Volumetric Attention for 3D Medical Image Segmentation and Detection

Xudong Wang[1,2*], Shizhong Han[1], Yunqiang Chen[1],
Dashan Gao[1], and Nuno Vasconcelos[2]

[1] 12 Sigma Technologies, San Diego, USA
[2] Dept. of Electrical and Computer Engineering, Univ. of California, San Diego, USA
{xuw080,nuno}@ucsd.edu {Shan,yunqiang,dgao}@12sigma.ai

Abstract A volumetric attention(VA) module for 3D medical image segmentation and detection is proposed. VA attention is inspired by recent advances in video processing, enables 2.5D networks to leverage context information along the z direction, and allows the use of pretrained 2D detection models when training data is limited, as is often the case for medical applications. Its integration in the Mask R-CNN is shown to enable state-of-the-art performance on the Liver Tumor Segmentation (LiTS) Challenge, outperforming the previous challenge winner by 3.9 points and achieving top performance on the LiTS leader board at the time of paper submission. Detection experiments on the DeepLesion dataset also show that the addition of VA to existing object detectors enables a 69.1 sensitivity at 0.5 false positive per image, outperforming the best published results by 6.6 points.

Keywords: Volumetric Attention · 3D Images · LiTS · DeepLesion.

1 Introduction

A natural solution to 3D medical image segmentation and detection problems is to rely on 3D convolutional networks, such as the 3D U-Net of [5] or the extended 2D U-Net of [15]. However, current GPU memory limitations prevent the processing of 3D volumes with high resolution. This is problematic, because the use of low-resolution volumes leads to low precision or miss-detection of small lesions and tumors and blur in lesion mask predictions, especially on boundaries. Hence, there is a need to trade-off the spatial resolution of each 2D slice for the number of slices processed. This implies a trade-off between the precision with which segmentation or detection can be performed and the amount of contextual information, in the z direction, that can be leveraged. A popular solution is to a use a 2D network to segment or detect the structures of interest in 2D or 2.5D slices and then concatenate the results to build a 3D segmentation mask or bounding box.

Christ et al. proposed a 2D U-Net for liver and tumor segmentation, followed by a conditional random field for segmentation refinement [4]. Li et al. proposed a hybrid Dense 2D/3D UNet of three-stages [13]. They found that a pre-trained 2D model

*This work was fully conducted during the internship in 12 Sigma Technologies, USA.

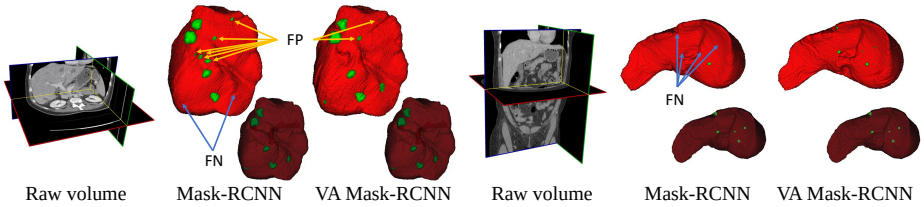

Figure 1: Comparison of 3D segmentations by the Mask-RCNN and the proposed VA Mask-RCNN on the LiTS `val` set. Red denotes segmented liver, green segmented lesions. 3D ground truth is shown on the bottom right, with liver in dark red and lesions in dark green. Left: while the Mask-RCNN misses two lesions (false nagative, FN) and has six false positive (FP) instances, the VA Mask-RCNN detects all lesions with only two FPs. Right: VA Mask RCNN detects 5 very small lesions, 4 of which are missed by the Mask-RCNN. These examples illustrate how the VA module *both* enhances small lesion prediction and enables the network to avoid false positives. (best viewed in color)

can significantly boost performance of their network. Han proposed a 2.5D (adjacent slices) residual U-Net for liver lesion segmentation [9]. These approaches are limited by the lack of contextual information. Since even human experts need to inspect multiple slices to reach confident assessments of confusing lesions, this is likely to upper bound their performance. To address this problem, Yan et al. [18] proposed a 3D context enhanced region-based CNN. However, their method is based on a region proposal network (RPN) and cannot be implemented as a single-stage detector, such as SSD and YOLO, or a segmentation network, such as U-Net, without an RPN component. Furthermore, because only the feature map derived from a central image is processed by the RPN to generate proposals, the proposal generation process has no access to 3D context. Given that missed proposals can not be recovered, this places an upper bound on detection performance.

