ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240483

CPNet: 3D Semantic Relation and Geometry Context Prior Network for Multi-Organ Segmentation

Yuzhu Ji[1,*], **Mingshan Sun**[2,**], **Yiqun Zhang**[1] **and Haijun Zhang**[3]

[1]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
[2]CVTE Research, Guangzhou, China
[3]Harbin Institute of Technology, Shenzhen, China

Abstract. Automatic multi-organ segmentation of the abdominal region is a critical yet challenging task in computer-aided medical image analysis. Recent advances in CNN- and Transformer-based encoder-decoder models tend to implicitly learn context features by using enhanced effective receptive fields to capture local and global range dependencies. However, due to the complex anatomical structure, those models cannot recover the anatomical topology properly and result in broken organs with inaccurate semantic labels. Therefore, in this paper, by considering the anatomy priors of multi-organs, we propose a Context Prior Network, namely CPNet, which integrates the 3D context semantic relations and geometry priors as explicit anatomical constraints. Specifically, a Semantic Relation Prior Propagation (SRPP) module is designed to propagate the semantic relations between voxels progressively. Moreover, a Multiple Context Prior Prediction (MCPP) module is adopted to preserve the accurate shape and topology by recovering 3D contours and surface normal. Experimental results demonstrate our proposed model outperforms state-of-the-art models for multi-organ segmentation on Abdomen CT and MRI datasets, especially for recovering organs with correct semantic labels and anatomical structures.

1 Introduction

Automatic multi-organ segmentation is crucial in facilitating computer-aided medical image analysis and assisting disease diagnosis [22]. It aims at accurately grouping voxels into multiple organ regions and recovering the typologies of the organs. In practice, automatic multi-organ segmentation can not only reduce the laboriousness of manual annotations of physicians and radiologists but also offer potential for clinical applications, such as surgery planning, radiation therapy planning, and morphology assessment [29, 22].

Thanks to the powerful generalization abilities of deep neural networks, Convolution Neural Network (CNN) and Vision Transformers (ViT) based encoder-decoder models [44, 6, 3, 13, 14] have been widely applied to medical image segmentation tasks and achieved remarkable performance. Specifically, CNN-based 2D segmentation models can be applied to 3D volume segmentation via slice-by-slice prediction. Many representative CNN architectures for semantic segmentation and medical image segmentation, including fully convolutional network (FCN) [36], U-Net [35], DeepLab series [4, 5, 39], etc., have significantly inspired the development of new models for multi-organ segmentation [40, 3]. However, the lack of 3D context information modeling limits the performance in capturing complex 3D structures. Therefore, CNN-based 3D segmentation models are proposed. Those models regarded 3D volumes as input and achieved better performance by adopting 3D convolutional layers that can provide more effective 3D receptive fields [6, 24].

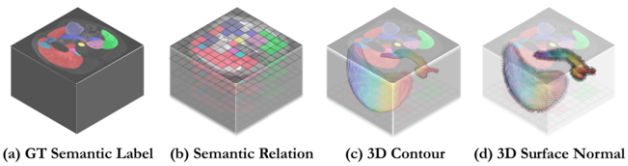

Figure 1. Multi-context priors derived from the (a) ground-truth (GT) semantic labels. *Semantic relations* (b) are extracted according to semantic labels of neighbours surrounding a central voxel. *Geometry priors* can be denoted by 3D contour (c) and surface normal (d) of abdominal organs.

Generally, CNN-based models are well-suited for learning local context features but may not be good at capturing global context due to the limited effectiveness of receptive fields. Therefore, to break such limitations, models with residual or skip connections [35, 32, 41], deformable convolutions [16], larger kernels [16], multi-level feature fusion [3, 19], and attention mechanism [34, 2] are advanced for producing rich context features. In contrast, ViT-based models [8, 27, 3, 1, 13, 26, 14] are proposed to pattern the long-range dependencies by utilizing transformer blocks, which are tailored to establish the global context relations within a sequence of tokens of 3D patches.

Albeit CNN- and ViT-based models tend to implicitly learn the organ structures by leveraging context features covered by local receptive fields and global dependencies, these models cannot properly recover the shape and lead to segmenting 3D volume into wrong organs with broken regions and incorrect topology. In practice, the structure of an organ and its surrounding tissues should be taken into consideration to draw a pathological diagnosis. On the other hand, anatomical prior knowledge of abdominal organs is also crucial to annotating voxels with correct labels by a radiologist. However, a key issue of recent data-driven models is that the *semantic relations* and *geometry* priors of the anatomical structure of abdominal organs cannot be accurately captured by identifying organs voxel-wisely.

To solve the above issues, this work presents a Context Prior Network, namely CPNet, which integrates the 3D semantic context relations and geometry priors (See Figure 1) as explicit anatomical con-

* Corresponding Author. Email: yuzhu.ji@gdut.edu.cn
** Corresponding Author. Email: mingshine.sun@gmail.com

straints for multi-organ segmentation. To achieve this, a Semantic Relation Prior Propagation (SRPP) module is designed to capture the semantic relations between the organ voxel and its context by injecting the semantic relation cues into the multi-scale feature maps progressively. In addition, a Multi-Context Prior Prediction (MCPP) module is proposed to encourage the model to recover the accurate shape and topology of the organs by using 3D contours and surface normal. Experimental results on both Abdomen CT [29] and MRI datasets [22] demonstrate our proposed model outperforms state-of-the-art models for multi-organ segmentation.

Our main contributions can be summarized as follows:

- This paper proposes to integrate multiple context priors in terms of anatomy structures and topology for multi-organ segmentation by explicitly modeling the semantic relations and geometries.
- A context prior propagation module is proposed to progressively propagate surrounding semantic relation priors into multi-scale features. It effectively alleviates the issues of inaccurate segmentation results caused by the complex context of anatomical structures and peripheral organs.
- We develop a multi-contextual relation prediction module to preserve more accurate topology and complete shape of multi-organs by leveraging 3D contour and surface normal as geometry priors.
- Experimental results demonstrate that our model can outperform state-of-the-art methods for multi-organ segmentation on the Abdomen CT and MRI datasets [29, 22], especially for recovering the connectivity and consistency in terms of semantic label and shape for different modalities.

2 Related Work

In this section, we mainly review recent advancements in deep encoder-decoder models for medical image segmentation. These models can be roughly divided into two categories based on whether the encoder-decoder models are convolution or transformer-based.

