Lesion Detection by Efficiently Bridging 3D Context

Zhishuai Zhang[1],Yuyin Zhou[1],Wei Shen[1],Elliot Fishman[2],Alan Yuille[1]

[1]The Johns Hopkins University, Baltimore, MD 21218, USA
[2]The Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA

Abstract. Lesion detection in CT (computed tomography) scan images is an important yet challenging task due to the low contrast of soft tissues and similar appearance between lesion and the background. Exploiting 3D context information has been studied extensively to improve detection accuracy. However, previous methods either use a 3D CNN which usually requires a sliding window strategy to inference and only acts on local patches; or simply concatenate feature maps of independent 2D CNNs to obtain 3D context information, which is less effective to capture 3D knowledge. To address these issues, we design a hybrid detector to combine benefits from both of the above methods. We propose to build several light-weighted 3D CNNs as subnets to bridge 2D CNNs' intermediate features, so that 2D CNNs are connected with each other which interchange 3D context information while feed-forwarding. Comprehensive experiments in DeepLesion dataset show that our method can combine 3D knowledge effectively and provide higher quality backbone features. Our detector surpasses the current state-of-the-art by a large margin with comparable speed and GPU memory consumption.

1 Introduction

Lesion detection is an essential task for clinical applications such as computer-aided diagnosis. With the emergence of modern CNNs, object detection in 2D natural images has been developed quickly and achieves promising performance [1, 5–7]. However, it is still unclear how to adapt these algorithms into CT scans effectively. The main gap is how to efficiently involve 3D context information into these detectors. This problem has attracted many research attentions [2, 4, 10], due to its importance for the success of lesion detection.

Current solutions come in two folds. One uses fully 3D connected CNNs, which can directly exploit 3D knowledge, for detection and classification. However, due to GPU memory limit, it can only be performed on small patches in a sliding-window fashion [4] or on small-patch candidates generated by a 2D detector [2], leading to high time complexity. It is also unable to make use of ImageNet pretraining, thus only achieves inferior lesion detection accuracy as reported in [10]. To alleviate the issues of 3D CNNs, other studies are exploring how to combine 2D CNN features from consecutive CT slices for classification and regression, so as to better utilize the 3D context information. [10]

followed R-FCN [1] which used a Region Proposal Network (RPN) to predict suspicious regions and a Region Classification Network (RCN) to further classify and regress those suspicious regions. [10] proposed to concatenate backbone feature maps from neighboring CT slices to feed into RCN, in order to gather 3D information in the RCN subnet. Under this pipeline, a backbone network can take the whole CT scans as input which can be trained in an end-to-end manner, from ImageNet pretrained weights. However, the backbone networks are still independent 2D CNNs, and no 3D information can be aggregated until the final backbone features are computed. Another problem is that the central CT slice and the contextual CT slices share the same 2D CNN weights, which may be less optimal since we expect to distill different and complementary knowledge from those different slices.

We propose a hybrid detector combining advantages of fully 3D connected CNN detectors (strong knowledge of 3D context) [2, 4] and 2D CNN concatenated detectors (efficiency and ability to use ImageNet pretrained weights) [10]. Similar to [10], we use 2D CNNs for CT slices at different axial locations as our backbone. However, as discussed before, this is less optimal since these isolated 2D CNNs cannot extract and exploit 3D context information. To address this problem, we propose light-weighted 3D CNN subnets named 3D Fusion Modules (3DFMs) to bridge those 2D CNNs, allowing information flow from different slices. These subnets connect the internal layers of 2D CNNs, so that each 2D CNN can distill knowledge from its neighbor 2D CNNs, to exploit 3D information and focus on different knowledge. The main difference between [10] and our method is that in [10], the 3D context information is not exploited in the layers before the RCN, and the RCN cannot fully utilize 3D context since the its input features only have high-level semantics without low-level details, and the RCN has a very shallow structure which is incapable of learning rich 3D information; on the contrary, in our method, the 3D information is exploited gradually throughout our backbone CNNs, and 3DFMs learn 3D information at low-level, mid-level and high-level layers. Our design breaks the isolation among 2D CNNs, and enables them to distill different knowledge from different input slices, thus the backbone provides stronger features with richer 3D context encoded.

3DFMs introduce few parameters and small computation overhead, while greatly improving the detection accuracy. Experiments on DeepLesion [11] show our hybrid detector significantly improves the sensitivities at every false positive (FP) rate and on every lesion type. With 27 CT scan slices as input, hybrid detector improves the average sensitivities by 1.4 and the sensitivity at $\frac{1}{8}$ FP per image by 2.7. Our method surpasses [10] and achieves a new state-of-the-art.