In this work, we propose to address these limitations with ideas inspired by recent video processing work, where a similar problem is posed by the need to trade off the modeling of long-range dependencies between video frames and the spatial resolution of each frame. The proposed approach is inspired by [17], which added a non-local network to a 3D convolutional network (C3D/I3D) for video classification, using a space-time dependency/attention mechanism. We generalize this method into a flexible and computationally efficient Volumetric Attention (VA) module, which sequentially infers 3D enhanced attention maps along two separate dimensions, channel and spatial. The attention maps produced by this module are multiplied by the input feature map to enable adaptive feature refinement, using a 2D network. Similar to [12], global spatial pooling and global channel pooling are used to reduce computational cost.

The VA module has several interesting properties. First, it enables the processing of high spatial resolution images, while leveraging contextual information over multiple slices of the 3D CT volume. Second, it can be combined with any CNN architecture, including one-stage and two-stage detectors and segmentation networks. Third, it is computationally efficient, due to extensive use of spatial and channel pooling. Fourth, because the VA module can operate on image sub-regions, it can also benefit RPN networks. Fifth, since VA can be used with 2D networks, it can leverage pre-trained 2D

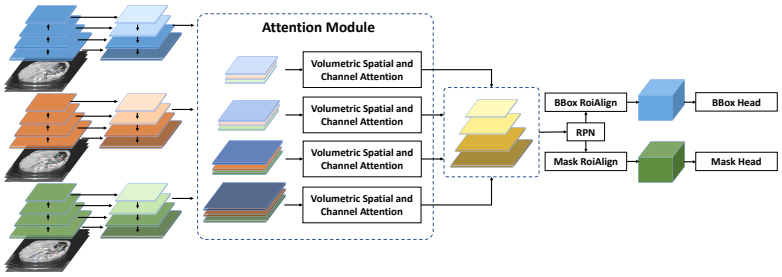

Figure 2: Architecture of the Volumetric Attention(VA) Mask-RCNN. Three continuous 2.5D images, each composed of 3 adjacent slices, are shown as example.

CNN weights for transfer learning. The proposed VA attention module is implemented within the Mask-RCNN, leading to an architecture denoted the VA Mask-RCNN. As illustrated in Fig.1, this not only reduces segmentation false positives, but also enables the retrieval of very small lesions that are missed by the Mask-RCNN model. The VA Mask-RCNN is shown to obtain state-of-the-art performance, 74.1 dice per case, on the LiTS liver tumor segmentation challenge `test` set, significantly outperforming (3.9 points) the winner of last year's challenge. It is the top method on the challenge leaderboard at the time of submission of this paper. To assess the generalization ability of the VA Mask-RCNN to 3D CT volumes, we have also performed experiments on the DeepLesion dataset. The VA Faster-RCNN achieved a sensitivity of 69.1 at 0.5 false positives(FPs)/image, outperforming the best published results by 6.6 points.

2 Volumetric Attention

The overall architecture of the VA Mask R-CNN is shown in Fig.2. The VA attention module operates on the Mask R-CNN feature pyramids extracted from a *target* 2.5D image, where detection takes place, and neighboring *contextual* 2.5D images. The 2.5D images are each composed of 3 adjacent slices. The attention module has three components: bag of long-range features, volumetric channel attention, and volumetric spatial attention. Unlike the self-attentive feature map of [17], VA uses long-range features from neighboring slices, which are combined with the feature map of the target slice to generate spatial and channel attention responses. A detailed scheme of the attention module is given in Fig. 3. We next discuss the three components combined with Mask-RCNN in detail.

