This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

XNet: Wavelet-Based Low and High Frequency Fusion Networks for Fully- and Semi-Supervised Semantic Segmentation of Biomedical Images

Yanfeng Zhou[1,2] Jiaxing Huang[1,2] Chenlong Wang[3] Le Song[3*] Ge Yang[1,2*]
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]BioMap Research, California, USA
{zhouyanfeng2020, huangjiaxing2021}@ia.ac.cn, {chenlong, songle}@biomap.com,
ge.yang@ia.ac.cn

Abstract

Fully- and semi-supervised semantic segmentation of biomedical images have been advanced with the development of deep neural networks (DNNs). So far, however, DNN models are usually designed to support one of these two learning schemes, unified models that support both fully- and semi-supervised segmentation remain limited. Furthermore, few fully-supervised models focus on the intrinsic low frequency (LF) and high frequency (HF) information of images to improve performance. Perturbations in consistency-based semi-supervised models are often artificially designed. They may introduce negative learning bias that are not beneficial for training. In this study, we propose a wavelet-based LF and HF fusion model XNet, which supports both fully- and semi-supervised semantic segmentation and outperforms state-of-the-art models in both fields. It emphasizes extracting LF and HF information for consistency training to alleviate the learning bias caused by artificial perturbations. Extensive experiments on two 2D and two 3D datasets demonstrate the effectiveness of our model. Code is available at https://github.com/Yanfeng-Zhou/XNet.

1. Introduction

Semantic segmentation is a fundamental task in biomedical image analysis, where the goal is to assign a class label to each pixel. Methods for semantic segmentation of biomedical images based on convolutional neural networks (CNNs) have achieved remarkable success [42, 47, 23, 8]. Some studies extend these methods to 3D and achieve promising results on volumetric segmentation [35, 10, 62, 37]. Recently, the combination of transformers and CNNs has become popular [5, 54, 53, 20, 54]. Transformers can capture long-range dependencies [50, 13, 30] to compensate for the limited receptive fields of CNNs.

However, most of the existing methods focus on model architecture to better extract features [67, 38, 65]. Few methods explore the intrinsic LF and HF information of images that may be useful for segmentation [58, 49].

For semantic segmentation of biomedical images, supervised methods require large-scale labeled images, which are costly and time-consuming to produce. To alleviate this problem, researchers propose semi-supervised methods that learn with a small number of labeled images and a substantial number of unlabeled images [48, 51, 59]. The common solutions include adversarial training [36, 46], pseudo-labeling [16, 60, 56] and consistency regularization [6, 12]. Consistency regularization is currently the best performing method [32, 34], it perturbs input images, intermediate features or output predictions, allowing models to learn consistency from the perturbation [26, 39, 34].

However, current perturbation strategies are artificially designed, such as rotation [26], noise addition [39], distance mapping [32] and dropout [39], etc. They may introduce negative learning bias, such as segmenting noisy images is equivalent to learning an extra denoising task.

Furthermore, fully- and semi-supervised semantic segmentation are regarded as two different research fields. Unified models that simultaneously achieves state-of-the-art remain limited.

To solve the above problems, we propose a wavelet-based LF and HF fusion model XNet. XNet can simultaneously realize fully-supervised learning based on LF and HF information fusion, and semi-supervised learning based on LF and HF outputs consistency. To be specific, we use wavelet transform to generate LF and HF images and feed them into XNet. XNet fuses their LF and HF information and then generates dual-branch segmentation predictions.

[*] Corresponding author.

For supervised learning, segmentation predictions absorb the complete LF and HF information of raw images. For semi-supervised learning, dual-output pays different attention to LF and HF information leading to consistency differences. These differences are used for training on unlabeled images.

Motivation. For semantic segmentation problem, the HF information generally represents image details, while the LF information are often abstract semantics (the LF and HF images in Figure 2 intuitively show their differences). The strategy of extracting and fusing different frequency information can help model better focus on LF semantics and HF details to improve performance. Furthermore, our model uses wavelet transform to generate LF and HF images for consistency difference-based semi-supervised learning. These consistency differences stem from the different focus on LF and HF information, which alleviates the learning bias caused by artificial design.

Our contributions are summarized as follows:

- We propose a low and high frequency fusion model XNet, which achieves state-of-the-art in both fully- and semi-supervised semantic segmentation of biomedical images simultaneously.
- XNet uses wavelet transform to generate LF and HF images for consistency learning, which can alleviate the learning bias caused by artificial perturbations.
- Extensive benchmarking on two 2D and two 3D public biomedical datasets confirms the effectiveness of XNet.

2. Related Work

Fully-Supervised Semantic Segmentation of Biomedical Images. With the rise of deep learning, CNNs have been widely used in semantic segmentation [57, 22, 15, 61], such as FCN [31], DeepLab v3+ [8], etc. For biomedical images, efficient encoder-decoder architecture achieves superior performance [63, 23], such as UNet [42], UNet++ [67], UNet 3+ [21], etc. Furthermore, researchers extend this architecture to 3D to meet the needs for volumetric segmentation. [35] proposes a 3D fully CNN VNet. [10] extends UNet to 3D. ConResNet [62] is proposed inter slice context residual learning. Recently, incorporating transformer with encoder-decoder architecture has impressive results [54, 66], such as SwinUNet [5], TransBTS [53], UNETR [20], etc. These models capture both long-range dependencies and local features to improve performance.

Semi-Supervised Semantic Segmentation of Biomedical Images. To alleviate the lack of labeled images, semi-supervised semantic segmentation of biomedical images has become a key approach [29, 24, 27]. Currently the dominant strategies include adversarial training [36, 46], pseudo-labeling [16, 60, 56], and consistency regularization [6, 12]. Adversarial training use generative adversarial networks [19] to continuously improve the performance of both generator that generates segmentation predictions and discriminator that judges the authenticity of predictions. Pseudo-labeling utilizes high confidence predictions to improve model performance. Consistency regularization-based methods have better performance [32, 33, 34]. They utilize unlabeled images by enforcing consistency between different predictions. DTC [32] proposes a dual-task consistency network that predicts segmentation maps and geometry-aware level set representations. TCSMv2 [26] utilizes transformation consistency to allow the network to generate consistent predictions for differently perturbed inputs. [34] proposes an uncertainty rectified pyramid consistency (URPC) strategy.