2.1 CNN-based Encoder-Decoder Models

In recent years, CNN-based encoder-decoder models have rapidly developed and achieved remarkable performance in biomedical image segmentation. Among them, Ronneberger *et al*. [35] initially proposed UNet, a symmetric encoder-decoder model that integrates multi-level features with skip connections. Due to its powerful segmentation capability, U-Net and its variants have gained significant attention for solving computer-aided medical image analysis tasks, including automatic multi-organ segmentation [16], brain tumour segmentation [32, 40], coronary artery segmentation [42, 15], etc.

Specifically, Zhou *et al*. [45] presented UNet++ with a nested UNet architecture to fuse multi-level features with dense skip connections Huang *et al*. [18] proposed the UNet 3+ model to learn hierarchical representations by aggregating the full-scale feature maps with dense skip connections. However, initial 2D-based U-Net models typically perform full volumetric segmentation by processing the 3D data slice-by-slice. To capture context features of 3D volume, 2.5D and 3D approaches are proposed to integrate features from neighbouring slices in the 3D space. Çiçek *et al*. [6] developed 3D U-Net by replacing 2D convolution with its 3D convolution and achieved dense segmentation on volumetric data. In addition, Li *et al*. [25] proposed a hybrid architecture H-DenseUNet and achieved efficient segmentation of liver and tumour from CT images. It aims to incorporate intra- and inter-slice contextual information by cascading 2D and 3D convolution networks. After that, Oktay *et al*. [34] proposed an Unet model with attention gates, namely AG-UNet. An attention gate layer is designed to encourage the model to select useful features for segmentation. Unlike the above models that introduced (dense) skip connections, residual connections, and attention mechanism into the original UNet, Isensee *et al*. [20] proposed a self-configuring method, namely nnUNet, to achieve automatic and adaptive configuration of training the UNet models. Techniques in terms of choosing network architecture, data pre-processing, post-processing, empirical settings during training and inference procedures, optimizers, etc., can be automatically adopted based on the given medical image dataset [20].

2.2 Transformer-based Models

Recent years have witnessed the rapid development of the Vision Transformer (ViT) model [8] and its applications for medical image segmentation [3, 1, 13]. Thanks to the powerful ability to capture long-range dependencies, ViT models have been explored for implicitly pattern global context features within the volumetric data and excepted to break the limitations in extracting insufficient context features within local receptive fields covered by convolutional kernels. Specifically, Valanarasu *et al*. [38] proposed MedT with a gated axial-attention transformer block. Chen *et al*. [3] proposed TransUNet, which is a hybrid framework that combines the Transformer block with the CNN-based UNet. Huang *et al*. [19] proposed another hybrid CNN and Transformer architecture, namely ScaleFormer to aggregate and distribute inter- and intra-scale local-global features by using transformer blocks. Lin *et al*. [2] proposed to introduce a dual-scale encoding mechanism into the hierarchical Swin Transformer UNet architecture. In contrast, Cao *et al*. [1] offered SwinUNet, a pure Transformer-based UNet model built upon the hierarchical Swin-Transformer.

Due to runtime efficiency and computational complexity, the above Transformer-based models are generally worked on 2D slices or 2D medical images. In contrast, Wang *et al*. [40] proposed TransBTS for brain tumour segmentation. It integrates a Transformer into a 3D CNN architecture to model local and global context features. Hatamizadeh *et al*. [14] proposed UNETR, a transformer-based UNet tailored to perform segmentation on 3D volumetric data directly. In addition, inspired by self-supervised learning, Tang *et al*. [13] proposed a 3D Transformer-based model, namely SwinUNETR with a pre-trained Swin Transformer backbone. Zhou *et al*. [43] proposed nnFormer to learn the 3D volumetric context representations by designing a local and global volume-based self-attention mechanism. Recently, the state space model (SSM) and its variants [9, 11, 12, 10] were proposed to capture the log-range dependencies more efficiently by using different scanning strategies, it becomes competitive with CNN and Transformer-based models and has been applied in natural language processing and computer vision tasks [21, 33]. In particular, Ma *et al*. [30] introduced SSMs in the multi-organ segmentation task to establish long-term dependencies between organs.

However, the above models are inclined to implicitly realize the anatomical organ structures by modeling local and global context features following a data-driven training pipeline. Some recent works proceed to introduce topological constraints to regularize the learning process of models for coronary artery segmentation [42] and natural image segmentation [17]. Nevertheless, due to the complex geometry of the organs, the anatomical constraints are not well perceived, but they can be regarded as strong guidance for segmentation in practice. Therefore, we propose to introduce anatomy structure and topology

