DetectDiffuse: Aggregation- and Attention-driven Universal Lesion Detection with Multi-scale Diffusion Model

Xinyu Li[1][0009−0008−7731−2498], Danni Ai[1*][0000−0002−2285−0570], Jingfan Fan[1][0000−0003−4857−6490], Tianyu Fu[2][0000−0002−9808−960X], Hong Song[3], Deqiang Xiao[1], and Jian Yang[1][0009−0002−8914−4619]

[1] Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Photonics, Bejing Institute of Technology, No 5 Zhongguancun South Street, Haidian District, 100081, Beijing, China.
danni@bit.edu.cn
https://www.inavilab.com/
[2] School of Medical Technology, Bejing Institute of Technology, No 5 Zhongguancun South Street, Haidian District, 100081, Beijing, China.
[3] School of Computer Science and Technology, Bejing Institute of Technology, No 5 Zhongguancun South Street, Haidian District, 100081, Beijing, China.

Abstract. Automated Universal Lesion Detection (ULD) based on computed tomography (CT) images provides physicians with rapid and objective information regarding lesion locations and shapes. However, it is difficult to detect universal lesions in various regions because of the disparity in lesion sizes and the grayscale variation present in CT images. In this paper, we propose DetectDiffuse, a multi-scale diffusion model driven by feature aggregation and 3D attention. First, we utilize the diffusion model to generate noisy detection boxes, incorporating a scale factor to simulate lesions at different scales and mitigate detection errors. Second, we develop a Neighborhood Aggregation (NA) module to enhance the model's capability to distinguish between lesioned and normal tissues. This module aggregates features within and around detection boxes, reducing false detections caused by significant grayscale differences in lesions. Third, we propose a 3D Stripe Attention (SA) module leveraging dimensional disambiguation. This module uses an attention mechanism to extract information across different dimensions of CT images more effectively. We performed comparison experiments on five datasets, the results show that the proposed method outperforms the 12 compared state-of-the-art methods, and improves the performance by 5.82% compared with the best method.

Keywords: Universal Lesion Detection · 3D Stripe Aggregation · Neighbor boxes Aggregation.

1 Introduction

Computed tomography images can quickly scan the human body and are widely used in the diagnosis of diseases such as tumors, trauma and infection [19]. In

clinical practice, physicians primarily rely on manually examining CT images layer by layer, search for abnormal regions that differ from their prior knowledge bank of "healthiness" when mentally segmenting the lesions. It is a more common scenario where physicians do not segment the contours of lesions explicitly but assess the condition of the lesions implicitly [17]. However, it is a time consuming process, prone to inter- and intra-human error. Therefore, automatically detecting lesions via computer assistance is crucial for supporting physicians in disease diagnosis. Researchers have developed a lot of studies [20, 27, 5, 7] on automated identification of lesions in specific organs and achieved good results. However, several challenges must be addressed to achieve universal lesion detection. First, lesions in different organs have different shapes and sizes, it is difficult to design detection boxes for lesions with variable shapes. Second, lesions in different parts of the CT image have different grayscales and may overlap with the grayscale distributions of normal tissues in other parts. The two challenges mentioned above together lead to a susceptibility to false detections in ULD tasks.

To achieve universal lesion detection in CT images, most of the early studies [10, 22, 18] focused on localizing lesions by modifying the backbone networks of anchor-based detection frameworks such as Faster R-CNN [12], which utilize hand-designed anchor boxes for lesion detection. However, the size of lesions in images of different parts of the body and different stages of the disease has large differences. Manually designed anchor boxes are typically effective only for detecting moderately sized lesions and may fail to detect smaller lesions in the early stages or larger lesions in the later stages. To avoid the above problems, anchor-free-based [28, 8] detection frameworks have been proposed [11, 15]. These frameworks localize lesions by predicting their center points and achieve better detection across lesions of varying sizes. These methods primarily rely on convolutional networks to extract features along the vertical axis of the CT image, which limits their ability to model relationships between different layers. Some approaches [21, 9] have attempted to address this by incorporating Transformer or self-attention mechanisms [3, 29], which improve detection by modeling long-range feature dependencies while preserving the convolutional networks' ability to capture local features. However, due to the computational demands of Transformers, these methods often suffer from slow training and inference speeds.