Wavelet-Based DNNs for Semantic Segmentation. Based on powerful frequency and spatial representation capabilities, wavelet transform has been incorporated with DNNs and some methods have been explored in semantic segmentation [58, 49, 64, 44, 18, 28]. The common strategies include using wavelet transform as pre- or post-processing [49, 58] and replacing some layers of CNNs (such as up- and down-sampling) with wavelet transform [18, 64]. However, most of them are only suitable for specific segmentation objects, which limits their generalization and application. [1] proposes a symmetric CNN enhanced by wavelet transform (Aerial LaneNet) for lane-marking semantic segmentation in aerial imagery. CWNN [14] uses wavelet constrained pooling layers to replace the conventional pooling for synthetic aperture radar image segmentation. WaveSNet [25] uses wavelet transform to extract image details during down-sampling and uses inverse transform to recover details during up-sampling. In contrast, we use wavelet transform to generate LF and HF images as dual-branch input to extract LF and HF features. We compare our model with previous wavelet-based models in Section 4.4 and show superior performance of our model.

3. Method

In this section, we give an overview of the proposed model XNet in Section 3.1. Then we analyze the role of wavelet transform and propose a method for generating LF and HF images in Section 3.2. We further introduce LF and HF fusion module in Section 3.3. Finally, we analyze the feasibility of XNet on fully- and semi-supervised learning in Section 3.4.

3.1. Overview

Figure 1 shows an overview of the proposed model XNet which consists of four modules, including LF encoder, HF encoder, LF and HF fusion module and dual-branch decoder. LF and HF encoders could extract semantic features and detailed features from LF and HF images, respectively.

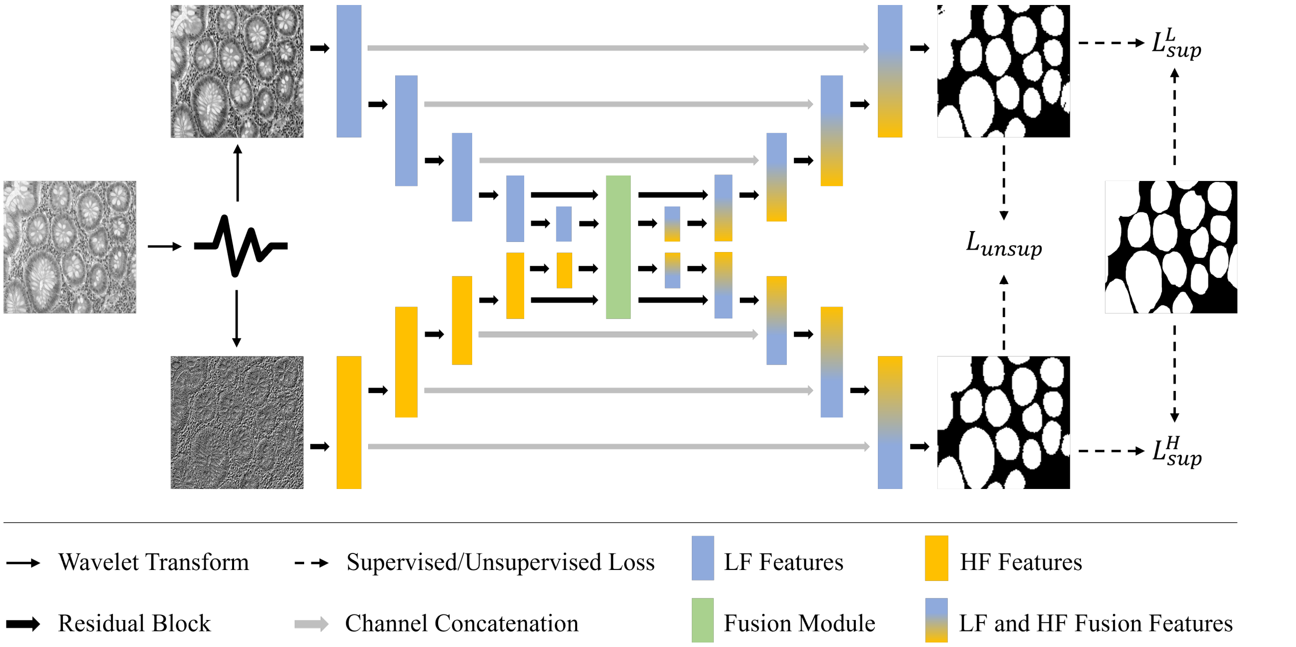

Figure 1. Overview of proposed model XNet. Blue and orange represent LF and HF encoder, respectively. Green represents fusion module. Mixed colors represent dual-branch decoder. XNet learns from unlabeled images by minimizing L_{unsup} and learns from labeled images by minimizing L^L_{sup}, L^H_{sup} and L_{unsup}.

Fusion module fuses them to generate fusion features with both LF semantics and HF details. Dual-branch decoder uses fusion features to output segmentation predictions.

Training process is also shown in Figure 1. By wavelet transform of the raw images, we acquire the corresponding LF and HF images and then input them into LF and HF encoders to generate LF and HF features, respectively. These features are fused in fusion module and then fed into decoder to generate segmentation predictions of LF and HF branch, respectively. For supervised training, model is optimized by minimizing supervised loss and dual-output consistency loss on labeled images. For semi-supervised training, model is optimized by minimizing supervised loss on labeled images and dual-output consistency loss on unlabeled images. Thus, no matter fully- or semi-supervised training, the total loss function L_{total} is defined as:

$$L_{total} = L_{sup} + \lambda L_{unsup}, \tag{1}$$

where L_{sup} is supervised loss, L_{unsup} is unsupervised loss, i.e., dual-output consistency loss, λ is a weight to control the balance between L_{sup} and L_{unsup}. To be specific, the supervised loss L_{sup} consists of LF supervised loss $L^L_{sup}(\cdot)$ and HF supervised loss $L^H_{sup}(\cdot)$. L_{sup} is defined as:

$$L_{sup} = L^L_{sup}(p^L_i, y_i) + L^H_{sup}(p^H_i, y_i), \tag{2}$$

where p^L_i and p^H_i represent LF and HF segmentation predictions of the i-th image, respectively. y_i represents ground truth of the i-th image. The unsupervised loss L_{unsup} is achieved by cross pseudo supervision (CPS) loss [9]: Use One branch prediction as pseudo-label to supervise the other branch, and vice versa. L_{unsup} is defined as:

$$L_{unsup} = L^L_{unsup}(p^L_i, \hat{p}^H_i) + L^H_{unsup}(p^H_i, \hat{p}^L_i), \tag{3}$$

where $L^L_{unsup}(\cdot)$ and $L^H_{unsup}(\cdot)$ represent LF and HF unsupervised loss, respectively. $\hat{p}^L_i$ and $\hat{p}^H_i$ represent LF and HF pseudo-label generated by p^L_i and p^H_i, respectively (we adopt a simple strategy for pseudo-label generation: label the pixel as the class with the highest confidence prediction.).

In this study, $L^L_{sup}(\cdot)$, $L^H_{sup}(\cdot)$, $L^L_{unsup}(\cdot)$ and $L^H_{unsup}(\cdot)$ all use dice loss [35]. We choose the branch that performed better in training stage as the final outputs during inference.

3.2. Wavelet Transform

2D (3D) images are essentially 2D (3D) discrete non-stationary signals, containing different frequency ranges and spatial locations information. Wavelet transform can effectively preserve these information while decomposing them.

To be specific, take 2D images as an example. We use wavelet transform to decompose raw images into LF, horizontal HF, vertical HF and diagonal HF components (LL, HL, LH and HH). They respectively save LF and different HF information of raw images. We represent LF images L with LF components and represent HF images H as the sum of HF components in different directions. L and H are defined as:

$$L = LL, \tag{4}$$

$$H = HL + LH + HH. \tag{5}$$

L and H are shown in Figure 2. We can see H emphasize details, while L focus on semantics.

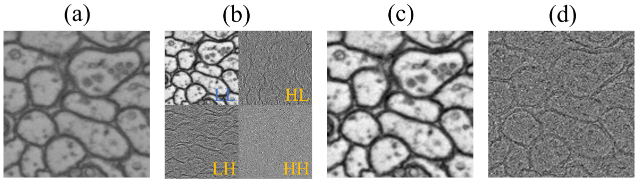

Figure 2. Taking CREMI [17] as an example, visualize LF and HF results. (a) Raw image. (b) Wavelet transform results. (c) LF image. (d) HF image.

Why use wavelet transform? Compared to other methods (such as Fourier transform), wavelet transform is an efficient way to generate L and H. Using L as input, XNet can focus more on LF semantics, because L has less noise and details. In contrast, H has more noise but clearer object boundaries, which can help model focus more on HF details. Furthermore, using L and H for semi-supervised training, the consistency difference comes from intrinsic LF and HF information of images, which can alleviate the learning bias caused by artificial perturbations.

3.3. LF and HF Fusion Module

The architecture of LF and HF fusion module is shown in Figure 3. Using LF and HF features as inputs, fusion module uses 3×3 convolutions to acquire same size, upsampling or down-sampling features and concatenates their channels. Then the channel-concatenated features are input to 1×1 transition convolutions to generate LF and HF fusion features.

Why use fusion module? The fusion module can fuse LF and HF features into complete features. Without fusion, each branch would lack semantics or details, which are detrimental to segmentation. We demonstrate the separation-fusion X-shaped network architecture is the key to improve performance in ablation studies of Section 4.5.

3.4. Feasibility of Fully- and Semi-Supervision

For biomedical images, we assume that the raw image I consists of LF features F_L, HF features F_H, LF additive noise N_L and HF additive noise N_H. Therefore, I is defined as:

$$I = F_L + F_H + N_L + N_H. \tag{6}$$

We make the assumption because related studies have shown that noise in biomedical images is generally additive [3, 2, 43, 41]. For semantic segmentation problem, accurate segmentation requires LF semantics (such as shape, color, etc.) and HF details (such as edges, textures, etc.).

Wavelet transform W can decouple image I to generate LF image L and HF image H:

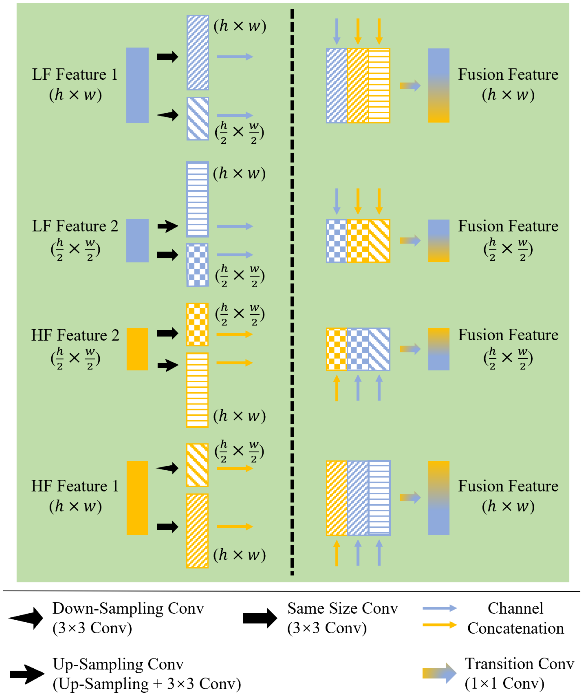

Figure 3. Architecture of LF and HF fusion module. Same Size Conv represents the output and input features have same size. Down-Sampling Conv reduce the size of output features by half. Up-Sampling Conv doubles the size of output features. Transition Conv uses channel-concatenated features as input and outputs fusion features.

$$\begin{aligned} L, H &= W(I), \\ L &= F_L + N_L, \\ H &= F_H + N_H. \end{aligned} \tag{7}$$

LF and HF encoders E_L, E_H extract F_L and F_H from L and H, respectively:

$$\begin{aligned} F_L &= E_L(L), \\ F_H &= E_H(H). \end{aligned} \tag{8}$$

Fusion module M fuses F_L and F_H to acquire the complete features F_M:

$$F_M = M(F_L, F_H). \tag{9}$$

For supervised learning, decoding complete information can acquire segmentation predictions. For semi-supervised learning, because each decoding branch pays different attention to LF and HF information, there are differences in LF semantics and HF details between the predictions of dual-branch decoder. These differences can be used for semi-supervised training based on consistency regularization.