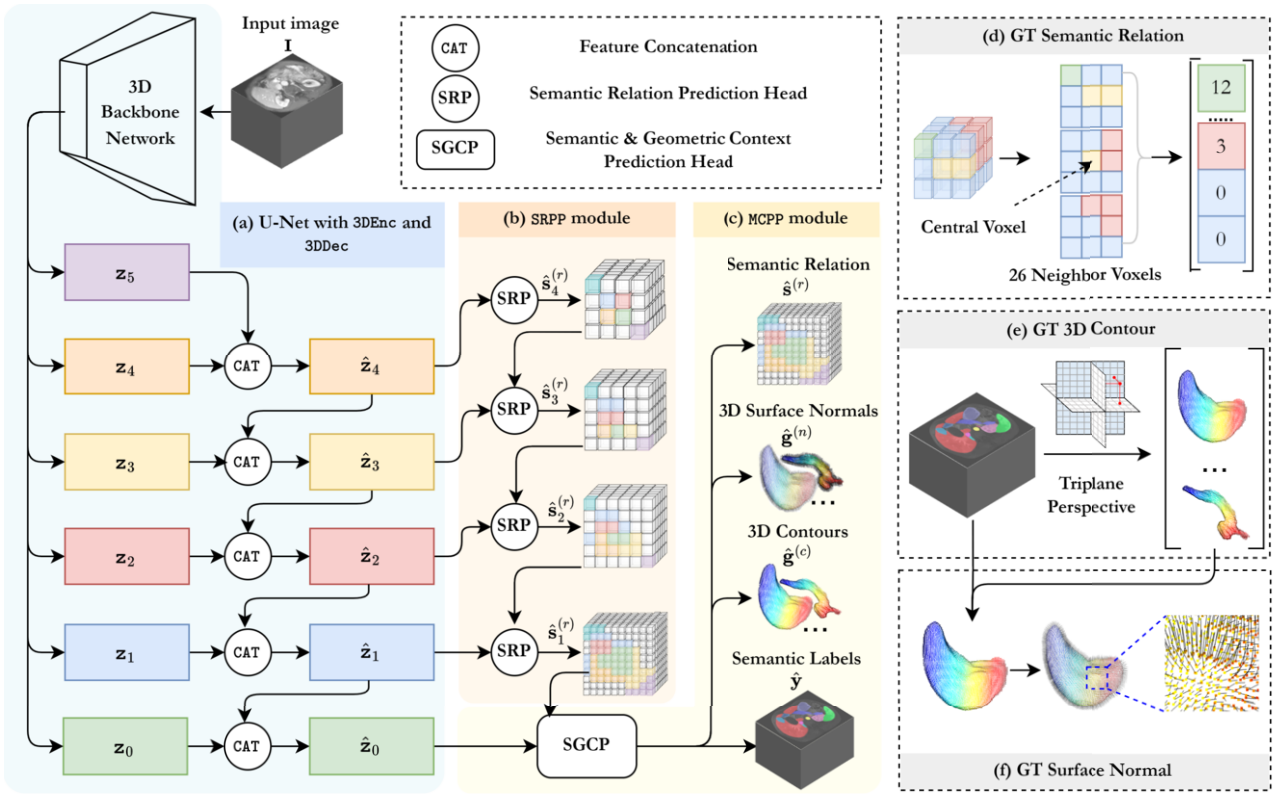

Figure 2. The overview of our proposed CPNet. The input medical data is encoded by using a 3D backbone network. Multi-scale features are integrated and decoded under a 3D U-Net Encoder-Decoder framework (a). A SRPP module (b) is designed to predict and propagate semantic relations (d) to multi-scale features. A MCPP module (c) is developed to predict semantic segmentation with correct 3D contour (e) and surface normal (f).

priors, including spatial semantic relation, 3D contour, and surface normal to facilitate multi-organ segmentation by comprehensively considering organ shapes and their relations.

3 Method

3.1 Overview

Let $\mathbf{I}$ denotes the input 3D CT or MRI abdominal volume data, our main goal is to predict the accurate semantic labels and recover the correct geometry of multiple organs. To this end, we developed CPNet by introducing multiple context priors in terms of organ structures and topology as anatomical constraints. The overview of our proposed CPNet architecture is illustrated in Figure 2. It consists of three main modules: (a) A 3D U-Net for encoding and fusing multi-scale features, (b) A Semantic Relation Prior Propagation (SRPP) module to inject context semantic relations into multi-scale features progressively, and (c) A Multi-context Prior Prediction (MCPP) module for recovering the correct geometry by using 3D contour and surface normal. To construct multiple context priors of anatomical structures of organs, we extract 3D semantic relations (d) according to 26 connected neighbourhoods surrounding a central voxel. In addition, 3D contours (e) are generated by slicing the ground-truth semantic labels from tri-plane perspectives and the surface normal (f) of the 3D shape is computed consequentially.

3.2 Multi-scale 3D Feature Extraction and Fusion

Recent works have demonstrated the effectiveness of UNet architecture for medical image segmentation. As a widely used encoder-decoder model, a 3D U-Net architecture is adapted for multi-scale feature extraction and fusion in our proposed CPNet. As shown in Figure 2(a), CPNet inherits a basic 3D U-Net network from [20], including a 3D backbone network as an encoder with 3D convolution blocks, a decoder for iteratively upsampling and fusing multi-scale features by using skip connections. This can be expressed as:

$$[\hat{\mathbf{z}}_0, \hat{\mathbf{z}}_1, \cdots, \hat{\mathbf{z}}_l] = \texttt{3DDec}(\texttt{3DEnc}(\mathbf{I})), \tag{1}$$

where $\hat{\mathbf{z}}_i$ denotes the i-th level of features produced by the decoder 3DDec. $\mathbf{I}$ denotes the input 3D CT or MRI data volume with resolution resolution $D \times H \times W$, where D, H and W are the spatial depth, height and width, respectively. In practice, high-level feature $\hat{\mathbf{z}}_l$ models the high-level semantic and coarse instance-level location of each organ in the 3D volumetric space, while low-level feature $\hat{\mathbf{z}}_0$ contains more detailed local information which can be used to recover fine shapes and structures of multiple organs. We fuse high- and low-level features by concatenating two adjacent scales of features produced by the encoder 3DEnc and decoder 3DDec parts.

3.3 Progressive Semantic Relation Prior Propagation

Since the core of our model is to introduce context priors as explicit anatomical constraints, it aims to encourage the model to learn to recover semantic labels with correct organ structures and topology. The semantic relation can be regarded as a strong context prior knowledge of the layout of multi-organs. To achieve this, the semantic relation prior is extracted by recording the semantic labels of the neighbour voxels surrounding a central voxel. Figure 2(d) illustrates the construction of semantic relation for a voxel. It is defined by the current

voxel and its 26 neighbours within a $3 \times 3 \times 3$ receptive field. Therefore, for each position, the semantic relation prior is expanded by semantic labels of the central voxel and its neighbour voxels. For different levels of features, the semantic relation priors indicate different scales of relations. Intuitively, the high-level semantic relations indicate global anatomical structures of abdominal organs, such as the location, distributions, and layouts of the anatomy priors. The low-level semantic relations indicate the fine-grained local patterns of the inter- and intra-organ semantic context relations. Therefore, as shown in Figure 2(b), to take full use of the local and global semantic relation priors, we propose to progressively propagate the semantic relations to multi-level features for predicting the coarse-to-fine semantic context priors under the deeply supervised framework. Concretely, we adopted a semantic relation prediction head in each layer of the decoder. The i-th level of the decoded feature is concatenated with the semantic relation predicted by the previous $i-1$-th level. After that, a $1 \times 1 \times 1$ convolutional layer is applied to obtain the current layer semantic relation feature. The computation of the progressive semantic relation prior propagation (SRPP) module can be formulated as:

$$\hat{\mathbf{s}}_i^{(r)} = \begin{cases} \mathtt{P}_i^{(r)}(\mathtt{CONV}_i(\mathtt{CAT}(\hat{\mathbf{z}}_{i-1}, \hat{\mathbf{s}}_{i-1}^{(r)}))) & \text{if } i = 1, 2, \cdots, l-1 \\ \mathtt{P}_i^{(r)}(\mathtt{CONV}_i(\hat{\mathbf{z}}_i)) & \text{if } i = l \end{cases}, \tag{2}$$

where $\mathtt{P}_i^{(r)}$ denotes the i-th prediction head of the i-th level of features. $\hat{\mathbf{s}}_i^{(r)}$ refers the i-th semantic relation prediction with spatial size $\mathbb{R}^{\hat{c} \times d \times h \times w}$, where $\hat{c}$ denotes the number of voxels within the neighbourhoods.