2.1 Bag of Long-range Features

To account for dependencies along the z direction of the 3D CT volume, the VA Mask R-CNN complements the target 2.5D image, with neighboring images, both above and below the target image. These are denoted as contextual images. The features extracted from these images are concatenated for each level of the spatial pyramid, according to

$$\mathbf{X}^i_{long} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N] \in \mathbb{R}^{N \times C^i \times H^i \times W^i}, \quad (1)$$

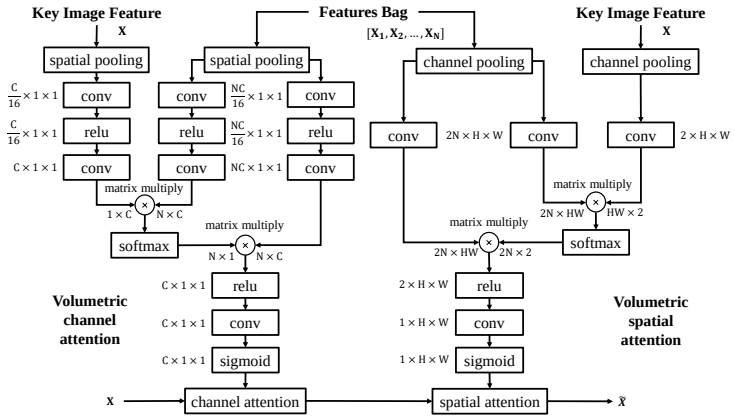

Figure 3: Volumetric Spatial and Channel Attention Module. N is the bag size, C, H, W the feature map channel size, height and width, respectively. Spatial and channel pooling are used to reduce computation.

where i is the pyramid level, $C^i \times H^i \times W^i$ its dimensions (chanel, height, and width, respectively), $\mathbf{X}^i_{long}$ the corresponding bag of long-range features, and N the number of contextual images. The features $\mathbf{X}_k$ are sorted by the order of the corresponding images along the z direction of the 3D volume.

2.2 Volumetric Channel Attention

This attention mechanism is inspired by that of [12,17]. The bag of features $\mathbf{X}_{long} \in \mathbb{R}^{N\times C\times H\times W}$ and corresponding target image feature map $\mathbf{X}_{tgt} \in \mathbb{R}^{C\times H\times W}$ are each subject to a global average pooling operator $\mathbf{F}^c_{avg}$. Following [12], computation is reduced by replacing the linear embedding layer of the original non-local blocks of [17] by two 1×1 convolutional layers with reduction ratio of 16. This is implemented as $\mathbf{F}^c_{emb}(\mathbf{X}) = W_2\delta(W_1\mathbf{F}^c_{avg}(\mathbf{X}))$, where $W_1 \in \mathbb{R}^{\frac{C}{16}\times C}$, $W_2 \in \mathbb{R}^{C\times\frac{C}{16}}$ and δ is the RELU function. The slices attention signal is finally computed with a softmax

$$\mathbf{S}^c_{att} = \text{softmax}(\mathbf{F}^c_{emb}(\mathbf{X}_{tgt}) \cdot \mathbf{F}^c_{emb}(\mathbf{X}_{long})) \in \mathbb{R}^{1\times N} \tag{2}$$

along dimension N, where $\mathbf{F}^c_{emb}(\mathbf{X}_{tgt}) \in \mathbb{R}^{1\times C}$, $\mathbf{F}^c_{emb}(\mathbf{X}_{long}) \in \mathbb{R}^{C\times N}$ and $\cdot$ refers to matrix multiplication. The slice attention signal $\mathbf{S}^c_{att}$ is then applied to $\mathbf{F}^c_{emb}(\mathbf{X}_{long}) \in \mathbb{R}^{N\times C}$ according to $\mathbf{S}^c_{att} \cdot \mathbf{F}^c_{emb}(\mathbf{X}_{long})$ and this is followed by a relu layer, a 1×1 conv layer and a sigmoid layer, to learn a nonlinear interaction $\mathbf{S}_c \in \mathbb{R}^{C\times 1\times 1}$ between channels. Then channel-wise multiplication is applied on $\mathbf{X}_{tgt} \in \mathbb{R}^{C\times H\times W}$.

