaries in the contexts of shadow artifacts via semi-supervised shadow-aware network with boundary refinement, by adding shadow imitation regions to the original images and design shadow-masked tranformer blocks to perceive missing anatomy. A densely connected 3D pyramidal dilated convolutuion network [104] is proposed with sequential cross-frame uncertainty guidance to exploit the longitudinal information and perceive size-varied vessel regions for intravascular ultrasound sequence segmentation. All of these boundary correction schemes can explore the transcending boundary edge contextual information to filter out the unreliable boundary or edge predictions in intermediate feature maps and multi-scale consistency segmentation.

2) Architecture: We categorize the architectures into Encoder-Decoder architecture [19], [20], [27], [28], detection-based segmentation architecture [19], [27], [33]–[35] and generative models based architecture [47].

(a) Encoder-Decoder architecture [19], [20], [27], [28]. As the most popular segmentation architecture, the Encoder-Decoder architecture first encodes the input data into the latent feature, and then decodes the latent feature into the output segmentation mask.

The abovementioned segmentation building blocks can be interconnected flexibly within the Encoder-Decoder architecture. A classic CNN example of an Encoder-Decoder architecture is the U-Net [1], which is comprised of convolutional neu-

ral networks, pooling/upsampling layers, and skip-connection layers. U-Net stands out as the most widely adopted biomedical segmentation model due to its versatility and simplicity [77], [79], [80], [105]–[107]. Other segmentation building blocks in the preceding sub-section except CNNs are often incorporated into U-Net-like networks to enhance performance. For example, DAEFF-Net [4] with attention mechanism, CE-Net [3] with multi-scale mechanism, and CDM [7] with both CNN and transformer encoders.

Moreover, there are also modifications to add elaborate structure-level elements and holistically connect them to basic U-Net structure. For example, RF-Net [6] designed two sequential U-Net with recurrent initial segmentation residual representation and feedback transmission to enhance the segmentation confidence of missing/ambiguous boundary pixels with uncertainty rectification. SMU-Net [15] designed two parallel U-Net shaped paths and a straight middle path to additionally utilize background saliency maps as input to improve foreground segmentation performance.

(b) Detection-based segmentation architecture [19], [27], [33]–[35]. Different from the Encoder-Decoder architecture, detection-based segmentation architecture explicitly integrates the procedures of detection and segmentation. As two-stage framework, Mask R-CNN [30] firstly regressed the object boundary box through Faster R-CNN, and then added a segmentation branch. In Mask R-CNN based medical image segmentation, Lian et al. [108] detected anatomical parts and diseases proposals and then mined structure-aware relationship for the detection and segmentation of thoracic diseases. Ding et al. [109] improved the Mask R-CNN architecture for ultrasound nerve segmentation through multi-scale mechanism, attention mechanism, and upsampling skip-connection. Furthermore, box annotations [35] are required as coarse mask generation for pseudo mask labels in weakly supervised instance segmentation, then these labels are utilized as training samples for the self-training instance segmentation stage. As one-stage frameworks, PolarMask [31], [110] via polar coordinate and contour proposal networks [32]–[34] via contour modeling utilized the regressed shape representation for the simultaneous object detection and segmentation. Compared with the Encoder-Decoder segmentation architecture, detection-based segmentation under the guidance of region proposal and shape representation are more computationally efficient, suitable for instance segmentation, and beneficial for context relationship expression, while the Encoder-Decoder segmentation architecture has the advantage of fine segmentation.

(c) Generative model architecture [36], [47]. Different from the above discriminative models that directly characterize the conditional probability $P(Y|X = x)$, where X is the observable variable, Y is the target variable, and x is an observation, generative models formulate the joint data probability distribution $P(X; Y)$, with the advantage of better task understanding and uncertainty expression [36], [47]. In regard to the taxonomy and principle of generative models, we refer the readers to [36]. Generative adversarial networks consist of a generator for distribution formulating and a discriminator for reality judgement. The key is to guide the weight updating of the generator for representative features under the feedback from the discriminator. Ruan et al. [111] and Mahmood et al. [112] used adversarial training to distinguish tumours from cysts and distinguish overlapped nuclei, respectively. Diffusion model [113] is very popular recently, including noise adding based forward diffusion and denoising based backward diffusion. Wu et al. [114], [115] applied diffusion model with denoising effective network designs into medical image segmentation and achieved state-of-the-art performance. Compared with discriminative models, generative models incorporate more comprehensive distribution and can generate new data to assist target tasks, while discriminative models extract more practical features.

B. Loss Function

Loss function in deep learning characterizes the difference between the predicted result and the expected ground truth / state, and then the difference is minimized by updating the deep learning model parameters through the back-propagation algorithm, resulting in meaningful feature representation.

The basic segmentation-task oriented loss functions consist of region based losses (such as dice coefficient loss and intersection over union loss), distribution based loss (such as cross-entropy loss and focal loss), boundary based loss, and compound loss [28], [37], [38]. For boundary loss functions for better boundary segmentation performance, Kervadec et al. [116] designed a differentiable integral of non-symmetric L_2 distance metric over the contours, while Du et al. [38] paid attention to the boundary pixels gotten through dilation and erosion operations in the similar form of dice coefficient loss. Region based, distribution based and boundary loss functions can be easily combined together to form compound loss function. We refer the readers to [37] for more details about medical segmentation losses.

The deep learning model design and training strategies might interact with the loss function to achieve a holistic segmentation design. Generative models introduce the adversarial loss in generative adversarial networks, and the noise prediction loss in diffusion models. Semi-supervised learning introduces the consistency loss, while unsupervised learning utilizes clustering loss, reconstruction loss, or contrastive loss [39], [40]. In transfer learning, multi-task solution [43] might simultaneously utilize segmentation and classification loss function, while domain adaption tries to minimize the domain gap between the source domain and the target domain through loss function such as maximum mean discrepancy [44]. In disentangled representation learning, loss function, such as latent regression loss [45], can be used to improve the latent representation. Further, due to their complexity, domain adaptation and disentangling representation learning involve different aspects of deep learning, such as generative models, segmentation task and semi-/self-/un-supervised strategies, to boost their performance, and thus use combined loss functions. In these cases, it is important to balance various loss functions to get good model performance.

C. Training Strategies

Training strategies are designed for better performance, training efficiency, generalization, stability, or interpretability. Considering the focus and length of the paper, we simply mention the multi-task learning. Multi-task learning jointly learns related tasks to extract task-shared features and improve task-specific features, improving efficiency, generalization and performance of the target task or all tasks [42], [43]. To achieve the multi-task feature learning, the model architecture defines the information flow according to the connection order of task structures, and the architectures can be categorized into cascaded, parallel, interacted and hybrid [43]. We refer the readers to [42], [43] for detailed information. We have introduced related works of multi-task learning of ultrasound denoising and segmentation in the Introduction.

Some other training strategies are marked in Figure 1. We refer the readers to corresponding references in Figure 1 for detailed knowledge.

D. Summarization

Typical ultrasound breast and echocardiography segmentation algorithms are summarized in Table I. It can be seen that many works hypothesized that deep learning segmentation framework can reduce speckle noise effect. Holistic segmentation building blocks, including attention mechanism [4], [17], [63]–[67], [7], [16], multi-scale mechanism [17], [68] are the most popular strategies. Moreover, boundary correction and refinement [6], [17], [71] is popular in breast ultrasound segmentation, while motion enhanced representation [72], [73] and multi-task learning (chamber-view classification, quantification, uncertainty estimation) [66], [67], [72], [73] is popular in echocardiography segmentation.

Table I additionally show challenges and solutions to labeled data limitation [39], [68], [75], uncertainty estimation [67], [117] and label coherence [65], [118]. While not extensively examined in this work, they are also hot topics in deep learning ultrasound segmentation.

III. Ultrasound Denoising

denoising methods can be classified into four main categories [119]–[123]: spatial domain filtering, transform domain filtering, deep learning filtering, and hybrid methods. Spatial domain filtering includes local adaptive, non-local means (NLM), PDE (partial differential equation), and total variation (TV) based methods. All of these abovementioned denoising methods are summarized in Table II.

A. Spatial Domain Filtering

(a) Local adaptive filtering. Local adaptive filter rectifies a pixel referring to local statistics, including weighted average value [124], [144], [145], median value [125] and extreme value [126]. These filters have the advantage of simple principle and fast speed, but the performance heavily depends on local window size.

(b) Non-local means filtering. Different from the local filters that focus on local similarity, NLM utilizes self-similarity in a larger window. Measured as Euclidean intensity distance between reference blocks and the selected block, self-similarity measure is used to perform a weighted average among central pixels of reference blocks to obtain the new intensity for the central pixel of the selected block [127]. Optimal Bayesian NLM (OBNLM) [128] introduces NLM for ultrasound denoising by designing a new similarity measure called Pearson distance. OBNLM can well preserve structural information with suitable parameters, but it has relevant limitations when dealing the problems of high computational cost, optimal parameter selection and imperfect similarity measurement. To improve OBNLM's similarity measure, Zhan et al. [129] refined common distance through principal component analysis. To tackle bias due to speckle noise, Sudeep et al. [130] proposed an unbiased NLM method that estimated and subtracted independent bias signal using the maximum likelihood method. According to [123], the groundbreaking NLM principle has stimulated important progress by being integrated with transform domain based or TV based methods.

(c) PDE and TV based methods. Spatial domain methodology has seen an increasing research attention and dominant performance improvement from PDE-based methods [131], [133]–[135], [146] and TV [135], [147], [148] methods. PDE-based methods mainly apply the anisotropic diffusion (AD) [131], [133], [134], [146] under the guidance of diffusion coefficients [131], encouraging internal diffusion in similar regions and inhabiting interaction between different regions. The diffusion coefficient consists of three parts, boundary edge information, noise information and a non-negative decreasing function, keeping a balance between edge preservation and noise removal. Various AD-based denoising methods have been proposed by implementing different edge detection and detail-preserving strategies. The edge and noise information of classical AD methods has been shown in Table II. Specifically, Gabor-based anisotropic diffusion method [146] exploited Gabor transform to detect tissue edges and enhance the discrimination between the edges and the noise for improving the diffusion and despeckling performance. Sudeb et al. [134] represented edge information in another PDE-based method, which injected the past edge information into the diffusion and preserved fine features. TV based methods mainly focus on fidelity term and regularization term [135] to use smooth and regularization prior. Mei et al. [50] utilized phase congruence-based edge significance measure called phase asymetry to adaptively detect edge features, and integrated AD with TV in order to leverage the strengths of AD in homogeneous regions and TV in the proximity of features. All of these AD methods have relatively low computational cost and perform well at low-to-moderate noise levels, but some of them may fail in noise reduction at high noise levels, especially on noisy edges with small diffusion coefficient.

B. Transform Domain Filtering

Transform domain filtering uses various basis functions to represent ultrasound images, in which wavelet based methods are the most widely explored [123], and the core idea is to remove the noise-related coefficients in transform domain, for example, by the means of thresholding methods. For detailed

TABLE II: Summary of the abovementioned denoising methods.

