MedCenterDet: A Center-based Object Detection Framework for 3D Medical Image

Qiang Zeng[1,2], Ao Shen[4], Xiangtong Du[5], Xiaoyu Xu[1,2], Shaohua Zhi[1,2], Wufeng Xue[6], and Fei Pan[1,2,3]

[1] Wu Jieh Yee School of Interdisciplinary Studies, Lingnan University, Hong Kong SAR, China; fei.pan@ln.edu.hk

[2] Lingnan University Shenzhen Research Institute, Shenzhen, 518063, Guangdong, China

[3] Hong Kong Centre for Cerebro-cardiovascular Health Engineering, Hong Kong SAR, China; fpan@hkcoche.org

[4] College of Information Science and Engineering, Hohai University, Changzhou, 213200, Jiangsu, China

[5] School of Medical Imaging, Xuzhou Medical University, Xuzhou, 221004, Jiangsu, China

[6] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Guangdong, China

Abstract. Precise object localization in three-dimensional (3D) medical imaging is critical for clinical diagnosis, yet voxel-level segmentation remains labor-intensive and sensitive to boundary ambiguity. While object detection offers a more efficient alternative, current methods heavily rely on predefined anchors, which require extensive hyperparameter tuning and exponentially increase computational complexity within 3D search spaces. In this paper, we propose MedCenterDet, a novel anchor-free, center-based object detection framework tailored for 3D medical images. Our approach directly predicts object centers via heatmap regression and analytically decodes bounding box dimensions using an anisotropic Gaussian target generation strategy, inherently capturing morphological shape priors. To enhance feature discriminative power in dense 3D volumes, we introduce a counterfactual attention learning (CAL) module. By contrasting factual target-structure regions with counterfactual noise interventions, this module explicitly optimizes attention quality, forcing the network to focus on clinically relevant patterns. A subsequent two-stage refinement module leverages multi-scale features to correct spatial misalignments and ensure geometric precision. We comprehensively evaluated MedCenterDet on two public datasets. Experimental results demonstrate that our framework achieves state-of-the-art performance across both benchmarks, while maintaining a significantly lower parameter count compared to established methods. The code is available at https://github.com/WiDayn/MedCenterDet.

Keywords: center-based object detection · medical imaging · deep learning · anchor-free detection

1 Introduction

In medical image analysis, precise target localization is a prerequisite for numerous clinical applications. While segmentation provides pixel-level masks, annotating 3D volumes is labor-intensive and suffers from inter-observer variability due to the infiltrative boundaries of lesions [9,11]. Object detection offers an efficient alternative by requiring only bounding boxes. Such boxes and center points can also serve as compact region-of-interest descriptors for downstream segmentation, classification, and longitudinal follow-up, reducing the search space of subsequent analysis. However, prevalent anchor-based methods (e.g., nnDetection [3], MedYOLO [19]) rely heavily on predefined anchor hyperparameters. In 3D data, applying multiple anchors per voxel exponentially increases the search space, computational burden, and optimization difficulty.

Anchor-free methods eliminate manual tuning [20,4] but face severe convergence and regression challenges within the dense, unstructured 3D search space. Recent works mitigate this by reintroducing anchor mechanisms [23] or utilizing restrictive geometric priors (e.g., spheres) [15], which limits their flexibility across morphologically diverse targets. More recent 3D medical detection studies have explored transformer-based query refinement for lymph nodes [21], efficient 3D attention for lung nodule detection [1], and self-training for universal lesion detection in computed tomography (CT) [7]. These advances highlight the value of 3D context modeling, but many still rely on task-specific query, attention, or proposal-mining designs.

To achieve precise localization without the optimization burden of vast search spaces or rigid priors, we propose MedCenterDet, a novel anchor-free framework. Inspired by the center-based paradigm [6,25], it directly predicts target centers and decodes size information via heatmap regression. By restricting feature extraction to predicted centers, this approach resolves the 3D search space dilemma and inherently yields clinically valuable center point coordinates.

In summary, our contributions are: (1) We introduce MedCenterDet, an anchor-free center-based 3D detection framework that effectively mitigates dense search space challenges without predefined anchors. (2) We propose an anisotropic Gaussian target generation method to explicitly encode and decode object spatial extents from heatmaps. (3) We design a CAL module to explicitly optimize attention quality and heatmap prediction. (4) Extensive experiments demonstrate that MedCenterDet achieves state-of-the-art (SOTA) results on public 3D datasets with fewer parameters.