The segmentation predictions of LF and HF branches are defined as:

$$P_L, P_H = D(F_M), \tag{10}$$

where P_L and P_H represent LF and HF segmentation predictions, D represents dual-branch decoder.

Method	Model	GlaS				CREMI			
		JI ↑	DC ↑	95HD ↓	ASD ↓	JI ↑	DC ↑	95HD ↓	ASD ↓
Fully-Supervised (100%)	UNet [42]	81.54	89.83	8.82	1.72	75.47	86.02	5.62	1.06
	UNet++ [67]	81.92	90.06	9.96	1.89	75.89	86.30	4.56	0.85
	Att-UNet [38]	81.41	89.75	10.53	1.99	**76.87**	**86.92**	3.92	0.74
	Aerial LaneNet [1]	70.08	82.41	19.29	3.83	67.04	80.27	6.73	1.31
	MWCNN [28]	74.11	85.13	15.40	3.19	67.23	80.40	7.02	1.36
	HRNet-W18 [47]	**82.68**	**90.52**	9.71	1.90	74.06	85.10	**3.77**	**0.68**
	Res-UNet [63]	79.23	88.41	10.90	2.17	74.32	85.27	4.79	0.90
	WDS [44]	-	-	-	-	63.74	77.86	6.42	1.29
	U^2-Net [40]	76.87	86.92	12.61	2.58	74.23	85.21	4.92	0.89
	U^2-Net* [40]	81.14	89.59	9.93	1.86	71.94	83.68	5.91	1.08
	UNet 3+ [21]	79.83	88.78	10.81	2.12	74.92	85.66	4.24	0.78
	UNet 3+ w/ DS [21]	80.25	89.04	12.14	2.26	74.23	85.21	4.51	0.87
	SwinUnet [5]	73.06	84.43	14.14	3.21	63.37	77.58	7.92	1.53
	Yin *et al.* [58]	74.29	85.25	13.08	2.79	56.07	71.85	5.21	1.22
	WaveSNet [25]	78.52	87.97	11.36	2.19	70.67	82.82	5.06	1.01
	nnUNet [23]	80.66	89.29	11.07	2.04	73.95	85.02	5.32	1.11
	nnUNet‡ [23]	82.34	90.31	**8.74**	**1.69**	76.08	86.42	5.08	1.09
	XNet	**84.77**	**91.76**	**7.87**	**1.55**	**79.23**	**88.41**	**3.66**	**0.61**
Semi-Supervised (20%+80%)	MT [48]	76.41	86.62	13.28	2.65	75.58	86.09	5.60	1.10
	EM [51]	76.81	86.88	12.28	2.54	73.24	84.55	6.64	1.28
	UAMT [59]	76.55	86.72	13.43	2.73	74.04	85.08	5.71	1.10
	CCT [39]	77.60	87.39	11.23	2.27	**75.74**	**86.20**	6.93	1.31
	CPS [9]	**80.46**	**89.17**	**10.56**	**2.08**	74.87	85.63	6.47	1.25
	URPC [34]	76.84	86.91	10.97	2.31	74.70	85.22	**4.42**	**0.89**
	CT [33]	79.02	88.28	12.02	2.33	73.43	84.68	6.33	1.23
	XNet	**80.89**	**89.44**	**9.86**	**2.07**	**76.28**	**86.54**	**4.19**	**0.76**

Table 1. Comparison with fully- and semi-supervised state-of-the-art models on GlaS and CREMI test set. Semi-supervised models are based on UNet. DS indicates deep supervision. * indicates lightweight models. ‡ indicates training for 1000 epochs. - indicates training failed. **Red** and **bold** indicate the best and second best performance.

To sum up, XNet can be used in both fully- and semi-supervised learning. Figure 4 shows topological flow chart of segmentation process of XNet.

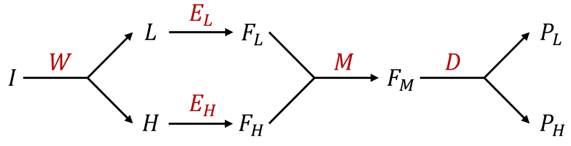

Figure 4. Topological flow chart of segmentation process.

4. Experiments

4.1. Datasets

To evaluate our model, we conduct experiments on two 2D datasets (GlaS [45] and CREMI [17]) and two 3D datasets (LA [55] and LiTS [4]).

GlaS. This is a gland segmentation dataset, including 165 Hematoxylin and Eosin (H&E) stained images of benign and malignant tissue. The image size is 775×522. The number of training and test images are 85 (37 benign, 48 malignant) and 80 (37 benign, 43 malignant), respectively.

CREMI. This is a electron microscopy dataset for neuronal membrane segmentation. It consists of three images stacks for three different types of neurons. Each stack consists of 125 slices of size 1250×1250. We use the first two stacks for training and the third stack for testing. Due to large image size is not convenient for training, we use a sliding window to crop the raw images to 256×256. Finally, we get 3575 images for training and 3075 images for testing.

LA. This is a left atrial segmentation dataset from 2018 Atrial Segmentation Challenge. It consists of 100 3D MRI images, with a resolution of 0.625×0.625×0.625mm. Following [59, 32], we use 80 images for training and 20 images for testing.

LiTS. This is a liver and tumor segmentation dataset from 2017 Liver Tumor Segmentation Challenge. It consists of 131 3D CT images. Following [52], we use 100 images for training and 31 images for testing.