Category	Basic principle	Examples	Description	Advantage	Disadvantage
Local adaptive filtering	Rectifies a pixel referring to local statistics based on local similarity.	Lee [124]	Weighted average value	Simple principle and fast speed.	The performance heavily depends on local window size.
		Loupas et al. [125]	Median value		
		Tay et al. [126]	Extreme value		
Non-local means filtering	Updates pixel by performing weighted average on the center pixels of non-local blocks with self-similarity measure.	NLM [127]	The original NLM work	NLM stimulates important progress. OBNLM can well preserve structur-8.96606preserve	

It's difficult for a trained denoising model to adapt to the scenarios of different noise levels. In scenarios of low noise level, challenge mainly lies in feature preservation especially in edge detail preservation; while for a high noise level scenario, challenge lies in both noise reduction and feature preservation. To adapt to different scenarios in different noise levels, some methods add different levels of artificial noises to the training data [139], [156], while Lan et al. [137] trained different models for different noise levels, and when tackling a new speckled image, they firstly estimated the noise level and then selected corresponding model for denoising. In a more integrated way for an addictive Gaussian noise, Soh et al. [140] utilized noisy image's prior from variational auto-encoder to facilitate image restoration at different noise levels. Due to the non-symmetrical and spatially correlated characteristics of the speckle noise [61], modeling the non-Gaussian speckle noise circumstances at different noise levels is a very challenging topic that is not touched in existing deep learning models.

Once trained, deep learning based denoising can be fast, however, the model training is relatively complex, and as mentioned above, the final performance is greatly affected by training data and treatment measures of different noise levels.

D. Hybrid Denoising Methods

The advantages and limitations of all abovementioned denoising methods have been summarized correspondingly. Hybrid methods combine different methods to synthesize different advantages and overcome corresponding limitations [53], [148]. Karamjeet et al. [157] successively utilized the local and non-local filters to achieve a better performance trade-off between noise reduction and feature preservation. Gilboa et al. [158] combined non-local operators with TV to improve the texture preservation ability of PDE based methods, which was introduced to tackle multiplicative speckle noise using split Bregman iterations in [141], [159]. As a milestone of denoising, BM3D [142] and SARBM3D [143] combined non-local block matching with transform domain thresholding for a first denoising, and then used the Wiener filter based on the matching result of the first step to recover feature details. Deep learning denoising models that combine traditional denoising with deep learning knowledge for detail-preserving domain adaptation have been introduced in Section III-C. Hybrid denoising methods can achieve better denoising performance with a high computational cost.

We refer the interested reader to a broad view of Gaussian noise based image filtering, image synthesis and regularizing general inverse problems in survey work [150]. However, handling non-Gaussian speckle noise removal and image segmentation in a holistic view is still underexplored.

IV. Materials and Methods

In this section, we describe the datasets, denoising method, deep learning based segmentation framework, training setup, segmentation metric [160], and statistical analysis for our experiments about the denoising effect of deep learning based ultrasound image segmentation.

TABLE III: Dataset size and dataset split

Dataset	Amount of training images	Amount of validation images	Amount of testing images	Ratio for split
Dataset 1	378	126	126	6:2:2
Dataset 2	147	16	None	10-folder cross-validation [54]
Dataset 3	14930	2576	2554	5.86:1:1 [163]
Dataset 4	962	324	312	6:2:2

A. Datasets

According to [13], [14], ultrasound imaging is used to examine many body parts, such as breast, prostate, heart, liver, nerve, fetus, and so on. Considering efficiency, reproducibility, and the lack of public ultrasound datasets, we choose two representative applications, breast ultrasound imaging and echocardiography imaging. In this way, our experiment is based on four public and two self-collected ultrasound datasets. Following are detailed descriptions.

1) Dataset 1: a breast dataset, BUSI[1] [161], collected in Baheya Hospital with LOGIQ E9 and LOGIQ E9 Agile. Images with one tumour delineated are used in our work.

2) Dataset 2: a breast dataset, Dataset B[2] [162], collected from the UDIAT Diagnostic Centre with Siemens ACUSON Sequoia C512 system and 17L5 HD linear array transducer.

3) Dataset 3: a big dataset of echocardiography videos, Echodynamic[3] [163], collected from Stanford University Hospital with a resolution of 112 × 112. As with [163], we mixed end-systolic and end-diastolic frames in both training and testing.

4) Dataset 4: an echocardiography dataset, Camus[4] [164], collected from the University Hospital of St Etienne with GE Vivid E95 ultrasound scanner and GE M5S probe and with a resolution of 512 × 512. As with [164], our experiment uses good and medium quality cases, but excludes poor cases. We segment the left ventricle endocardium. We mix four-chamber and two-chamber frames, as well as end-systolic and end-diastolic frames.

5) External Datasets for Test: We additionally collected 115 echocardiography images from the center for cardiovascular medicine of Shanghai chest hospital in Shanghai city and 47 breast ultrasound images from Haimen district traditional Chinese medical hospital in Nantong city of Jiangsu province for external cross-dataset testing.

Detailed dataset split of Dataset 1-4 for training, validation, and testing is shown in Table III.

B. Ultrasound Denoising

As mentioned in the Section III, thousands of denoising methods can be grouped into four main categories and totally six categories. Due to the impracticality of testing all denoising methods individually, we selected one or two representative methods with public source code for each category. Specifically, we choose Lee's filter (Lee) [124], optimized Bayesian non-local means (OBNLM) [128], detail-preserving anisotropic diffusion (DPAD) [133], phase asymmetry ultrasound denoising with fractional anisotropic diffusion and total

[1] https://scholar.cu.edu.eg/?q=afahmy/pages/dataset
[2] http://www2.docm.mmu.ac.uk/STAFF/M.Yap/dataset.php
[3] https://echonet.github.io/dynamic/
[4] https://www.creatis.insa-lyon.fr/Challenge/camus/

variation (PFDTV) [50], generalized likelihood estimation method (GLM) [136] based on wavelet, denoising convolutional neural networks (DnCNN) [165] and block-matching 3-D algorithm for complex speckle noise removal in synthetic aperture radar (SARBM3D) [143].

Denoising results after necessary parameter tuning are shown in Figure 2. The parameters in experiments are typically initialized based on the recommendations of the original authors in their respective works [166], [167]. Subsequently, the despeckling results of 20 images on each dataset were meticulously evaluated by two experienced medical professionals, who assessed the performance in aspect of speckle noise removal and detail preservation. Modifications to the denoising parameters were implemented in accordance with expert guidance and feedback to optimize denoising performance tailored to the specific dataset and

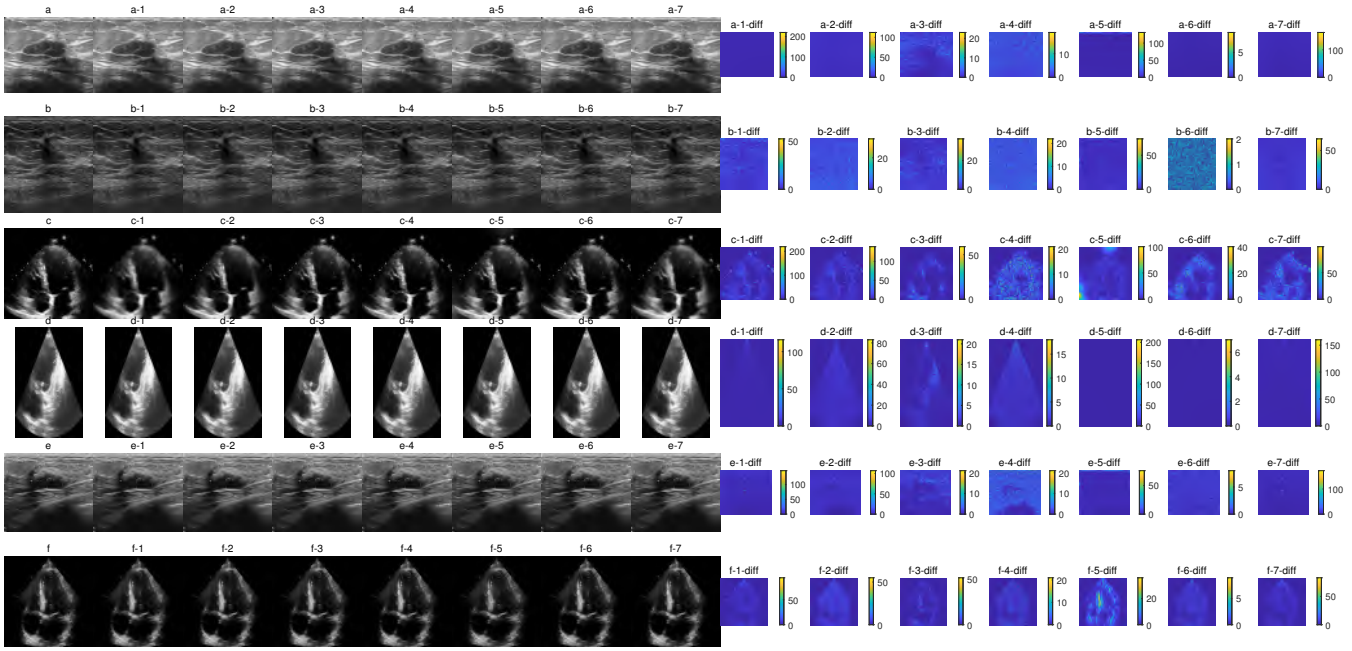

Fig. 2: Denoising results after necessary parameter tuning to avoid artifact. On the left of the figure, a-f represent Dataset 1-4, external breast ultrasound dataset, and external echocardiography dataset, respectively. And 1-7 represent despeckling results of Lee, OBNLM, DPAD, PFDTV, GLM, DnCNN, SAR-BM3D, respectively. Lee: window size = 3. OBNLM: M = 7, alpha = 3, h = 0.7, offset = 100. DPAD: (time step and iterations) breast ultrasound datasets, 0.2 and 100; Dataset 3 and external echocardiography dataset, 0.1 and 30; Dataset 4, 0.02 and 30. PFDTV: Δt = 0.15, s = 15, k0 = 20, alpha = 1.2, niter = 8. GLM: threshold factor = 3, window size = 3, decomposition scale = 4. DnCNN: has been described in the context. SAR-BM3D: number of looks = 1, decomposition level = 3, block(window) size 8 × 8, search area size 39 × 39. The right side of the figure shows absolute difference maps between the despeckled results and original images.

in which, T/F and P/N indicate the amount of pixels which are predicted consistently/inconsistently with the ground truth and positive/negative respectively. The combination of T/F and P/N represents a logical "AND".

The equations for 95% HD and ASSD are as follows:

$$\begin{aligned} 95\%\ HD &= max\{95\%sup_{x\in X}(d(x;Y)); 95\%sup_{y\in Y}(d(X;y))\} \\ ASSD &= \frac{1}{n_X + n_Y}\left(\sum_{x\in X} d(x;Y) + \sum_{y\in Y} d(X;y)\right) \\ d(x;Y) &= \inf_{y\in Y} d(x;y) \\ d(x;y) &= \|x - y\|_2 \end{aligned} \tag{3}$$

in which, x and y represent single point on the surface of segmentation mask, X and Y represent the surface of segmentation mask with many points, n_X and n_Y represent the amount of points on the surface X and Y respectively, inf gets minimal value from following expression, and $95\%sup$ gets 0.95 quantile from following expression.

Surface dice coefficient measures the surface overlap of two masks at a clinically accepted distance tolerance, which is set as 1 pixel in our experiments.

F. Statistical Analysis

Friedman test [175]–[177] and Nemenyi post hoc test [177], [178] are utilized to evaluate the performance difference among all 8 versions of datasets (the original dataset and 7 despeckled datasets). When Friedman test results in $p < 0.05$ that indicates a statistically significant difference, Nemenyi post hoc test will then be run to show the detailed difference. In Nemenyi post hoc test, if the average rank difference between two methods exceeds the critical difference, then there may exists statistically significant performance difference between the two methods. Friedman test and Nemenyi post hoc test can be respectively done by scipy.stats.friedmanchisquare[7] and Orange.evaluation.compute_CD[8].