2 Approach

2.1 Overview Pipeline of Our Detector

The backbone of our detector is shown in Figure 1. Following [10], to make use of ImageNet pretrained weights, we combine 3 adjacent CT slices into a 3-channel image like a natural image, to feed to VGG16 [9], which serves as the backbone

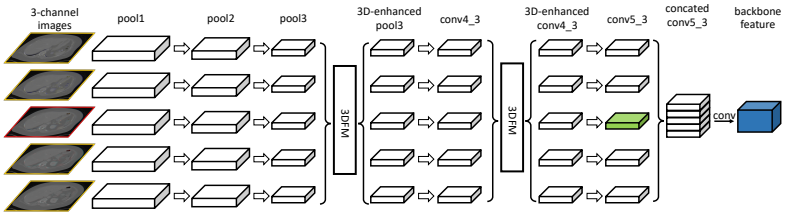

Fig. 1: Backbone of our hybrid lesion detector. Different rows illustrate different 2D CNNs for the corresponding images. The ground-truth boxes are labelled in the central image (with red boundary) and other 3-channel images (with yellow boundary) are served as 3D context. The central `conv5_3` feature (marked in green) is used in RPN and the fused feature (marked in blue) is used in RCN. Best view in color.

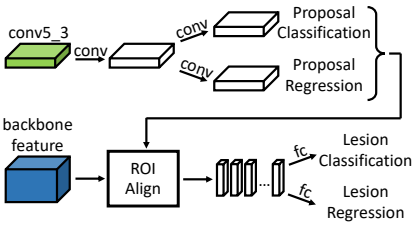

Fig. 2: RPN (in the top row) and RCN (in the bottom row) sub-networks.

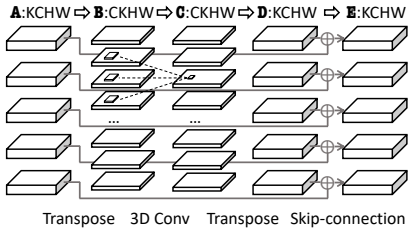

Fig. 3: 3D Fusion Module. K is 5 in this example. See Subsection 2.2 for details.

2D CNN of our detector. When considering more 3D context, we combine context slices into 3-channel images and feed them to different VGG16 branches. Each VGG16 branch takes a 3-channel image as input, and generates a `conv5_3` feature map as output. The `conv5_3` feature from the central slice (marked in green) is used in the Region Proposal Network (RPN) to generate proposals, and the concatenation of `conv5_3` features from all slices (marked in blue) is used in the Region Classification Network (RCN) to classify and regress proposals. However, unlike [10], where 2D CNNs feed-forward isolatedly, we use a novel and efficient 3D Fusion Module (3DFM) to bridge internal features from different 2D CNNs to build a hybrid backbone. The hybrid detector backbone can better exploit 3D context and make different 2D CNNs to learn different patterns, while utilizing ImageNet pretrained weights. Details of 3DFM are discussed in Subsection 2.2.

Given the backbone of our hybrid lesion detector, we follow [7] to employ an RPN and an RCN to generate and classify proposals. As Figure 2 shows, we use the `conv5_3` feature of the central branch (marked in green) to generate proposals, and use ROIAlign [3] to generate features from the concatenated feature of different branches (marked in blue), for each proposal. Finally, those features are used to classify and regress the proposals, and generate lesion detection results.

2.2 3D Fusion Module

3D context has been shown to be extremely important to detect objects in CT scan images [2, 4, 10]. However, existing methods to utilize 3D information are

either memory expensive and only able to process small 3D patches, or inefficient which naively concatenate features from different slices. In this paper, we propose an efficient and computation cheap 3D Fusion Module (3DFM), as shown in Figure 3, to combine 3D context information in the backbone 2D CNNs.

3DFM takes internal features ($A_i \in \mathbb{R}^{C\times H\times W}$, C, H and W are the channel, height and width of the feature map) from the backbone CNNs as inputs, as shown in the first column in Figure 3. Given K input images, there will be K intermediate features, and each of them is generated from a 3-channel CT image as shown in Figure 1. We first concatenate them to build a 4D tensor $\mathbf{A} \in \mathbb{R}^{K\times C\times H\times W}$, and transpose it make the channel to be the first dimension ($\mathbf{B} \in \mathbb{R}^{C\times K\times H\times W}$), as shown in the second column in Figure 3. A 3D convolution is used to gather 3D context information to generate a 3D fused feature map $\mathbf{C} \in \mathbb{R}^{C\times K\times H\times W}$. The kernel size is $3\times 1\times 1$ corresponding to the K, H and W dimensions, so we are utilizing the context along the axial direction by convolving across neighbor slices. We use $3\times 1\times 1$ instead of $3\times 3\times 3$ because the context along the other two directions is already considered in the 2D convolutions in the backbone CNN, and thus we only need to consider the axial direction to reduce computation/memory overhead. Finally $\mathbf{C}$ is transposed backed to $K\times C\times H\times W$ as $\mathbf{D}$, and the sum of $\mathbf{A}$ and $\mathbf{D}$ (noted as $\mathbf{E}$) is split to K feature maps with shape $C\times H\times W$, which are used in the backbone 2D CNNs for future processing.