Why choose them? These datasets contain four modalities: light microscopy, electron microscopy, MRI and CT. They also contain 2D and 3D images, respectively. Evaluating

Method	Model	LA				LiTS			
		JI ↑	DC ↑	95HD ↓	ASD ↓	JI ↑	DC ↑	95HD ↓	ASD ↓
Fully-Supervised (100%)	VNet [35]	73.94	85.02	10.63	3.04	70.13	78.07	42.17	11.38
	UNet 3D [10]	82.67	90.51	6.13	1.80	78.63	**86.21**	23.00	8.32
	Res-UNet 3D	85.13	91.97	4.96	1.48	52.88	61.18	61.67	21.34
	ESPNet 3D [37]	66.14	79.62	25.13	5.06	-	-	-	-
	DMFNet [7]	80.43	89.16	8.96	2.26	76.77	84.44	22.73	8.51
	ConResNet [62]	**85.60**	**92.24**	**4.27**	**1.39**	77.22	84.45	**20.65**	7.56
	CoTr [54]	81.42	89.76	7.83	2.06	75.16	83.03	27.30	8.00
	TransBTS [53]	78.97	88.25	9.71	2.57	76.22	83.92	27.74	**7.51**
	UNETR [20]	79.66	88.68	9.04	2.44	71.22	78.75	37.42	10.11
	nnUNet [23]	83.14	90.79	6.33	1.90	78.23	85.52	42.78	11.35
	nnUNet‡ [23]	84.43	91.56	5.51	1.52	**78.71**	86.02	27.26	7.78
	XNet 3D	**86.58**	**92.81**	**3.89**	**1.30**	**80.92**	**87.95**	**18.50**	**5.74**
Semi-Supervised (20%+80%)	MT [48]	78.58	88.01	7.06	2.12	72.60	80.38	**27.46**	10.25
	EM [51]	77.91	87.58	7.78	2.28	-	-	-	-
	UAMT [59]	78.85	88.17	6.88	2.09	74.78	82.48	26.98	10.59
	CCT† [39]	82.04	90.13	6.79	1.90	73.92	81.56	**25.03**	11.28
	CCT* [39]	77.91	87.58	7.63	2.35	70.81	78.91	27.90	10.44
	DTC [32]	76.68	86.80	10.30	2.76	**74.53**	**82.50**	35.94	12.35
	DTC* [32]	76.10	86.43	9.71	2.68	68.11	76.15	50.05	13.23
	CPS† [9]	**82.78**	**90.58**	6.24	**1.79**	71.63	79.26	28.94	**9.45**
	CPS* [9]	79.06	88.31	7.26	2.16	69.34	77.21	40.85	11.99
	URPC [34]	79.98	88.88	7.57	2.12	-	-	-	-
	CT† [33]	81.59	89.86	**6.10**	1.91	71.57	78.95	47.09	13.48
	CT* [33]	78.86	88.18	9.06	2.45	68.96	76.69	58.68	15.29
	XNet 3D	**83.54**	**91.03**	**6.00**	**1.76**	**75.74**	**83.27**	36.88	**9.26**

Table 2. Comparison with fully- and semi-supervised state-of-the-art models on LA and LiTS test set. Due to GPU memory limitations, some semi-supervised models using smaller architectures, † and * indicate models are based on lightweight 3D UNet (half of channels) and VNet, respectively. ‡ indicates training for 1000 epochs. - indicates training failed. **Red** and **bold** indicate the best and second best performance.

model performance on multiple modalities and different dimensions is more representative and convincing.

4.2. Evaluation

We use Jaccard index (JI), Dice coefficient (DC), 95th percentile Hausdorff distance (95HD), and average surface distance (ASD) as performance metrics to evaluate segmentation results. JI and DC emphasize pixel-wise accuracy, while 95HD and ASD emphasize boundary accuracy. These metrics are widely used for benchmarking performance of biomedical image segmentation.

4.3. Implementation Details

We implement our model using PyTorch. Training and inference of all models are performed on four NVIDIA GeForce RTX3090. We use SGD with momentum to train models, the momentum is set at 0.9 and the weight decay is set at 0.00005. The number of epochs is set at 200. The learning rate decays by 0.5 every 50 epochs. The weight λ for unsupervised loss function increases linearly with epoch, $\lambda = \lambda_{max} * \frac{epoch}{max\_epoch}$.

For 2D datasets, we use flip, rotation, transposition for data augmentation and input images are resized to 128×128 during training and inference. For GlaS, the initial learning rate is set at 0.5, λ_{max} is set at 5, batch size is set at 2. For CREMI, the initial learning rate is set at 0.5, λ_{max} is set at 1, batch size is set at 16.

For 3D datasets, we use flip, bias field, noise and blur for data augmentation, the initial learning rate is set at 0.1, batch size is set at 1. For inference, we use a sliding window strategy with overlap ratio of 0.5 and the max connected component in predictions as the final segmentation results. For LA, following [59, 32], we apply the same preprocessing methods, patch size is set at 96×96×80, λ_{max} is set at 5. For LiTS, following [52], we use the soft tissue CT window range of [-100, 250] HU and crop the images centering at liver regions, patch size is set at 112×112×32, λ_{max} is set at 0.5.

For semi-supervised segmentation, we report the performance of all models trained with 20% labeled images and 80% unlabeled images, which is the common semi-supervised experimental partition.

4.4. Comparison with State-of-the-art Models

We compare XNet extensively with previous models both on fully- and semi-supervised semantic segmentation. Due to other state-of-the-art models have different experimental setup, we reimplement them for a fair comparison.

2D Fully-Supervision. We compare XNet with supervised state-of-the-arts, including UNet [42], UNet++ [67], Att-UNet [38], HRNet-W18 [47], Res-UNet [63], U^2-Net [40], UNet 3+ [21], SwinUnet [5] and nnUNet [23]. We also compare our model with previous wavelet-based models, including Aerial LaneNet [1], MWCNN [28], WDS [44], Yin *et al*. [58], WaveSNet [25].

