APAUNet: Axis Projection Attention UNet for Small Target in 3D Medical Segmentation

Yuncheng Jiang[1,2,3,⋆], Zixun Zhang[1,2,3,⋆], Shixi Qin[1,2,3], Yao Guo[4], Zhen Li[2,1,3,†], and Shuguang Cui[2,1,5]

[1] FNii, CUHK-Shenzhen, Guangdong, China
{yunchengjiang@link.,zixunzhang@link.,lizhen@}cuhk.edu.cn
[2] SSE, CUHK-Shenzhen, Guangdong, China
[3] SRIBD, CUHK-Shenzhen, Guangdong, China
[4] Shanghai Jiao Tong Univerisity, Shanghai, China
[5] Pengcheng Laboratory, Shenzhen,Guangdong, China

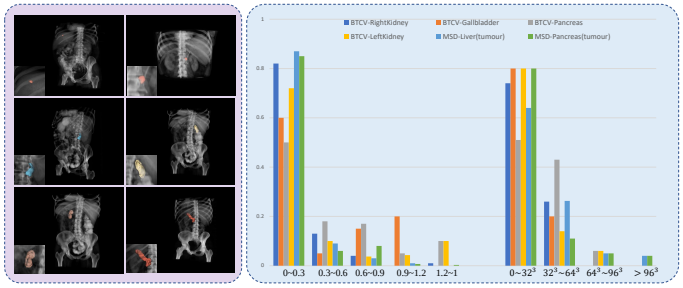
Abstract. In 3D medical image segmentation, small targets segmentation is crucial for diagnosis but still faces challenges. In this paper, we propose the **A**xis **P**rojection **A**ttention UNet, named **APAUNet**, for 3D medical image segmentation, especially for small targets. Considering the large proportion of the background in the 3D feature space, we introduce a projection strategy to project the 3D features into three orthogonal 2D planes to capture the contextual attention from different views. In this way, we can filter out the redundant feature information and mitigate the loss of critical information for small lesions in 3D scans. Then we utilize a dimension hybridization strategy to fuse the 3D features with attention from different axes and merge them by a weighted summation to adaptively learn the importance of different perspectives. Finally, in the APA Decoder, we concatenate both high and low resolution features in the 2D projection process, thereby obtaining more precise multi-scale information, which is vital for small lesion segmentation. Quantitative and qualitative experimental results on two public datasets (BTCV and MSD) demonstrate that our proposed APAUNet outperforms the other methods. Concretely, our APAUNet achieves an average dice score of 87.84 on BTCV, 84.48 on MSD-Liver and 69.13 on MSD-Pancreas, and significantly surpass the previous SOTA methods on small targets.

Keywords: 3D medical segmentation · Axis Projection Attention.

1 Introduction

Medical image segmentation, which aims to automatically and accurately diagnose lesion and organ regions in either 2D or 3D medical images, is one of the critical steps for developing image-guided diagnostic and surgical systems. In practice, compared to large targets, such as organs, small targets like tumors

⋆ Equal contributions. Code is available at github.com/zx33/APAUNet.
† Corresponding Author



(a) Target shape distribution (b) Target size distribution

Fig. 1. Target shape samples and size distribution of MSD and synapse multi-organ segmentation dataset. (a) 6 example organs from synapse multi-organ segmentation dataset. (b) The target size distribution. The x-axis is the target size interval, and the y-axis is the proportion (%) of corresponding samples to the whole dataset. The left part shows the relative proportion (%) of the target size to the whole input, while the right part shows the absolute size of the target with a interval step of 32 voxels. It can be observed that the relative target sizes of most samples in all the 6 categories are less than 0.6% with various shapes.

or polyps are more important for diagnosis, but also prone to be ignored. In this paper, we focus on 3D medical image segmentation (CT/MRI), with an emphasis on small lesions. This task is challenging mainly due to the following two aspects: 1) severe class imbalance of foreground (lesions) and background (entire 3D scans); 2) large variances in shape, location, and size of organs/lesions.

Recent progress in medical image segmentation has mainly been based on UNet [1], which applies a U-shaped structure with skip-connections to merge multi-scale features. However, due to the inductive bias of the locality of the convolutions, the U-shaped networks still suffer from the limited representation ability. Some studies utilized a coarse-to-fine segmentation framework [2]. These approaches refine the final segmentation in the fine stage, by shrinking input features to the region of interest (ROI) predicted in the coarse stage. Also, instead of using vanilla 3D convolutions, some works tried to explore a 2.5D fashions [3], which performed 2D convolutions on xy-axis at the low-level layers of the network and 3D convolutions on the high-level layers. Other works attempted to use an ensemble of 2D and 3D strategies, which fuses the 2D predictions from different views with 3D predictions to get better results [4] or refine the 2D predictions using 3D convolutions [5]. Besides, inspired by the great success of Transformers, some works explored the feasibility of applying self-attention into medical images by integrating CNN-based architectures with Transformer-like modules [6,7,8] to capture the patch-level contextual information.