2.3 Volumetric Spatial Attention

The volumetric spatial attention module uses max and average pooling to shrink feature maps along the channel dimension, concatenating them into two channel feature maps

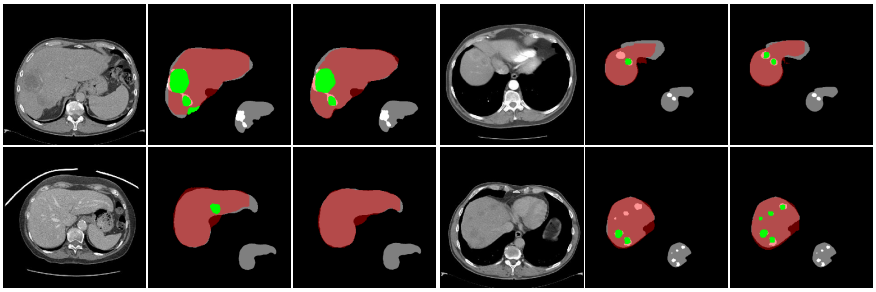

Figure 4: 2D visualization of segmentations by Mask-RCNN and VA Mask R-CNN on LiTS `val` set. Segmented liver is shown in red and lesions in green. Zoomed out ground truth masks are shown on bottom right, with liver in gray and lesions in white. The VA Mask-RCNN produces smoother segmentation boundaries and lower FP and miss rates. In the top left, the gallbladder area is easily confused with the lesion area. VA Mask-RCNN leverages contextual slices to remove this FP. (best viewed in color and zoom in for details)

$\mathbf{F}^s_{pool}(\mathbf{X}) = [\mathbf{F}^s_{max}(\mathbf{X}), \mathbf{F}^s_{avg}(\mathbf{X})] \in \mathbb{R}^{2\times H\times W}$. An embedding function is then implemented as $\mathbf{F}^s_{emb}(\mathbf{X}) = W\mathbf{F}^s_{pool}(\mathbf{X})$, where W is a learned convolutional weight layer. The slice attention signal is finally computed with a softmax

$$\mathbf{S}^s_{att} = \text{softmax}(\mathbf{F}^s_{emb}(\mathbf{X}_{tgt}) \cdot \mathbf{F}^s_{emb}(\mathbf{X}_{long})) \in \mathbb{R}^{1\times N} \tag{3}$$

along dimension N. A spatial attention map $S_s \in \mathbb{R}^{1\times H\times W}$ is then generated with an architecture similar to that of Section 2.2 and element-wise multiplied with $\mathbf{X}_{tgt}$.

3 Experiments

The volumetric attention was evaluated on two public datasets, Liver Tumor Segmentation (LiTS) [1] and DeepLesion[19]. All experiments used a PyTorch implementation [2] of the Mask-RCNN and Faster R-CNN. Unless otherwise noted, all hyperparameters are as in [14] for the Faster-RCNN and [10] for the Mask-RCNN.

3.1 Datasets and Evaluation

LiTS is a dataset of liver lesions, including 131 training and 70 test CT scans, acquired in six different clinical sites. Lesion segmentation performance is evaluated and ranked by the Dice coefficient per volume, averaged over all test cases. For additional insight on the quality of the segmentation, we also break down the average Dice/lesion per lesion size: the coefficients measured for small (diameter $<$ 15 mm), medium (diameter between [15mm, 30mm] and large (diameter $>$ 30mm) legions are denoted as Dice$_s$, Dice$_m$ and Dice$_l$ respectively. DeepLesion is a dataset with a larger variety of lesions, including 33,688 bookmarked radiology images from 10,825 studies of 4,477 unique patients. For each bookmarked image, a bounding box is generated to indicate the location of each lesion. We use the official split (70% training, 15% validation and 15% test) at the patient level, for training and testing. For consistency with prior art, detection results are evaluated with the False Positives (FPs) per Image metric.

3.2 LiTS Experiments

Pre-processing: For 3D liver/lesion detection and segmentation, we stack three adjacent axial slices into a 3-channel image and apply the Mask-RCNN to detect and segment the liver/lesion for the center slice. 3D segmentation results are then obtained by stacking the masks generated for all slices. The Mask-RCNN is trained to detect both liver and lesions, to enable the removal of false lesions outside the liver by simply computing the logical AND of the predicted liver and lesion masks. Since the focus of this task is on the liver and lesions, the CT scan's Hounsfield unit (HU) is clamped between [-200, 300] and normalized to a floating point between [0, 1]. Each slice is scaled to 1024×1024 pixels and the slice-thickness resampled to 1.5mm.