In this paper, we propose a multi-scale diffusion model driven by feature aggregation and striped attention for universal lesion detection in CT images. First, we employ the multi-scale forward diffusion process to generate noise detection boxes following a uniform Gaussian distribution for an image sequence composed of both the target and reference images. To simulate lesions of various sizes, we apply a random scale factor, further increasing the scale variance among the noise detection boxes. After feature extraction, a neighbor aggregation module is introduced to help the network differentiate between the features inside the detection boxes and the surrounding tissue features. Subsequently, the 3D stripe attention module disassembles the features into three independent dimensions, performing attention computation along these directions to aggregate

3D information from multiple perspectives. Finally, the lesion detection decoder utilizes reverse diffusion to obtain accurate lesion detection results.

Our contribution can be summarized as follows: firstly, we propose a multi-scale diffusion model as the framework for the proposed method. Among it, the neighbor-hood feature aggregation module is proposed. The problem of large differences in the scale and gray scale distribution of lesions in different parts of the body is effectively addressed by aggregating the tissue features around the detection boxes. Secondly, we introduce a 3D stripe attention module. This module can model 3D data along three different directions for long distance feature modeling, efficiently utilizing the spatial information inherent in 3D data, thereby enhancing the network's performance in lesion detection. Thirdly, experimental results on five different datasets demonstrate that our proposed method significantly improves detection performance compared to other ULD methods, achieving a detection sensitivity of 84.71% and an mAP of 63.89%, which are 3.08% higher than the current leading method.

2 Method

2.1 Overview

We propose the multi-scale diffusion model driven by feature aggregation and striped attention as shown in Fig. 1. First, a multi-scale diffusion model [4] is used, which covers lesions of various scales by setting scale factors γ and arbitrarily distributing the size of the boxes. Afterward, in the multi-scale forward diffusion process, two modules are proposed: neighborhood box aggregation (NA) module and 3D stripe attention (SA) module. In NA module, the region of interest (RoI) features of the noise boxes are used to compare with the features that surround it. By suppressing RoI features that are similar to the surrounding features and enhancing those that differ more significantly, the framework is better equipped to distinguish between lesions and normal tissues. Subsequently, the RoI features are passed to the SA module, which disassembles the features along three directions and computes attention. By stacking features from reference layers in different directions into the detection layer, weak lesions are enhanced, and non-lesion feature representations are suppressed. Finally, these features are fed into the lesion detection decoder to obtain the final detection results.

2.2 Neighbor Boxes Aggregation Module

To assist the network in distinguishing between lesions and the surrounding normal tissues and organs, it is crucial to compare the features within the detection box to those of the surrounding normal tissues. This comparison also enhances both the detection rate of the proposed method across various lesions and its generalization performance across different modalities. To this end, we propose a neighborhood boxes aggregation module, which reduces false-positive detections caused by tissues with similar imaging features to lesions by projecting the surrounding neighborhood features into the detection box in an attention-weighted

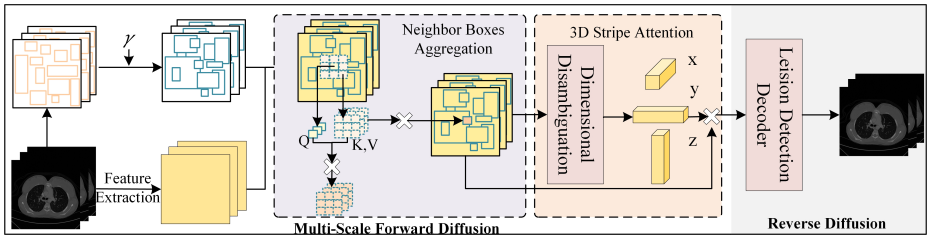

Fig. 1. Overview of the proposed framework.

manner. By comparing the features within the detection box to those of the surrounding tissues, the module effectively mitigates false positives due to tissues that closely resemble lesion imaging characteristics.