3DFM is flexible and can be inserted anywhere in the backbone CNNs to fuse the 3D information. In our detector, we insert 3DFMs in a sparse manner: only at the `pool3` and `conv4_3` layers in VGG16, as in Figure 1. These 3DFMs will combine those independent 2D VGG16 branches into a sparsely bridged 3D CNN, which will serve as the backbone CNN of our detector. Extensive experiments show our design is light-weighted and takes very little computation/memory overhead, while effectively exploiting 3D context knowledge and improving the accuracy significantly.

3 Experiments

3.1 Implementation Details

Our hybrid detector is implemented with Tensorflow. We use VGG16 as our backbone CNN, and remove the `pool4` layer to keep the output resolution to be $\frac{1}{8}$ of the input image. We take the same CT scan image preprocessing as in [10], which rescales the CT intensity to 0-255, resizes the images and clips the black border. We use the horizontal flip data augmentation which is very common for object detection. For each sample, we take adjacent 3, 9, 15, 21 or 27 CT slices to generate 1, 3, 5, 7 or 9 input images with 3 channels each, to evaluate the efficacy of hybrid detector at different 3D context richness levels, and to make a fair comparison with the state-of-the-art 3DCE [10]. For the training, we use a batch size of 2, and train the hybrid detector for 120k iterations. The initial learning rate is 10^{-3} and is reduced to 10^{-4} after the first 90k iterations. We take the official `train`/`test` subsets to train and report accuracy. Comprehensive experiments and ablation studies are reported in the following subsections.

Table 1: Performance (%) on the official `test` split for DeepLesion dataset. $0.125, 0.25, \cdots, 16$ represent the number of FPs per image.

Settings	0.125	0.25	0.5	1	2	4	8	16	AVG@$\frac{1}{8}$:8
Baseline - 3 slices	31.52	43.95	57.19	68.51	77.47	83.59	87.77	90.66	64.28
3DCE [10] - 9 slices	-	-	59.32	70.68	79.09	84.34	87.81	89.62	-
Baseline - 9 slices	35.48	48.84	62.42	73.06	80.73	85.82	89.22	91.21	67.94
Hybrid - 9 slices	38.25	50.66	62.97	73.20	80.66	85.80	89.04	91.21	68.65
Baseline - 15 slices	37.53	51.23	63.97	74.53	81.39	86.15	89.37	91.28	69.17
Hybrid - 15 slices	40.33	53.01	65.26	75.78	82.44	86.84	89.76	**91.69**	70.49
Baseline - 21 slices	38.81	52.32	64.93	75.25	82.19	86.61	89.44	91.25	69.93
Hybrid - 21 slices	40.74	53.80	66.06	75.66	82.60	86.88	89.79	91.62	70.79
3DCE [10] - 27 slices	-	-	62.48	73.37	80.70	85.65	89.09	91.06	-
Baseline - 27 slices	38.43	52.09	65.03	75.10	81.88	86.05	89.10	91.05	69.67
Hybrid - 27 slices	**41.12**	**53.83**	**66.32**	**76.27**	**82.89**	**87.01**	**89.84**	**91.69**	**71.04**

3.2 Experimental Results

To evaluate the efficacy of our method, we conduct extensive experiments on DeepLesion [11]. Following the metric used in LUNA challenge [8], we compute the sensitivity at 7 pre-defined false positive (FP) per image rates: $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, 4 and 8 FPs per sample, as well as the average sensitivity at these 7 pre-defined FP rates. We also compute the sensitivity at the FP per image rate of 16, to compare with the 3DCE [10]. For all our baselines and hybrid detectors, we train and evaluate for four times, and report the average performance, to alleviate the randomness caused by initialization and training data shuffling.