Although previous methods have achieved remarkable progress, there are still some issues: 1) The 2.5D or the ensemble of 2D and 3D methods still suffer from the limited representation ability since the 2D phases only extract features from two axes while ignoring the information from the other axis, which worsens the

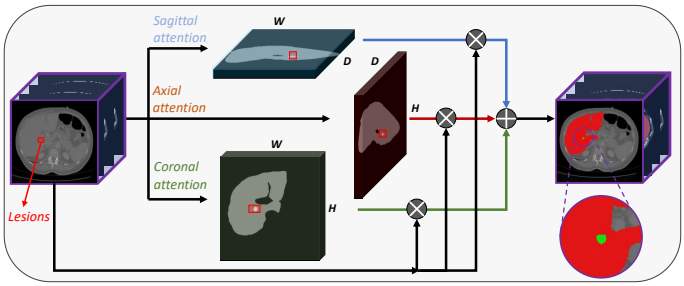


Fig. 2. In our APAUNet, we first project the 3D features into three orthogonal 2D planes to capture local contextual attentions from the three 2D perspectives, and then fuse them with the original 3D features. Finally, we adaptively fuse the features by weighted summation.

final segmentation prediction. Also, the two-stage designs are difficult for end-to-end training and require more computational resources. 2) Transformer-like models require higher computational cost on self-attention, and thus, have limited applications in 3D scenarios. Moreover, these models only learn the attentive interactions between patches, yet ignoring the local pattern inside the patch. 3) In addition, the imbalance between target and background has been ignored, which is vital for 3D medical segmentation. As shown in Fig. 1, on MSD chanllenge and BTCV datasets, the majority samples of the tumour target and small organ target are smaller than 0.6% to the whole 3D scans with various shapes.

In this paper, we propose an Axis Projection Attention (APA) UNet, named APAUNet, which utilizes an orthogonal projection strategy and a dimension hybridization strategy to overcome the aforementioned challenges. Specifically, our APAUNet follows the established design of 3D-UNet but replaces the main and functional component, the 3D convolution based encoder/decoder layers, with our APA encoder/decoder modules. In the *APA encoder*, the initial 3D feature maps are projected to three orthogonal 2D planes, i.e., *sagittal*, *axial*, and *coronal* views. Such a projection operation could mitigate the loss of critical information for small lesions in 3D scans. For instance, the original foreground-background area ratio of 3D features is $O(1/n^3)$ before the projection, but after projection, the ratio can be promoted to $O(1/n^2)$. Afterwards, we extract the local contextual 2D attention along the projected features to perform the asymmetric feature extraction and fuse them with the original 3D features. Eventually, the fused features of three axes are summed as the final output by three learnable factors, as shown in Fig. 2. Correspondingly, our *APA decoder* follows the same philosophy as the APA encoder but takes input features from two resolution levels. In this way, the decoder can effectively leverage the contextual information of multi-scale features. Furthermore, we also utilize an oversampling strategy to ensure the occurrence of foregrounds in each batch during the training process.

In summary, our contributions are in three-fold: (1) We propose the Axis Projection Attention UNet. APAUNet utilities the orthogonal projection strat-

egy to enhance the asymmetric projection attention and feature extraction. (2) We introduce a novel dimension hybridization strategy to fuse 2D and 3D attention maps for better contextual representation in both encoder and decoder blocks. Besides, we further leverage a multi-resolution fusion strategy into decoder blocks for context enhancement. (3) Extensive experiments on Synapse multi-organ segmentation (BTCV) [9] and Medical Segmentation Decathlon (MSD) challenge [10] datasets demonstrate the effectiveness and efficiency of our APAUNet, especially on small targets.

2 Related Work

2.1 CNN-based Medical Image Segmentation

CNNs, serving as the standard model of medical image segmentation, have been extensively studied in the past. The typical U-shaped network, U-Net [1], which consists of a symmetric encoder and decoder network with skip-connections, has become a common choice for medical image analysis. Afterwards, different variations of U-Net were proposed, such as Res-UNet [11], and Dense-UNet [12]. Besides, there are also some studies using AutoML to search for UNet architectures or an ensemble of 2D and 3D features, e.g., C2FNAS [13] uses a two stages NAS to search for the 3D architecture, and [4] utilizes meta learner to learn the ensemble of 2D and 3D features. Although these architectures have achieved remarkable progress in various 2D and 3D medical image segmentation tasks, they lack the capability to learn the global context and long-range spatial dependencies, even though followed by down-sampling operations. Thus, it leads to the degraded performance on the challenging task of small lesion segmentation.

2.2 Attention Mechanism for Medical Imaging

Attention mechanisms have been widely applied to segmentation networks, which can be categorized into two branches. The first branch is the hard attention, which typically uses a coarse-to-fine framework for segmentation tasks. [2] exploited two parallel FCNs to first detect the ROI of input features, then conducted the fine-grained segmentation over these cropped ROI patches for volumetric medical image segmentation. RA-UNet [14] introduced a residual attention module that adaptively combined multi-level features, which precisely extracted the liver region and then segmented tumours in this region. However, these hard attention methods usually need extensive trainable parameters and can be difficult to converge, which are not efficient for 3D medical segmentation tasks. The second branch is the adoption of the self-attention mechanism. One of the early attempts was the Attention U-Net [15], which utilized an attention gate to suppress irrelevant regions of the feature map while highlighting salient features. UTNet [6] adopted efficient self-attention encoder and decoder to alleviate the computational cost for 2D medical image segmentation. UNETR [7] further employed a pure transformer by introduceding a multi-head self-attention mechanism into the 3D-UNet structure, taking advantage of both Transformers and

CNNs to learn the sequential global features. Nonetheless, considering the class imbalance issue of small lesions and large variants of organs, the methods mentioned above are not effective enough. To this end, our work aims at developing an efficient approach, thoroughly taking advantage of the attention mechanism for specifically small lesion segmentation in 3D scans.

