

RESEARCH ARTICLE

Position-Aware Anti-Aliasing Filters for 3D Medical Image Analysis

STANLEY T. YU¹ AND HONG-YU ZHOU², (Member, IEEE)¹Stanford Online High School, Redwood City, CA 94063, USA²Department of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong

Corresponding author: Hong-Yu Zhou (whuzhouhongyu@gmail.com)

ABSTRACT Maximum pooling, average pooling, and strided convolution are three widely adopted down-sampling approaches in deep learning based 3D medical image analysis. However, these methods have their own pros and cons. Maximum pooling and strided convolution are advantageous in capturing the discriminative features but often lead to the aliasing problem. In comparison, average pooling anti-aliases the representations but produces less discriminative representations. To address such shortcoming, anti-aliased maximum pooling (MaxBlurPool) uses low-pass filters to mitigate the aliasing effect. However, these filters are designed to be fixed, making it difficult to adapt to various spatial positions. In this paper, we propose position-aware anti-aliasing filters (PASS) to learn spatially adaptive low-pass filters. Compared to maximum pooling, PASS integrates a one-layer local attention module, whose computational cost is minimal. Thus, PASS can be incorporated into existing network architecture with minor efforts. In comparison to previous anti-aliased counterparts, PASS brings consistent and clear performance gains on brain tumor segmentation, pulmonary nodule detection, and cerebral hemorrhage detection. Besides, PASS also greatly improves the model robustness under adversarial attack.

INDEX TERMS Down-sampling, anti-aliasing, attention mechanism, medical imaging.


I. INTRODUCTION

Down-sampling has been a fundamental component of digital signal processing. Based on this knowledge, modern deep convolutional neural networks (DCNNs) employ multiple down-sampling layers to perform rate reduction on images, where the spatial resolution and number of channels of feature maps is gradually reduced and increased, respectively. With this compression process, we can obtain semantically rich representations from the high-level layers of DCNNs, which are often more generalizable than low-level features. In 3D medical image analysis with CNNs, maximum pooling and average pooling are two widely adopted down-sampling methods. Specifically, maximum pooling aims to summarize the most activated presence of features by calculating the maximum value within a fixed small region. In comparison, average pooling aggregates the average values of different feature patches. As another option, convolution can also be used to conduct down-sampling, where we increase the stride

of convolution. The resulting down-sampling convolution operation is often named as strided convolution.

As aforementioned, maximum and average pooling adopt two different ways to implement down-sampling, and they too face different problems accordingly. Maximum pooling captures the most predominant parts, which makes the produced features (cf. Fig. 1b) discriminative. Nonetheless, maximum pooling layers in DCNNs inevitably result in aliasing results because of the preserved high-frequency signals. This characteristic makes high-level semantic representations sensitive to small shifts [1]. In comparison, average pooling is anti-aliased, and thus helps preserve the shift invariance in DCNNs. However, the outputs of average pooling are often less discriminative compared to maximum pooling, as shown in Fig. 1c. As a result, the performance of average pooling is often inferior to that of maximum pooling in a range of tasks [2]. For strided convolution, Zhang [1] pointed out that it also suffers from the same issue as maximum pooling does.

On the other hand, applying low-pass filtering is the default solution to anti-alias in traditional signal processing. Inspired by such phenomena, anti-alias maximum pooling

The associate editor coordinating the review of this manuscript and approving it for publication was Ramakrishnan Srinivasan .

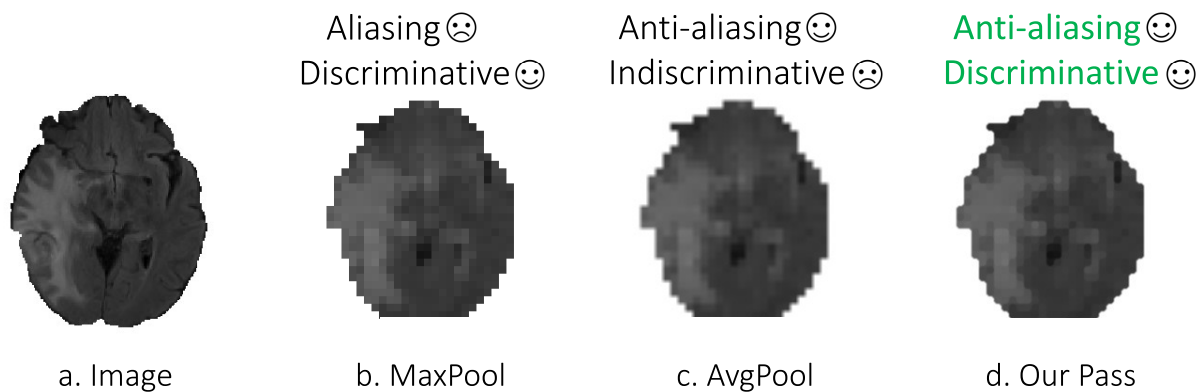


FIGURE 1. Processed results using maximum pooling (MaxPool), average pooling (AvgPool), and our methodology (PASS).