2 Methodology

Feature representation learning To establish a robust feature representation for 3D object detection, we first employ a 3D backbone network to extract multi-scale features from the input volume, denoted as S_2 through S_5. These symbols denote four pyramid stages with progressively lower spatial resolution and richer semantics; the corresponding fused feature levels are denoted as N_2

through N_5. To enhance these features and integrate high-level semantics with low-level details, we directly adopt the channel attention (CA) and selective feature fusion (SFF) modules from [5], as shown in Fig. 1(f, d).

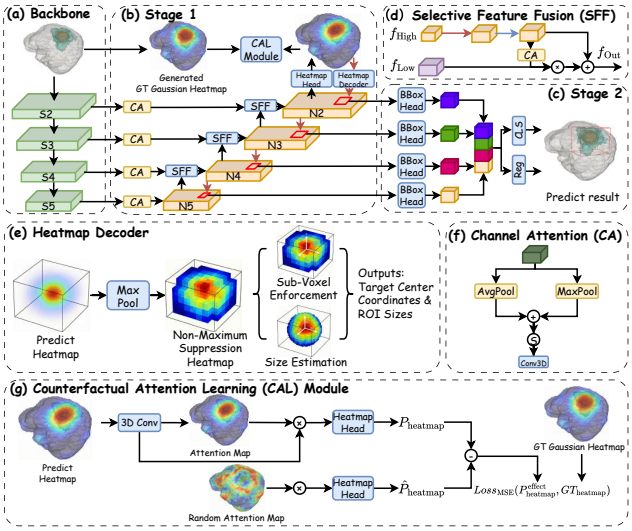

Fig. 1: Overview of MedCenterDet. The backbone and Stage 1 fuse multi-scale features with the CA/SFF modules and predict center heatmaps, while the CAL module regularizes attention with a random attention intervention. The heatmap decoder generates coarse proposals, and Stage 2 refines their confidence scores and 3D boxes.

The overall workflow is stage-level sequential: an input volume is encoded by the backbone, Stage 1 fuses multi-scale features and predicts center heatmaps, the decoder converts the top-K peaks from each of the C channels into $K \times C$ coarse proposals, and Stage 2 refines these proposals in parallel.

Drawing inspiration from the counterfactual attention learning framework proposed by Rao *et al.* [17], we incorporate a 3D CAL module (Fig. 1(g)) to explicitly supervise the attention distribution of our model. The CAL module guides the network to focus on factual target-structure regions by leveraging counterfactual causality to quantify and optimize attention quality. Specifically, given the input feature maps F and the factual attention map A derived via a 3D convolution layer, we employ an intervention operation $do(\cdot)$ to generate counterfactual attention maps $\bar{A}$ by replacing the factual weights with random noise from a uniform distribution. Following Pearl's intervention notation, $do(A = \bar{A})$ means that the learned dependency between the input feature and attention map is cut and the attention map is externally set to random weights. The final effective heatmap prediction $P^{\text{effect}}_{\text{heatmap}}$ represents the counterfactual-inspired attention regularization of the attention mechanism, defined as the difference

between the factual prediction and the counterfactual outcome, as shown in Eq. (1).

$$P_{\text{heatmap}}^{\text{effect}} = \mathcal{H}(F \odot A) - \mathcal{H}\left(F \odot do(A = \bar{A})\right), \tag{1}$$

where $\mathcal{H}(\cdot)$ denotes the operation of the heatmap head, and $\odot$ represents element-wise multiplication. By subtracting the counterfactual prediction from the factual prediction, $P_{\text{heatmap}}^{\text{effect}}$ isolates the contribution of the learned attention. Finally, to force the model to distinguish between meaningful target patterns and random noise, we optimize the network by minimizing the mean squared error (MSE) loss between this effective prediction and the ground truth Gaussian heatmap G. The loss function is defined as shown in Eq. (2). During inference, the trained attention layer is directly utilized to modulate feature maps, thereby incurring negligible additional parameters and computational overhead.

$$\mathcal{L}_{\text{CAL}} = \left| P_{\text{heatmap}}^{\text{effect}} - G_{\text{heatmap}} \right|_2^2. \tag{2}$$

Anisotropic heatmap generation and decoding To model 3D geometric variations, we represent target structures as anisotropic Gaussian distributions where axial spreads correlate with physical dimensions. For a ground truth box centered at (c_x, c_y, c_z), the target value for voxel (x, y, z) is generated via Eq. (3).

$$Y_{x,y,z} = \exp\left(-\left(\frac{(x-c_x)^2}{2\sigma_x^2} + \frac{(y-c_y)^2}{2\sigma_y^2} + \frac{(z-c_z)^2}{2\sigma_z^2}\right)\right), \tag{3}$$

where $(\sigma_x, \sigma_y, \sigma_z)$ are the object-adaptive standard deviations [12].