2.3 Small Target Segmentation

Segmentation of small objects, with limited available features and imbalanced sample distribution, is a common challenge in computer vision tasks, especially for medical images. Many previous studies have explored the solutions for natural images. For example, improved data augmentation methods [16,17] were proposed to increase the diversity of data and enhance the model generalization ability. In addition, advanced feature fusion techniques were adopted to better capture the small objects [18]. However, few works investigate this issue for medical images, and the obtained performance is usually less superior. For instance, [19] utilized a reinforcement learning (RL) based search method to construct the optimal augmentation policy for small lesions segmentation. UNet 3+ [20] exploited the full-scale skip-connections to extract small-scale features efficiently. C2FNAS [13] and DINTS [21] used Neural Architecture Search (NAS) to find an appropriate network for small features extraction. MAD-UNet [22] stacked multi-scale convolutions with attention modules to reduce the effects of intra-class inconsistency and enrich the contextual information. Nonetheless, these approaches are neither superior in performance nor computationally efficient, which motivates us to make an in-depth investigation of small object segmentation in 3D medical scans.

3 Methodology

Fig. 3 illustrates the overall architecture of our APAUNet. Following the idea of 3D-UNet, our APAUNet consists of several axis projection attention (APA) encoder/decoder blocks with five resolution steps. In this section, we first introduce the macro design of the APA encoder/decoder in Sec. 3.1, and then dig into the detailed block design in Sec. 3.2.

3.1 Axis Projection Attention (APA) Encoder and Decoder

The axis projection attention (APA) Encoder aims to extract contextual information of multiple resolution levels from different perspectives. The structure of APA Encoder is depicted in Fig. 3 (left part).

In practice, given a 3D medical image $I \in \mathbb{R}^{C \times H \times W \times D}$, the APA Encoder extracts multi-level features at different resolution scales $X^i \in \mathbb{R}^{C_i \times \frac{H}{2^i} \times \frac{W}{2^i} \times \frac{D}{2^i}}$. For the i-th level, the input feature X^i is fed to three *Internal Encoder (IE) blocks* in parallel to capture the contextual attention from three different perspectives. In order to capture the features of small targets more effectively, the original

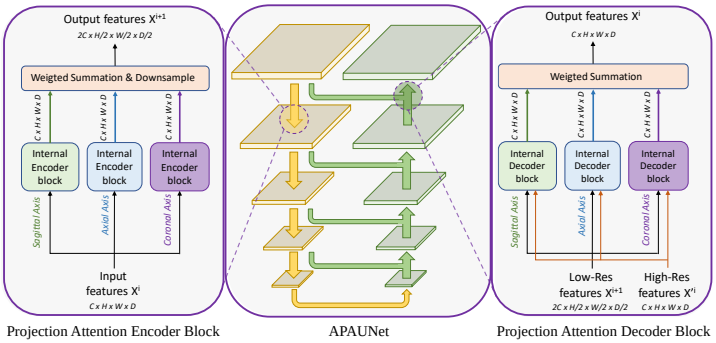


Fig. 3. The overview of APAUNet. Taking 3D scans as the input, our APAUNet exploits the 3D-UNet as the backbone. The yellow and green arrows denote the axis projection attention (APA) encoder and decoder blocks, respectively. For the left part, within one APA encoder block, we use three parallel internal encoder (IE) to construct the 2D attention on each projected orthogonal planes to filter the redundant background features and then adaptively fuse the features from the three perspectives. Correspondingly, for the right part, each APA decoder block merges the multi-scale features progressively to generate segmentation results.

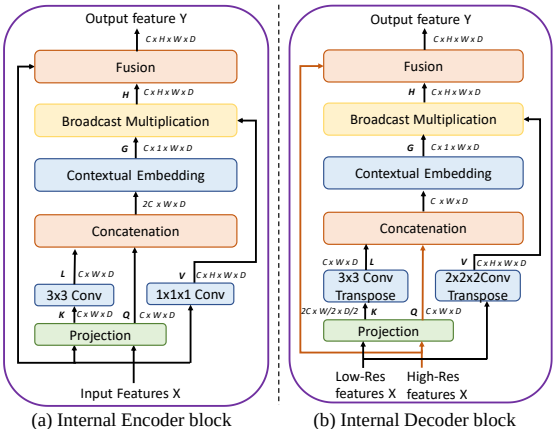
3D features will be projected to three orthogonal 2D planes to extract the 2D spatial attention. And then, the learned 2D attentions and 3D feature maps are aggregated to enhance the feature representation. After obtaining fused features of the three axes, we use three learnable parameters $\beta_{i,a}$ to obtain the weighted summation of the enhanced features:

$$Y^i = \sum_{a=1}^{3} \beta_{i,a} \cdot \mathrm{IE}(X^i, a), \sum_{a=1}^{3} \beta_{i,a} = 1 \tag{1}$$

where $\mathrm{IE}(\cdot)$ is the operation of IE block and a denotes the three orthogonal axes. This aggregation function can further help the network adaptively learn the importance of different projection directions to achieve asymmetric information extraction. Afterwards, another $1 \times 1 \times 1$ convolution and $2 \times 2 \times 2$ average pooling are applied to perform the down-sampling operation to get X^{i+1}, the input features of next level.