(i.e., MaxBlurPool) [1] incorporated gaussian blur into archetypical maximum pooling, leading to enhanced shift invariance and improved task performance. However, the blur filters in MaxBlurPool are designed to be fixed across different spatial positions, resulting in the inflexibility to handle various contents. To mitigate this problem, we introduce position-aware anti-aliasing filters (PASS) to learn adaptive blur filters based on local contents. Compared to MaxBlurPool, PASS introduces a plug-and-play local attention module ahead of low-pass filters to adaptively change filter values based on local input features. The proposed attention module comprises only one convolutional layer and its computational cost is minimal. From Fig. 1d, we see that our PASS integrates the advantages of maximum and average pooling, producing anti-aliased and discriminative representations.

We validate the effectiveness of PASS on a range of medical imaging tasks, which include brain tumor segmentation (BraTS [3] dataset), pulmonary nodule detection (LUNA [4]), and cerebral hemorrhage detection (this is an in-house dataset). In experiments, we show that the proposed PASS dramatically boosts the performance over both typical pooling approaches (maximum pooling, average pooling, and strided convolution) and previous anti-aliased methods by obvious margins. Here are some noteworthy results. Compared to maximum pooling, our PASS boosts the overall performance by 3.1 and 3.8 percents on brain tumor segmentation and cerebral hemorrhage detection, respectively. Moreover, PASS consistently outperforms previous anti-aliasing counterparts by over 1 percent on all three medical imaging tasks, demonstrating the generalization ability of PASS.

To summarize, our paper has the following core contributions:

- 1) We introduce a new down-sampling component, named PASS, to adaptively anti-alias medical image representations by taking into account the local contents in 3D volumes with a local attention module.
- 2) PASS is computationally efficient, including only one convolution layer. In practice, PASS can be easily

integrated into existing 3D medical imaging models as a plug-and-play component.

- 3) We perform extensive experiments on various medical imaging tasks to validate the effectiveness of PASS. The experimental results show that PASS can outperform typical pooling schemes and advanced anti-aliasing methodologies by observable and consistent margins.

II. RELATED WORK

It is common practice to apply low-pass filtering before down-sampling to avoid aliasing in digital signal processing. Inspired by this operation, the initial convolutional neural network [5] proposed to use average pooling for down-sampling. However, Scherer *et al.* [2] implemented different sub-sampling methods on a variety of tasks and drew a conclusion: maximum pooling operation significantly outperforms other sub-sampling operations. Consequently, modern DCNNs [6], [7], [8], [9] mostly adopted maximum pooling as the default pooling methodology in the network architecture for performance boosts while ignoring the impact of the aliasing emerged in maximum pooling. On the other hand, anti-aliasing has been an important direction in medical image analysis. Nonetheless, almost all related work [10], [11], [12], [13] focused on anti-aliasing the input images instead of latent representations, prevent their incorporation into modern DCNNs.

There are a number of papers investigating how to mitigate the aliasing effect produced by maximum pooling. Malinowski *et al.* [14] introduced a smoothness regularization term that in conjuncture with learnable pooling regions to alleviate the aliasing problem, which improved the performance on object and event recognition tasks. Hénaff and Simoncelli [15] proposed L_2 pooling based on the Nyquist theorem to avoid aliasing artifacts in learned representations. Azulay and Weiss [16] pointed out that aliasing adversely affects the invariance characteristic of DCNNs and presented two ways to mitigate this problem: (i) anti-aliasing the intermediate representations and (ii) increasing