For any noise box, take 8 rectangular boxes around it with exactly the same shape as its neighboring boxes, and the coordinates of the kth detection box are:

$$B_{t,i}^{k} = [x_1 - m^k w, y_1 - n^k d, x_2 - m^k w, y_2 - n^k d] \tag{1}$$

where $w = x_2 - x_1$, $d = y_2 - y_1$ are the length and width of the current noise detection box, respectively. m^k, $n^k \in -1, 0, 1$ indicates the positional relationship between the neighborhood box and the noise detection box.

Form a sequence of features $\{X_{t,i}, X_{t,i}^1, \ldots, X_{t,i}^k, \ldots, X_{t,i}^8\} \in \mathbb{R}^{w \times d}$, which corresponding to the noise box and the neighboring boxes. Then vectorize them to obtain the corresponding RoI feature vector set $\{V_{t,i}, V_{t,i}^1, \ldots, V_{t,i}^k, \ldots, V_{t,i}^8\} \in \mathbb{R}^{w \times d \times C}$, where C is the number of feature channels. Finally, a linear projection is performed to characterize the token sequence of neighborhood RoI features:

$$\begin{aligned} Z &= \mathcal{F}_{proj}\left(\left[V_{t,i}, V_{t,i}^1, \ldots, V_{t,i}^k, \ldots, V_{t,i}^8\right]\right) \\ &= \left[V_{t,i}, V_{t,i}^1, \ldots, V_{t,i}^k, \ldots, V_{t,i}^8\right] \boldsymbol{E} \end{aligned} \tag{2}$$

in which $\boldsymbol{E} \in \mathbb{R}^{(w \times h \times C) \times D}$ is the learnable projection transform and D is the output channel, this operation is realized by 1×1 convolution in this paper. Then the group of vectors is fed into the multi-head self-Attention (MSA) Block to compute the attention, which consists of the MHA mechanism with Layer Normalization (LN) and uses residual concatenation to obtain the augmented features, as represented by $Z' = \mathrm{LN}(\mathrm{MSA}(Z) + Z)$. The feature vector corresponding to the noise box is used as query, and the rest of the feature vectors are used as the key and value inputs of the MSA, respectively. After residual linking with the original feature vectors, the enhanced feature vectors are remodeled into the enhanced feature map X' containing the neighboring information by layer normalization. The enhanced features are then weighted based on their difference from the original features via a dynamic convolution block. This block consists of Dynamic Convolution (DC), Layer Normalization, and Fully Connected (FC) layers. Residual connections are employed to obtain the final enhanced features, as represented by $X' = LN\big(FC\big(DC(X, X')\big) + X\big)$.

At this stage, the noise box is an enhanced feature that has incorporated the aggregation of neighboring features. If the features in the noise box are significantly different from the neighbors, X' will basically maintain the original state. Conversely, if the feature in the noise box are similar to those of the neighbors, X' will be weakened through weighting, thereby preventing false-positive detections. Consequently, this module relies solely on the features extracted by the feature extractor, ensuring robust generalization performance. As long as the feature extractor identifies a region with significant differences, it can signal to the lesion detection decoder that this region is highly likely to contain a lesion.

2.3 3D Stripe Attention Module

Fully extracting three-dimensional information from CT images enhances the network's ability to understand the morphological features of lesions. To achieve this, we propose a 3D stripe attention module that employs 1D convolution to model features in long sequences from three directions, enabling efficient and lightweight attention weighting. The realization of this module is shown in Fig. 2.