Similarly, the APA Decoder modules are used to extract and fuse multi-resolution features to generate segmentation results. The detailed design is shown in Fig. 3 (right part). The APA Decoder block has a similar structure to the APA Encoder block, but takes two features with different resolutions as the input.

To be more specific, features of both low and high resolutions are fed into the APA Decoder block simultaneously, where the low resolution feature maps $X^{i+1} \in \mathbb{R}^{2C \times \frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}}$ from the APA Decoder at the $(i+1)$-th level and the high resolution feature maps $X'^{i} \in \mathbb{R}^{C \times H \times W \times D}$ from the APA Encoder at the i-th level. Afterwards, the *Internal Decoder (ID) block* aggregates both high and low resolution features to generate 2D contextual attentions from the three



(a) Internal Encoder block (b) Internal Decoder block

Fig. 4. The detailed structure of Internal Encoder/Decoder (IE/ID) block used in the APA Encoder/decoder along the *sagittal* (H) view respectively. (a) The IE block. (b) The ID block. They share a similar structure design while the ID block takes two inputs from different resolution levels.

perspectives, then the 3D feature maps are fused with the 2D attentions to obtain 3D contextualized features, similar to that in the APA Encoder. To this end, the small-scale foreground information is better preserved, avoiding losing the crucial features. Finally, we take the weighted summation of three 3D contextualized features as the output features to next level.

3.2 Internal Encoder and Decoder Blocks

Inspired by the Contextual Transformer (CoT) [23], we adopt 3×3 convolutions in our internal blocks to mine the 2D context information and follow the design of the attention mechanism in CoT blocks. Meanwhile, we introduce several refinement strategies to tackle the challenges of 3D medical images, especially for small lesions. The detailed structure of internal blocks are illustrated in Fig. 4.

Orthogonal Projection Strategy. In the IE block, to better filter the irrelevant background and amplify the critical information of small lesions, we first project the input 3D features to three 2D planes. In particular, the 3D input feature $X \in \mathbb{R}^{H \times W \times D \times C}$ is projected to the *sagittal*, *axial* and *coronal* planes of the Cartesian Coordinate System to generate keys (K) and queries (Q), whereas the values (V) keep the 3D shape. Taking the *sagittal* view as an example, the input X is projected to 2D to get keys and queries: $K, Q \in \mathbb{R}^{C \times W \times D}$, while $V \in \mathbb{R}^{C \times H \times W \times D}$ is obtained by a single $1 \times 1 \times 1$ convolution. Here we adopt the summation of global average pooling (GAP) and global max pooling (GMP) along the desired axis (H in this case) as the projection operator:

$$K = Q = \mathrm{GAP}_H(X) + \mathrm{GMP}_H(X) \tag{2}$$

The projected K, Q are then used to learn different 2D contextual attentions along three orthogonal dimensions to better capture the key information of small lesions.

Dimension Hybridization Strategy. After the orthogonal projection, a 3×3 group convolution with a group size of 4 is employed on K to extract the local attention $L \in \mathbb{R}^{C \times W \times D}$, which contains the local spatial representation related to the neighboring keys. After that, we concatenate the local attention L with Q to further obtain the attention matrix $G \in \mathbb{R}^{C \times 1 \times W \times D}$ by two consecutive 1×1 2D convolutions and dimension expend:

$$L = h(\text{GConv}_{3\times3}(K)) \tag{3}$$

$$G = \text{Unsqueeze}(\text{Conv}_{1\times1}(h(\text{Conv}_{1\times1}([L, Q])))) \tag{4}$$

where GConv is the group convolution, $h(\cdot)$ denotes the normalization activation function, $[\cdot, \cdot]$ denotes the concatenation operation, and Unsqueeze$(\cdot)$ is the dimension expend operation. The attention matrix G encodes not only the contextual information within isolated query-key pairs but also the attention inside the keys.

Next, based on the 2D global attention G, we calculate the hybrid attention map $H \in \mathbb{R}^{H \times W \times D \times C} = V \odot G$, where $\odot$ denotes the broadcast multiplication, rather than the local matrix multiplication operation used in the original CoT blocks. This is because we empirically discovered that the local matrix multiplication operation does not contribute to any performance improvement under our 2D-3D hybridization strategy and consumes more GPU memory during training. Finally, the obtained hybrid attention maps are fused with the input feature X by a selective attention [24] to get the output features Y. The complete structure of the IE block is illustrated in Fig. 4(a).