To transform the dense heatmap predictions into a sparse set of continuous geometric parameters, we design a differentiable decoding strategy inspired by anchor-free detection paradigms [25]. First, we identify coarse integer peaks $\mathbf{v} = (z, y, x)$ using a 3D Max-Pooling operation acting as an non-maximum suppression (NMS) filter [25,12]. Second, to mitigate quantization errors inherent to the discrete grid, we recover the sub-voxel center $\hat{\boldsymbol{\mu}}$ by applying a discrete central difference approximation around the local maximum [24], efficiently estimating the continuous peak offset as shown in Eq. (4). Third, we estimate the anisotropic spread $\hat{\boldsymbol{\sigma}}$ via the weighted second central moment within a local window $\mathcal{W}$, formulated in Eq. (5). The derived $\hat{\boldsymbol{\sigma}}$ serves as a high-recall, coarse geometric prior for the second-stage refinement to crop the target region of interests (RoIs).

$$\hat{\mu}_d = v_d + \text{clip}\left(\frac{\mathbf{H}(\mathbf{v} + \mathbf{e}_d) - \mathbf{H}(\mathbf{v} - \mathbf{e}_d)}{2}, -0.5, 0.5\right), \quad \forall d \in \{x, y, z\} \tag{4}$$

$$\hat{\sigma}_d = \sqrt{\frac{\sum_{\mathbf{r} \in \mathcal{W}} \mathbf{H}(\mathbf{v} + \mathbf{r}) \cdot r_d^2}{\sum_{\mathbf{r} \in \mathcal{W}} \mathbf{H}(\mathbf{v} + \mathbf{r}) + \epsilon}}, \quad \forall d \in \{x, y, z\} \tag{5}$$

Here, $\mathbf{e}_d$ is the unit vector along axis d, $\mathbf{r}$ indexes offsets inside the local window $\mathcal{W}$, and r_d is the d-th component of $\mathbf{r}$.

Second-stage refinement We employ a second-stage refinement module to regress the final precise bounding boxes. Leveraging the coarse target centers $\hat{\boldsymbol{\mu}}$ derived from the decoding process, we project these 3D coordinates onto the multi-level feature pyramid $\{N_2, N_3, N_4, N_5\}$ to index corresponding feature vectors via trilinear interpolation. These multi-scale features are processed by independent Refinement Heads and subsequently aggregated via a final Conf/Reg Module (Fig. 1(d)). The total loss $\mathcal{L}_{\text{total}}$ is defined in (Eq. (6)).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{heatmap}} + \mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{cal}} \tag{6}$$

Here, $\mathcal{L}_{\text{heatmap}}$ computes the MSE between the predicted and ground truth heatmaps, i.e., $\mathcal{L}_{\text{heatmap}} = \|P_{\text{heatmap}} - G_{\text{heatmap}}\|_2^2$ for the raw heatmap branch, whereas $\mathcal{L}_{\text{cal}} = \|P_{\text{heatmap}}^{\text{effect}} - G_{\text{heatmap}}\|_2^2$ supervises the counterfactual effective prediction. $\mathcal{L}_{\text{conf}}$ employs Focal Loss ($\eta = 2.0$) [13] for proposal confidence p_t from the CLS/confidence branch. $\mathcal{L}_{\text{reg}}$ optimizes geometric precision using a combination of Smooth L_1 [8] and Generalized intersection over union (IoU) Loss [18], weighted by 10 and 1, respectively. Finally, $\mathcal{L}_{\text{cal}}$ imposes explicit supervision on the counterfactual attention module (Eq. (2)).

3 Experiments

Data preparation We evaluated our method on two public datasets. From the brain tumor segmentation (BraTS) 2021 Task 1 dataset [2,16], we utilized 1,251 fast fluid attenuated inversion recovery (FLAIR) magnetic resonance (MR) scans. From the Total Segmentator dataset [22], we utilized 1,430 CT scans, from which we extracted annotations for liver, bladder, and kidneys. We extracted the minimum 3D bounding box that encompasses all combined masks for BraTS, and derived individual boxes for each label in Total Segmentator (merging both kidneys into a single class). All input volumes for our proposed method employed conventional 3D data augmentation techniques. For all comparative baselines, we adhered to their respective recommended data processing pipelines. All baselines were run with their public implementations and recommended/default configurations unless a parameter was required to match the same input resolution and evaluation protocol. A five-fold cross-validation strategy was applied to both datasets.