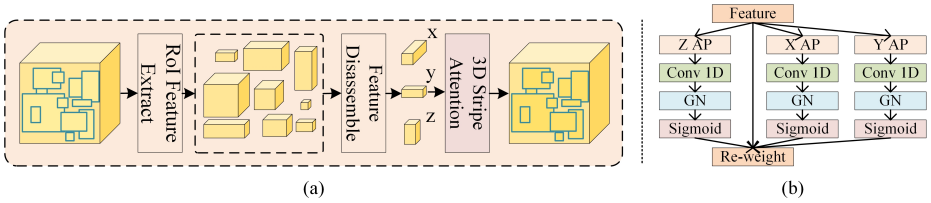

Fig. 2. (a) The flowchart of 3D feature enhancement, (b) the structure of SA module.

For the input image I_{det} and the reference layer I_{ref}, it is first reduced to a 3D image by feature stitching:

$$I_{3D} = concate\left(I_{ref,1}, ..., I_{ref,\frac{h}{2}}, I_{det}, I_{ref,\frac{h}{2}+1}, ..., I_{ref,h}\right) \tag{3}$$

where h denotes the total number of layers of the reference layer, and to ensure that the layer to be detected is in the center of the 3D image, h is taken as an even number. And then the 3D RoI feature $X_{3D} \in \mathbb{R}^{w \times d \times h \times C}$ is taken out from the corresponding feature map of the 3D image according to the 3D RoI composed of noise detection boxes in each layer. To extract the spatial features, stripe pooling in three directions is applied to X_{3D} to disentangle the dimensions and thus aggregate the spatial information from different directions. At this stage, the features in the 3D image are pooled by strips, disassembling the dimension into a sequence of features in three directions to capture the information contained in the long feature sequence. We employ 1D convolution with a kernel size of 3 to extract information in these three directions, thereby enhancing the network's ability to embed the information necessary for 3D target localization. Given that

3D images typically use a small batch size, we implement Group Normalization (GN) to mitigate the reduction in generalization ability caused by changes in data distribution during training. Additionally, we use the Sigmoid function as the activation function to introduce nonlinearity, thereby enhancing the module's learning and expressive capabilities. Finally, the attention weights computed for the different directions are weighted into the original image. To assist the lesion detection decoder in predicting the lesion location, the 3D feature Y_{3D} is disassembled along h and reduced to a 2D feature Y_{det}:

$$\left(Y_{ref,1}, \ldots, Y_{ref,\frac{h}{2}}, Y_{det}, Y_{ref,\frac{h}{2}+1}, \ldots, Y_{ref,h}\right) = split(Y_{3D}, h) \tag{4}$$

3 EXPERIMENTS

3.1 Experimental Setup

Datasset. In this paper, the DeepLesion dataset [24] is used as a validation of the performance of the proposed DetectDiffuse framework on ULD task. The lesions include lung nodules, enlarged lymph nodes, liver tumors, and so on. In this paper, we use the official dataset division method, 70%, 15%, and 15% of the dataset are used for training, validation, and testing, respectively.
Implementation detail. The DetectDiffuse framework was run on Ubuntu 22.02 and trained using a NVIDIA GeForce RTX 3090 GPU. During training, the backbone network ResNet50 is initialized with pretrained weights based on ImageNet-1K dataset, and the rest is initialized with weights using the Xavier initialization method [6]. To optimize the network weights, an AdamW optimizer with an initial learning rate of 2.5×10^{-5} and weight decay of 10^{-4}.
Evaluation criteria. In order to validate the performance of DetectDiffues on ULD task, the results are evaluated using the sensitivity at False Positive Per Image (FPPI) of 0.5 and 1 as well as the mean Average Precision at 50% IoU threshold (mAP@50).