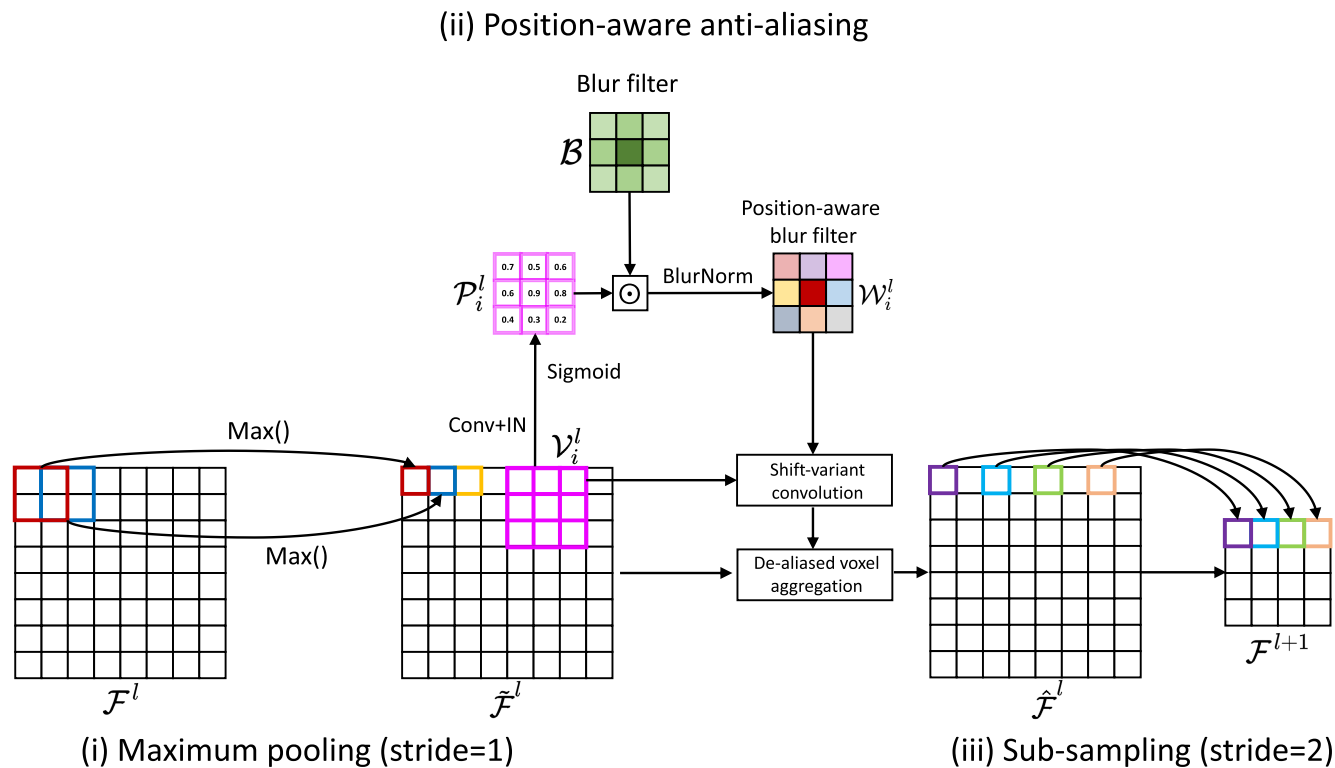


FIGURE 2. Methodology overview. PASS decomposes the down-sampling process into three steps: (i) maximum pooling (stride=1), (ii) position-aware anti-aliasing, (iii) sub-sampling (stride=2). Note that the above two figures are just 2D illustrations, whereas we conduct 3D operations in practice.

data augmentation. Lee *et al.* [17] utilized a gate function to control the mix of maximum and average pooling. The proposed gated pooling improved both the task performance and prediction stability. Zhang [1] pointed out that maximum pooling deteriorates the shift invariance in modern DCNNs and proposed to integrate typical maximum pooling with gaussian blur filters, i.e., MaxBlurPool. The experimental results showed that MaxBlurPool brings observable performance gains on natural image classification while resulting in significant improvements of model robustness against shift perturbations. Meanwhile, Singh *et al.* [18] investigated the impact of MaxBlurPool in lung tuberculosis detect, showing MaxBlurPool is beneficial under different detection architecture. Vyas and Liao [19] utilized deep learning based image segmentation to anti-alias seismic data. However, above approaches failed to address the spatially adaptive problem in anti-aliasing approaches, which may lead to inflexible anti-aliased image representations because of fixed blur filters. Instead, we introduce position-aware anti-aliasing filters to address the locality problem in anti-aliasing.

III. METHODOLOGY

Fig. 2 presents an overview of proposed PASS. In practice, PASS decomposes the down-sampling process into three separate steps. First, we perform maximum pooling with stride=1 to preserve the most discriminative representations from the feature map. Then, we apply position-aware anti-aliasing to

the result of maximum pooling. Specifically, we first compute a local attention map based on the local features. Next, we calculate Hadamard product of the local attention matrix and a pre-defined gaussian blur filter, whose result is passed to the blur normalization to ensure it is a low-pass filter. We then convolve the feature map produced from maximum pooling (stride=1) with the obtained position-aware blur filter. Finally, we sub-sample the convolved result to acquire the down-sampled feature map. In the following, we will describe each step in details.

Step (i): Non-strided maximum pooling (stride=1). Most maximum pooling operations employ a stride of 2, which can be decomposed into two procedures: maximum pooling with stride=1 (i.e., non-strided maximum pooling) and sub-sampling with stride=2. Suppose $\mathcal{F}^l \in \mathbb{R}^{H \times W \times D \times C}$ denotes the input features to the l -th layer. Step (i) can be summarized as follows:

$$\tilde{\mathcal{F}}^l = \text{MaxPool}_{2,1}(\mathcal{F}^l), \tag{1}$$

where subscripts $\{2, 1\}$ denotes the kernel size and stride of maximum pooling, respectively. The goal of non-strided maximum pooling is to preserve the discriminative features in feature maps.

Step (ii): Position-aware anti-aliasing. As aforementioned, maximum pooling inevitably produces aliasing effects because of the maximization operation. Although the obtained features are discriminative, they also adversely

affect the shift invariance characteristic of DCNNs. In practice, the learned high-level representations with rich semantics may vary a lot with a small shift in the input [1], which severely deteriorates the performance and robustness.

We mitigate the above problem by applying position-aware anti-aliasing to the result of non-strided maximum pooling. Suppose the feature map $\tilde{\mathcal{F}}^l$ contains N overlapping local volumes, where $N = H \times W \times D \times C$ denotes the number of positions in $\tilde{\mathcal{F}}^l$, where H, W, D are the three dimensions of the 3D feature map, and C is the number of channels in the feature map. We use $\mathcal{V}_i^l \in \mathbb{R}^{K \times K \times K}$ to denote the local volume centered at the i -th position, $i \in \{0, 1, \dots, N - 1\}$. Then, we apply a series of operations, including convolution, instance normalization [20], and the sigmoid function to \mathcal{V}_i^l , which are expressed as follows:

$$\mathcal{P}_i^l = \text{Sigmoid-IN-Conv}_{3,1}(\mathcal{V}_i^l), \quad (2)$$

where subscripts $\{3, 1\}$ refer to the kernel size and stride of convolution, respectively. $\mathcal{P}_i^l \in \mathbb{R}^{K \times K \times K}$ denotes the position-wise weight matrix for the fixed blur filter $\mathcal{B} \in \mathbb{R}^{K \times K \times K}$. The blur filter \mathcal{B} is initialized as the multivariate gaussian distribution, which can be formalized as follows:

$$\mathcal{B}[j, k, m] = e^{-\frac{j^2+k^2+m^2}{2\sigma^3}}, \quad (3)$$

where j, k, m are indices whose range is $[-\lfloor \frac{K}{2} \rfloor, \lfloor \frac{K}{2} \rfloor]$. σ is set to 0.9.