Benchmark results: To evaluate performance on LiTS, the feature bag size of (1) was set to 9, the weights of the feature extractor and RPN copied from detectors pre-trained on the MS-COCO and DeepLesion datasets, and the smallest image scale set to 1024. Table 1 presents a copy of the LiTS leaderboard, at the time of submission of this paper. All algorithms are evaluated on the LiTS `test` set. The VA Mask R-CNN achieves state-of-the-art performance, with 74.10 dice per case. This outperforms the previous LiTS challenge winner by 3.9 points and the best published results by 1.9 points.

Team	Model	Dice per case
3D U-Net(Ours) [5]	3D U-Net	55.0
G. Chlebus [3]	2D U-Net	65.0
E. Vorontsov et al. [16]	2D + 3D FCN	65.0
Y. Yuan [20]	Deconv-Conv Net	65.7
X. Han [9]	2D U-Net	67.0
LeHealth	-	70.2
Mask-RCNN(Ours)[10]	Mask-RCNN	70.3
X. Li et al.[13]	H-DenseUNet	72.2
VolumetricAttention	VA Mask-RCNN	**74.1**

Table 1: Comparison with LiTS Challenge leaderboard, as of July 1st, 2019

Ablation study and evaluation: To better understand the proposed architecture, the LiTS dataset was split, using 75% of the `train` data to create a training set and the remaining 25% as a `val` set for a local ablation study. Table 2 summarizes the resulting dice per volume, averaged over all cases, and dice per cases, averaged over small, medium and large lesions. All these are control experiments, all hyper-parameters and settings remaining the same as in the benchmark experiments, unless otherwise noted.

Benefits of VA attention: Three conclusions can be drawn from Table 2a. First, the 2D approaches outperform the 3D U-Net, even before addition of the VA attention module. This shows that 2D networks are at least competitive for 3D mask segmentation. Since the Mask-RCNN achieved the best performance on these experiments, we use it as base model in the remainder of the paper. It should, however, be pointed out that VA could equally be combined with the 2D U-Net. Second, the addition of the VA module further increases performance, increasing the Dice coefficient per case by 4.1 points. Third, this gain is especially large for small lesions. Note how the lack of contextual information along the z direction severely compromises the small lesion performance of the mask R-CNN. Fig.4 illustrates how VA attention enables the Mask R-CNN to reject confusing FP lesions and produce smoother segment boundaries.

Influence of pre-training: [11] claims that ImageNet pre-training does not improve accuracy of networks trained with as few as 10k COCO images. As shown in Table 2b, this does not hold for medical imaging where, due to the difficulties of collecting and labeling datasets, few datasets have 10k examples. Furthermore, while MS-COCO has $\sim$5 objects/image, this number is much smaller for medical image datasets. For LiTS

	Dice	Dice_s	Dice_m	Dice_l
3D U-Net	35.3	17.0	39.2	61.3
2D U-Net	48.8	39.7	58.2	68.3
Mask-RCNN	56.7	44.3	70.6	78.4
Ours	**60.8**	**52.2**	**71.4**	**78.7**

(a) Dice comparison.

	Pre-training dataset		
+ImageNet	✓	✓	✓
+MS-COCO		✓	✓
+DeepLesion			✓
Dice per case	60.8	61.9	**63.3**

(b) Pre-training dataset.

Scale	Dice	Dice_s	Dice_m	Dice_l
512	50.2	35.8	65.1	77.9
800	61.1	52.1	71.6	79.3
1024	**63.3**	**54.3**	**73.7**	**80.3**
1333	**63.5**	**54.8**	**73.5**	**80.4**

(c) Influence of image scales.

	Dice	Dice_s	Dice_m	Dice_l
Baseline	56.7	44.3	70.6	78.4
+channel att	61.5	52.2	72.7	78.7
+spatial att	**63.3**	**54.3**	**73.7**	**80.3**

(d) Influence of VA modules.