Table 1. Comparison of Universal Lesion Detection (ULD) methods

Methods	3DCE	A3D	MULAN	DSA	DKMA	SATr	DiffULD	Ours
FPPI@0.5	62.48	74.10	76.10	77.38	78.10	81.02	77.84	**81.27**
FPPI@1	73.37	81.81	82.50	84.06	85.26	86.64	84.57	**86.84**
mAP@50	41.23	46.31	46.69	53.11	52.83	52.98	52.66	**54.94**

3.2 Comparison with ULD Methods

To demonstrate the effectiveness of DetectDiffuse in extracting 3D information and capturing lesion features, we compare DetectDiffuse with seven representative ULD methods, namely, 3DCE [22], A3D [25], MULAN [23], DSA-ULD [15],

DKMA-ULD [14], SATr [9] and DiffULD [26]. The quantitative results of the seven methods and DetectDiffuse are shown in Table 1. The experiment results have showed that DetectDiffuse maintains the best performance on all measures.

The qualitative results of the comparison experiment are shown in Fig. 3. In the figure, green boxes indicate correct detection, yellow boxes indicate false positive detection, and red dotted boxes indicate false negative detection.

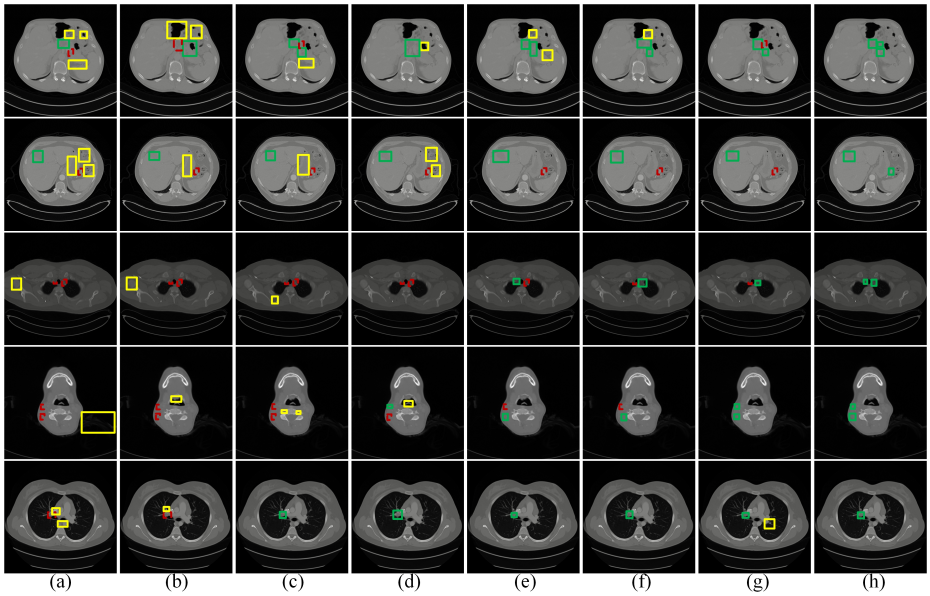

Fig. 3. Qualitative results of ULD Methods. In the figure, (a)-(h) represent 3DCE, A3D, MULAN, DSA-ULD, DKMAULD, SATr, DiffULD and Detectdiffuse respectively.

As a ULD method, we believe that it should have the ability to detect never-before-seen lesions even when faced with them. To this end, we performed zero-shot experiments on four additional datasets: BraTS2021 [1], COVID-19-20 [13], LiTS [2], and Task08 [16]. For the above datasets, we extracted the lesion-containing anatomical slices through registered segmentation masks, precise delineation of pathological regions via bounding box coordinates derived from binary lesion maps and converting it to MS COCO format.

Fig. 4 shows the qualitative results of DetectDiffuse on four zero-shot datasets and the box plots of the sensitivities of egiht ULD methods. It can be found that DetectDiffuse is able to detect most of the lesions with minimal False detection.