Next, we calculate the Hadamard product of the local weight matrix \mathcal{P}_i^l and blur filter \mathcal{B} :

$$\mathcal{Q}_i^l = \mathcal{P}_i^l \odot \mathcal{B}, \quad (4)$$

where \odot stands for the Hadamard product operator, and $\mathcal{Q}_i^l \in \mathbb{R}^{K \times K \times K}$. To ensure \mathcal{Q}_i^l is a low-pass filter, we apply blur normalization to \mathcal{Q}_i^l :

$$\mathcal{W}_i^l = \text{BlurNorm}(\mathcal{Q}_i^l). \quad (5)$$

In practice, we found the softmax function works well for blur normalization. However, other normalization methods may have similar effects.

Then, we apply shift-variant convolution to feature map $\tilde{\mathcal{F}}^l$ by first computing the following dot product of the normalized position-aware blur kernel \mathcal{W}_i^l and the local 3D volume \mathcal{V}_i^l to obtain an anti-aliased value v_i^l at every position,

$$v_i^l = \mathcal{W}_i^l \cdot \mathcal{V}_i^l, \quad (6)$$

and then aggregating the N anti-aliased values, i.e. $\{v_0^l, \dots, v_{N-1}^l\}$, into the anti-aliased feature map $\hat{\mathcal{F}}^l \in \mathbb{R}^{H \times W \times D \times C}$.

Step (iii): Sub-sampling (stride=2). Finally, we apply a sub-sampling operation with stride=2 to reduce the spatial dimension of $\hat{\mathcal{F}}^l$, resulting in the output feature map $\mathcal{F}^{l+1} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times C}$:

$$\mathcal{F}^{l+1} = \text{Sub-sampling}_2(\hat{\mathcal{F}}^l). \quad (7)$$

\mathcal{F}^{l+1} serves as the input to the next layer in DCNNs.

Training loss function. For brain tumor segmentation, the training loss function is a weighted summation of the cross-entropy loss and the dice loss. Specifically, the cross-entropy loss is formulated as:

$$\mathcal{L}_{ce} = -\frac{1}{U} \sum_{u=1}^U \sum_{z=1}^Z y_u^z \log p_u^z, \quad (8)$$

where Z stands for the number of categories, U denotes the number of voxels in the predicted segmentation mask, y_u^z denotes the ground-truth binary label of category z at the u -th voxel, and p_u^z is the corresponding predicted probability of category z .

On the basis of the above notations, the dice loss is formulated as follows:

$$\mathcal{L}_{dice} = 1 - \frac{2 \sum_{u=1}^U \sum_{z=1}^Z y_u^z p_u^z}{\sum_{u=1}^U \sum_{z=1}^Z y_u^z + \sum_{u=1}^U \sum_{z=1}^Z p_u^z}. \quad (9)$$

The training loss function for segmentation is a weighted combination of the cross-entropy loss and the dice loss. The cross-entropy loss measures the pixel-level classification accuracy while the dice loss alleviates the data imbalance problem:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{ce} + \beta \mathcal{L}_{dice}. \quad (10)$$

In practice, we set α and β to 0.2 and 1.0, respectively. For pulmonary nodule detection and cerebral hemorrhage detection, we only use the cross-entropy loss.

IV. EXPERIMENTS

A. DATASETS

We evaluate PASS on three medical imaging tasks, which are brain tumor segmentation, pulmonary nodule detection, and cerebral hemorrhage detection. For brain tumor segmentation, we made experiments on the well-established BraTS-2018 [3] dataset, which comprises 351 magnetic resonance imaging (MRI) scans of the human brain. There are three classes in BraTS-2018: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). As for pulmonary nodule detection, we used the LUNA-16 [4] dataset, which consists of 888 annotated thoracic computed tomography (CT) scans. In LUNA-16, radiologists made a total of 5,855 annotations, where only nodules $\geq 3\text{mm}$ were categorized as relevant lesions. Each nodule annotation was checked by at least one radiologist. The in-house dataset for cerebral hemorrhage detection comprises 1,486 brain CT volumes, which are used to analyse the cause of cerebral hemorrhage. The evaluation metrics are dice score, AUC and accuracy on BraTS-2018, LUNA-16 and the in-house cerebral hemorrhage dataset, respectively. We repeat each experiment 5 times and report their average results.

B. BASELINES

We compare PASS against five baselines: maximum pooling (MaxPool), average pooling (AvgPool), strided convolution

(StridedConv), gated pooling (GTPool) [17], and MaxBlurPool [1]. GTPool is built on top of maximum and average pooling, where a gate operation is proposed to control the mix of maximum and average pooling. MaxBlurPool inserts a fixed blur filter between non-strided maximum pooling and sub-sampling. Note that the initial implementations of GTPool and MaxBlurPool are 2D-based. In our experiments, we implement 3D versions of them.