3.4 Multiple Context Prior Prediction

As mentioned above, our key insight is to accurately recover the semantics and geometry of abdominal organs by explicitly introducing anatomical constraints. To achieve this, a Multiple Context Prior Prediction (MCPP) module is presented to accurately predict the shape and topology of the organs by using 3D contours and surface normal (See Figure 2(c)). Specifically, we designed an additional 3D contour prediction head to identify the discontinuity between the voxels of an organ and its surrounding tissues. In addition, an extra surface normal prediction head is introduced to discriminate the intra- and inter-class voxels around the 3D surface of an organ and recover the correct topology.

In practice, low-level feature contains more detailed local information. It is more suitable for realizing correct geometry with fine shapes and structures. Therefore, the low-level feature $\hat{\mathbf{z}}_0$ is fed into the MCPP module to simultaneously predict semantic labels and geometry:

$$\left[\hat{\mathbf{s}}^{(r)}, \hat{\mathbf{g}}^{(n)}, \hat{\mathbf{g}}^{(c)}, \hat{\mathbf{y}}\right] = \mathtt{MCPP}\left(\mathtt{CAT}(\mathbf{s}_1^{(r)}, \hat{\mathbf{z}}_0)\right), \tag{3}$$

where $\hat{\mathbf{g}}^{(n)}$ and $\hat{\mathbf{g}}^{(c)}$ denote the predicted surface normal and contour respectively. $\hat{\mathbf{s}}^{(r)}$ represents the final predicted semantic relation and $\hat{\mathbf{y}}$ is the predicted semantic label. The structure of the MCPP module consists of four separate prediction heads for semantic relation prediction, contour prediction, normal vector regression, and semantic segmentation, respectively.

To generate the ground truth (GT) of the 3D contour and surface normal shown in Figure 2 (e) and (f), we first slice the 3D segmentation labels from the tri-plane perspectives in terms of axial, sagittal and coronal views. Specifically, for each view, boundaries of 2D semantic masks are extracted slice-by-slice and the corresponding index numbers of the slice are recorded. After that, the GT 3D contour $\mathbf{g}^{(c)}$ can be obtained by gathering all the boundary points w.r.t. axial, sagittal and coronal views according to the recorded 3D coordinates. In addition, for calculating the GT 3D surface normal $\mathbf{g}^{(n)}$, we follow the approach presented in Open3D [44]. Concretely, a KD-Tree searching algorithm is first used to retrieve the nearest points in a searching ball with a specific radius for each point on the contour. Then, a covariance analysis method is adopted to estimate a normal vector based on the nearest neighbours. For clarity, we refer the readers to supplementary meterial [23] for more details.

It is worth noting that some of the estimated normal vectors may point toward the inside of the organ according to the default settings of randomly choosing normal candidates. Therefore, to eliminate such confusing cases, we reverse the directions of those normal vectors so that all the estimated normal vectors point to the outside of the organ. In practice, we identify the direction of normal vectors by the semantic labels of a surface point and its nearest neighbour that the normal vector pointing to. If the semantic labels are the same, the direction of a normal vector is towards the inside of the organ and should be reversed.

3.5 Loss Functions

Multiple loss functions are used in training our model to predict the semantic label and recover the correct geometry of multi-organs. Specifically, we used a compound loss function of the cross-entropy loss CELoss and dice loss DiceLoss for semantic segmentation by following [20]:

$$\mathcal{L}_{seg} = \lambda_{ce}\mathtt{CELoss}(\mathbf{y}, \hat{\mathbf{y}}) + \lambda_{dice}\mathtt{DiceLoss}(\mathbf{y}, \hat{\mathbf{y}}), \tag{4}$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ denote the GT semantic label and predicted semantic label, respectively. λ_{ce} and λ_{dice} are the weights for each loss. In addition, a relation-balanced cross-entropy loss RCELoss is adopted to predict and propagate semantic relations to multi-scale features progressively:

$$\mathcal{L}_{rce} = \sum_{i=0}^{l} \mathtt{RCELoss}(\mathbf{s}^{(r)}, \hat{\mathbf{s}}_i^{(r)}), \tag{5}$$

where $\mathbf{s}^{(r)}$ and $\hat{\mathbf{s}}_i^{(r)}$ represent the GT semantic relation and predicted relation, respectively. i denotes the i-th side-way output and l is the number of levels of semantic relation predictions. We design the RCELoss by reweighting the losses w.r.t. three types of difficulties of different semantic relations. For more details, we refer the readers to the supplementary document [23].

Moreover, to recover the geometry of multi-organs, we adopted cross-entropy loss CELoss and L1 loss L1Loss to penalize the wrong boundary points and inaccurate normal vectors. Specifically, for 3D contour prediction, we introduce the cross-entropy loss to encourage the model to produce accurate contours:

$$\mathcal{L}_{gce} = \mathtt{CELoss}(\mathbf{g}^{(c)}, \hat{\mathbf{g}}^{(c)}), \tag{6}$$

where $\mathbf{g}^{(r)}$ and $\hat{\mathbf{g}}^{(r)}$ are GT 3D contours and predicted contours of multi-organs. Finally, an L1Loss is adopted to regress surface normal, which can be defined as:

$$\mathcal{L}_{gn} = \mathtt{L1Loss}(\mathbf{g}^{(n)}, \hat{\mathbf{g}}^{(n)}), \tag{7}$$

where $\mathbf{g}^{(n)}$ and $\hat{\mathbf{g}}^{(n)}$ are GT 3D surface normal vectors and predicted normal vectors of multi-organs. The overall loss can be defined as:

$$\mathcal{L} = \lambda_{seg}\mathcal{L}_{seg} + \lambda_{rce}\mathcal{L}_{rce} + \lambda_{gce}\mathcal{L}_{gce} + \lambda_{gn}\mathcal{L}_{gn}, \tag{8}$$

where λ_{seg}, λ_{rce}, λ_{gce}, and λ_{gn} are the loss weights to trade-off different terms.

Table 1. Organ-wise quantitative results of 3D segmentation on Abdomen CT dataset. The best result for each class is bolded.