Paired t-test and Wilcoxon signed-ranks test are used to find the performance difference between the original dataset and a specifically despeckled dataset. Paired t-test is commonly used to evaluate segmentation difference between different algorithms [179]–[181], however, according to [182], Wilcoxon signed-ranks test is more sensitive than the paired t-test. We apply both paired t-test and Wilcoxon signed-ranks test to show the relative amplitude of difference. Implemented by scipy.stats.ttest_ind and scipy.stats.wilcoxon, these paired tests can directly indicate a difference of greater or less.

V. Experimental Results

The experimental results consist of segmentation metrics, statistical analysis results, and some representative visualized

[7] https://docs.scipy.org/doc/scipy
[8] https://orange3.readthedocs.io/en/latest/

TABLE IV: Training details for different segmentation tasks

	Tasks	Data augmentation	Input size	Input channel	Batch size	Total training epochs	Learning rate scheduler
Dataset 1	U-Net	a	256	1	4	200	c
	SK-U-Net	a	256	1	4	100	d
	CE-Net	a	256	1	4	80	d
	RF-Net	a	256	1	4	80	d
	CDM	a	256	1	4	200	c
	Multi-task	a	256	1	2	200	c
Dataset 2	U-Net SK-U-Net	a	256	1	4	200	c
	CE-Net	a	256	1	4	80	d
	RF-Net	a	256	1	4	80	d
	CDM	a	256	1	4	200	c
	Multi-task	a	256	1	2	200	c
Dataset 3	U-Net DAEFF-Net	b	112	3	32	25	e
	RF-Net	b	128	3	32	25	e
	CDM	b	112	3	32	25	e
	U-Net (semi-supervision)	b	112	3	32	100	f
	Multi-task	b	112	3	8	40	e
Dataset 4	U-Net DAEFF-Net	b	512	1	4	40	e
	RF-Net	b	512	1	4	25	e
	CDM	b	512	1	4	25	e
	Multi-task	b	256	1	2	40	e

a. random rotation in a range, random shift, random scaling, and random rotation of multiples of 90°
b. random rotation in a range, random shift, and random scaling
c. constant learning rate of 0.0001
d. initial learning rate of 0.0001, halves when the loss does not drop for 7 consecutive epochs, with the least learning rate of 5e-6
e. initial learning rate of 0.0001, halves at 40%, 60% and 80% of the training process
f. initial learning rate of 0.0001, halves when the loss does not drop for 5 consecutive epochs, with the least learning rate of 1e-6

segmentation results. **For clarity, we place all detailed experimental results in the Appendix A/B and analyse the results in this section.** Firstly, we introduce the meaning of statistical analysis results in Table A1-A14, where Table A1-A4 is for fully-supervised intra-dataset testing, Table A5 is for semi-supervised intra-dataset testing, Table A6-A13 is for fully-supervised cross-dataset testing, and Table A14 is for multi-task intra-dataset testing. Secondly, we analyse Friedman test results. Thirdly, we analyse paired test results on intra- and cross-dataset testing. Fourthly, we show some visualized segmentation results in Figure B1-B7 to illustrate the consistency of visualization results and statistical test results. Finally, we summarize the overall denoising performance.

A. Meaning of Tables

Segmentation metrics, Friedman test results and paired test results are shown in Table A1-A13.

1) Friedman Test and Nemenyi post hoc Test: Friedman test is applied based on median metrics in the table. The average

both test, respectively. For a specific combination of dataset, model, and denoising method, if there are more than 3 of 7 segmentation metrics evaluated as statistically better/worse than the original dataset, then this combination will be set as the shadings of light/dark grey.

The paired test results in different rows describe the performance difference of a despeckled dataset compared with the original dataset. The paired test results in the original column describe the performance difference of a model compared with U-Net. Particularly, the paired test results in the original column of Table A5 describe the performance difference of semi-supervised U-Net compared with the fully supervised U-Net in Table A3.

B. Result and Analysis of Friedman Tests and Nemenyi post hoc Tests

For the Friedman tests in Table A1-A14 indicating statistically significant difference, Nemenyi post hoc tests are applied in Figure A1, A2 and A3 to show specific differences. Particularly, we additionally show Nemenyi post hoc tests for all the intra-dataset and cross-dataset testing results in Figure A1(c) and A2(k), respectively. For example, Figure A1(a) shows that OBNLM and DnCNN despeckled Dataset 2 are statistically significantly better than DPAD despeckled Dataset 2.

To sum up, no denoising method performs statistically better than using the original dataset directly. DPAD [133] and Lee have a relatively better overall rankings than other denoising methods in the intra-dataset tesing (Figure A1(c)) and cross-dataset testing (Figure A2(k)), respectively.

C. Result and Analysis of Paired Tests

1) Intra-dataset Testing: All paired test results for intra-dataset testing of denoising pre-processing are shown in Table A1-A5. Analysis of these paired tests can be concluded in Table V.

As a whole, good effect from denoising pre-processing is unstable and slight (if exists) in various medical ultrasound deep learning scenarios.

The **unstability** in Dataset 1 and Dataset 2 with CDM network: in Dataset 1, CDM network has good despeckling effect, while the despeckling effect disappears when dealing with Dataset 2. The unstability in Dataset 4 with U-Net, DAEFF-Net, RF-Net: using U-Net has no denoising effect, and using DAEFF-Net has good denoising effect, while no denoising effect is achieved when using RF-Net. The reason maybe that U-Net is not sensitive to small boundary and detail changes brought by denoising; while DAEFF-Net is superior to U-Net in boundary refinement due to the attention mechanism, and can sense subtle boundary changes from denoising; while RF-Net already has boundary/residual refinement mechanism in its own model architecture, and so that subtle boundary changes from denoising is of little importance. There are three cases with good denoising effect in conditions of fixed dataset and model: (a) Dataset 1 with CDM network; (b) semi-supervision on Dataset 3 with smaller dataset size than the fully supervised training and higher noise levels than Dataset 4; (c) Dataset 4 with large image size, medium dataset size and slightly better performed DAEFF-Net. However, for each dataset, when using the best-performing model (see the specific metrics, RF-Net for Dataset 1, CDM for Dataset 2, U-Net for Dataset 3, RF-Net for Dataset 4), denoising brings no good effect.

The **improvement** brought by denoising is much weaker than the good ones by stronger model and more training data. It can be seen in the original columns of Table A1-A5 that most improvement by stronger model and more training data can be indicated by both paired t-test and Wilcoxon signed-ranks test, while statistically and significantly better denoising effect indicated by both tests only appears in semi-supervised U-Net on PFDTV despeckled Dataset 3.

Considering the single denoising method, DPAD [133] performs the best. Only DPAD gets non-worse performance in all cases, and DPAD despeckled datasets perform statistically better than original datasets in 6 of the total 19 cases.

Regarding the multi-task learning of denoising and segmentation with SFS block [5] based feature fusion (in Table A14), we have seen the obvious performance degradation compared with the holistic single-task segmentation learning.

2) Cross-dataset Testing: Firstly, in cross-dataset testings (Table A6-A13), severe heterogeneity brings a large performance drop compared with intra-dataset testings, for example, U-Net on intra Dataset 1 testing gets the dice of 0.9210(0.0539) (median metircs and interquartile range), while U-Net on cross testing of testing Dataset 2 with models trained based on Dataset 1 gets the dice of 0.8781(0.3334).

Secondly, as a whole, denoising performs badly compared with using the original dataset directly, with 72 ✔, 66 ● and 114 ☒ in Table VI, reflecting unstable denoising effect to some extent. Good denoising effect in some cases of intra-dataset testing has been weakened or lost in cross-dataset testing. Lee has the highest proportion of cases with good denoising effect, with 21 ✔, 4 ● and 11 ☒.

Thirdly, cross-dataset ultrasound denoising improvement is weak and far from being able to alleviate the performance degradation due to the severe heterogeneity in cross-dataset deep learning of ultrasound segmentation, and the performance metrics are still very low compared with intra-dataset testing.

In summary, it is impossible to get stable and considerable segmentation improvement from denoising in the training and testing stages for cross-dataset testing purpose. More emphasis should be placed on a generalized self-supervised deep learning model, larger heterogeneous datasets, strong self-correcting learning framework for a wide variety of heterogeneity problems [62], [183]–[185].

D. Visualization Results

To illustrate the overall similar performance of different denoising methods in intra-dataset testing indicated by Friedman tests, visualized segmentation results of U-Net on Dataset 1-4 are shown in Figure B1-B4. Further, to illustrate the statistically better or worse performance indicated by paired tests on more than 3 segmentation metrics, visualized segmentation results of 3 tasks are shown in Figure B5-B7.

TABLE V: Intra-dataset testing analysis. The PSNR is calculated by comparing the original dataset with the DPAD denoised dataset, and lower PSNR means higher noise level. The dataset size indicates the size of fully supervised training data. The arrows after the model name, " (#); ⇑⇑ (⇓⇓); * (+) (has the same meaning with section V-A), indicate the performance of this model is statistically significantly different from corresponding fully-supervised U-Net. The signs of ✔, ● and ☒ represent good, no and bad denoising effect, respectively.

Dataset				Model	Denoising effect (22 ✔, 87 ●, 24 ☒)						
Dataset name	PSNR	Image size	Dataset size		Lee	OBNLM	DPAD	PFDTV	GLM	DnCNN	SARBM3D
Dataset 1	43.8604	256	medium (378)	U-Net	☒	●	●	●	☒	●	●
				SK-U-Net	●	●	●	✔	●	●	☒
				CE-Net⇑	●	●	●	●	●	●	☒
				RF-Net⇑	●	●	●	●	●	●	●
				CDM⇑⇑	✔	✔	✔	●	●	●	✔
Dataset 2	39.9187	256	small (147)	U-Net	●	●	●	●	☒	●	●
				SK-U-Net	●	●	●	●	●	✔	●
				CE-Net⇑	●	✔	●	●	●	●	●
				RF-Net⇑	●	●	●	●	●	●	●
				CDM⇑	●	●	●	●	●	●	●
Dataset 3	40.3899	112	large (14930)	U-Net	☒	☒	✔	●	☒	●	☒
				DAEFF-Net⇓	☒	☒	✔	●	☒	☒	☒
				RF-Net⇓	✔	●	✔	●	☒	✔	●
				CDM⇓	●	☒	●	●	☒	✔	☒
			small (200)	U-Net⇓	✔	☒	✔	✔	☒	✔	☒
Dataset 4	53.606	512	medium (962)	U-Net	●	●	●	●	●	●	●
				DAEFF-Net⇑	✔	✔	✔	●	✔	●	✔
				RF-Net⇑	●	●	●	●	●	●	●
				CDM⇑	☒	☒	●	☒	●	●	●

(1) Considering all 133 cases as a whole, good denoising effect is unstable and slight compared with good model effect, with 22 ✔, 87 ● and 24 ☒.
(2) Three cases with good denoising effects are: (a) Dataset 1 with CDM network;
(b) the semi-supervision on Dataset 3 with smaller dataset size than the fully supervised training, while also having higher noise level than Dataset 4;
(c) Dataset 4 with large image size, medium dataset size and slightly better performed DAEFF-Net.
(3) Considering single denoising method, DPAD performs the best.
(4) The improvement brought by denoising is much weaker than the improvements by stronger models and more training data.
(5) For each dataset, when using the best-performing model (RF-Net, CDM, U-Net, RF-Net for Dataset 1-4), despeckling brings no good effect.