C. IMPLEMENTATION DETAILS

We implement PASS using PyTorch [21]. For fairness, we carefully tune hyper-parameters for each dataset, where baselines and our PASS share the same training protocol. We save the checkpoint which produces the lowest loss value and use it for testing.

1) MODEL DESIGN

For the image segmentation task, 3D U-Net [22] and Attention U-Net [23] are respectively used as the segmentation network, where we replace the maximum pooling layer with different down-sampling layers (i.e., down-sampling baselines and our PASS) to investigate their impacts. Likewise, we replace the maximum pooling layer in 3D ResNet-18 [24], [25] with different down-sampling layers for the classification task.

2) BRAIN TUMOR SEGMENTATION (BraTS-2018)

We build segmentation models using 3D U-Net [22] and Attention U-Net [23]. Attention U-Net is included to investigate the compatibility between PASS and existing attention modules in DCNNs. We use Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) as the optimizer, and set the weight decay to $1e-5$. The initial learning rate is $1e-3$. The input image size is $160 \times 160 \times 128$. We train each model for 300 epochs, where the learning rate is divided by 10 every 100 epochs. The data augmentation strategies include random crop, random rotation, and random flip. The batch size is 4. As mentioned earlier, the loss function is a weighted combination of the cross-entropy loss and the dice loss. We randomly split the dataset into training (70%), validation (10%) and test sets (20%). Instance normalization is used as the default normalization method.

3) PULMONARY NODULE DETECTION (LUNA-16)

3D ResNet-18 [24], [25] is used as the backbone network in this task. During the training stage, we apply random crop around nodules to produce positive samples. The ratio between the numbers of positive and negative samples is 1:2. We resize all samples to $48 \times 48 \times 48$. The loss function is the cross entropy loss. Adam ($\beta_1=0.9$, $\beta_2=0.999$) is used as the optimizer, and the initial learning rate is set to $1e-4$. The weight decay of Adam is $1e-5$. We do not run a fixed number of training epochs on LUNA-16. Specifically, we do not stop the training process until the validation loss does not decrease for up to 30 epochs. The resulting total number of training epochs is 210. The learning rate is divided

by 10 every 40 epochs. Random crop, random rotation, and random flip are employed as data augmentation strategies. The batch size is 128. We randomly split the dataset into training (70%), validation (10%) and test sets (20%). Batch normalization is used as the default normalization method.

4) CEREBRAL HEMORRHAGE DETECTION

As in the task of pulmonary nodule detection, we also adopt 3D ResNet-18 [24], [25] as the backbone. The input image size is $256 \times 256 \times 30$. The loss function is the cross entropy loss. The default optimizer is SGD, where we set the momentum to 0.9 and the weight decay to $1e-4$. The initial learning rate is $1e-2$. We train each model for 200 epochs, and the learning rate is divided by 10 every 30 epochs. The data augmentation strategies include random crop, random rotation, and random flip. The batch size is 64. We randomly split the dataset into training (70%), validation (10%) and test sets (20%). Batch normalization is again used as the default normalization method.

D. COMPARISONS WITH THE STATE OF THE ART

1) BRAIN TUMOR SEGMENTATION

Experimental results using different down-sampling methodologies are presented in Table 1. Comparing MaxPool with AvgPool, we see that maximum pooling is more advantageous in segmenting the whole tumor (WT) while AvgPool performs better on the tumor core (TC) and enhancing tumor (ET). Considering WT is much larger than TC and ET, the above comparisons demonstrate the aliasing adversely affects the segmentation of small objects. By comparing StridedConv with MaxPool, we find that these two down-sampling approaches display similar performance, both of which outperform AvgPool on the segmentation of WT while showing slightly worse results on TC and ET. This phenomenon verifies the conclusion provided by [1], which is StridedConv and MaxPool have similar characteristics.

For state-of-the-art approaches, GTPool surpasses MaxPool, AvgPool, and StridedConv on all three tumor categories by obvious margins. This is consistent with the superiority of the gate function used in GTPool, which incorporates the advantages of MaxPool and AvgPool by mixing their outputs. Similar to GTPool, MaxBlurPool also integrates a low-pass filter with maximum pooling for anti-aliasing while maintaining the discriminative features. Compared to GTPool, MaxBlurPool achieves consistent improvements on three tumor classes. When we replace MaxBlurPool with our PASS, we observe very obvious improvements, especially in TC and ET which are smaller and thus harder to segment compared to WT. These improvements reflect that the anti-aliased yet discriminative representations may aid the discovery of small objects. Besides, we find PASS is complementary to Attention U-Net even if Attention U-Net employs multiple attention modules in the network. The reason behind might be that the attention modules of Attention U-Net mainly lie

TABLE 1. Brain tumor segmentation on BraTS-2018. WT, TC and ET stand for whole tumor, tumor core and enhancing tumor, respectively. \uparrow means higher is better. Best results are bolded. P-value is calculated between the mean dice scores of our PASS and MaxBlurPool.