	DSC(%)↑						NSD(%)↑					
	nnU-Net	SegResNet	UNETR	SwinUNETR	U-Mamba	Ours	nnU-Net	SegResNet	UNETR	SwinUNETR	U-Mamba	Ours
liver	97.06	95.21	90.47	93.02	**97.13**	93.25	**95.82**	91.57	81.25	82.97	95.92	91.36
right kidney	87.38	82.95	71.05	75.40	86.53	**89.03**	86.21	80.27	66.09	70.53	86.40	**88.06**
spleen	91.62	86.46	78.69	84.02	93.57	**93.81**	90.38	84.21	73.47	80.20	**93.61**	93.57
pancreas	83.62	74.40	60.10	69.73	**86.51**	86.59	91.39	83.53	68.59	71.91	94.02	**94.49**
aorta	96.18	94.91	89.16	93.83	95.70	**97.13**	97.45	95.58	87.01	93.40	97.27	**98.59**
inferior vena cava	88.39	83.74	75.92	82.50	88.22	**89.46**	87.76	82.42	71.99	78.34	87.84	**89.20**
right adrenal gland	**82.31**	72.77	63.73	74.66	81.40	80.09	**93.18**	85.79	76.15	87.03	92.36	90.00
left adrenal gland	79.21	68.51	47.50	67.67	83.23	**81.75**	88.84	79.45	58.66	78.88	92.58	**91.42**
gallbladder	72.93	66.78	53.21	57.25	74.36	**79.33**	73.31	64.11	47.47	53.20	74.67	**80.61**
esophagus	**86.06**	79.27	69.29	78.53	85.23	82.58	**93.08**	87.06	78.26	86.63	91.49	89.55
stomach	89.02	82.07	71.28	76.56	89.31	**89.38**	90.33	83.02	69.03	69.81	90.76	**91.03**
duodenum	75.61	63.88	49.51	60.42	77.94	**78.41**	88.30	79.61	69.32	75.31	89.30	**90.60**
left kidney	**90.50**	79.49	67.24	73.67	89.63	89.26	**90.35**	76.86	63.24	67.99	90.19	88.62
Avgerage	86.15	79.26	68.24	75.94	86.83	**86.95**	89.72	82.58	70.04	76.63	90.49	**90.78**

Table 2. Organ-wise quantitative results of 3D segmentation on Abdomen MRI dataset. The best result for each class is bolded.

	DSC(%)↑						NSD(%)↑					
	nnU-Net	SegResNet	UNETR	SwinUNETR	U-Mamba	Ours	nnU-Net	SegResNet	UNETR	SwinUNETR	U-Mamba	Ours
liver	97.35	96.54	93.44	96.27	97.32	**97.50**	97.47	95.90	89.05	94.79	97.53	**97.83**
right kidney	**96.25**	93.90	82.66	93.55	95.94	95.96	97.47	95.88	82.28	93.90	**97.66**	97.61
spleen	91.31	90.23	86.17	91.34	**93.83**	93.00	92.01	90.69	82.80	90.19	**94.17**	93.86
pancreas	86.39	82.49	72.71	77.50	86.50	**87.48**	95.53	92.28	82.68	87.42	95.65	**96.78**
aorta	93.27	93.31	85.24	89.07	92.43	**95.02**	95.72	95.56	86.52	90.55	94.79	**97.41**
inferior vena cava	81.92	82.79	72.86	78.44	83.22	**83.54**	86.98	87.62	76.05	82.71	88.12	**88.30**
right adrenal gland	62.38	60.66	45.39	55.40	62.87	**63.94**	80.47	77.78	62.57	73.16	80.10	**81.54**
left adrenal gland	70.21	67.79	47.35	47.86	**72.25**	72.21	84.74	83.01	61.43	62.55	85.88	**86.17**
gallbladder	77.20	78.72	48.33	65.37	82.84	**85.28**	75.51	75.81	40.95	61.85	82.79	**85.13**
esophagus	74.89	74.08	56.55	64.77	79.55	**79.60**	90.71	89.76	73.95	80.38	93.87	**94.48**
stomach	82.60	76.09	66.97	71.98	82.57	**83.76**	85.95	79.89	69.19	75.53	85.97	**87.60**
duodenum	70.08	66.34	48.78	59.21	**73.09**	72.48	88.92	87.20	74.95	81.48	90.78	**91.01**
left kidney	96.37	96.09	86.22	92.66	96.49	**96.57**	98.05	97.98	84.79	93.77	**98.42**	98.26
Avgerage	83.09	81.46	68.67	75.65	84.53	**85.10**	89.96	88.41	74.40	82.18	91.21	**92.00**

4 Experiments

In this section, we start by presenting the datasets, evaluation metrics, and implementation details. Following that, we discuss the experimental results obtained from comprehensive comparison and ablation studies to demonstrate the effectiveness of our proposed CPNet.

4.1 Datasets and Evaluation Metrics

To verify the effectiveness of our method on various types of multi-organ segmentation datasets, we evaluated CPNet on the Abdomen CT and MRI datasets. The Abdomen CT dataset was provided by MICCAI 2022 FLARE Challenge [29] for the segmentation of 13 abdominal organs. It contained 50 CT scans collected from the MSD Pancreas dataset [37] used for training and another 50 scans from different medical centers [7] for testing. The annotations of the training set were from the AbdomenCT-1K dataset [28], while annotations of the testing set were provided by the challenge organizers. The Abdomen MRI dataset was released by MICCAI 2022 AMOS Challenge [22], we followed settings of U-Mamba [30] that merge the 40 training scans and 20 validation scans for training, and 50 extra annotated MRI scans were used for testing. As with the Abdomen CT dataset, the same set of 13 organs was selected to train the model and facilitate the comparison of abdominal organ segmentation based on different modalities. Following [31], we adopted two widely used metrics, *i.e*. the Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) to evaluate the segmentation performance.

4.2 Implementation Details

We implement CPNet based on nnU-Net [20] framework and we follow the same hyperparameter setting and training policy of nnU-Net. During training, the CPNet is optimized by using an SGD optimizer with the initial learning rate 5×10^{-3} and weight decay 3×10^{-5}. The model is trained for 1,000 epochs. For each iteration, a batch of two volumes are randomly selected and cropped to resolution $40 \times 224 \times 192$ (Abdomen CT) and $48 \times 160 \times 224$ (Abdomen MRI) from the original data. For loss weights in Eq. (8), λ_{seg}, λ_{rce}, λ_{gce}, and λ_{gn} are set to 1.0. Moreover, we set the relation-balanced CE loss weight according to the type of each voxel. Specifically, for easy cases in the background and intra-organ regions, we set the loss weights in `RCELoss` to 0.1 and 0.3, respectively. For hard cases, *i.e*. voxels are surrounded by different classes, the loss weight is set to 0.6. Moreover, to augment the diversity of the training data volume, we follow the medical image data preprocessing steps presented in [20]. To generate the GT surface normal, we set the radius of the searching ball to 10 and the maximum number of nearest neighbours to 100. All the experiments are implemented using PyTorch on a workstation with one NVIDIA GeForce RTX 4090 (24G) GPU.