TABLE VI: Cross-dataset testing analysis. The arrows before and after the model name indicate the performance of this model compared with U-Net on intra-dataset and cross-dataset testing studies respectively.

Intra-dataset training and testing				Cross-dataset testing		Models	Denoising effect (72 ✔, 66 ●, 114 ☒)						
Dataset name	PSNR	Image size	Good denoising effect	Dataset name	PSNR		Lee	OBNLM	DPAD	PFDTV	GLM	DnCNN	SARBM3D
Dataset 1	43.8604	256	✔	Dataset 2	39.9187	U-Net	☒	☒	☒	☒	●	●	☒
						SK-U-Net⇓	✔	✔	✔	✔	✔	✔	✔
						⇑CE-Net⇑	✔	●	☒	●	●	●	●
						⇑RF-Net⇑	●	●	☒	●	●	●	●
						⇑⇑CDM⇑	✔	✔	●	✔	●	✔	✔
				External bus	40.6974	U-Net	☒	●	●	✔	●	✔	●
						SK-U-Net⇓	✔	●	●	✔	●	☒	●
						⇑CE-Net⇑	✔	☒	☒	☒	☒	☒	✔
						⇑RF-Net⇑	✔	●	☒	●	●	●	✔
			✔			⇑⇑CDM⇑	✔	●	☒	●	☒	●	✔
Dataset 2	39.9187	256		Dataset 1	43.8604	U-Net	☒	☒	☒	☒	☒	☒	☒
						⇑SK-U-Net⇑	☒	☒	☒	☒	☒	☒	☒
						⇑CE-Net⇑	✔	✔	✔	●	✔	●	✔
						⇑RF-Net⇑	✔	●	☒	✔	✔	✔	☒
						⇑CDM⇑	✔	●	☒	✔	✔	●	✔
				External bus	40.6974	U-Net	☒	☒	☒	☒	☒	☒	●
						SK-U-Net⇑	☒	☒	☒	☒	☒	●	●
						⇑CE-Net	✔	●	●	●	✔	●	✔
						⇑RF-Net⇓	✔	☒	☒	●	●	●	●
						⇑CDM⇓	✔	●	☒	☒	●	●	✔
Dataset 3	40.3899	112	DPAD	Dataset 4	53.606	U-Net	✔	✔	✔	✔	✔	✔	✔
			DPAD			⇓DAEFF-Net⇑	✔	☒	✔	☒	☒	☒	☒
			DPAD			⇓RF-Net⇑	✔	☒	✔	☒	☒	☒	☒
						⇓CDM⇑	✔	☒	☒	☒	☒	☒	☒
			DPAD	External echo	44.5495	U-Net	☒	☒	●	☒	☒	☒	☒
			DPAD			⇓DAEFF-Net⇓	●	☒	☒	☒	☒	☒	☒
			DPAD			⇓RF-Net⇓	●	☒	✔	☒	☒	☒	●
						⇓CDM⇓	●	☒	☒	☒	☒	☒	☒
Dataset 4	53.606	512	✔	Dataset 3	40.3899	U-Net	✔	✔	✔	✔	✔	✔	✔
						⇑DAEFF-Net⇑	✔	☒	●	●	☒	●	☒
						⇑RF-Net⇑	☒	✔	☒	☒	●	☒	☒
						⇑CDM⇑	☒	☒	☒	☒	☒	☒	☒
			✔	External echo	44.5495	U-Net	✔	✔	✔	✔	☒	✔	✔
						⇑DAEFF-Net⇑	✔	●	●	●	●	●	●
						⇑RF-Net⇑	☒	●	☒	●	☒	☒	☒
						⇑CDM⇑	☒	☒	☒	☒	☒	✔	☒

(1) Considering all 252 cases as a whole, good despeckling effect is unstable and slight compared with good model effect, with 72 ✔, 66 ●, 114 ☒.
(2) From the lines where intra-dataset testing with good denoising effect in the table,
it can be seen that good denoising effect in some cases of intra-dataset testing has been weakened or lost in cross-dataset testing.
(3) Considering the single denoising method, Lee has the best performance, with 21, 4 and 11 cases of good, no and bad denoising effects, respectively.
Good denoising effect is unstable in cross-dataset testing compared with using the original dataset directly.

TABLE VII: Time cost (s) for 7 denoising methods and 6 datasets

	Lee	OBNLM	DPAD	PFDTV	GLM	DnCNN	SARBM3D
Dataset 1	0.042	5.841	2.246	3.425	276.026	0.034	68.452
Dataset 2	0.010	3.839	1.322	2.276	6.071	0.023	4.761
Dataset 3	0.028	0.568	0.101	0.309	1.339	0.056	7.026
Dataset 4	0.026	10.556	1.172	6.623	6.628	0.008	127.991
External bus	0.013	4.149	1.342	2.329	122.605	0.025	50.619
External echocardiography	0.026	4.206	1.203	2.143	28.997	0.024	54.578

E. Overall Denoising Performance

In regard to the speed of method, we have used Lee [124] and DnCNN [165] methods that have the shortest running time, while also implementing DPAD [133] as a relatively faster method. The average time cost for 6 datasets and 7 denoising methods is shown in Table VII. Among these data, the running time for DnCNN is calculated on a computer with Nvidia GeForce RTX 3090, Intel Core i9-10900K CPU, and 64 GB DDR4 memory, and the rest is calculated on a computer with Matlab R2021a, Intel Core i7-8700K CPU, and 16 GB memory. The order from fast to slow is Lee, DnCNN, DPAD, PFDTV, OBNLM, GLM, and SARBM3D.

Regarding the denoising effect, we have described that DPAD [133] has the best performance in intra-dataset testing, and Lee [124] has relatively better performance in cross-dataset testing. **Considering both speed and denoising effects, we recommend DPAD and Lee for relatively better denoising hyper-parameter.**

VI. Discussion

A comprehensive denoising-segmentation cross benchmarking is designed to assess whether denoising is necessary for deep learning ultrasound segmentation, when acquiring a performance balance between denoising and segmentation in achieving not only the speckle noise robustness but also in improving the segmentation performance. We comprehensively cover three denoising-segmentation solutions, some state-of-the-art denoising and deep learning models, segmentation metrics, statistical analysis, as well as ultrasound deep learning scenarios and additionally eliminate the nondeterminism effect in the deep learning training. However, there are some limitations in our method. Firstly, we overlook the situation where speckle pattern is a beneficial feature [186]–[189] since speckles are produced by microstructure-like scatterers with a size smaller than the wavelength of ultrasonic pulse waves in living soft tissues [190], [191]. Speckle pattern has been used to represent echocardiography motion [192], [193], calculate breast tumour classification feature [194], [195]. Secondly, our experiment is mostly based on random 6:2:2 dataset split due to the running efficiency and the certain rationality of random 6:2:2 dataset split. We simply use cross-validation on the smallest Dataset 2.

Generally, the segmentation performance improvement from denoising pre-processing is more unstable and slighter (if exists) compared with the improvement from holistic deep learning segmentation framework. It might be noted that the denoising effect is similar to the unstable effect of deep learning training hyper-parameter that can yield unpredictable outcomes—either small positive/negative or neutral effect.

However, the context-aware and self-correcting capability of holistic deep segmentation framework is contemplated as a major performance boost for ultrasound image segmentation. Specifically, data-driven and label-guided deep learning approaches have demonstrated the ability to acquire robust features and reduce speckle noise. Techniques like attention mechanisms, transformers, multi-scale architectures have been proven to be valuable building blocks for image segmentation. Additionally, task-adaptive strategies such as boundary/residual correction and refinement and the incorporation of motion-enhanced feature representation and multi-analysis task-aware learning have gained popularity in model design. Moreover, small detail enhancement and adjustment at the boundaries of input data typically can be efficiently adapted and/or corrected by the holistic deep learning system, where **the system's self-correction and segmentation refinement are partially achieved by the framework components (Figure 1) and can be exploited by contextual perception (attention, transformer, multi-scale architectures and etc.) and uncertainty rectification (e.g., boundary/residual correction and refinement) towards the fine segmentation.**

Our experiments also demonstrate that multi-task denoising-segmentation with SFS block [5] based feature fusion does cause a segmentation performance degradation. This issue is directly due to the introduction of interference from the denoising sub-network into the segmentation sub-network within the SFS block-based feature fusion. While the possible root causes consist of imperfect denoising ground truth, sub-optimal multi-task feature fusion, and the inherent unreliability of the multi-task denoising-segmentation paradigm, which compels low-level image processing tasks to complement mid- and high-level image processing tasks.

Further experiments are required to pinpoint the exact cause of degraded multi-task denoising-segmentation performance. Nevertheless, it is theoretically probable that significant effort will be expended with the expectation of sub-optimal results. Firstly, acquiring perfect ground truth for deep learning-based ultrasound denoising has long been a challenging issue. Secondly, some multi-task learning have their limitations since the cross-task gap may exist due to the different reasoning and input-output workflows. Furthermore, different task-aware learning simply focused on the structural representation of task-specific entities, it may miss some important task-agnostic semantic information or some unstructured representation in the contextual emergence for the overall performance boosting.

In the deep learning segmentation review section, we have summarized various research directions, in which good feature representation, generalization, stability, and interpretability have become more and more popular recently. For example, some researches published in top journals are very concerned about generalization topics [183], [185], [196], [197] that are closely related to domain adaptation and disentangling representation learning. Continuing the practice of denoising for speckle removal contradicts the more favorable path toward the generalization for a wide variety of heterogeneity problems. Furthermore, the emergence of robust universal segmentation

model, such as Segment Anything [198]–[200] and camouflaged object segmentation [201]–[203] in open-world deep learning [204], [205], has brought a new era of universal and interactive image understanding that might be generally and contextually immune to the complex noises. It will take a long time to generate truly universal image understanding without manual point/ROI prompt or fine-tuning on specific task, however, the exploratory work such as Segment Anything greatly enhanced the confidence of domain adaptation and heterogeneity solutions that also should be robust to the complex speckle noise.

VII. Conclusion

We have fully established a denoising-segmentation cross benchmarking from the unified perspective of segmentation and denoising, and gives a certain conclusion about the denoising effect in deep learning based ultrasound image segmentation. The holistic deep framework outperforms despeckling pre-processing and multi-task denoising-segmentation methods in addressing the challenge of speckle noise and improving segmentation performance.

Regarding denoising as a kind of hyper-parameter in deep learning framework, a broader perspective is to unify contextual perception and uncertainty rectification in a more integrated segmentation framework with simultaneous speckle robustness, whatever the source of every individual uncertainty in single-task deep learning. It is then a holistic context-aware and self-correcting deep learning segmentation problem. The proposed cross benchmarking work can be used as an important reference for researches in the field of deep learning, medical ultrasound image analysis, synthetic aperture radar [206], [207] or optical coherent imaging [208], [209].

Acknowledgment

The authors thank all cited authors for providing the source code used in this work and the anonymous reviewers for their valuable comments on this paper.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] M. Byra, P. Jarosik, A. Szubert, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O'Boyle, C. Comstock, and M. Andre, "Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network," *Biomedical Signal Processing and Control*, vol. 61, p. 102027, 2020.

[3] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.