	Dice	Dice_s	Dice_m	Dice_l
Baseline	56.7	44.3	70.6	78.4
RPN	**63.3**	**54.3**	**73.7**	**80.3**
RCNN	61.3	51.7	71.8	79.9

(e) Influence of VA location.

# Slices	Dice	Dice_s	Dice_m	Dice_l
$9(3 \times 3)$	61.7	52.2	71.6	79.5
$21(3 \times 7)$	62.5	52.6	72.2	79.8
$27(3 \times 9)$	**63.3**	**54.3**	**73.7**	80.3
$33(3 \times 11)$	63.1	53.6	73.4	**80.6**

(f) Influence of number of slices.

Table 2: Evaluation on LiTS `val` set, in terms of dice per volume, averaged over all cases, and dice per lesions, averaged over small, medium and large lesions.

the number is smaller than 1, especially when the 3D volume is split into 2D slices and these are considered different examples. Table 2b shows that, in this case, ImageNet pre-training still has an important role in combating overfitting. Adding MS-COCO to the pre-training dataset further improves performance by 1.1 points. This is mostly because the COCO tasks encourage the network to more accurately localize objects. Finally, due to the non-trivial domain shift between MS-COCO and LiTS, further pre-training on DeepLesion improves performance by an additional 1.4 points.

Image scales. Table 2c shows that larger image scales lead to better performance, especially for small lesions. However, performance saturates at a scale of 1333 pixels. This is only marginally better than a scale of 1024 but requires substantially more memory. For this reason, a scale of 1024 is adopted in the remainder of the paper.

Spatial vs. Channel Attention: to compare the relative importance of the two attention mechanisms, the two modules were incrementally added to the 2D Mask-RCNN, with the results of Table 2d. These experiments use 9 slices. The addition of channel attention enhances performance by more than 4 points, and the subsequent addition of spatial attention increases performance by another 1.8 points. In summary, both attention mechanisms are important.

Location of attention module: the VA module can be added as shown in Fig.2, i.e. to the last stage of feature extraction, before the RPN, or after the bounding box ROI align and mask ROI align steps, i.e. before the RCNN. Table 2e shows that attention is more effective if introduced before the RPN. While this improves performance by 6.6 Dice points per case, the gain is only 4.6 points when attention is introduced after the RCNN. This shows that 3D context is important for high quality proposal generation. Since only RPN detected ROIs are used to crop feature maps, addition of attention after the RPN only improves the ability to reject FPs. In this case, attention cannot improve the retrieval of lesions that are otherwise missed.

Feature bags size: Table 2f compares the network performance as the feature bag size. While dice per case increases with feature bag size, the small and medium lesion performance starts to worsen beyond a bag size of 9. We note that for applications sensitive to inference time, smaller bag size may be preferable.

3.3 Extension Experiments on DeepLesion

Model	Backbone	0.5	1	2
Faster-RCNN[8]	VGG-16	56.9	67.3	75.6
R-FCN[6]	VGG-16	55.7	67.3	75.4
Improved R-FCN [6]	VGG-16	56.5	67.7	76.9
Data-level fusion, 11 slices	VGG-16	58.5	70.0	77.9
3-DCE,9 Slices[18]	VGG-16	59.3	70.7	79.1
3-DCE,27 Slices[18]	VGG-16	62.5	73.4	80.7
Faster-RCNN+VA, 9 Slices	ResNet50	**67.6**	**75.6**	**82.5**
Deformable Faster-RCNN+VA	ResNet50	**69.1**	**77.9**	**83.8**

Table 3: Sensitivity(%) at 0.5, 1 and 2 FPs per image on the DeepLesion `test` set.

Model	Backbone	1 FPs	AP_{50}
Faster-RCNN[8]	ResNet152	77.4	64.9
Faster-RCNN[8]	ResNet101	75.1	61.8
Faster-RCNN[8]	ResNet50	73.4	60.0
Deformable Faster-RCNN[7]	ResNet50	76.3	62.4
Faster-RCNN+VA	ResNet50	75.6	63.0
Deformable Faster-RCNN+VCA	ResNet50	76.8	63.8
Deformable Faster-RCNN+VSA	ResNet50	76.9	64.1
Deformable Faster-RCNN+VA	ResNet50	**77.9**	**65.0**

Table 4: Sensitivity (%) at 1 FPs/image and AP_{50} on the DeepLesion `test` set.