Experimental setup and evaluation metrics All experiments were conducted using a single NVIDIA GeForce RTX 3090 GPU with 24 GB of video memory. The training batch size was set to 4 for the majority of the models. For the ResNet-50 variant, the batch size was adjusted to 3 to accommodate its higher GPU memory requirements. For network optimization, we employed the AdamW optimizer with an initial learning rate of 1×10^{-3}. To dynamically adjust the learning rate during training, we utilized the ReduceLROnPlateau scheduler with a reduction factor of 0.7 and a patience of 25 epochs. The training process was configured to stop when the learning rate decayed to a minimum threshold of 1×10^{-6}.

We adopt the standard common objects in context (COCO)-style metrics [14], including average precision $(AP)_{50}$, AP_{75}, and $mAP_{50:95}$. To assess robustness across object sizes, we stratified instances into three equipopulous categories based on volume quantiles of the ground truth. Specifically, ground-truth box volume $V = w \times h \times d$ was sorted within each dataset and split by the 33.3% and 66.7% quantiles into Bottom, Middle, and Top groups; AP_{50} and average recall (AR) were then computed on each group using the same COCO-style matching protocol.

Implementation details For the ground truth generation, the bounding box dimensions (w, h, d) are mapped to the Gaussian standard deviations via $\sigma_i = \max(s_i/3, 3)$ for $s_i \in \{w, h, d\}$, ensuring the Gaussian response accurately reflects the object boundary. The multi-class heatmaps are organized as a C-channel output tensor with a sigmoid activation, where C represents the number of target categories. During the decoding phase, a $5 \times 5 \times 5$ Max-Pooling operation serves as the NMS filter. For the moment-based size estimation in Eq. (5), the local window $\mathcal{W}$ is defined as a $7 \times 7 \times 7$ neighborhood. Finally, we select the Top-K ($K = 10$) candidate centers independently from each of the C channels, resulting in a total of $K \times C$ proposals.

Table 1: Comparison with existing approaches on BraTS 2021 and Total Segmentator. Methods marked with † denote variants without the CAL module. Best results: first , second , third (mean ± std, best in bold).

Method	**Para.**	**BraTS 2021**			**Total Seg.**		
		AP_{50}	AP_{75}	$mAP_{50:95}$	AP_{50}	AP_{75}	$mAP_{50:95}$
nnDetection [3]	19.4M	79.3±2.0	33.7±2.5	37.2±2.3	35.5±1.0	17.5±2.5	18.3±1.5
RetinaNet [10]	51.1M	86.5±1.5	45.5±2.1	46.1±1.4	52.6±1.2	25.7±2.4	24.3±1.5
Focused Dec. [23]	43.4M	80.1±1.8	34.7±1.7	32.9±2.0	68.5±1.1	37.4±1.6	38.4±1.8
MedYOLO-S [19]	15.7M	85.9±1.7	42.9±2.3	43.2±1.7	48.2±1.6	24.5±1.3	25.6±0.5
-M [19]	49.6M	80.5±2.4	33.5±2.2	34.8±1.8	46.2±0.6	24.2±1.7	25.2±1.3
-L [19]	113.5M	0.1±0.5	0.00±0.0	0.00±0.0	64.3±1.3	29.9±1.0	32.7±0.7
Ours-EffNet-S†	2.4M	88.2±2.6	41.5±1.6	39.9±2.2	37.9±2.2	12.7±2.1	16.5±1.7
Ours-EffNet-L†	3.4M	90.3±2.0	50.0±1.2	45.3±0.9	42.8±2.0	27.0±2.0	24.3±1.7
Ours-Res18†	39.7M	90.3±0.6	50.6±2.4	50.6±1.9	75.1±2.7	45.4±0.8	42.1±1.6
Ours-Res50†	135.7M	90.1±1.6	52.9±1.1	51.0±0.5	76.4±0.6	45.9±2.0	43.1±0.7
Ours-EffNet-S	2.4M	90.0±1.9	51.7±2.1	47.2±1.1	47.3±2.2	16.3±1.6	21.5±1.3
Ours-EffNet-L	3.4M	90.6±1.7	**55.9±0.7**	**51.4±0.7**	**78.3±1.3**	40.8±1.2	41.3±1.4
Ours-Res18	39.7M	89.8±2.4	54.8±1.6	49.8±2.5	76.4±1.6	47.0±1.2	42.6±1.6
Ours-Res50	135.7M	**90.9±1.8**	54.5±2.0	51.1±0.7	78.3±1.3	**47.7±1.2**	**43.8±1.9**

Comparison with existing methods and ablation study We compare our approach against several representative 3D medical object detection frameworks: nnDetection [3], RetinaNet [10], Focused Dec. [23], and the MedYOLO series [19]. As shown in Table 1, our method consistently outperforms the baseline approaches across both benchmarks.