[4] L. Guo, B. Lei, W. Chen, J. Du, A. F. Frangi, J. Qin, C. Zhao, P. Shi, B. Xia, and T. Wang, "Dual attention enhancement feature fusion network for segmentation and quantitative analysis of paediatric echocardiography," *Medical Image Analysis*, vol. 71, p. 102042, 2021.

[5] Z. Huang, R. Zhao, F. H. Leung, S. Banerjee, T. T.-Y. Lee, D. Yang, D. P. Lun, K.-M. Lam, Y.-P. Zheng, and S. H. Ling, "Joint spine segmentation and noise removal from ultrasound volume projection images with selective feature sharing," *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1610–1624, 2022.

[6] K. Wang, S. Liang, and Y. Zhang, "Residual feedback network for breast lesion segmentation in ultrasound image," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 471–481.

[7] H. Wu, X. Huang, X. Guo, Z. Wen, and J. Qin, "Cross-image dependency modelling for breast ultrasound segmentation," *IEEE Transactions on Medical Imaging*, 2023.

[8] H. V. Pham, S. Qian, J. Wang, T. Lutellier, J. Rosenthal, L. Tan, Y. Yu, and N. Nagappan, "Problems and opportunities in training deep learning software systems: An analysis of variance," in *Proceedings of the 35th IEEE/ACM international conference on automated software engineering*, 2020, pp. 771–783.

[9] P. Nagarajan, G. Warnell, and P. Stone, "Deterministic implementations for reproducibility in deep reinforcement learning," *arXiv preprint arXiv:1809.05676*, 2018.

[10] Z. Jiang, S. E. Salcudean, and N. Navab, "Robotic ultrasound imaging: State-of-the-art and future perspectives," *Medical Image Analysis*, vol. 89, p. 102878, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136184152300138X

[11] N. Masoumi, H. Rivaz, I. Hacihaliloglu, M. O. Ahmad, I. Reinertsen, and Y. Xiao, "The big bang of deep learning in ultrasound-guided surgery: A review," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 70, no. 9, pp. 909–919, 2023.

[12] Y. Wang, X. Ge, H. Ma, S. Qi, G. Zhang, and Y. Yao, "Deep learning in medical ultrasound image analysis: A review," *IEEE Access*, vol. 9, pp. 54 310–54 324, 2021.

[13] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, and T. Wang, "Deep learning in medical ultrasound analysis: a review," *Engineering*, vol. 5, no. 2, pp. 261–275, 2019.

[14] A. Carovac, F. Smajlovic, and D. Junuzovic, "Application of ultrasound in medicine," *Acta Informatica Medica*, vol. 19, no. 3, p. 168, 2011.

[15] Z. Ning, S. Zhong, Q. Feng, W. Chen, and Y. Zhang, "Smu-net: Saliency-guided morphology-aware u-net for breast lesion segmentation in ultrasound image," *IEEE Transactions on Medical Imaging*, vol. 41, no. 2, pp. 476–490, 2021.

[16] C. Zhao, W. Chen, J. Qin, P. Yang, Z. Xiang, A. F. Frangi, M. Chen, S. Fan, W. Yu, X. Chen *et al.*, "Ift-net: Interactive fusion transformer network for quantitative analysis of pediatric echocardiography," *Medical Image Analysis*, vol. 82, p. 102648, 2022.

[17] C. Xue, L. Zhu, H. Fu, X. Hu, X. Li, H. Zhang, and P.-A. Heng, "Global guidance network for breast lesion segmentation in ultrasound images," *Medical image analysis*, vol. 70, p. 101989, 2021.

[18] H. Wu, J. Liu, F. Xiao, Z. Wen, L. Cheng, and J. Qin, "Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion," *Medical Image Analysis*, vol. 78, p. 102397, 2022.

[19] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[20] 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*, vol. 7, no. 6, pp. 545–569, 2023.

[21] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[22] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 4–24, 2020.

[23] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 5, pp. 1–32, 2021.

[24] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, 2022.

[25] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives," *Medical image analysis*, p. 102762, 2023.

[26] H. Xiao, L. Li, Q. Liu, X. Zhu, and Q. Zhang, "Transformers in medical image segmentation: A review," *Biomedical Signal Processing and Control*, vol. 84, p. 104791, 2023.

[27] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Processing*, vol. 16, no. 5, pp. 1243–1267, 2022. [Online].

Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ipr2.12419
[28] N. S. Punn and S. Agarwal, "Modality specific u-net variants for biomedical image segmentation: a survey," *Artificial Intelligence Review*, pp. 1–45, 2022.
[29] Y. Gong, H. Zhu, J. Li, J. Yang, J. Cheng, Y. Chang, X. Bai, and X. Ji, "Sccnet: Self-correction boundary preservation with a dynamic class prior filter for high-variability ultrasound image segmentation," *Computerized Medical Imaging and Graphics*, vol. 104, p. 102183, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895611123000010
[30] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
[31] E. Xie, W. Wang, M. Ding, R. Zhang, and P. Luo, "Polarmask++: Enhanced polar representation for single-shot instance segmentation and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5385–5400, 2022.
[32] E. Upschulte, S. Harmeling, K. Amunts, and T. Dickscheid, "Contour proposal networks for biomedical instance segmentation," *Medical image analysis*, vol. 77, p. 102371, 2022.
[33] Y. Yang, W. Yu, H. Du, L. Ling, Q. Feng, S. Tu, and W. Yang, "Coupled contour regression for efficient delineation of lumen and external elastic lamina in intravascular ultrasound images," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2023.
[34] S. Bohlender, I. Oksuz, and A. Mukhopadhyay, "A survey on shape-constraint deep learning for medical image segmentation," *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 225–240, 2023.
[35] T.-T. Zhang, H. Shu, Z.-R. Tang, K.-Y. Lam, C.-Y. Chow, X.-J. Chen, A. Li, and Y.-Y. Zheng, "Weakly supervised real-time instance segmentation for ultrasound images of median nerves," *Computers in Biology and Medicine*, vol. 162, p. 107057, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S001048252300522X
[36] J. M. Tomczak, *Deep generative modeling*. Springer, 2022.
[37] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. L. Martel, "Loss odyssey in medical image segmentation," *Medical Image Analysis*, vol. 71, p. 102035, 2021.
[38] J. Du, K. Guan, P. Liu, Y. Li, and T. Wang, "Boundary-sensitive loss function with location constraint for hard region segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2022.
[39] G.-J. Qi and J. Luo, "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
[40] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical Image Analysis*, p. 102444, 2022.
[41] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
[42] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.
[43] Y. Zhao, X. Wang, T. Che, G. Bao, and S. Li, "Multi-task deep learning for medical image computing and analysis: A review," *Computers in Biology and Medicine*, p. 106496, 2022.
[44] H. Guan and M. Liu, "Domain adaptation for medical image analysis: a survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2021.
[45] X. Liu, P. Sanchez, S. Thermos, A. Q. O'Neil, and S. A. Tsaftaris, "Learning disentangled representations in the imaging domain," *Medical Image Analysis*, p. 102516, 2022.
[46] G. Menghani, "Efficient deep learning: A survey on making deep learning models smaller, faster, and better," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–37, 2023.
[47] Y. Chen, X.-H. Yang, Z. Wei, A. A. Heidari, N. Zheng, Z. Li, H. Chen, H. Hu, Q. Zhou, and Q. Guan, "Generative adversarial networks in medical image augmentation: a review," *Computers in Biology and Medicine*, p. 105382, 2022.
[48] M. A. Bansal, D. R. Sharma, and D. M. Kathuria, "A systematic review on data scarcity problem in deep learning: solution and applications," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–29, 2022.
[49] J. Wang, Y. Tang, Y. Xiao, J. T. Zhou, Z. Fang, and F. Yang, "Grenet: Gradually recurrent network with curriculum learning for 2-d medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
[50] K. Mei, B. Hu, B. Fei, and B. Qin, "Phase asymmetry ultrasound despeckling with fractional anisotropic diffusion and total variation," *IEEE Transactions on Image Processing*, vol. 29, pp. 2845–2859, 2020.
[51] J. Virmani, R. Agarwal *et al.*, "Assessment of despeckle filtering algorithms for segmentation of breast tumours from ultrasound images," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 1, pp. 100–121, 2019.
[52] A. Abbasian Ardakani, A. Bitarafan-Rajabi, A. Mohammadzadeh, A. Mohammadi, R. Riazi, J. Abolghasemi, A. Homayoun Jafari, and M. Bagher Shiran, "A hybrid multilayer filtering approach for thyroid nodule segmentation on ultrasound images," *Journal of Ultrasound in Medicine*, vol. 38, no. 3, pp. 629–640, 2019.
[53] H. G. Khor, G. Ning, X. Zhang, and H. Liao, "Ultrasound speckle reduction using wavelet-based generative adversarial network," *IEEE Journal of Biomedical and Health Informatics*, 2022.
[54] H. Lee, M. H. Lee, S. Youn, K. Lee, H. M. Lew, and J. Y. Hwang, "Speckle reduction via deep content-aware image prior for precise breast tumor segmentation in an ultrasound image," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2022.
[55] H. Chen, Y. Wang, J. Shi, J. Xiong, J. Jiang, W. Chang, M. Chen, and Q. Zhang, "Segmentation of lymph nodes in ultrasound images using u-net convolutional neural networks andnetwY6(l)]TJbo0(")-5enerati6(netw16(c

images," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023.
[70] W. Qi, H. C. Wu, and S. C. Chan, "Mdf-net: A multi-scale dynamic fusion network for breast tumor segmentation of ultrasound images," *IEEE Transactions on Image Processing*, vol. 32, pp. 4842–4855, 2023.
[71] R. Huang, M. Lin, H. Dou, Z. Lin, Q. Ying, X. Jia, W. Xu, Z. Mei, X. Yang, Y. Dong *et al.*, "Boundary-rendering network for breast lesion segmentation in ultrasound images," *Medical image analysis*, vol. 80, p. 102478, 2022.
[72] H. Wei, J. Ma, Y. Zhou, W. Xue, and D. Ni, "Co-learning of appearance and shape for precise ejection fraction estimation from echocardiographic sequences," *Medical Image Analysis*, vol. 84, p. 102686, 2023.
[73] W. Xue, H. Cao, J. Ma, T. Bai, T. Wang, and D. Ni, "Improved segmentation of echocardiography with orientation-congruency of optical flow and motion-enhanced segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6105–6115, 2022.
[74] Y. Chen, C. Zhang, C. H. Ding, and L. Liu, "Generating and weighting semantically consistent sample pairs for ultrasound contrastive learning," *IEEE Transactions on Medical Imaging*, 2022.
[75] X. Cao, H. Chen, Y. Li, Y. Peng, S. Wang, and L. Cheng, "Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation," *IEEE transactions on medical imaging*, vol. 40, no. 1, pp. 431–443, 2020.
[76] A. Gilbert, M. Marciniak, C. Rodero, P. Lamata, E. Samset, and K. Mcleod, "Generating synthetic labeled data from existing anatomical models: an example with echocardiography segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2783–2794, 2021.
[77] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020.
[78] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82 031–82 057, 2021.
[79] M. Xian, Y. Zhang, H. Cheng, F. Xu, B. Zhang, and J. Ding, "Automatic breast ultrasound image segmentation: A survey," *Pattern Recognition*, vol. 79, pp. 340–355, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320318300645
[80] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 137–178, 2021.
[81] J. Jiang, Y. Guo, Z. Bi, Z. Huang, G. Yu, and J. Wang, "Segmentation of prostate ultrasound images: The state of the art and the future directions of segmentation algorithms," *Artif. Intell. Rev.*, vol. 56, no. 1, p. 615–651, apr 2023. [Online]. Available: https://doi.org/10.1007/s10462-022-10179-4
[82] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability*, vol. 13, no. 3, p. 1224, 2021.
[83] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
[84] D. Mahapatra, A. Poellinger, and M. Reyes, "Interpretability-guided inductive bias for deep learning based medical image," *Medical image analysis*, vol. 81, p. 102551, 2022.
[85] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville, "Reseg: A recurrent neural network-based model for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 41–48.
[86] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, "Scene labeling with lstm recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3547–3555.
[87] W. Xuan and G. You, "Detection and diagnosis of pancreatic tumor using deep learning-based hierarchical convolutional neural network on the internet of medical things platform," *Future Generation Computer Systems*, vol. 111, pp. 132–142, 2020.
[88] R. Li, K. Li, Y.-C. Kuo, M. Shu, X. Qi, X. Shen, and J. Jia, "Referring image segmentation via recurrent refinement networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5745–5753.
[89] T. Wollmann, M. Gunkel, I. Chung, H. Erfle, K. Rippe, and K. Rohr, "Gruu-net: Integrated convolutional and gated recurrent neural network for cell segmentation," *Medical image analysis*, vol. 56, pp. 68–79, 2019.
[90] Y. Meng, H. Zhang, Y. Zhao, D. Gao, B. Hamill, G. Patri, T. Peto, S. Madhusudhan, and Y. Zheng, "Dual consistency enabled weakly and semi-supervised optic disc and cup segmentation with dual adaptive graph convolutional networks," *IEEE transactions on medical imaging*, 2022.
[91] N. Gaggion, L. Mansilla, C. Mosquera, D. H. Milone, and E. Ferrante, "Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: Applications to chest x-ray analysis," *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 546–556, 2023.
[92] H. Jia, H. Tang, G. Ma, W. Cai, H. Huang, L. Zhan, and Y. Xia, "A convolutional neural network with pixel-wise sparse graph reasoning for covid-19 lesion segmentation in ct images," *Computers in Biology and Medicine*, vol. 155, p. 106698, 2023.
[93] Y. Lu, Y. Chen, D. Zhao, B. Liu, Z. Lai, and J. Chen, "Cnn-g: Convolutional neural network combined with graph for image segmentation with theoretical analysis," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 3, pp. 631–644, 2020.
[94] Q. Ma, S. Zhou, C. Li, F. Liu, Y. Liu, M. Hou, and Y. Zhang, "Dgrunit: Dual graph reasoning unit for brain tumor segmentation," *Computers in Biology and Medicine*, vol. 149, p. 106079, 2022.
[95] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
[96] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 531–11 539.
[97] M. Jiang and B. Chiu, "A dual-stream centerline-guided network for segmentation of the common and internal carotid arteries from 3d ultrasound images," *IEEE Transactions on Medical Imaging*, vol. 42, no. 9, pp. 2690–2705, 2023.
[98] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[99] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
[100] P. Wang, J. Peng, M. Pedersoli, Y. Zhou, C. Zhang, and C. Desrosiers, "Shape-aware joint distribution alignment for cross-domain image segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 8, pp. 2338–2347, 2023.
[101] D. Mishra, S. Chaudhury, M. Sarkar, and A. S. Soin, "Ultrasound image segmentation: A deeply supervised network with attention to boundaries," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 6, pp. 1637–1648, 2019.
[102] H. J. Lee, J. U. Kim, S. Lee, H. G. Kim, and Y. M. Ro, "Structure boundary preserving segmentation for medical image with ambiguous boundary," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[103] F. Chen, L. Chen, W. Kong, W. Zhang, P. Zheng, L. Sun, D. Zhang, and H. Liao, "Deep semi-supervised ultrasound image segmentation by using a shadow aware network with boundary refinement," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2023.
[104] M. Xia, H. Yang, Y. Huang, Y. Qu, G. Zhou, F. Zhang, Y. Wang, and Y. Guo, "3d pyramidal densely connected network with cross-frame uncertainty guidance for intravascular ultrasound sequence segmentation," *Physics in Medicine & Biology*, vol. 68, no. 5, p. 055001, feb 2023. [Online]. Available: https://dx.doi.org/10.1088/1361-6560/acb988
[105] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.
[106] D. Gut, Z. Tabor, M. Szymkowski, M. Rozynek, I. Kucybała, and W. Wojciechowski, "Benchmarking of deep architectures for segmentation of medical images," *IEEE Transactions on Medical Imaging*, 2022.
[107] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.
[108] J. Lian, J. Liu, S. Zhang, K. Gao, X. Liu, D. Zhang, and Y. Yu, "A structure-aware relation network for thoracic diseases detection and

segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 8, pp. 2042–2052, 2021.
[109] Y. Ding, I. Member, Q. Yang, Y. Wang, D. Chen, Z. Qin, and J. Zhang, "Mallesnet: A multi-object assistance based network for brachial plexus segmentation in ultrasound images," *Medical Image Analysis*, vol. 80, p. 102511, 2022.
[110] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, "Polarmask: Single shot instance segmentation with polar representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 193–12 202.
[111] Y. Ruan, D. Li, H. Marshall, T. Miao, T. Cossetto, I. Chan, O. Daher, F. Accorsi, A. Goela, and S. Li, "Mb-fsgan: Joint segmentation and quantification of kidney tumor on ct by the multi-branch feature sharing generative adversarial network," *Medical image analysis*, vol. 64, p. 101721, 2020.
[112] F. Mahmood, D. Borders, R. J. Chen, G. N. McKay, K. J. Salimian, A. Baras, and N. J. Durr, "Deep adversarial training for multi-organ nuclei segmentation in histopathology images," *IEEE transactions on medical imaging*, vol. 39, no. 11, pp. 3257–3267, 2019.
[113] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
[114] J. Wu, H. Fang, Y. Zhang, Y. Yang, and Y. Xu, "Medsegdiff: Medical image segmentation with diffusion probabilistic model," *arXiv preprint arXiv:2211.00611*, 2022.
[115] J. Wu, R. Fu, H. Fang, Y. Zhang, and Y. Xu, "Medsegdiff-v2: Diffusion based medical image segmentation with transformer," *arXiv preprint arXiv:2301.11798*, 2023.
[116] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," *Medical image analysis*, vol. 67, p. 101851, 2021.
[117] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information fusion*, vol. 76, pp. 243–297, 2021.
[118] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical image analysis*, vol. 65, p. 101759, 2020.
[119] T. Joel and R. Sivakumar, "An extensive review on despeckling of medical ultrasound images using various transformation techniques," *Applied Acoustics*, vol. 138, pp. 18–27, 2018.
[120] N. Biradar, M. L. Dewal, and M. K. Rohit, "Speckle noise reduction in b-mode echocardiographic images: A comparison," *IETE Technical Review*, vol. 32, no. 6, pp. 435–453, 2015.
[121] S. V. M. Sagheer and S. N. George, "A review on medical image denoising algorithms," *Biomedical signal processing and control*, vol. 61, p. 102036, 2020.
[122] A. E. Ilesanmi and T. O. Ilesanmi, "Methods for image denoising using convolutional neural network: a review," *Complex & Intelligent Systems*, vol. 7, no. 5, pp. 2179–2198, 2021.
[123] B. Goyal, A. Dogra, S. Agrawal, B. S. Sohi, and A. Sharma, "Image denoising review: From classical to state-of-the-art approaches," *Information fusion*, vol. 55, pp. 220–244, 2020.
[124] J.-S. Lee, "Digital image enhancement and noise filtering by use of local statistics," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 165–168, 1980.
[125] T. Loupas, W. McDicken, and P. L. Allan, "An adaptive weighted median filter for speckle suppression in medical ultrasonic images," *IEEE transactions on Circuits and Systems*, vol. 36, no. 1, pp. 129–135, 1989.
[126] P. C. Tay, C. D. Garson, S. T. Acton, and J. A. Hossack, "Ultrasound despeckling for contrast enhancement," *IEEE Transactions on Image Processing*, vol. 19, no. 7, pp. 1847–1860, 2010.
[127] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 60–65.
[128] P. Coupé, P. Hellier, C. Kervrann, and C. Barillot, "Nonlocal means-based speckle filtering for ultrasound images," *IEEE transactions on image processing*, vol. 18, no. 10, pp. 2221–2229, 2009.
[129] Y. Zhan, M. Ding, L. Wu, and X. Zhang, "Nonlocal means method using weight refining for despeckling of ultrasound images," *Signal Processing*, vol. 103, pp. 201–213, 2014.
[130] P. Sudeep, P. Palanisamy, J. Rajan, H. Baradaran, L. Saba, A. Gupta, and J. S. Suri, "Speckle reduction in medical ultrasound images using an unbiased non-local means method," *Biomedical Signal Processing and Control*, vol. 28, pp. 1–8, 2016.
[131] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 12, no. 7, pp. 629–639, 1990.
[132] Y. Yu and S. T. Acton, "Speckle reducing anisotropic diffusion," *IEEE Transactions on image processing*, vol. 11, no. 11, pp. 1260–1270, 2002.
[133] S. Aja-Fernández and C. Alberola-López, "On the estimation of the coefficient of variation for anisotropic diffusion speckle filtering," *IEEE Transactions on Image Processing*, vol. 15, no. 9, pp. 2694–2701, 2006.
[134] S. Majee, R. K. Ray, and A. K. Majee, "A new non-linear hyperbolic-parabolic coupled pde model for image despeckling," *IEEE Transactions on Image Processing*, vol. 31, pp. 1963–1977, 2022.
[135] S. K. Jain and R. K. Ray, "Non-linear diffusion models for despeckling of images: achievements and future challenges," *IETE Technical Review*, vol. 37, no. 1, pp. 66–82, 2020.
[136] A. Pizurica, W. Philips, I. Lemahieu, and M. Acheroy, "A versatile wavelet domain noise filtration technique for medical imaging," *IEEE transactions on medical imaging*, vol. 22, no. 3, pp. 323–331, 2003.
[137] Y. Lan and X. Zhang, "Real-time ultrasound image despeckling using mixed-attention mechanism based residual unet," *IEEE Access*, vol. 8, pp. 195 327–195 340, 2020.
[138] X. Feng, Q. Huang, and X. Li, "Ultrasound image de-speckling by a hybrid deep network with transferred filtering and structural prior," *Neurocomputing*, vol. 414, pp. 346–355, 2020.
[139] O. Karaoğlu, H. Ş. Bilge, and İ. Uluer, "Removal of speckle noises from ultrasound images using five different deep learning networks," *Engineering Science and Technology, an International Journal*, vol. 29, p. 101030, 2022.
[140] J. W. Soh and N. I. Cho, "Variational deep image restoration," *IEEE Transactions on Image Processing*, vol. 31, pp. 4363–4376, 2022.
[141] F. Dong, H. Zhang, and D.-X. Kong, "Nonlocal total variation models for multiplicative noise removal using split bregman iteration," *Mathematical and Computer Modelling*, vol. 55, no. 3-4, pp. 939–954, 2012.
[142] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.
[143] S. Parrilli, M. Poderico, C. V. Angelino, and L. Verdoliva, "A nonlocal sar image denoising algorithm based on llmmse wavelet shrinkage," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 606–616, 2011.
[144] D. T. Kuan, A. A. Sawchuk, T. C. Strand, and P. Chavel, "Adaptive noise smoothing filter for images with signal-dependent noise," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 165–177, 1985.
[145] V. S. Frost, J. A. Stiles, K. S. Shanmugan, and J. C. Holtzman, "A model for radar images and its application to adaptive digital filtering of multiplicative noise," *IEEE Transactions on pattern analysis and machine intelligence*, no. 2, pp. 157–166, 1982.
[146] Q. Zhang, H. Han, C. Ji, J. Yu, Y. Wang, and W. Wang, "Gabor-based anisotropic diffusion for speckle noise reduction in medical ultrasonography," *J. Opt. Soc. Am. A*, vol. 31, no. 6, pp. 1273–1283, Jun 2014. [Online]. Available: https://opg.optica.org/josaa/abstract.cfm?URI=josaa-31-6-1273
[147] Z. Jin, J. Wang, L. Min, and M. Zheng, "An adaptive total generalized variational model for speckle reduction in ultrasound images," *Journal of the Franklin Institute*, vol. 359, no. 15, pp. 8377–8394, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0016003222005531
[148] X. Liu and W. Lian, "Non-convex high-order TV and L_0-norm wavelet frame-based speckle noise reduction," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 12, pp. 5174–5178, 2022.
[149] S. V. Mohd Sagheer and S. N. George, "A review on medical image denoising algorithms," *Biomedical Signal Processing and Control*, vol. 61, p. 102036, 2020.
[150] M. Elad, B. Kawar, and G. Vaksman, "Image denoising: The deep learning revolution and beyond—a survey paper," *SIAM Journal on Imaging Sciences*, vol. 16, no. 3, pp. 1594–1654, 2023.
[151] G. Fracastoro, E. Magli, G. Poggi, G. Scarpa, D. Valsesia, and L. Verdoliva, "Deep learning methods for synthetic aperture radar image despeckling: An overview of trends and perspectives," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 2, pp. 29–51, 2021.
[152] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, "Deep learning on image denoising: An overview," *Neural Networks*, vol. 131, pp. 251–275, 2020.