3.3 Ablation Experiment

We design ablation experiments to verify the importance of the NA module and the SA module. After removing the above two modules, the baseline method can

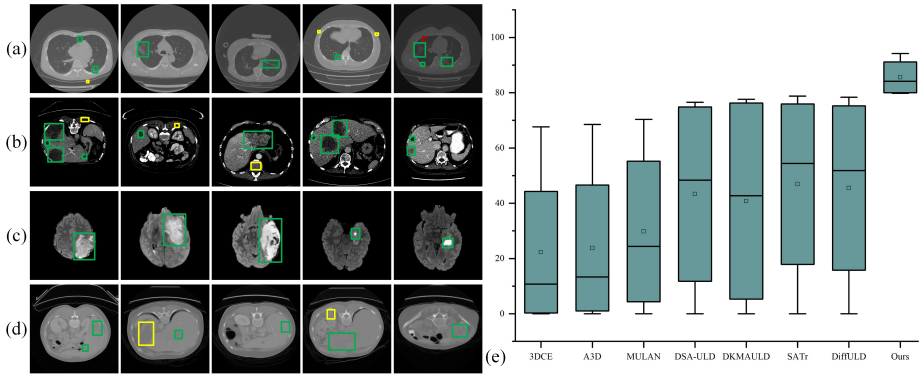

Fig. 4. Zero-shot validation results on different datasets. (a)-(d) denote the qualitative results of DetectDiffuse on COVID-19-20, LiTS, BraTS2021 and Task08 dataset respectively. (e) Boxplots of different methods in zero-shot datasets.

be obtained. The method of ablating the NA module is denoted by W/o NA; and the method of ablating the SA module is denoted by W/o SA. To further validate the ability of the SA module to acquire 3D information, we replaced the module with other lesion detection methods utilizing 3D information, respectively called W/ Conv (using 3D Convolution), W/ Dense (using dense net), and W/ Attn (using multi-head self-attention). The quantitative results of lesion detection in DeepLesion dataset for different ablation methods and DetectDiffuse are shown in Table 2. As can be seen from the table, the proposed method has the best performance, proves the validity of the proposed NA and SA module.

Table 2. Ablation experiment

Methods	Baseline	W/o SA	W/o NA	W/ Conv	W/ Dense	W/ Attn	Ours
FPPI@0.5	74.67	79.35	77.41	75.63	76.98	76.92	**81.27**
FPPI@1	79.91	85.26	84.53	83.19	83.85	84.17	**86.84**
mAP@50	51.17	52.60	52.97	51.47	51.68	52.31	**54.94**

4 Conclusion

In this paper, we find that existing universal lesion detection methods only consider the 3D information in the vertical direction of the CT image, which leads to underutilization of spatial information to produce FN detection. To address this problem, we propose a 3D stripe attention module to fully utilize the spatial information in a dimensional disassembly-weight imposition manner. To address FP detection due to lesion imaging approximating normal tissues, a neighbor box

aggregation module is developed in this paper. The feature aggregation assists the network to understand the difference between the features in the detection box and the surrounding background features to further improve the performance of lesion detection. Experiments on the DeepLesion dataset demonstrate that our method achieves the best generalized lesion detection performance.

Acknowledgments. This study was supported by National Natural Science Foundation of China (62025104, 82330061, 62331005)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
2. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
4. Chen, S., Sun, P., Song, Y., Luo, P.: Diffusiondet: Diffusion model for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19830–19843 (2023)
5. Dou, Q., Chen, H., Yu, L., Qin, J., Heng, P.A.: Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection. IEEE Transactions on Biomedical Engineering **64**(7), 1558–1567 (2016)
6. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
7. He, L., Pan, Y., Jin, W., Tan, R., Xue, Y., Sun, D., Zhang, J., Xiang, P., Fang, Q., Wang, Y., et al.: Soft robots with cy5: An "intake and work" imaging technique for intraoperative navigation of gastric lesion. Cyborg and Bionic Systems **6**, 0212 (2025)
8. Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., Shi, J.: Foveabox: Beyound anchor-based object detection. IEEE Transactions on Image Processing **29**, 7389–7398 (2020)
9. Li, H., Chen, L., Han, H., Kevin Zhou, S.: Satr: Slice attention with transformer for universal lesion detection. In: International conference on medical image computing and computer-assisted intervention. pp. 163–174. Springer (2022)
10. Li, Z., Zhang, S., Zhang, J., Huang, K., Wang, Y., Yu, Y.: Mvp-net: multi-view fpn with position-aware attention for deep universal lesion detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22. pp. 13–21. Springer (2019)