Multi-Resolution Fusion Decoding. In the ID block, in order to better obtain the multi-scale contextual information from the multi-resolution features, we integrate the upsampling operation into the attention extraction process. The design of ID block is similar to the IE block introduced above but takes two inputs from different resolution levels, where the high resolution features $X' \in \mathbb{R}^{C \times H \times W \times D}$ from the encoder are taken to produce the queries and the low resolution features from the previous decoder $X \in \mathbb{R}^{2C \times \frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}}$ are taken to generate the keys and upsampled values. Then a $3 \times 3 \times 3$ transpose convolution is applied to upsample the keys to obtain the local attention L. Thus the decoder could fully capture the hybrid attention from various scales. The subsequent contextual extraction operations are similar to the IE block. And finally, we fuse the hybrid attention maps with the high resolution features X' to generate the output features Y. The detailed design of the ID block is illustrated in Fig. 4(b).

3.3 Loss Function

Following the previous methods [25,26], we jointly use the Dice loss and the Cross Entropy loss to optimize our network. Specifically,

$$\mathcal{L}_{total} = \mathcal{L}_{dice} + \mathcal{L}_{CE} \tag{5}$$

$$\mathcal{L}_{dice} = 1 - \frac{2}{C}\sum_{j=1}^{C}\frac{\sum_{i=1}^{N} X_{ij}Y_{ij}}{\sum_{i=1}^{N} X_{ij} + \sum_{i=1}^{N} Y_{ij}} \tag{6}$$

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}(Y_{ij}\log X_{ij}) \tag{7}$$

where N denotes the total number of voxels, C is the number of target classes, while X and Y are the softmax output of the prediction and the one-hot ground truth, respectively.

4 Experiments

4.1 Datasets and Implementation Details

Our APAUNet is evaluated on two 3D segmentation datasets: Synapse multi-organ segmentation dataset (BTCV) [9] and Medical Segmentation Decathlon (MSD) challenge [10]. The synapse multi-organ segmentation dataset includes 30 cases of CT scans. Following the settings of [27,28], we split 18 cases for model training while the rest 12 cases are used as test set. We report the 95% Hausdorff Distance (HD95) and Dice score (DSC) on 8 organs (aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas and stomach) following [27,28]. While for MSD challenge we choose task 3 (Liver Tumours CT segmentation) and task 7 (Pancreas Tumour CT segmentation), since they contains more small targets. The Liver dataset contains 131 3D CT volumes with ground truth labels of both small liver tumours and large liver regions. While for the Pancreas dataset, it contains 282 3D CT volumes with labels of medium pancreas and small pancreas tumour regions. Since the MSD online test server is temporarily closed for submissions recently, we use the 5-fold cross-validation to evaluate our APAUNet on MSD datasets. Besides, we also reproduce the the comparison methods on MSD with the same settings if the results on these datasets are not reported/evaluated in their original papers.

We implement our APAUNet on PyTorch and MONAI. For data preprocessing on MSD, we employ the method introduced in nnU-Net [29], which crops the zero values and re-samples all cases to the median voxel spacing of their respective dataset. Besides, the third-order spline interpolation is used for image data and the nearest-neighbor interpolation for the corresponding segmentation mask. During the training stage, we use a batch size of 2 and an input patch size of ($96 \times 96 \times 96$), where the patches are sampled from the original volumes. Specifically, to address the problem that the foreground targets are hard to be randomly cropped, we adopt an over-sampling strategy as introduced in [29] to enforce that more than half of the samples in a batch contain the foreground targets. All the patches are randomly flipped as data augmentation. We use the

Table 1. Segmentation results on the synapse multi-organ segmentation dataset. Note: Aor: Aotra, Gall: Gallbladder, LKid: Kidney (Left), RKid: Kidney (Right), Liv: Liver, Pan: Pancreas, Spl: Spleen, Sto: Stomach. All the results are obtained from the corresponding papers, except nnUNet.

Method	Average		Aor	Gall	LKid	RKid	Liv	Pan	Spl	Sto
	HD95	DSC								
TransUNet	32.62	77.49	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
CoTr	27.38	78.08	85.87	61.38	84.83	79.36	94.28	57.65	87.74	73.55
UNETR	23.87	79.57	89.99	60.56	85.66	84.80	94.46	59.25	87.81	73.99
MISSFormer	19.20	81.96	86.99	68.65	85.21	82.00	94.41	65.67	**91.92**	80.81
nnFormer	15.80	86.56	92.13	70.54	86.50	86.21	96.88	83.32	90.10	**86.83**
nnUNet	13.69	86.79	**93.20**	71.50	84.39	**88.36**	**97.31**	82.89	91.22	85.47
APAUNet	**11.26**	**87.84**	92.88	**75.26**	**88.47**	87.80	95.33	**85.47**	90.88	86.59

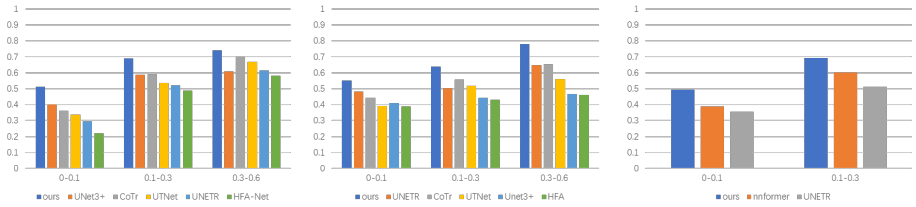
Table 2. Segmentation results (Dice score) on the MSD Liver and Pancreas Tumour datasets. The results of nnU-Net and C2FNAS are from the MSD leaderboard, while other methods are reproduced/implemented by ourselves on fold-1, including our APAUNet. (* denotes the average score with the standard deviations of 5-fold cross-validation.) Latency (s) is measured by a sliding window inference on GPU.