In Table 1, we can see XNet achieves competitive performance on GlaS and CREMI. For GlaS, XNet improves the optimal results by 2.09% in JI, 1.24% in DC, 0.87 pixels in 95HD and 0.14 pixels in ASD. For CREMI, XNet improves the optimal results by 2.36% in JI, 1.49% in DC, 0.11 pixels in 95HD and 0.07 pixels in ASD. Table 1 also shows that XNet has better generalizability than previous wavelet-based models (Aerial LaneNet [1], MWCNN [28], WaveSNet [25], etc.).

2D Semi-Supervision. We compare XNet with semi-supervised state-of-the-arts, including MT [48], EM [51], UAMT [59], CCT [39], CPS [9], URPC [34], CT [33].

Table 1 illustrates the comparison results. For GlaS, XNet outperforms all the other models, particularly outperforms the suboptimal result by 0.43% in JI, 0.27% in DC, 0.7 pixels in 95HD and 0.01 pixels in ASD. For CREMI, XNet achieves improvement by 0.54% in JI, 0.34% in DC, 0.23 pixels in 95HD and 0.13 pixels in ASD than the suboptimal result.

3D Fully-Supervision. We compare 3D XNet with supervised state-of-the-arts, including VNet [35], 3D UNet [10], 3D Res-UNet, 3D ESPNet [37], DMFNet [7], ConResNet [62], CoTr [54], TransBTS [53], UNETR [20] and nnUNet [23].

From Table 2, we can see XNet outperforms previous state-of-the-art models by a large margin. To be specific, XNet improves the optimal results by 0.98% in JI, 0.57% in DC, 0.38 voxels in 95HD and 0.09 voxels in ASD for LA, 2.21% in JI, 1.74% in DC, 2.15 voxels in 95HD and 1.77 voxels in ASD for LiTS, respectively.

3D Semi-Supervision. We use 3D architectures to extend 2D semi-supervised models to 3D, including MT [48], EM [51], UAMT [59], CCT [39], CPS [9], URPC [34], CT [33]. We also show DTC [32] performance. The 3D semi-supervised results are shown in Table 2. As previous experiments, XNet also shows superior performance.

4.5. Ablation Studies

To verify effectiveness of each component, we perform the following ablation studies on GlaS.

Wavelet	JI ↑	DC ↑	95HD ↓	ASD ↓
Haar	79.70	88.70	**9.30**	**2.05**
Dmey	73.95	85.02	14.40	2.96
Coif 1	78.19	87.76	10.78	2.32
Bior 1.5	78.71	88.09	11.35	2.29
Bior 2.4	78.21	87.77	12.24	2.52
Db 2	**80.89**	**89.44**	9.86	2.07

Table 3. Comparison of different wavelet bases in semi-supervised training.

Comparison of Wavelet Bases. Table 3 shows the performance of different wavelet bases in semi-supervision, including Haar, Dmey, Daubechies 2 (Db 2), Coiflets 1 (Coif 1), Biorthogonal 1.5 (Bior 1.5) and Biorthogonal 2.4 (Bior 2.4). We find that Db 2 wavelet has better pixel-wise accuracy, while Haar wavelet has better boundary accuracy. Based on the above experiments, we apply Db 2 as wavelet basis to related experiments in Table 1 and Table 2.

# LF Fusion	# HF Fusion	JI ↑	DC ↑	95HD ↓	ASD ↓
0 (w/o fusion)	0 (w/o fusion)	66.78	80.08	19.19	4.26
1	1	75.44	86.00	13.09	2.72
1	2	77.08	87.06	12.63	2.62
2	1	76.91	86.95	12.38	2.60
2	2	**80.89**	**89.44**	**9.86**	**2.07**
2	3	79.14	88.35	10.78	2.20
3	2	78.56	87.99	12.24	2.36
3	3	78.39	87.88	10.87	2.24

Table 4. Comparison of different numbers of LF and HF fusion features in LF and HF fusion module in semi-supervised training.

Comparison of The Number of Fusion Features. The comparison results are shown in Table 4. Without fusing any LF and HF features, the model performance is negatively impacted. Once introducing LF and HF fusion module, the XNet performance can be greatly improved. For example, compared to no fusing, using 1 LF and 1 HF fusion features improves performance by 8.66% in JI. From Table 4, we can also see the number of LF and HF fusion features are 2 and 2 to get the best performance. It may be because too many fusion features generate redundant information, which are bad for model training. We finally set the number of LF and HF fusion features are 2 and 2, respectively, and apply it to related experiments in Table 1 and Table 2.

Comparison of Dataset Partition. From Table 5, we find that XNet has good performance in both fully- and semi-supervised training with different partition. For supervised

Labeled	Unlabeled	JI ↑	DC ↑	95HD ↓	ASD ↓
10%	0%	59.18	74.36	27.90	7.16
	90%	75.63	86.13	14.28	2.78
20%	0%	69.84	82.24	17.31	3.87
	40%	74.19	85.19	13.58	2.92
	60%	77.46	87.30	11.82	2.54
	80%	80.89	89.44	9.86	2.07
30%	0%	72.06	83.76	15.39	3.33
	70%	81.39	89.74	9.98	2.02
100%	0%	84.77	91.76	7.87	1.55

Table 5. Performance of different dataset partition strategies.

training with 100% labeled images, XNet achieves 84.77% in JI. Using 10%, 20% and 30% labeled images and the rest of unlabeled images, semi-supervised training improves the supervised baseline by 16.45%, 11.05% and 9.33% in JI, respectively.

Table 5 also shows that with the increase of unlabeled images, XNet has better performance. Training with 20% labeled images and 40%, 60% and 80% unlabeled images, the supervised baseline is improved by 4.35%, 7.62% and 11.05% in JI, respectively.

Model	Params	MACs	JI ↑	DC ↑	95HD ↓	ASD ↓
UNet$^+$	264M	109G	82.73	90.55	9.32	1.77
Res-UNet$^+$	222M	279G	81.05	89.53	9.87	1.99
WaveSNet$^+$	349M	64G	78.56	87.99	13.27	2.46
XNet	326M	83G	**84.77**	**91.76**	**7.87**	**1.55**
UNet	35M	16G	81.54	89.83	8.82	1.72
XNet$^-$	82M	21G	83.49	91.00	8.52	**1.64**
XNet^{--}	20M	5G	**83.71**	**91.13**	**8.49**	1.69

Table 6. Comparison of model sizes and computational cost in fully-supervised training. $^+$ indicates increasing the number of convolutions and channels, $^-$ using half of channels, $^{--}$ using a quarter of channels.