[153] J. Liu, C. Li, L. Liu, H. Chen, H. Han, B. Zhang, and Q. Zhang, "Speckle noise reduction for medical ultrasound images based on cycle-consistent generative adversarial network," *Biomedical Signal Processing and Control*, vol. 86, p. 105150, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809423005839

[154] Y. Lei, R. L. Qiu, T. Wang, W. J. Curran, T. Liu, and X. Yang, "Chapter 7 - generative adversarial networks for medical image synthesis," in *Biomedical Image Synthesis and Simulation*, ser. The MICCAI Society book Series, N. Burgos and D. Svoboda, Eds. Academic Press, 2022, pp. 105–128. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128243497000141

[155] M. Alamir and M. Alghamdi, "The role of generative adversarial network in medical image analysis: An in-depth survey," *ACM Comput. Surv.*, vol. 55, no. 5, dec 2022. [Online]. Available: https://doi.org/10.1145/3527849

[156] P. Kokil and S. Sudharson, "Despeckling of clinical ultrasound images using deep residual learning," *Computer Methods and Programs in Biomedicine*, vol. 194, p. 105477, 2020.

[157] K. Singh, S. K. Ranade, and C. Singh, "A hybrid algorithm for speckle noise reduction of ultrasound images," *Computer methods and programs in biomedicine*, vol. 148, pp. 55–69, 2017.

[158] G. Gilboa and S. Osher, "Nonlocal operators with applications to image processing," *Multiscale Modeling & Simulation*, vol. 7, no. 3, pp. 1005–1028, 2009.

[159] S. Liang, F. Yang, T. Wen, Z. Yao, Q. Huang, and C. Ye, "Nonlocal total variation based on symmetric kullback-leibler divergence for the ultrasound image despeckling," *BMC Medical Imaging*, vol. 17, no. 1, pp. 1–12, 2017.

[160] D. Hirling, E. Tasnadi, J. Caicedo, M. V. Caroprese, R. Sjögren, M. Aubreville, K. Koos, and P. Horvath, "Segmentation metric misinterpretations in bioimage analysis," *Nature methods*, July 2023. [Online]. Available: https://doi.org/10.1038/s41592-023-01942-8

[161] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[162] M. H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, A. K. Davison, and R. Marti, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1218–1226, 2017.

[163] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley *et al.*, "Video-based ai for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 2020.

[164] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier *et al.*, "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2198–2210, 2019.

[165] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[166] J. Zhang, G. Lin, L. Wu, C. Wang, and Y. Cheng, "Wavelet and fast bilateral filter based de-speckling method for medical ultrasound images," *Biomedical Signal Processing and Control*, vol. 18, pp. 1–10, 2015.

[167] J. Zhang, C. Wang, and Y. Cheng, "Comparison of despeckle filters for breast ultrasound images," *Circuits, Systems, and Signal Processing*, vol. 34, no. 1, pp. 185–208, 2015.

[168] D. Feng, W. Wu, H. Li, and Q. Li, "Speckle noise removal in ultrasound images using a deep convolutional neural network and a specially designed loss function," in *International Workshop on Multiscale Multimodal Medical Imaging*. Springer, 2019, pp. 85–92.

[169] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2. IEEE, 2001, pp. 416–423.

[170] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[171] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, 2007, pp. 60–69.

[172] D. Al-Karawi, S. Al-Zaidi, N. Polus, and S. Jassim, "Machine learning analysis of chest ct scan images as a complementary digital test of coronavirus (covid-19) patients," *MedRxiv*, 2020.

[173] pytorch, "Reproducibility," https://pytorch.org/docs/stable/notes/randomness.html, accessed August 9, 2022.

[174] ——, "Using deterministic algorithms," https://pytorch.org/docs/stable/generated/torch.use_deterministic_algorithms.html, accessed August 9, 2022.

[175] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the american statistical association*, vol. 32, no. 200, pp. 675–701, 1937.

[176] ——, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940.

[177] Z.-H. Zhou, "Model selection and evaluation," in *Machine Learning*. Springer, 2021, pp. 25–55.

[178] P. B. Nemenyi, *Distribution-free multiple comparisons.* Princeton University, 1963.

[179] Y. Zhu, R. Wei, G. Gao, L. Ding, X. Zhang, X. Wang, and J. Zhang, "Fully automatic segmentation on prostate mr images based on cascaded fully convolution network," *Journal of Magnetic Resonance Imaging*, vol. 49, no. 4, pp. 1149–1156, 2019.

[180] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic regression-hsmm-based heart sound segmentation," *IEEE transactions on biomedical engineering*, vol. 63, no. 4, pp. 822–832, 2015.

[181] S. Y. Shin, S. Lee, I. D. Yun, and K. M. Lee, "Deep vessel segmentation by learning graphical connectivity," *Medical image analysis*, vol. 58, p. 101556, 2019.

[182] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine learning research*, vol. 7, pp. 1–30, 2006.

[183] B. Billot, C. Magdamo, Y. Cheng, S. E. Arnold, S. Das, and J. E. Iglesias, "Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain mri datasets," *Proceedings of the National Academy of Sciences*, vol. 120, no. 9, p. e2216399120, 2023. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2216399120

[184] M. Zhang, L. Qu, P. Singh, J. Kalpathy-Cramer, and D. L. Rubin, "Splitavg: A heterogeneity-aware federated deep learning method for medical imaging," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 9, pp. 4635–4644, 2022.

[185] C. Liang, B. Cheng, B. Xiao, Y. Dong, and J. Chen, "Multilevel heterogeneous domain adaptation method for remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[186] J. A. Noble and D. Boukerroui, "Ultrasound image segmentation: a survey," *IEEE Transactions on medical imaging*, vol. 25, no. 8, pp. 987–1010, 2006.

[187] Y. Chen, Z. Xiong, Q. Kong, X. Ma, M. Chen, and C. Lu, "Circular statistics vector for improving coherent plane wave compounding image in fourier domain," *Ultrasonics*, vol. 128, p. 106856, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0041624X22001627

[188] S. Goudarzi, A. Basarab, and H. Rivaz, "A unifying approach to inverse problems of ultrasound beamforming and deconvolution," *IEEE Transactions on Computational Imaging*, vol. 9, pp. 197–209, 2023.

[189] L. S. Nguon and S. Park, "Extended aperture image reconstruction for plane-wave imaging," *Ultrasonics*, vol. 134, p. 107096, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0041624X23001725

[190] D. Hyun, L. L. Brickson, K. T. Looby, and J. J. Dahl, "Beamforming and speckle reduction using neural networks," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 66, no. 5, pp. 898–910, 2019.

[191] D. Hyun, A. Wiacek, S. Goudarzi, S. Rothlübbers, A. Asif, K. Eickel, Y. C. Eldar, J. Huang, M. Mischi, H. Rivaz, D. Sinden, R. J. G. van Sloun, H. Strohm, and M. A. L. Bell, "Deep learning for ultrasound image formation: Cubdl evaluation framework and open datasets," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 68, no. 12, pp. 3466–3483, 2021.

[192] V. Grau, H. Becher, and J. A. Noble, "Registration of multiview real-time 3-d echocardiographic sequences," *IEEE transactions on medical imaging*, vol. 26, no. 9, pp. 1154–1165, 2007.

[193] K. de Waal and N. Phad, "Speckle tracking echocardiography in newborns," *hemodynamics and cardiology*, pp. 219–233, 2018.

[194] W. K. Moon, C.-M. Lo, C.-S. Huang, J.-H. Chen, and R.-F. Chang, "Computer-aided diagnosis based on speckle patterns in ultrasound

images," *Ultrasound in medicine & biology*, vol. 38, no. 7, pp. 1251–1261, 2012.

[195] R.-F. Chang and C.-M. Lo, "Computer-aided diagnosis for b-mode, elastography and automated breast ultrasound," in *International Workshop on Digital Mammography*. Springer, 2014, pp. 9–15.

[196] S. Gao, H. Zhou, Y. Gao, and X. Zhuang, "Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability," *Medical Image Analysis*, vol. 89, p. 102889, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523001494

[197] T. Peng, C. Wang, C. Tang, Y. Gu, J. Zhao, Q. Li, and J. Cai, "A multi-center study of ultrasound images using a fully automated segmentation architecture," *Pattern Recognition*, vol. 145, p. 109925, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320323006234

[198] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023.

[199] J. Ma and B. Wang, "Segment anything in medical images," *arXiv preprint arXiv:2304.12306*, 2023.

[200] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: An experimental study," *Medical Image Analysis*, vol. 89, p. 102918, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523001780

[201] J. Pei, T. Cheng, D.-P. Fan, H. Tang, C. Chen, and L. Van Gool, "Osformer: One-stage camouflaged instance segmentation with transformers," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 19–37. [Online]. Available: https://doi.org/10.1007/978-3-031-19797-0_2

[202] W. Zhou, Y. Cai, L. Zhang, W. Yan, and L. Yu, "Utlnet: Uncertainty-aware transformer localization network for rgb-depth mirror segmentation," *IEEE Transactions on Multimedia*, pp. 1–11, 2023.

[203] X. Yan, M. Sun, Y. Han, and Z. Wang, "Camouflaged object segmentation based on matching–recognition–refinement network," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023.