Backbone	Down-sampling	Dice (\uparrow)			
		Mean	WT	TC	ET
3D U-Net	MaxPool	82.9	87.5	82.6	78.5
	AvgPool	82.7	86.7	82.8	78.6
	StridedConv	82.8	87.6	82.5	78.4
	GTPool	83.9	88.6	83.5	79.6
	MaxBlurPool	84.4	89.1	84.2	79.9
	Our PASS	86.0	89.9	86.5	81.7
	P-value	6.25e-3			
Attention U-Net	MaxPool	84.1	88.8	83.6	80.0
	AvgPool	84.2	88.3	83.9	80.4
	StridedConv	84.2	89.0	83.5	80.2
	GTPool	84.8	89.9	84.4	80.2
	MaxBlurPool	85.6	90.1	85.3	81.4
	Our PASS	86.9	90.8	86.9	82.9
	P-value	7.32e-3			

TABLE 2. Pulmonary nodule detection on LUNA-16. \uparrow means higher is better. The best result is bolded. P-value is calculated between our PASS and MaxBlurPool.

Down-sampling	MaxPool	AvgPool	StridedConv	GTPool	MaxBlurPool	Our PASS	P-value
AUC (\uparrow)	97.6	97.2	97.4	98.0	98.3	99.3	9.34e-3

in the decoder branch while our PASS layers are all in the encoder.

2) PULMONARY NODULE DETECTION

From Table 2, we see that MaxPool performs better than AvgPool on LUNA-16. This is because the lung nodules from LUNA are mostly larger than 3 mm, making them easier to recognize even though there exist severe aliasing in high-level semantics. Interestingly, we find that GTPool provides a 0.4-percent improvement over MaxPool, showing that introducing anti-aliasing is still beneficial even to large objects. Comparing MaxBlurPool to GTPool, we observe that these two down-sampling methods achieve comparable performance. Our PASS provides the consistent and obvious performance improvements over all baselines. Specifically, PASS surpasses MaxBlurPool by 1 percent. Considering baselines all display quite high performance ($>97\%$), we believe about 1-percent improvement is already convincing enough to validate the effectiveness of proposed PASS on LUNA-16. The above comparisons show that introducing position-aware anti-aliasing to down-sampling is beneficial and necessary for pulmonary nodule detection.

3) CEREBRAL HEMORRHAGE DETECTION

Compared to pulmonary nodule detection, detecting cerebral hemorrhage is harder as the hemorrhagic spot is much smaller and thus more difficult to find. As shown in Table 3, AvgPool performs slightly better than MaxPool. This is consistent with the segmentation performance of ET on BraTS-2018,

indicating that aliasing may adversely affect the discovery of small objects as the learned high-level semantics is highly influenced by aliasing. Again, StridedConv displays similar performance as MaxPool does. By integrating the advantages of maximum and average pooling, GTPool brings a 0.5-percent improvement over AvgPool. MaxBlurPool obviously surpasses GTPool by incorporating an explicit low-pass filter into maximum pooling. Once again our PASS surpasses MaxBlurPool by 2 percents, again verifying the advantage of learning position-aware anti-aliased representations.

4) STATISTICAL SIGNIFICANCE

A t-test validation is conducted on all three datasets. We compute p -values between the best and the second best results. Specifically, on BraTS-2018, we calculate two p -values based on mean dice scores of 3D U-Net and Attention U-Net. The p -values on brain tumor segmentation, pulmonary nodule detection, and in-house cerebral hemorrhage detection are 6.25e-3 (3D U-Net)/7.32e-3 (Attention U-Net), 9.34e-3, and 3.74e-3, respectively. All p -values are smaller than 0.01, indicating that performance improvements brought by our PASS are statistically significant at the 1% significance level.

E. ABLATION STUDIES

In this section, we conduct ablation studies to investigate the impacts of different modules in PASS. All ablative experiments were performed on brain tumor segmentation (BraTS-2018).