Method	Liver			Pancreas			Latency
	Organ	Cancer	Average	Organ	Cancer	Average	
UNet3+	87.32	70.10	78.71	75.30	46.00	60.65	39.65
HFA-Net	92.65	65.83	79.24	80.10	45.47	62.78	24.80
UTNet	92.51	69.79	81.15	80.30	49.68	64.99	18.16
UNETR	90.52	66.47	78.50	80.47	51.30	65.88	13.46
CoTr	90.20	69.88	80.04	77.9	50.20	64.05	21.77
nnU-Net	95.24	**73.71**	**84.48**	79.53	52.27	65.90	-
C2FNAS	94.91	71.63	83.27	80.59	52.87	66.73	-
APAUNet	**96.10**	72.50	84.30	**83.05**	**55.21**	**69.13**	**13.36**
APAUNet*	95.56 ± 0.35	71.99 ± 0.81	83.78 ± 0.35	82.29 ± 1.73	54.93 ± 1.64	68.61 ± 1.63	-

SGD with an initial learning rate of 0.1, a momentum of 0.9, and a weight decay of 4e-5 with cosine annealing. The models are trained for 500 epochs in total with a 50-epochs warm-up. All the models are trained on an NVIDIA V100 GPU.

4.2 Comparisons with Other Methods

We first evaluate our APAUNet against several state-of-the-art methods, including TransUNet [28], CoTr [8], UNETR [7], MISSFormer [27], and nnFormer [30]. The results are shown in Tab. 1. Our APAUNet achieves the best average dice score of 87.84 and the best average HD95 score of 11.26, and outperforms the second best method nnFormer by 4.54 and 1.28 on these two metrics respectively. For the class-wise results, our APAUNet also achieves the best dice scores in five of the categories and surpasses the second best method by $0.75 \sim 4.72$, while is only 1.55, 1.04 and 0.24 lower on Liver, Spleen and Stomach.



(a) MSD-Liver (b) MSD-Pancreas (c) BTCV-Gallbladder

Fig. 5. Statistical results of small cases on MSD and BTCV datasets. Note that since most targets are concentrated in the range of 0-0.3%, in order to show the results more clearly, we divide 0-0.3% into two intervals of 0-0.1% and 0.1-0.3%. And also, the split validation set of BTCV does not contain any targets with size of 0.3-0.6%.

While on MSD datasets, we compare our APAUNet with HFA-Net [31], UTNet [6], UNETR [7], UNet3+ [20], CoTr [8], C2FNAS [13] and nnUNet [29]. Since the MSD online leaderboard submission is temporarily closed at the time of the experiments, we focus on comparing our algorithm with the first 5 methods which are reproduced under the same settings, and also compare the 5-fold results with C2FNAS and nnUNet for reference. Tab. 2 shows the overall results of these methods. On Liver Tumour dataset, our APAUNet achieves a dice score of 96.10 on organ segmentation and 72.50 on cancer segmentation, which outperforms the second best method by 3.45 (HFA-Net) and 2.40 (UNet3+) on the dice score respectively. Also, the average dice score of our APAUNet is 84.30, which is 3.15 higher than UTNet. Similarly, the results of our APAUNet consistently outperform those of other methods on Pancreas dataset. Comparing with C2FNAS and nnUNet, our APAUNet surpasses both methods on all scores except Liver-cancer and Liver-average, where nnUNet uses additional test time augmentation, model ensembling and post-processing for better results. The results demonstrate the effectiveness of our APAUNet. More detailed results of 5-fold cross validation are shown in appendix.

4.3 Results on Small Cases

To verify the effectiveness of our APAUNet on small cases, we count the results of small targets with size less than 0.6% on Liver, Pancreas and BTCV. Fig. 5 shows the performance of our APAUNet and comparison methods. We observe that our APAUNet surpasses the second best results by $3.9 \sim 12.6$ in absolute dice scores. Concretely, on the interval of $0 \sim 0.1\%$, our APAUNet achieves dice scores of 51.3, 55.1 and 49.4 on three datasets respectively, and has an improvement of 27.9%, 14.3% and 27.1% compared to the second best methods. Similarly on the interval of $0.1 \sim 0.3\%$, the dice scores of our APAUNet are 16.6%, 14.9% and 15.2% higher than the second best methods respectively. These exciting results further highlights the superiority of our APAUNet, especially on small target segmentation.