4.3 Main Results

We performed experiments on the Abdomen CT and MRI datasets by comparing our CPNet with five state-of-the-art (SOTA) models for multi-organ segmentation, including nnU-Net [20], SegResNet [32], UNETR [14], SwinUNETR [13] and U-Mamba [30]. For a fair comparison, these models are implemented based on the nnU-Net and trained with the default settings of training policies and hyperparameters released by the authors.

Evaluation results on Abdomen CT dataset. Table 1 summarized the quantitative comparison results w.r.t. different organs on the Abdomen CT dataset. It shows that our model outperforms SOTA methods by achieving 86.95% on the Average DSC and 90.78%

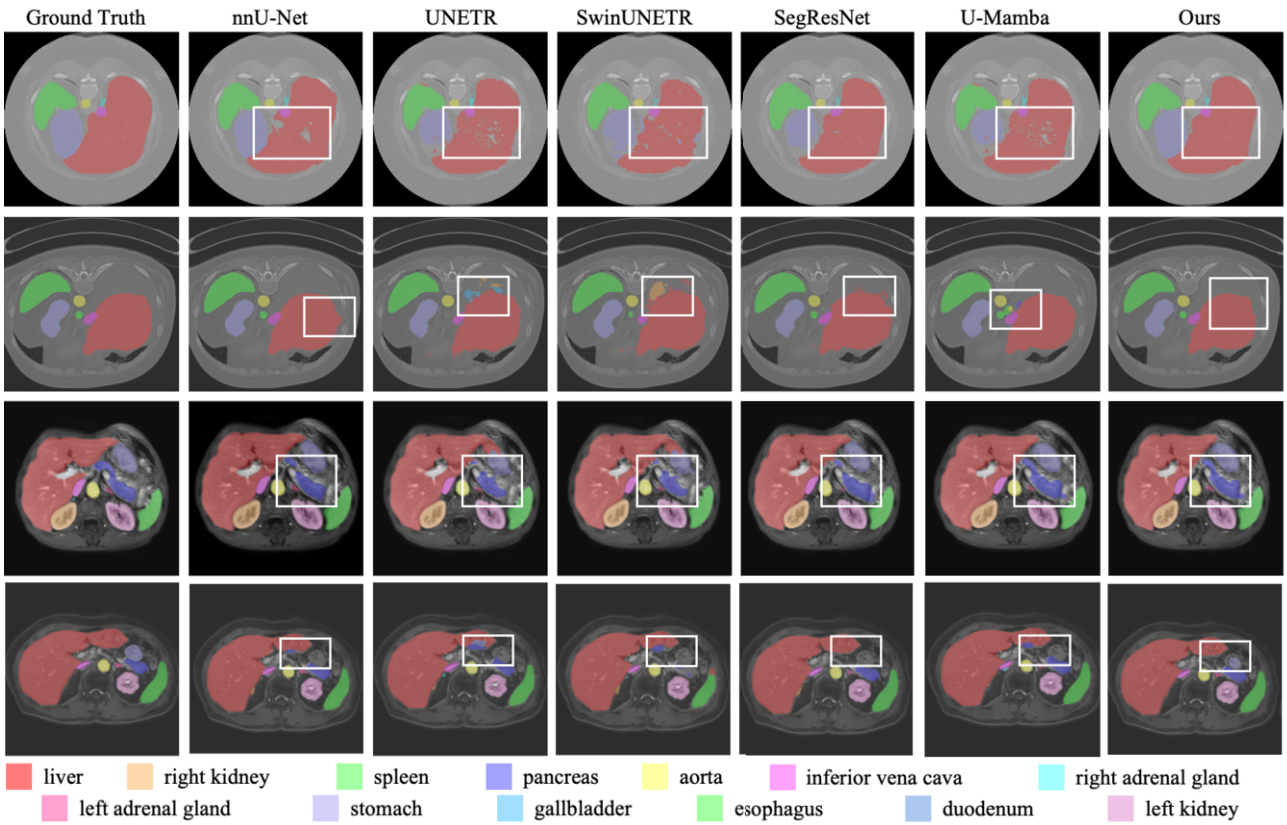

Figure 3. Qualitative results of our proposed CPNet on the Abdomen CT (the first two lines) and MRI (the last two lines) slices.

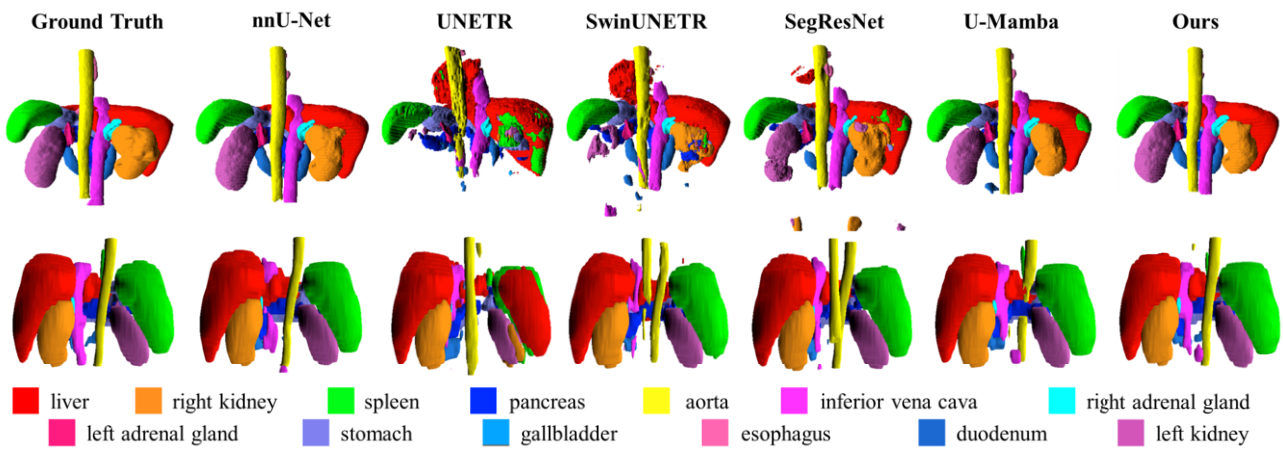

Figure 4. Qualitative results of our proposed CPNet on the abdomen CT (the first row) and MRI (the second row) volumes.

on the Average NSD, respectively. Specifically, in comparison with the baseline nnU-Net, our model improves the Average DSC and NSD by 0.80% and 1.06%, respectively. It demonstrates the effectiveness of introducing semantic relation and geometry priors to improve the segmentation performance. On the other hand, for SOTA transformer- and Mamba-based models, which aim to learn the implicit long-range contextual dependencies, our model still achieves 0.12% and 0.29% improvement in comparison with U-Mamba on the Average DSC and NSD, respectively. However, it is worth noting that our model is built upon the nnU-net with basic 3D convolutional blocks but no bells and whistles modules, such as the feature pyramid network (FPN) and its variants, feature fusion with attention mechanism, non-local blocks for capturing long-range dependencies, etc. We believe our model can be further improved by integrating the implicit local-global context information.