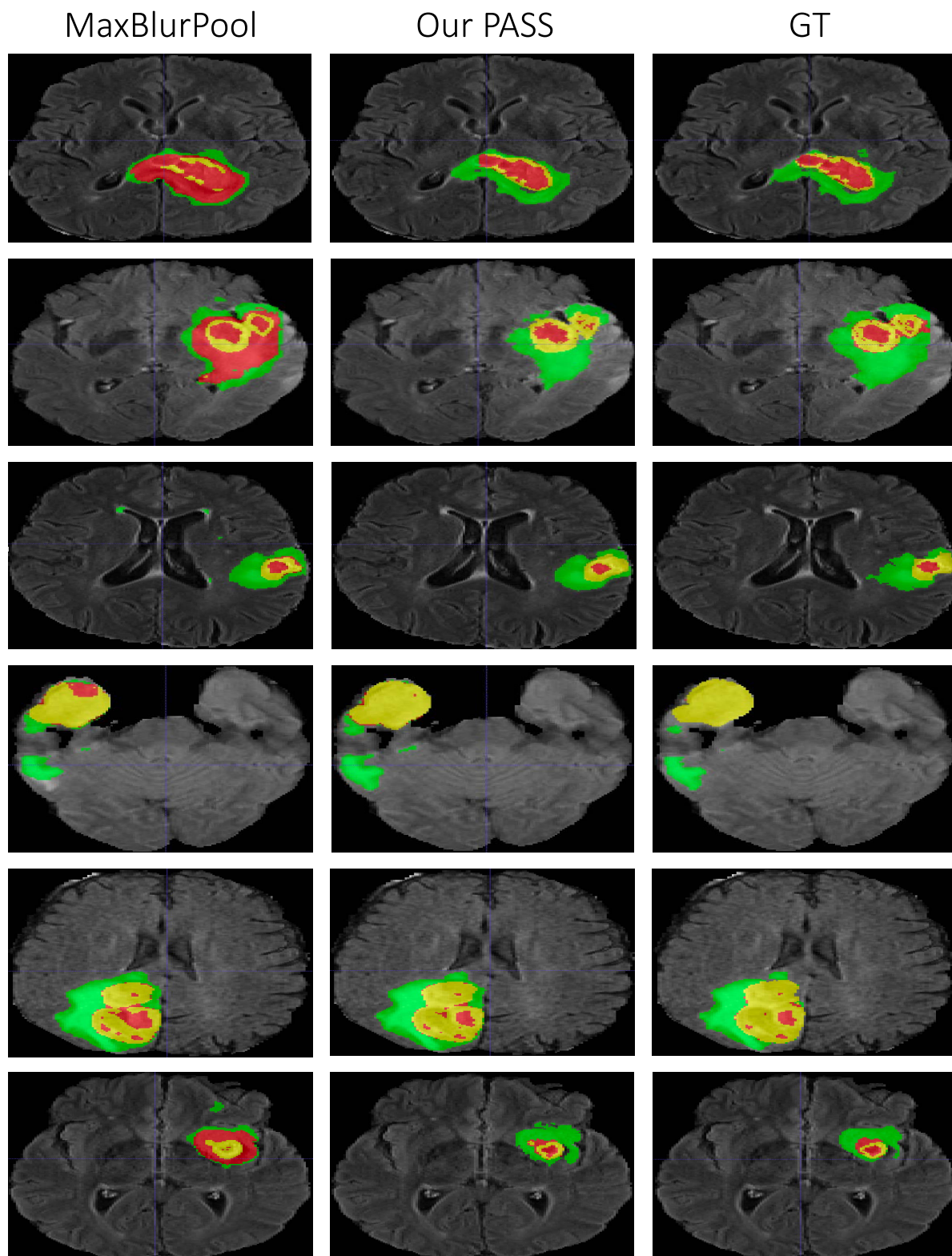


FIGURE 3. Visualization of tumor segmentation using 3D U-Net. The whole tumor (WT) includes a union of green, yellow and red labels, the tumor core (TC) is a union of red and yellow, and the enhancing tumor (ET) is shown in yellow.

Table 4 presents the experimental results. Compared to maximum pooling (row 0), adding the blur filter \mathcal{B} brings

a 0.5-percent improvement. Next, we investigate the influence of the convolution and instance normalization layers

TABLE 3. In-house cerebral hemorrhage detection. \uparrow means higher is better. The best results is bolded. P-value is calculated between our PASS and MaxBlurPool.

Down-sampling	MaxPool	AvgPool	StridedConv	GTPool	MaxBlurPool	Our PASS	P-value
Acc. (\uparrow)	87.6	88.0	87.8	88.5	89.4	91.4	3.74e-3

TABLE 4. Ablation study on brain tumor segmentation. The baseline is 3D U-Net with maximum pooling (row 0). The evaluation metric is mean dice score. Conv and IN represent the convolution layer and instance normalization layer, respectively. \mathcal{B} stands for the blur filter and K denotes the blur kernel size.

	Conv+IN	Sigmoid	\mathcal{B} ($K=3$)	\mathcal{B} ($K=5$)	BlurNorm	Mean Dice
0						82.9
1			✓			83.4
2	✓		✓			84.0
3	✓	✓	✓			85.2
4	✓	✓	✓		✓	86.0
5	✓	✓		✓	✓	85.7

TABLE 5. Task performance under adversarial attack.

	MaxPool	AvgPool	StridedConv	MaxBlurPool	our PASS
Pulmonary nodule detection (AUC)	31.7	35.0	29.8	41.5	47.3
Cerebral hemorrhage detection (Acc.)	25.2	29.1	25.4	35.7	40.6

(cf. Eq. 2). It is obvious that adding Conv+IN helps slightly boost the segmentation performance by 0.6 percents (cf row 2 in Table 4). Somewhat surprisingly, from row 3 in Table 4, we observe over 1-percent performance gain after adding the sigmoid to Conv+IN, implying the regularization effect of the sigmoid function to local attentions. Comparing row 3 with row 1, we observe about 1.8-percent improvement over the fixed gaussian blur filter. This improvement is brought by the proposed local attention module (without blur normalization), which verifies the necessity of introducing position-aware anti-aliasing. After adding BlurNorm to local attention, we further observe 0.8-percent improvement, which clarifies the effectiveness of blur normalization to ensure the low-pass property of the learned blur filter. Besides, we also studies the impact of enlarging the blur kernel size. In row 5, we find that changing the kernel size from 3 to 5 leads to a slight performance drop by 0.3 percents.

F. DISCUSSION

In this section, we first visually analyze the segmentation results of the brain tumor. Then, we add adversarial perturbations to the input and investigate the strength of PASS in enhancing the model robustness.