Comparison of Model Size and Computational Cost. To illustrate that the performance improvement comes from well-designed components rather than additional parameter increases. We compare the performance of models with similar size and computational cost in Table 6. To be specific, we expand UNet, Res-Unet and WaveSNet to a similar number of parameters (Params) and multiply-accumulate operations (MACs) as XNet. This has positive effects but cannot reach XNet performance. Furthermore, we reduce the number of channels of XNet to a half and a quarter. These lightweight networks still have superior performance, indicating that the various designs of XNet are the key to improving performance.

Effectiveness of Components. To demonstrate the improvement of different components, we conduct step-by-step ablation studies for fully- and semi-supervised segmentation, and the results are shown in Table 7. By using only raw images as input, we achieve 82.03% and 78.52% JI for fully- and semi-supervised baseline, respectively. Only using LF images, fully- and semi-supervised baseline drop 1.26% and 2.22%, respectively. Only using HF images, the model performance is very poor, dropping by 27.53% and 25.96%, respectively. By using LF and HF images as inputs but without fusion, fully- and semi-supervised baseline drop 6.21% and 11.74%, respectively. Fusing LF and HF information improves baseline by 2.74% and 2.37%, respectively.

Method	Raw	L	H	Fusion	JI ↑
Fully-Supervised	✓				82.03
		✓			80.77
			✓		54.50
		✓	✓		75.82
		✓	✓	✓	**84.77**
Semi-Supervised	✓				78.52
		✓			76.30
			✓		52.56
		✓	✓		66.78
		✓	✓	✓	**80.89**

Table 7. Ablation on effectiveness of various components, including LF images, HF images and fusion module.

Through the above results, we find that only using LF and HF images as input cannot achieve positive results. The fusion of LF and HF features is critical to improving performance. In other words, the X-shaped network architecture that separates and fuses LF and HF features is the most effective.

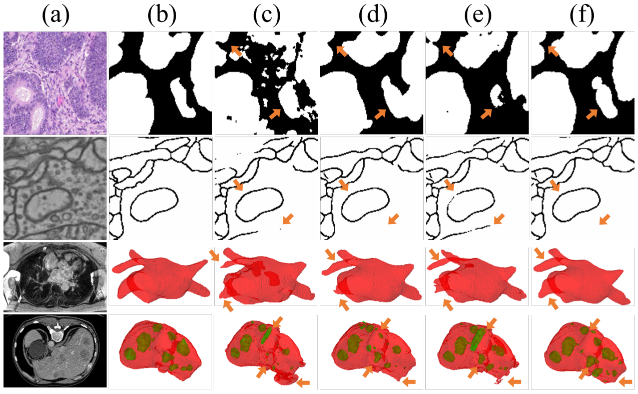

Figure 5. Qualitative results on GlaS, CREMI, LA and LiTS. (a) Raw images. (b) Ground truth. (c) MT. (d) Semi-supervised XNet (3D XNet). (e) UNet (3D UNet). (f) Fully-supervised XNet (3D XNet). The orange arrows highlight the difference among of the results.

4.6. Qualitative Results

Figure 5 shows some qualitative results of different models, XNet achieves higher accuracy on both fully- and semi-supervised segmentation. Furthermore, because of the efficient utilization of HF information, our model achieves better performance on edges and details of segmentation objects.

5. Limitations

Because XNet emphasizes HF information, when images hardly have HF information, XNet performance is negatively impacted. Figure 6 shows visual differences of HF images between ISIC-2017 [11] and CREMI. Compared to CREMI, the image shown from ISIC-2017 contains limited HF information. Table 8 compares XNet with fully- and semi-supervised baselines on ISIC-2017, XNet performance is lower than baselines.

Method	Model	ISIC-2017			
		JI ↑	DC ↑	95HD ↓	ASD ↓
Fully-Supervised	UNet [42]	74.49	85.38	9.96	4.03
	XNet	73.94	85.02	9.81	4.14
Semi-Supervised	MT [48]	72.42	84.00	11.55	4.39
	XNet	71.17	83.16	11.46	4.73

Table 8. Comparison of XNet and baseline models on ISIC-2017 test set. ISIC-2017 is a skin lesion segmentation dataset of dermoscopic images. It includes 2000 images for training and 750 images for testing.

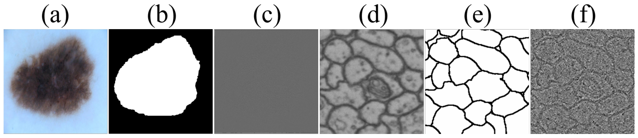

Figure 6. Visual comparison of HF images on ISIC-2017 and CREMI. (a) Raw image of ISIC-2017. (b) Ground truth of ISIC-2017. (c) HF image of ISIC-2017. (d) Raw image of CREMI. (e) Ground truth of CREMI. (f) HF image of CREMI.

6. Conclusion

We propose a wavelet-based low and high frequency fusion model XNet, which achieves state-of-the-art performance in both fully- and semi-supervised semantic segmentation of biomedical images. Extensive experiments on 2D and 3D datasets demonstrate the effectiveness of our proposed model. However, a limitation of XNet is that its performance may be negatively impacted when high frequency information is not available.

We believe that fully- and semi-supervised semantic segmentation models can and should be unified. We hope that our study may provide some examples and reflections for their unification.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China (grants 31971289, 91954201) and the Strategic Priority Research Program of the Chinese Academy of Sciences (grant XDB37040402).

References

[1] Seyed Majid Azimi, Peter Fischer, Marco Körner, and Peter Reinartz. Aerial lanenet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57(5):2920–2938, 2018.

[2] Isaac Bankman. *Handbook of medical image processing and analysis*. Elsevier, 2008.