Evaluation results on Abdomen MRI dataset. Organ-wise quantitative results on the Abdomen MRI dataset are listed in Table 2. It shows that our model outperforms SOTA methods by achieving 85.10% and 92% w.r.t. the Average DSC and NSD, respectively.

Similarly, significant improvement can be observed by comparing CPNet with nnU-Net. Specifically, our model gains 2.01% improvements on DSC and 2.04% on NSD. Moreover, compared to U-Mamba, our CPNet also shows 0.57% improvement on DSC and 0.79% improvement on NSD. Moreover, the performance improvements in terms of DSC and NSD on the MRI dataset are more significant than the CT dataset. This may be related to the diversity of the testing CT data collected from different medical centers. Additionally, experimental results for other metrics, *i.e*. Jaccard Index, 95% Hausdorff distance (HD95), and the runtime efficiency, can be found in the supplementary material [23].

Visual Comparison. To better visualize the multi-organ segmentation results, we present Figure 3 and Figure 4 to illustrate some qualitative segmentation results w.r.t. 2D (axial plane) and 3D views. In comparison to the SOTA methods, our proposed model shows strengths in three aspects: (1) Predicting accurate *semantic labels*. For results in Figure 3, the "liver" and "pancreas" are correctly segmented without broken regions. On the contrary, the Transformer- and Mamba-based models with the capability to capture long-range context dependencies produce over-segmentation results with intra-organ holes and outlier tissues. This may be related to invalid relations between the queries and keys were also picked up and learned using transformer or SSM blocks. In addition, such implicit long-range context relations may also be impacted by the noise and corrupt the segmentation results. (2) Recovering the correct *geometry*. It can be observed that the shapes of the abdomen organs can be well preserved and the connectivity and consistency with the anatomical structure are well-preserved, especially for organs "pancreas", "aorta", "right kidney", *etc*. (See Figure 4). (3) Handling medical data from different data centers and modalities, results in Figure 3 also indicate that our model can achieve better robustness on CT and MRI modalities in comparison with SOTA methods.

4.4 Ablation Study

Ablation on `SRPP` module. To analyze the effectiveness of our `SRPP` module (Sec.3.3), we used the nnU-Net as the *baseline*. The models with and without adding progressive propagation in the `SRPP` module are trained for ablation analysis. In addition, we also investigated the semantic context relation construction strategy w.r.t. different sizes of context scopes. The quantitative results are listed in Table 3. Specifically, the "SMP" denotes that the `SRPP` module progressively injects the predicted semantic relations. In contrast, "SM" means the counterpart without progressive propagation. "L" and "S" represent that the GT semantic relations are extracted according to 26- and 6-neighbors, respectively. It can be observed that adding semantic relations, *i.e*. "+SM-L", achieves minor improvement on the CT dataset but realizes a 0.99% improvement in DSC and 0.92% in NSD on the MRI dataset. In addition, the performance can be further improved by propagating context semantic relations progressively. The "+SMP-L" obtains 0.67% improvement in DSC and 0.92% in NSD on the CT dataset. In addition, it also gains 1.67% improvement in DSC and 1.56% in NSD on the MRI dataset compared with the baseline. Furthermore, by comparing the results of baseline models with "L" and "S" settings, it indicates that the larger the number of connected neighbourhoods, the better the performance.

Ablation on `MCPP` Module. We further ablated the `MCPP` module by introducing different combinations of geometry priors, *i.e*. 3D contour and surface normal. Quantitative results are reported in Table 4.

Table 3. Quantitative results of ablation study on `SRPP` module.

	Abdomen CT		Abdomen MRI	
	DSC(%)	NSD(%)	DSC(%)	NSD(%)
baseline	86.15	89.72	83.09	89.96
baseline+SM-S	86.34	90.19	84.08	90.88
baseline+SM-L	86.78	90.47	84.70	91.38
baseline+SMP-S	86.57	90.23	84.61	91.33
baseline+SMP-L	**86.82**	**90.64**	**84.76**	**91.52**

Table 4. Quantitative results of ablation study on `MCPP` module.

		Abdomen CT		Abdomen MRI	
Contour	Normal	DSC(%)	NSD(%)	DSC(%)	NSD(%)
✗	✗	86.15	89.72	83.09	89.96
✔	✗	86.35	90.11	83.66	90.52
✗	✔	86.47	90.15	83.48	90.60
✔	✔	**86.65**	**90.42**	**83.91**	**90.89**

Specifically, the nnU-Net trained without any geometry as supervision is regarded as the baseline model. We evaluated the performance of models trained by introducing 3D contour and surface normal separately and jointly. It shows that introducing 3D contour alone as the geometry prior brings a slight improvement in terms of Average DSC and NSD on the CT and MRI datasets. In addition, leveraging surface normal as supervision to recover the geometry achieves better performance in comparison with using 3D contour. The reason may be that the surface normal is derived from the surface point and its neighbours. In comparison to 3D contour, it can be regarded as a more strict constraint to encourage the model to recover the geometry with the correct neighbour coordinates. Furthermore, using both 3D contour and surface normal brings significant improvement w.r.t. 0.50% in DSC and 0.70% in NSD on the CT dataset, 0.70% improvement in DSC and 0.93% NSD on the MRI dataset, respectively.