[204] M. Mundt, Y. Hong, I. Pliushch, and V. Ramesh, "A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning," *Neural Networks*, vol. 160, pp. 306–336, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089360802300014X

[205] J. Parmar, S. Chouhan, V. Raychoudhury, and S. Rathore, "Open-world machine learning: Applications, challenges, and opportunities," *ACM Comput. Surv.*, vol. 55, no. 10, feb 2023. [Online]. Available: https://doi.org/10.1145/3561381

[206] C. Lin, C. Qiu, H. Jiang, and L. Zou, "A deep neural network based on prior-driven and structural preserving for sar image despeckling," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 6372–6392, 2023.

[207] S. Baraha and A. K. Sahoo, "Synthetic aperture radar image and its despeckling using variational methods: A review of recent trends," *Signal Processing*, vol. 212, p. 109156, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016516842300230X

[208] Y. Li, Y. Fan, and H. Liao, "Self-supervised speckle noise reduction of optical coherence tomography without clean data," *Biomed. Opt. Express*, vol. 13, no. 12, pp. 6357–6372, Dec 2022. [Online]. Available: https://opg.optica.org/boe/abstract.cfm?URI=boe-13-12-6357

[209] Q. Zhou, M. Wen, B. Yu, C. Lou, M. Ding, and X. Zhang, "Self-supervised transformer based non-local means despeckling of optical coherence tomography images," *Biomedical Signal Processing and Control*, vol. 80, p. 104348, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809422008023

APPENDIX

A. Experimental data

TABLE A1: Median metrics for intra-dataset testing of Dataset 1 (interquartile range is given in brackets). The meaning of paired test results is described in V-A and at the bottom of this table.

Friedman test's p-value: 0.5636 (>0.05) Nemenyi post hoc test: None		Original	Lee	OBNLM	DPAD	PFDTV	GLM	DnCNN	SARBM3D
U-Net	Accuracy	0.9942(0.0198)	0.9913(0.0201)↓	0.9921(0.0199)	0.9934(0.0157)	0.9926(0.0183)	0.9926(0.0219)↓	0.9931(0.0157)	0.9936(0.0173)
	Precision	0.9365(0.0701)	0.929(0.1181)↓	0.9385(0.0710)	0.9455(0.0703)	0.9329(0.0913)↓	0.941(0.0990)	0.9417(0.0729)	0.9293(0.0881)↓
	Recall	0.9303(0.0826)	0.9315(0.0629)	0.9367(0.0818)	0.9423(0.0843)	0.9472(0.0707)↑	0.9208(0.0752)↓	0.9393(0.0852)	0.9502(0.0808)↑
	Dice	0.9210(0.0539)	0.9183(0.0851)↓	0.9227(0.0648)	0.9306(0.0593)	0.9206(0.0645)	0.9197(0.0771)↓	0.9288(0.0610)	0.9216(0.0822)
	Surface Dice	0.5277(0.3484)	0.4459(0.3633)↓	0.4668(0.3510)	0.4974(0.3836)	0.4477(0.3490)↓	0.4787(0.3451)↓	0.4994(0.3769)	0.4822(0.3816)
	95% HD	5.0000(9.9864)	7.9431(36.7927)↓	6.6319(13.0685)	5.0000(9.1778)	5.0000(12.2315)	6.4031(20.2658)↓↓	6.0000(8.1716)	5.3852(12.5563)
	ASSD	1.8414(2.8360)	3.0468(7.9433)↓	2.4152(3.7626)	1.9146(2.6199)	2.1021(3.3257)	2.2256(3.9065)↓	2.2639(2.1248)	2.0469(3.8590)
SK-UNet	Accuracy	0.9931(0.0180)	0.9928(0.0176)	0.9940(0.0176)	0.9939(0.0177)	0.9938(0.0192)	0.9931(0.0173)	0.9934(0.0189)	0.9917(0.0190)↓
	Precision	0.9472(0.0720)↑	0.9257(0.0891)↓	0.9335(0.0803)↓	0.9365(0.0860)↓	0.9448(0.0713)	0.9269(0.0845)↓	0.9259(0.0879)↓	0.8637(0.1080)↓↓
	Recall	0.9326(0.1004)↓	0.9517(0.0620)↑	0.9501(0.0800)↑↑	0.9481(0.0988)↑↑	0.9458(0.0935)↑	0.9547(0.0715)↑↑	0.9607(0.0817)↑↑	0.9708(0.0559)↑↑

TABLE A10: Cross-dataset results: Testing Dataset 4 with models trained based on Dataset 3.

Friedman test's p-value: 6.9152e-11 (<0.05) Nemenyi post hoc test: Figure A2 (c)		Original	Lee	OBNLM	DPAD	PFDTV	GLM	DnCNN	SARBM3D
U-Net	Accuracy	0.9815(0.0189)	0.9809(0.0137)≈	0.9818(0.0140)≈	0.9847(0.0089)↑	0.9847(0.0103)≈	0.9847(0.0096)↑	0.9837(0.0118)≈	0.9801(0.0165)
	Precision	0.9529(0.0823)	0.9031(0.1072)↓	0.9530(0.0669)≈	0.9225(0.0816)↓	0.9571(0.0614)≈	0.9518(0.0645)≈	0.9550(0.0681)≈	0.8856(0.1427)↓
	Recall	0.8516(0.2245)	0.9137(0.1023)≈	0.8570(0.1379)≈	0.9264(0.0700)≈	0.8805(0.1081)≈	0.8849(0.0969)≈	0.8731(0.1364)≈	0.9169(0.0906)≈
	Dice	0.8827(0.1137)	0.8963(0.0776)≈	0.8910(0.0743)≈	0.9129(0.0499)≈	0.9080(0.0555)≈	0.9070(0.0509)≈	0.8995(0.0743)≈	0.8861(0.0749)≈
	Surface Dice	0.5317(0.3125)	0.5422(0.2728)≈	0.5254(0.2825)≈	0.6247(0.2511)≈	0.5983(0.2597)≈	0.6034(0.2577)≈	0.5741(0.2840)≈	0.5224(0.2735)
	95% HD	4.1139(6.2195)	4.0000(3.4031)≈	4.0000(2.7787)≈	3.1623(1.2947)≈	3.1623(1.6437)≈	3.1623(1.2947)≈	3.6056(2.1849)≈	4.2426(3.7445)≈
	ASSD	1.6812(1.6400)	1.6705(1.1630)≈	1.5934(1.0713)≈	1.3566(0.6497)≈	1.3746(0.7855)≈	1.3651(0.6883)≈	1.4769(0.9546)≈	1.7502(1.2153)≈
DAEFF-Net	Accuracy	0.9816(0.0108)≈	0.9829(0.0105)≈	0.9813(0.0125)↓	0.9826(0.0105)≈	0.9815(0.0115)↓	0.9815(0.0122)↓	0.9810(0.0112)↓	0.9813(0.0128)↓
	Precision	0.9537(0.0678)≈	0.9522(0.0661)↓	0.9425(0.0780)↓↓	0.9552(0.0626)≈	0.9594(0.0601)≈	0.9742(0.0523)≈	0.9605(0.0564)≈	0.9492(0.0694)↓
	Recall	0.8427(0.1232)≈	0.8616(0.1170)≈	0.8462(0.1113)≈	0.8499(0.1140)≈	0.8318(0.1268)↓	0.8197(0.1242)↓↓	0.8264(0.1219)↓	0.8416(0.1158)≈
	Dice	0.8851(0.0654)≈	0.8949(0.0586)≈	0.8844(0.0625)	0.8924(0.0592)≈	0.8836(0.0653)↓	0.8827(0.0682)↓	0.8813(0.0660)↓	0.8837(0.0638)↓
	Surface Dice	0.5166(0.2386)	0.5538(0.2441)≈	0.5112(0.2500)	0.5472(0.2383)≈	0.5275(0.2469)↓	0.5042(0.2620)↓	0.4929(0.2416)↓	0.5028(0.2492)↓
	95% HD	4.0000(2.0000)≈	3.9704(2.0000)≈	4.1170(2.3852)↓↓	4.0000(2.0000)≈	4.0000(2.0000)↓	4.0000(2.0990)	4.0000(1.8526)↓	4.0000(2.3566)↓
	ASSD	1.6542(0.8146)≈	1.5321(0.8077)≈	1.6925(0.9166)↓	1.5634(0.8178)≈	1.6834(0.8624)↓	1.6970(0.9371)↓	1.7333(0.8574)↓	1.6870(0.9272)↓
RF-Net	Accuracy	0.9846(0.0092)≈	0.9846(0.0090)≈	0.9813(0.0118)↓	0.9850(0.0088)≈	0.9838(0.0096)↓	0.9827(0.0118)↓	0.9836(0.0097)↓	0.9838(0.0105)↓
	Precision	0.9578(0.0581)≈	0.9360(0.0719)↓	0.9602(0.0584)≈	0.9474(0.0615)↓	0.9569(0.0600)↓	0.9733(0.0489)≈	0.9490(0.0633)↓	0.9445(0.0677)↓
	Recall	0.8785(0.0875)≈	0.9046(0.0801)≈	0.8575(0.1011)↓↓	0.8942(0.0826)≈	0.8699(0.0951)↓↓			

B. Visualized segmentation results

Fig. B1: Visualization of U-Net segmentation results on 8 versions of Dataset 1. Each row shows the original image, the segmentation results on the models trained from original dataset, despeckled datasets by Lee, OBNLM, DPAD, PFDTV, GLM, DnCNN and SARBM3D from left to right. The red and cyan contours are the ground truth and model results respectively. The samples in the first row perform relatively equally in 8 different versions. In other rows, the result with red rectangle border drawn has the best performance. It can be seen that each version of the dataset can perform the best in some samples, resulting in similar overall performance indicated by Friedman test.

Fig. B2: Visualization of U-Net segmentation results on 8 versions of Dataset 2, similar to the meaning of Figure B1.

Fig. B3: Visualization of U-Net segmentation results on 8 versions of Dataset 3, similar to the meaning of Figure B1.

Fig. B5: Visualization of DAEFF-Net segmentation results on the original Dataset 4 and the OBNLM despeckled version (for simplicity and clarity, only samples with an absolute difference of more than 15% in the surface dice metric are shown). Every three consecutive columns are the original image, the segmentation results from the original dataset, and from the OBNLM desepckled dataset. To compare the different segmentation performances on the original dataset and the OBNLM desepckled dataset, the small amounts of images above the red line are the despeckled results of original dataset that have better performances than those of OBNLM desepckled dataset, while the large amounts of images below the red line are the despeckled results of the OBNLM desepckled dataset that have better performance than those of original datasets. This large amount of images that are below the red-line illustrates a better performance of the OBNLM desepckled dataset than that of original dataset. This visualization results are consistent with the paired test results in Table A4 (the row of DAEFF-Net and the column of OBNLM).

Fig. B4: Visualization of U-Net segmentation results on 8 versions of Dataset 4, similar to the meaning of Figure B1.

Fig. B6: Visualization of CDM segmentation results on the original Dataset 4 and OBNLM despeckled version (for simplicity and clarity, only samples with an absolute difference of more than 3% in the dice metric are shown). The meaning is similar to Figure B5, illustrating the better performance of the original dataset in visualization results. The visualization results are consistent with the paired test results in Table A4 (the row of CDM and the column of OBNLM).

Fig. B7: Visualization of semi-supervised U-Net segmentation results on the original Dataset 3 and the PFDTV despeckled version (for simplicity and clarity, only samples with an absolute difference of more than 1 pixel in the ASSD metric are shown). The meaning is similar to Figure B5, illustrating the better performance of the PFDTV despeckled dataset in visualization results. The visualization results are consistent with the paired test results in Table A5 (the column of PFDTV).