The results on the official `test` set are shown in Table 1. We also compare with 3DCE which is the current state-of-the-art and already surpasses fully 3D connected detectors. 'Baseline' in the table is a Faster-RCNN [7] based detector with feature concatenation after the backbone CNN, and 'Hybrid' is 'Baseline' equipped with 3DFMs illustrated in Figure 3. We also plot the free-response receiver operating characteristic curves for our baseline and hybrid detectors in Figure 4. In the table and figure, we find that our hybrid detector with 3DFM is very effective in improving the detection quality. The sensitivity consistently goes up at all FP rate levels significantly with 27 slices as input, especially in the high precision case (*i.e.* fewer FPs per image). Our hybrid detector surpasses 3DCE greatly with the same `train`/`test` sets and achieves a new state-of-the-art.

3.3 Ablation Studies

Inference Speed and Memory Overhead Our 3D Fusion Modules (3DFM) efficiently combine 3D context information in the backbone. To quantitatively evaluate the computation/memory overhead, we run all our baselines and detectors on a machine with a single nVIDIA Titan Xp GPU. We report the total runtime for the official `test` set (4817 samples) and the max GPU memory consumed for inference. Results are shown in Table 2. Our 3DFMs introduce very small computation overhead and negligible GPU memory overhead. This verifies the efficiency of our method, which may be applied to more complex datasets.

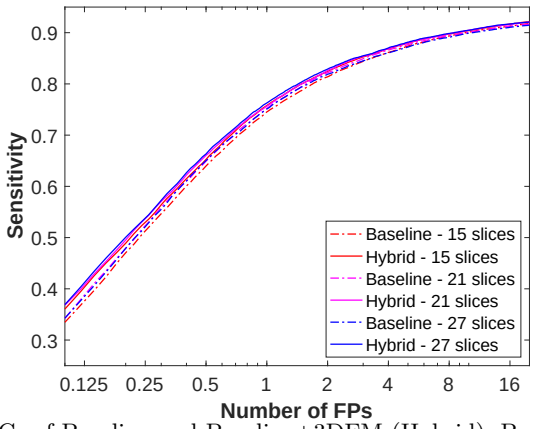

Fig. 4: FROCs of Baseline and Baseline+3DFM (Hybrid). Best view in color.

Table 2: Performance on the official `test` split for DeepLesion dataset.

Settings	AVG@$\frac{1}{8}$:8	Runtime (s)	FPS	Inference GPU memory (GB)
Baseline - 9 slices	67.94	246	19.58	0.455
Hybrid - 9 slices	68.65	256	18.82	0.459
Baseline - 15 slices	69.17	345	13.96	0.693
Hybrid - 15 slices	70.49	369	13.05	0.696
Baseline - 21 slices	69.93	452	10.66	0.930
Hybrid - 21 slices	70.79	479	10.06	0.934
Baseline - 27 slices	69.67	564	8.54	1.167
Hybrid - 27 slices	71.04	608	7.92	1.171

Architecture of 3DFM In this subsection, we compare our 3D Fusion Module with some other potential architectures combining 3D context information:

- Without Skip Connection: the 3D context information bridging module is the same as 3DFM (see Figure 3), but does not have the skip connection to combine the original backbone features with the 3D fused features.
- Without 3D Conv: the 3D context information bridging module concatenates the K backbone features with size of $C \times H \times W$ to a thicker tensor $KC \times H \times W$, and takes a 1×1 2D Conv to fuse information from different slices.

All experiments are conducted on the 27-slice inputs, and results are shown in Table 3. Both architectures described above achieve inferior performance: without skip connection, it has lower sensitivities at high FP levels even compared with our baseline detector; and using 2D Conv on a concatenated feature map leads to inferior sensitivities at all FP levels.

Number of 3DFMs 3DFMs can bridge the 3D context information in the 2D CNN backbones, and can be inserted anywhere in the 2D CNNs. In our detector, we insert 3DFMs at the `pool3` and `conv4_3` layers in VGG16 as in Figure 1. We also conduct diagnostic experiments by 1) inserting 3DFM at only `conv4_3` layer and 2) inserting 3DFMs at `pool2`, `pool3` and `conv4_3` layers. Results are shown in '3DFM@4' and '3DFM@234' of Table 3. Compared with '3DFM@4',

Table 3: Ablation of 3DFM architecture.

Setting	0.125	0.25	0.5	1	2	4	8	16	AVG@$\frac{1}{8}$:8
Baseline	38.43	52.09	65.03	75.10	81.88	86.06	89.10	91.05	69.67
Hybrid (Ours)	41.12	53.83	66.32	76.27	82.89	87.01	89.84	91.69	71.04
W/O Skip Connection	42.10	54.22	66.29	75.15	81.79	86.11	88.93	90.78	70.66
W/O 3D Conv	39.96	53.46	65.66	75.36	81.96	86.72	89.71	91.56	70.40
3DFM@4	40.56	53.37	65.62	75.91	82.49	86.77	89.56	91.57	70.62
3DFM@234	40.87	54.27	66.45	76.35	82.90	87.18	90.15	92.00	71.17

Table 4: Sensitivities of different types of lesion at 4 false positive per image. Our detector outperforms baseline on all 8 types.