11. Liu, Z., Xie, X., Song, Y., Zhang, Y., Liu, X., Zhang, J., Sheng, V.S.: Mlanet: Multi-layer anchor-free network for generic lesion detection. Engineering Applications of Artificial Intelligence **102**, 104255 (2021)
12. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1137–1149 (2016)
13. Roth, H.R., Xu, Z., Tor-Díez, C., Jacob, R.S., Zember, J., Molto, J., Li, W., Xu, S., Turkbey, B., Turkbey, E., et al.: Rapid artificial intelligence solutions in a pandemic—the covid-19-20 lung ct lesion segmentation challenge. Medical image analysis **82**, 102605 (2022)
14. Sheoran, M., Dani, M., Sharma, M., Vig, L.: Dkma-uld: domain knowledge augmented multi-head attention based robust universal lesion detection. arXiv preprint arXiv:2203.06886 (2022)
15. Sheoran, M., Dani, M., Sharma, M., Vig, L.: An efficient anchor-free universal lesion detection in ct-scans. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–4. IEEE (2022)
16. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
17. Sun, L., Wang, J., Huang, Y., Ding, X., Greenspan, H., Paisley, J.: An adversarial learning approach to medical image synthesis for lesion detection. IEEE journal of biomedical and health informatics **24**(8), 2303–2314 (2020)
18. Tang, Y.B., Yan, K., Tang, Y.X., Liu, J., Xiao, J., Summers, R.M.: Uldor: a universal lesion detector for ct scans with pseudo masks and hard negative example mining. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). pp. 833–836. IEEE (2019)
19. Townsend, D.W., Carney, J.P., Yap, J.T., Hall, N.C.: Pet/ct today and tomorrow. Journal of Nuclear Medicine **45**(1 suppl), 4S–14S (2004)
20. Wang, B., Qi, G., Tang, S., Zhang, L., Deng, L., Zhang, Y.: Automated pulmonary nodule detection: High sensitivity with few candidates. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 759–767. Springer (2018)
21. Wang, X., Han, S., Chen, Y., Gao, D., Vasconcelos, N.: Volumetric attention for 3d medical image segmentation and detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22. pp. 175–184. Springer (2019)
22. Yan, K., Bagheri, M., Summers, R.M.: 3d context enhanced region-based convolutional neural network for end-to-end lesion detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. pp. 511–519. Springer (2018)
23. Yan, K., Tang, Y., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., Summers, R.M.: Mulan: multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 194–202. Springer (2019)
24. Yan, K., Wang, X., Lu, L., Summers, R.M.: Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of medical imaging **5**(3), 036501–036501 (2018)

25. Yang, J., He, Y., Kuang, K., Lin, Z., Pfister, H., Ni, B.: Asymmetric 3d context fusion for universal lesion detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 571–580. Springer (2021)
26. Zhao, P., Li, H., Jin, R., Zhou, S.K.: Diffuld: diffusive universal lesion detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 94–105. Springer (2023)
27. Zhao, W., Jiang, D., Queralta, J.P., Westerlund, T.: Mss u-net: 3d segmentation of kidneys and tumors from ct images with a multi-scale supervised u-net. Informatics in Medicine Unlocked **19**, 100357 (2020)
28. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
29. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)