1) RESISTANCE TO ADVERSARIAL ATTACK

A black-box attacker [26] is used to evaluate the resistance to adversarial samples of different down-sampling methods. From Table 5, we see that our PASS is much more resistant to adversarial perturbations than different pooling methods and previous anti-aliasing approaches. For instance, PASS surpasses MaxBlurPool by 5.8 and 4.9 percents on

pulmonary nodule detection and cerebral hemorrhage detection, respectively. These comparisons further validate the anti-aliasing characteristic of PASS, which helps improve the model robustness. Additionally, we observe that average pooling performs better than maximum pooling and strided convolution, again indicating that anti-aliased features do help models to resist adversarial perturbations.

2) VISUAL ANALYSIS

We follow [27] to visualize the segmentation results in Fig. 3. We can see that our PASS greatly reduces small-sized false-positive predictions. For instance, MaxBlurPool produces lots of isolated noisy predictions because it cannot adaptively anti-alias different contents. In comparison, our PASS can greatly reduce false-positive segmentations. In addition, we see that PASS performs the best on the segmentation of the tumor core, which is consistent with the result reported in Table 1.

V. CONCLUSION

We propose Position-aware Anti-aliasing Filters (PASS) to adaptively anti-alias high-level representations with rich semantics. PASS introduces a position-aware local attention module to typical maximum pooling. PASS comprises only one convolutional layer, making it computationally efficient to replace existing down-sampling methods. Compared to typical pooling strategies and previous anti-aliasing counterparts, our PASS produces observable and consistent improvements on a variety of medical imaging tasks, including brain tumor segmentation, pulmonary nodule detection, and cerebral hemorrhage detection.

REFERENCES

- [1] R. Zhang, "Making convolutional networks shift-invariant again," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7324–7334.
- [2] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2010, pp. 92–101.
- [3] L. Weninger, O. Rippel, S. Koppers, and D. Merhof, "Segmentation of brain tumors and patient survival prediction: Methods for the brats 2018 challenge," in *Proc. Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, 2018, pp. 3–12.
- [4] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. Van Den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, and R. van der Gugten, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," *Med. Image Anal.*, vol. 42, pp. 1–13, Dec. 2017.
- [5] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, 1989, pp. 1–9.
- [6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.
- [10] C. Zhao, A. Carass, B. E. Dewey, J. Woo, J. Oh, P. A. Calabresi, D. S. Reich, P. Sati, D. L. Pham, and J. L. Prince, "A deep learning based anti-aliasing self super-resolution algorithm for MRI," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 100–108.
- [11] O. N. Romanyuk, S. V. Pavlov, O. V. Melnyk, S. O. Romanyuk, A. Smolarz, and M. Bazarova, "Method of anti-aliasing with the use of the new pixel model," in *Proc. SPIE*, vol. 9816, pp. 274–278, Dec. 2015.
- [12] C. Zhao, M. Shao, A. Carass, H. Li, B. E. Dewey, L. M. Ellingsen, J. Woo, M. A. Guttman, A. M. Blitz, M. Stone, and P. A. Calabresi, "Applications of a deep learning method for anti-aliasing and super-resolution in MRI," *Magn. Reson. Imag.*, vol. 64, pp. 132–141, Dec. 2019.
- [13] C. Zhao, B. E. Dewey, D. L. Pham, P. A. Calabresi, D. S. Reich, and J. L. Prince, "SMORE: A self-supervised anti-aliasing and super-resolution algorithm for MRI using deep learning," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 805–817, Mar. 2021.
- [14] M. Malinowski and M. Fritz, "Learning smooth pooling regions for visual recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2013, pp. 1–11.
- [15] O. J. Hénaff and E. P. Simoncelli, "Geodesics of learned representations," 2015, *arXiv:1511.06394*.
- [16] A. Azulay and Y. Weiss, "Why do deep convolutional networks generalize so poorly to small image transformations?" *J. Mach. Learn. Res.*, vol. 20, no. 184, pp. 1–25, 2019.
- [17] C.-Y. Lee, P. W. Gallagher, and Z. Tu, "Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2016, pp. 464–472.
- [18] J. Singh, A. Tripathy, P. Garg, and A. Kumar, "Lung tuberculosis detection using anti-aliased convolutional networks," *Proc. Comput. Sci.*, vol. 173, pp. 281–290, Jan. 2020.
- [19] M. Vyas and Q. Liao, "De-aliasing using the U-Net image segmentation algorithm," in *Proc. SEG Int. Exposit. Annu. Meeting*, 2020, pp. 1476–1480.
- [20] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.
- [21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.
- [22] O. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 424–432.
- [23] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.
- [24] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3154–3160.
- [25] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.
- [26] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2484–2493.
- [27] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *Proc. Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, 2018, pp. 311–320.

STANLEY T. YU has been with the Stanford Online High School, since 2017. He was at The University of Hong Kong Academy for the Talented. His research interests include machine learning based medical image analysis, image segmentation, lesion detection, and benign-malignant classification.

HONG-YU ZHOU (Member, IEEE) received the B.S. degree from Wuhan University, China, in 2015, and the M.S. degree from the Department of Computer Science and Technology, Nanjing University, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Computer Science, The University of Hong Kong. His research interest includes medical image analysis, where the main focus is on how to learn well-generalized representations for medical images.

...