Type	BN	AB	ME	LV	LU	KD	ST	PV
Baseline	72.69	84.07	87.27	90.04	89.70	85.73	76.99	83.50
Hybrid (Ours)	**73.84**	**84.63**	**88.43**	**91.14**	**90.50**	**86.16**	**77.91**	**85.64**

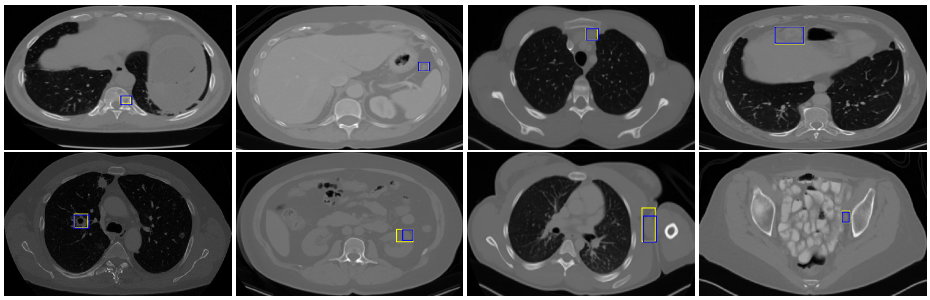

Fig. 5: Detection examples of eight types. Yellow and blue boxes are for ground-truth and detection result. All examples are detected by our hybrid detector while missed by our baseline detector.

adding another 3DFM at `pool3` significantly improve the performance from 70.62 to 71.04. However, adding an extra 3DFM at `pool2` will only give a marginal performance gain. For simplicity, we use only two 3DFMs in our final detector.

3.4 Analysis on different lesion types

We test our hybrid detector on different lesion types in DeepLesion [11]. There are 8 types of lesion labelled in the dataset, and the abbreviations are in the parentheses: bone (BN), abdomen (AB), mediastinum (ME), liver (LV), lung (LU), kidney (KD), soft tissue (ST) and pelvis (PV). In Table 4, we evaluate the sensitivities of our baseline detector and our hybrid detector equipped with 3DFM, at 4 FPs per image (27 slices). The results further confirms that our hybrid detector can improve the detection quality under all 8 lesion types, thus it is very general with consistent gains. We also show some qualitative results in Figure 5, where our baseline detector fails to detect the lesion, but the 3DFM equipped hybrid detector detects them with scores greater than 0.9 at 4 FP per image threshold. We observe that our detector is able to find difficult lesions such as small or low-contrast lesions.

4 Conclusions

We propose a hybrid detector which bridges 3D context information in 2D CNN backbones. Based on a baseline detector which takes adjacent CT scan images independently with the same 2D CNN, we enhance the backbone feature quality by fusing 3D context knowledge via 3DFMs. Extensive experiments have been conducted to show the efficacy of our hybrid detector, which improves the sensitivity at all false positive levels. The improvement is consistent under different settings (*e.g.*, number of input slices and lesion types). Qualitative analysis also suggests that our method outperforms the baseline method and remains valid even for some extremely difficult cases. Our approach surpasses existing methods and thus establishes a new state-of-the-art. The superior performance demonstrates its potential usage for different clinical applications.

Acknowledgements. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and NSFC No. 61672336.

References

1. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. In: NIPS. pp. 379–387 (2016)
2. Ding, J., Li, A., Hu, Z., Wang, L.: Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In: MICCAI. pp. 559–567. Springer (2017)
3. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2961–2969 (2017)
4. Liao, F., Liang, M., Li, Z., Hu, X., Song, S.: Evaluate the malignancy of pulmonary nodules using the 3d deep leaky noisy-or network. arXiv preprint arXiv:1711.08324 (2017)
5. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV. pp. 21–37 (2016)
6. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. pp. 779–788 (2016)
7. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015)
8. Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., van den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical image analysis **42**, 1–13 (2017)
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
10. Yan, K., Bagheri, M., Summers, R.M.: 3d context enhanced region-based convolutional neural network for end-to-end lesion detection. In: MICCAI. pp. 511–519. Springer (2018)
11. Yan, K., Wang, X., Lu, L., Zhang, L., Harrison, A.P., Bagheri, M., Summers, R.M.: Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In: CVPR. pp. 9261–9270 (2018)