To test the effectiveness of volumetric attention for the processing of 3D CT volume datasets, we performed some extension experiments on DeepLesion. This dataset enables the use of part of the 3D CT volume as context for 2D bounding box prediction on target slices. Since DeepLesion does not provide mask groundtruth, the VA module was implemented on Faster-RCNN-FPN and Deformable Faster-RCNN-FPN detectors, with ResNet50 backbones. As usual for DeepLesion, performance is evaluated with FPs per image. AP_{50} is also presented in Table4. All experiments in this section are based on training with the DeepLesion `train` and `val` sets, and testing on `test` set. Each 2.5D image is formed by concatenating 3 contiguous slices and scaled to 512$\times$ 512 pixels as in [19], the Faster-RCNN-FPN backbone is pretrained on ImageNet. Feature bag size is fixed to be 3, i.e. 9 slices.

Table 3 and Table 4, compare the proposed networks to several methods from the literature. The proposed networks achieve state of the art results, increasing sensitivity by 6.6 points at 0.5 FPs/image, 4.4 points at 1 Fp/image and 3.1 at 2 FPs/image. Table 4, shows that the proposed network with the ResNet50 backbone is comparable with the heavier Faster-RCNN with ResNet101 backbone. Independently adding Volumetric Channel Attention(VCA) and Volumetric Spatial Attention(VSA) to Deformable Faster-RCNN with ResNet50 can get 1.4 and 1.6 points performance increase separately, integrating VSA and VCA got 2.6 points performance improvement, this result is even slightly higher than much heavier Faster-RCNN with ResNet152 backbone.

4 Conclusion

In this paper, we proposed a volumetric attention module that enables 2.5D methods to leverage contextual information along the z direction and the use of pretrained 2D detection models when training data is limited, as is often the case for medical applications. VA can be combined with any CNN architecture, including one-stage and two-stage detectors and segmentation networks. It was shown that 2.5D networks with VA achieve state of the art results for *both* lesion segmentation and detection.

References

1. Bilic, P., et al.: The liver tumor segmentation benchmark (lits). arXiv:1901.04056 (2019)
2. Chen, K., et al.: mmdetection. https://github.com/open-mmlab/mmdetection (2018)

3. Chlebus, G., et al.: Neural network-based automatic liver tumor segmentation with random forest-based candidate filtering. arXiv:1706.00842 (2017)
4. Christ, P.F., et al.: Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In: MICCAI (2016)
5. Çiçek, Ö., et al.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: MICCAI (2016)
6. Dai, J., et al.: R-fcn: Object detection via region-based fully convolutional networks. In: NIPS (2016)
7. Dai, J., et al.: Deformable convolutional networks. ICCV (2017)
8. Girshick, R., et al.: Fast r-cnn. In: ICCV (2015)
9. Han, X.: Automatic liver lesion segmentation using a deep convolutional neural network method. arXiv:1704.07239 (2017)
10. He, K., et al.: Mask r-cnn. In: ICCV (2017)
11. He, K., et al.: Rethinking imagenet pre-training. arXiv:1811.08883 (2018)
12. Hu, J., et al.: Squeeze-and-excitation networks. In: CVPR (2018)
13. Li, X., et al.: H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging **37**(12), 2663–2674 (2018)
14. Lin, T.Y., et al.: Feature pyramid networks for object detection. In: CVPR (2017)
15. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
16. Vorontsov, E., et al.: Liver lesion segmentation informed by joint liver segmentation. In: ISBI (2018)
17. Wang, X., et al.: Non-local neural networks. In: CVPR (2018)
18. Yan, K., et al.: 3d context enhanced region-based convolutional neural network for end-to-end lesion detection. In: MICCAI (2018)
19. Yan, K., et al.: Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of Medical Imaging **5**(3), 036501 (2018)
20. Yuan, Y.: Hierarchical convolutional-deconvolutional neural networks for automatic liver and tumor segmentation. arXiv:1710.04540 (2017)