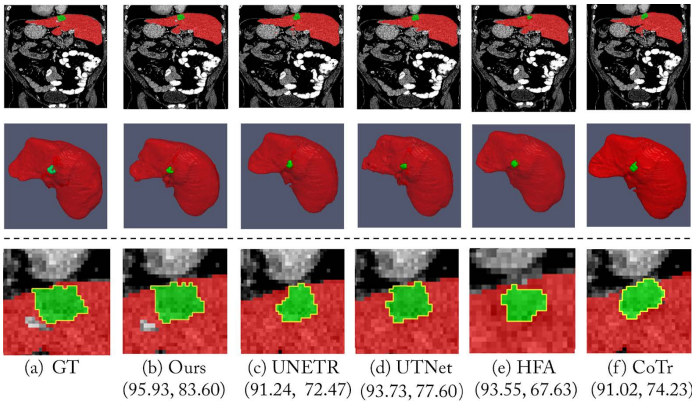


Fig. 6. Visualization of our proposed APAUNet and other methods on Liver. The upper two rows are the overall of segmentation results and the view of foreground targets, while the last row is an enlarged view of the segmentation results of small lesions. The dice scores $(organ, tumour)$ are given under each method.

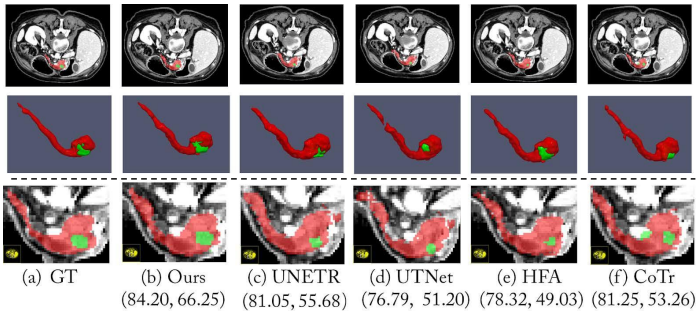


Fig. 7. The visualization results of our APAUNet and comparison methods on Pancreas.

4.4 Visualization Results

In this part, we visualize some segmentation results of our APAUNet and compare with UNETR, UTNet, CoTr and HFA-Net on the Liver Tumour dataset in Fig. 6 and on the Pancreas Tumour dataset in Fig. 7. It can be observed that all the methods have good performance on large organ segmentation with a dice score of over 90 on liver and over 75 on pancreas. While on small lesions, UNETR, UTNet, CoTr and HFA-Net could only locate the position and are still unsatisfactory in the segmentation of both the boundary and the size. In comparison, our APAUNet gets clearer and more accurate boundaries and sizes, further verifying the substantial advantage of APAUNet in medical image segmentation, especially for small targets.

Table 3. Ablation experiments on the 2D-3D hybridization strategy, projection operators and the weighted importance. Avg, Max and Conv refer to the avgpooling, maxpooling and convolution operations, respectively. It can be seen that the effectiveness of the design of our APAUNet.

Method	Liver			Pancreas		
	Organ	Cancer	Avg	Organ	Cancer	Avg
CoT-2D	82.66	59.08	70.87	74.30	43.28	58.79
CoT-3D	91.08	63.26	77.17	75.40	46.28	60.84
Conv	85.20	68.81	77.00	80.24	51.33	65.78
Max	87.34	69.95	78.65	80.47	53.28	66.87
Avg	95.89	71.68	83.79	82.39	**55.48**	68.94
Mean	95.63	71.55	83.59	81.67	54.30	67.99
APAUNet	**96.10**	**72.50**	**84.30**	**83.05**	**55.21**	**69.13**

4.5 Analysis and Ablation Studies

In this section, a variety of ablation studies are performed to thoroughly verify the design our APAUNet and validate the performance under different settings. More ablation results can be found in appendix.

2D-3D Hybridization Strategy To verify the effectiveness of our 2D-3D hybridization strategy, we conduct several ablation experiments. First, we replace the 2D operators in the original CoT block with 3D operators, which is a pure 3D version of CoTNet (CoT-3D). For the pure 2D version (CoT-2D), we project all the K, Q and V to 2D planes. The results are shown in Tab. 3. Compared with CoT-3D, our APAUNet achieves an improvement of 7.13 and 8.29 on average dice scores. Different from natural images, 3D medical CT scans contain more excessive background information that hinders the learning of contextual attention, while our 2D-3D hybridization strategy can filter the redundant information to enhance the performance. Also, APAUNet outperforms CoT-2D by a large margin, which indicates that it is suboptimal for the model to utilize only 2D attentions for 3D segmentation tasks.

Projection Operations For the choice of projection operators, we conduct several experiments to figure out the best choice in {avgpooling, maxpooling, convolution}. For the fair comparison of the channel-wise operation of pooling, here we use a depth-wise convolution to perform the projection operation. Experiment results are shown in Tab. 3. Unexpectedly, the results of using convolution are the worst, despite its learnable parameters that make it more flexible than the pooling operators. We conjecture that the convolution lacks the ability to capture important information and is easily disturbed by the noise due to the high proportion of background information in medical images, even though with large kernel sizes. Fortunately, pooling operators could directly process the global information. Specifically, avgpooling involves all parts of the image to get the

global statistics and maxpooling focuses only on the most salient part. By combining the complementarity of avgpooling and maxpooling, better performance can be achieved. Hence, we adopt the summation of avgpooling and maxpooling as the projection operation in our APA Encoder and Decoder.