[3] Mohit Bansal, Munesh Devi, Neha Jain, and Chinu Kukreja. A proposed approach for biomedical image denoising using pca_nlm. *International Journal of Bio-Science and Bio-Technology*, 6(6):13–20, 2014.

[4] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019.

[5] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

[6] Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. Factorised spatial representation learning: Application in semi-supervised myocardial segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 490–498. Springer, 2018.

[7] Chen Chen, Xiaopeng Liu, Meng Ding, Junfeng Zheng, and Jiangyun Li. 3d dilated multi-fiber network for real-time brain tumor segmentation in mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 184–192. Springer, 2019.

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[9] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021.

[10] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[11] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection:

A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[12] Wenhui Cui, Yanlin Liu, Yuxing Li, Menghao Guo, Yiming Li, Xiuli Li, Tianle Wang, Xiangzhu Zeng, and Chuyang Ye. Semi-supervised brain lesion segmentation with an adapted mean teacher model. In *International Conference on Information Processing in Medical Imaging*, pages 554–565. Springer, 2019.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Yiping Duan, Fang Liu, Licheng Jiao, Peng Zhao, and Lu Zhang. Sar image segmentation based on convolutional-wavelet neural network and markov random field. *Pattern Recognition*, 64:255–267, 2017.

[15] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9716–9725, 2021.

[16] Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Dmt: Dynamic mutual training for semi-supervised learning. *Pattern Recognition*, page 108777, 2022.

[17] J Funke, S Saalfeld, DD Bock, SC Turaga, and E Perlman. Miccai challenge on circuit reconstruction from electron microscopy images, 2016.

[18] Feng Gao, Xiao Wang, Yunhao Gao, Junyu Dong, and Shengke Wang. Sea ice change detection in sar images based on convolutional-wavelet neural networks. *IEEE Geoscience and Remote Sensing Letters*, 16(8):1240–1244, 2019.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing systems*, 27, 2014.

[20] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

[21] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020.

[22] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019.

[23] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[24] Qiangguo Jin, Hui Cui, Changming Sun, Jiangbin Zheng, Leyi Wei, Zhenyu Fang, Zhaopeng Meng, and Ran Su. Semi-supervised histological image segmentation via hierarchical consistency enforcement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–13. Springer, 2022.

[25] Qiufu Li and Linlin Shen. Wavesnet: Wavelet integrated deep networks for image segmentation. In *Pattern Recognition and Computer Vision: 5th Chinese Conference, PRCV 2022, Shenzhen, China, November 4–7, 2022, 2022, Proceedings, Part IV*, pages 325–337. Springer, 2022.

[26] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2020.

[27] Jinhua Liu, Christian Desrosiers, and Yuanfeng Zhou. Semi-supervised medical image segmentation using cross-model pseudo-supervision with shape awareness and local context constraints. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 140–150. Springer, 2022.

[28] Pengju Liu, Hongzhi Zhang, Wei Lian, and Wangmeng Zuo. Multi-level wavelet convolutional neural networks. *IEEE Access*, 7:74973–74985, 2019.

[29] Xiaofeng Liu, Fangxu Xing, Nadya Shusharina, Ruth Lim, CC Kuo, Georges El Fakhri, and Jonghye Woo. Act: Semi-supervised domain-adaptive medical image segmentation with asymmetric co-training. *arXiv preprint arXiv:2206.02288*, 2022.

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[32] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8801–8809, 2021.

[33] Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. *arXiv preprint arXiv:2112.04894*, 2021.

[34] Xiangde Luo, Guotai Wang, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Dimitris N Metaxas, and Shaoting Zhang. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, 80:102517, 2022.

[35] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric

medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[36] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.

[37] Nicholas Nuechterlein and Sachin Mehta. 3d-espnet with pyramidal refinement for volumetric brain tumor image segmentation. In *International MICCAI Brainlesion Workshop*, pages 245–253. Springer, 2018.

[38] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[39] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.

[40] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020.

[41] Jin Quan, William G Wee, and Chia Y Han. A new wavelet based image denoising method. In *2012 Data Compression Conference*, pages 408–408. IEEE, 2012.

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[43] John L Semmlow. *Biosignal and medical image processing*. CRC press, 2008.

[44] Pavel Sinha, Yimeng Wu, Ioannis Psaromiligkos, and Zeljko Zilic. Lumen & media segmentation of ivus images via ellipse fitting using a wavelet-decomposed subband cnn. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2020.

[45] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

[46] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015.

[47] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

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

[49] Kamini Upadhyay, Monika Agrawal, and Praveen Vashist. Wavelet based fine-to-coarse retinal blood vessel extraction using u-net model. In *2020 International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5. IEEE, 2020.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[51] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[52] Jianfeng Wang and Thomas Lukasiewicz. Rethinking bayesian deep learning methods for semi-supervised volumetric medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–190, 2022.

[53] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–119. Springer, 2021.

[54] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 171–180. Springer, 2021.

[55] Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical Image Analysis*, 67:101832, 2021.

[56] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4268–4277, 2022.

[57] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3684–3692, 2018.

[58] Xin Yin and Xiaoyang Xu. A method for improving accuracy of deeplabv3+ semantic segmentation model based on wavelet transform. In *Communications, Signal Processing, and Systems: Proceedings of the 10th International Conference on Communications, Signal Processing, and Systems, Vol. 2*, pages 315–320. Springer, 2022.

[59] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019.

[60] Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li. A simple baseline for semi-supervised semantic seg-

mentation with strong data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8229–8238, 2021.

[61] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *European conference on computer vision*, pages 173–190. Springer, 2020.

[62] Jianpeng Zhang, Yutong Xie, Yan Wang, and Yong Xia. Inter-slice context residual learning for 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(2):661–672, 2020.

[63] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

[64] Cheng Zhao, Bei Xia, Weiling Chen, Libao Guo, Jie Du, Tianfu Wang, and Baiying Lei. Multi-scale wavelet network algorithm for pediatric echocardiographic segmentation via hierarchical feature guided fusion. *Applied Soft Computing*, 107:107386, 2021.

[65] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[66] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.

[67] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2019.