5 Conclusions and Future work

In this paper, we proposed a context-prior network, dubbed CPNet, which aims to recover anatomical organ structure with accurate semantic labels and geometry for multi-organ segmentation. To achieve this, a Semantic Relation Prior Propagation (SRPP) module is proposed for more accurate semantic prediction by explicitly propagating semantic relation priors into multi-scale features. Moreover, a Multi-context Prior Prediction (MCPP) module is presented to recover the correct geometry of organs by using 3D contour and surface normal. Experimental results on the Abdomen CT and MRI datasets demonstrate our model can outperform state-of-the-art models, especially for recovering connectivity and consistency in terms of semantic labels and shapes. In the future, it is worth further investigations on model variations, such as combining implicit local-global context feature learning modules and exploring extra anatomical constraints as regularizers w.r.t. different segmentation tasks.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under grants: 62302104, 62476063 and 62102097, the Natural Science Foundation of Guangdong Province under grants: 2023A1515012884 and 2023A1515012855, the Science and Technology Program of Guangzhou under grant SL2023A04J01625.

References

[1] H. Cao, Y. Wang, et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In *ECCV*, pages 205–218, 2023.

[2] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. W. Kong. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Trans. Emerg. Top. Comput. Intell.*, 8(1):55–68, 2024.

[3] J. Chen, Y. Lu, et al. Transunet: Transformers make strong encoders for medical image segmentation. *CoRR*, 2021. URL https://arxiv.org/abs/2102.04306.

[4] L. 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 Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018.

[5] L. Chen, Y. Zhu, G. Papandreou, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, volume 11211 of *Lecture Notes in Computer Science*, pages 833–851. Springer, 2018.

[6] Ö. Çiçek, A. Abdulkadir, et al. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *MICCAI*, pages 424–432, 2016.

[7] K. W. Clark, B. A. Vendt, K. E. Smith, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging*, 26(6):1045–1057, 2013.

[8] A. Dosovitskiy, L. Beyer, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[9] A. Gu. *Modeling Sequences with Structured State Spaces*. Stanford University, 2023.

[10] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[11] A. Gu et al. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

[12] A. Gu et al. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021.

[13] A. Hatamizadeh, V. Nath, Y. Tang, et al. Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In *MICCAI*, volume 12962 of *Lecture Notes in Computer Science*, pages 272–284. Springer, 2021.

[14] A. Hatamizadeh, Y. Tang, V. Nath, et al. UNETR: transformers for 3d medical image segmentation. In *WACV*, pages 1748–1758. IEEE, 2022.

[15] S. He, Y. Ji, Y. Zhang, et al. Cfnet: A coarse-to-fine framework for coronary artery segmentation. In *Pattern Recognition and Computer Vision*, volume 14429 of *Lecture Notes in Computer Science*, pages 431–442. Springer, 2023.

[16] M. P. Heinrich, O. Oktay, and N. Bouteldja. Obelisk-net: Fewer layers to solve 3d multi-organ segmentation with sparse deformable convolutions. *Medical Image Anal.*, 54:1–9, 2019.

[17] X. Hu, F. Li, D. Samaras, and C. Chen. Topology-preserving deep image segmentation. In *NeurIPS*, pages 5658–5669, 2019.

[18] H. Huang, L. Lin, et al. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP*, pages 1055–1059, 2020.

[19] H. Huang, S. Xie, L. Lin, et al. Scaleformer: Revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation. In *IJCAI*, pages 964–971. ijcai.org, 2022.

[20] F. Isensee, Jaeger, et al. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18 (2):203–211, 2021.

[21] M. M. Islam and G. Bertasius. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022.

[22] Y. Ji, H. Bai, C. Ge, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. volume 35, pages 36722–36732, 2022.

[23] Y. Ji, M. Sun, Y. Zhang, and H. Zhang. Cpnet: 3d semantic relation and geometry context prior network for multi-organ segmentation, supplementary material. In *the 27-th Europen Conference on Artificial Intelligence, Santiago de Compostela, October 19-24*, 2024. URL https://github.com/AndrewChiyz/ECAI24-CPNet-m351.

[24] H. H. Lee, S. Bao, et al. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. 2022.

[25] X. Li, H. Chen, X. Qi, et al. H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from CT volumes. *IEEE Trans. Medical Imaging*, 37(12):2663–2674, 2018.

[26] A. Lin, B. Chen, J. Xu, et al. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Trans. Instrum. Meas.*, 71:1–15, 2022.

[27] Z. Liu, Y. Lin, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[28] J. Ma, Y. Zhang, S. Gu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):6695–6714, 2022.

[29] J. Ma, Y. Zhang, S. Gu, C. Ge, et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the FLARE22 challenge. *CoRR*, abs/2308.05862, 2023.

[30] J. Ma, F. Li, and B. Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *CoRR*, abs/2401.04722, 2024.

[31] L. Maier-Hein, A. Reinke, E. Christodoulou, et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. *CoRR*, abs/2206.01653, 2022.

[32] A. Myronenko. 3d MRI brain tumor segmentation using autoencoder regularization. In *MICCAI*, volume 11384 of *Lecture Notes in Computer Science*, pages 311–320. Springer, 2018.

[33] E. Nguyen, K. Goel, A. Gu, G. W. Downs, P. Shah, T. Dao, S. A. Baccus, and C. Ré. S4nd: Modeling images and videos as multidimensional signals using state spaces. *arXiv preprint arXiv:2210.06583*, 2022.

[34] O. Oktay, J. Schlemper, L. L. Folgoc, et al. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018.

[35] O. Ronneberger, P. Fischer, et al. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[36] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE TPAMI*, pages 640–651, 2017.

[37] A. L. Simpson, M. Antonelli, S. Bakas, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *CoRR*, abs/1902.09063, 2019.

[38] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *MICCAI*, volume 12901 of *Lecture Notes in Computer Science*, pages 36–46. Springer, 2021.

[39] H. Wang, Y. Zhu, B. Green, et al. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *ECCV*, volume 12349 of *Lecture Notes in Computer Science*, pages 108–126. Springer, 2020.

[40] W. Wang, C. Chen, et al. Transbts: Multimodal brain tumor segmentation using transformer. In *MICCAI*, pages 109–119, 2021.

[41] X. Xiao, S. Lian, et al. Weighted res-unet for high-quality retina vessel segmentation. In *ITME*, pages 327–331, 2018.

[42] X. Zhang, J. Zhang, L. Ma, et al. Progressive deep segmentation of coronary artery via hierarchical topology learning. In *MICCAI*, volume 13435 of *Lecture Notes in Computer Science*, pages 391–400. Springer, 2022.

[43] H. Zhou, J. Guo, Y. Zhang, et al. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Trans. Image Process.*, 32: 4036–4045, 2023.

[44] Q.-Y. Zhou, J. Park, and V. Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018.

[45] Z. Zhou, M. M. R. Siddiquee, et al. Unet++: A nested u-net architecture for medical image segmentation. In *MICCAI*, pages 3–11, 2018.