Weighted Importance of Projection Recall that there are three learnable weights during the fusion of features generated from different projection directions. Here, we conduct an experiment to study the effect of these weights. The results are shown in Tab. 3. Compared to the average mean, our APAUNet could gain an extra improvement of $0.47 \sim 1.38$ on both datasets with the learnable weighted importance, which shows the feasibility of our asymmetric feature extraction on different projection axes. At the same time, this interesting phenomenon shows that the amount of information contained in different perspectives in the 3D structure is different, which motivates us to discover better 2D-3D fusion strategies. More detailed experimental results can be found in the appendix.

5 Conclusion

In this paper, we propose a powerful network for the 3D medical segmentation task, called Axis Projection Attention Network (APAUNet). In order to deal with the highly imbalanced targets and background, we leverage an orthogonal projection strategy and dimension hybridization strategy to build our APAUNet. Extensive experiments on the BTCV and MSD datasets demonstrate the promising results on 3D medical segmentation tasks of our APAUNet, and the superior results on small targets show the effectiveness of our APA Encoder/Decoder blocks. In our experiments, we also show that the importance of different perspectives is different, which motivates us to look deeper into the projection strategy and dimension hybridization pattern. For the future work, we would like to enhance the performance of small target segmentation by exploring more advanced projection techniques and designing more efficient dimension fusion strategies.

Acknowledgements This work was supported in part by the National Key R&D Program of China with grant No.2018YFB1800800, by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen HK S&T Cooperation Zone, by NSFC-Youth 61902335, by Shenzhen Outstanding Talents Training Fund, by Guangdong Research Project No.2017ZT07X152 and No.2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No.2022B1212010001), by zelixir biotechnology company Fund, by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen, by Tencent Open Fund, and by ITSO at CUHKSZ.

References

1. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. (2015)
2. Zhu, Z., Xia, Y., Shen, W., Fishman, E., Yuille, A.: A 3d coarse-to-fine framework for volumetric medical image segmentation. In: 2018 International conference on 3D vision (3DV). (2018)
3. Wang, G., Shapey, J., Li, W., Dorent, R., Dimitriadis, A., Bisdas, S., Paddick, I., Bradford, R., Zhang, S., Ourselin, S., et al.: Automatic segmentation of vestibular schwannoma from t2-weighted mri by deep spatial attention with hardness-weighted loss. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2019) 264–272
4. Zheng, H., Zhang, Y., Yang, L., Liang, P., Zhao, Z., Wang, C., Chen, D.Z.: A new ensemble learning framework for 3d biomedical image segmentation. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. (2019) 5909–5916
5. Xia, Y., Xie, L., Liu, F., Zhu, Z., Fishman, E.K., Yuille, A.L.: Bridging the gap between 2d and 3d organ segmentation with volumetric fusion net. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2018) 445–453
6. Gao, Y., Zhou, M., Metaxas, D.N.: Utnet: a hybrid transformer architecture for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. (2021)
7. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. (2022)
8. Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging CNN and transformer for 3d medical image segmentation. CoRR **abs/2103.03024** (2021)
9. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. Volume 5. (2015) 12
10. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
11. Xiao, X., Lian, S., Luo, Z., Li, S.: Weighted res-unet for high-quality retina vessel segmentation. In: 2018 9th International Conference on Information Technology in Medicine and Education (ITME). (2018)
12. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging (2018)
13. Yu, Q., Yang, D., Roth, H., Bai, Y., Zhang, Y., Yuille, A.L., Xu, D.: C2fnas: Coarse-to-fine neural architecture search for 3d medical image segmentation. In: Proc. of CVPR. (2020)
14. Jin, Q., Meng, Z., Sun, C., Cui, H., Su, R.: Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans. Frontiers in Bioengineering and Biotechnology (2020)

15. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
16. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations. (2018)
17. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proc. of ICCV. (2019)
18. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence (2017)
19. Xu, J., Li, M., Zhu, Z.: Automatic data augmentation for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. (2020)
20. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., et al.: Unet 3+: A full-scale connected unet for medical image segmentation. In: Proc. of ICASSP. (2020)
21. He, Y., Yang, D., Roth, H., Zhao, C., Xu, D.: Dints: Differentiable neural network topology search for 3d medical image segmentation. In: Proc. of CVPR. (2021)
22. Li, W., Qin, S., Li, F., Wang, L.: Mad-unet: A deep u-shaped network combined with an attention mechanism for pancreas segmentation in ct images. Medical Physics (2021)
23. Li, Y., Yao, T., Pan, Y., Mei, T.: Contextual transformer networks for visual recognition. arXiv preprint arXiv:2107.12292 (2021)
24. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 510–519
25. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. (2016)
26. Xie, E., Wang, W., Wang, W., Sun, P., Xu, H., Liang, D., Luo, P.: Segmenting transparent object in the wild with transformer (2021)
27. Huang, X., Deng, Z., Li, D., Yuan, X.: Missformer: An effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162 (2021)
28. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. CoRR **abs/2102.04306** (2021)
29. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018)
30. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)
31. Zheng, H., Yang, L., Han, J., Zhang, Y., Liang, P., Zhao, Z., Wang, C., Chen, D.Z.: Hfa-net: 3d cardiovascular image segmentation with asymmetrical pooling and content-aware fusion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. (2019)