On the BraTS 2021 dataset, our lightweight variant, EffNet-L, achieves the highest performance, slightly surpassing even our heavier ResNet-50 counterpart. This suggests that for tasks with relatively homogeneous targets, the compact EffNet-L provides sufficient model capacity while mitigating the risk of overfitting associated with over-parameterized models. Consistent with [19], the MedYOLO-L failed to converge on this dataset, suggesting that large-scale models may be missing pertinent information. On the Total Segmentator dataset, which involves high anatomical diversity and dense multi-class targets, scaling up to the ResNet-50 backbone yields superior accuracy.

Table 2: Scale-stratified performance (AP_{50} and AR) on BraTS 2021 and Total Segmentator by ground-truth volume tertiles. Best results are highlighted as first, second, and third.

Method	**Bottom**		**Middle**		**Top**	
	AP_{50}	AR	AP_{50}	AR	AP_{50}	AR
BraTS 2021 [2]						
nnDetection [3]	74.1±3.5	76.5±4.1	83.1±1.3	82.4±1.7	80.6±1.8	81.1±2.2
Retina Net [10]	81.0±2.9	83.5±3.5	90.2±0.9	89.6±1.2	87.6±1.4	88.2±1.8
Focused Dec. [23]	75.3±3.1	77.9±3.9	84.2±1.5	83.8±1.7	81.7±2.0	82.1±2.2
MedYOLO-S [19]	78.2±3.0	80.7±3.6	88.0±1.1	87.6±1.5	85.1±1.7	85.7±1.9
-M [19]	75.1±3.7	77.4±4.3	84.1±1.4	83.7±1.6	82.1±1.8	82.4±2.1
-L [19]	0.00±0.0	0.00±0.0	0.30±0.8	0.50±0.9	0.00±0.0	0.00±0.0
Ours-EffNet-S	84.1±2.1	86.8±2.6	94.5±1.3	93.9±0.9	91.0±2.0	90.6±2.3
Ours-EffNet-L	84.1±1.6	86.8±1.3	94.6±1.4	**96.3±2.0**	**92.9±1.5**	**91.8±1.8**
Ours-Res18	84.8±2.7	**92.8±2.4**	89.6±1.9	**96.3±0.6**	88.9±1.0	83.5±1.7
Ours-Res50	**85.7±2.5**	91.6±1.3	**96.4±2.1**	95.1±1.4	88.5±0.7	**91.8±1.5**
Total Segmentator [22]						
nnDetection [3]	17.5±4.2	32.6±4.9	42.8±1.9	48.0±2.2	37.7±2.1	42.5±2.5
Retina Net [10]	27.7±3.9	42.9±4.5	60.1±1.5	62.7±2.0	55.5±1.8	60.8±2.4
Focused Dec. [23]	34.8±3.6	57.7±4.2	76.2±1.2	80.4±1.6	71.1±1.8	77.9±2.1
MedYOLO-S [19]	24.2±4.3	37.8±5.0	55.2±1.7	60.1±2.2	51.1±2.1	56.5±1.8
-M [19]	22.9±4.6	35.6±5.1	53.2±1.7	58.2±2.3	49.4±2.1	54.8±2.7
-L [19]	32.2±4.5	49.9±5.0	74.0±1.5	80.7±2.0	68.2±2.0	75.0±2.5
Ours-EffNet-S	23.2±3.8	39.5±4.5	54.1±1.6	60.9±1.9	50.3±2.0	57.7±2.3
Ours-EffNet-L	38.7±1.9	64.5±1.1	**85.7±0.5**	**91.5±1.5**	79.8±0.5	92.8±1.6
Ours-Res18	36.1±2.5	67.5±1.2	83.4±2.5	90.5±0.6	82.1±0.4	**93.2±0.4**
Ours-Res50	**39.9±2.6**	**68.0±2.6**	80.5±2.9	87.0±2.0	**82.9±1.0**	92.8±2.1

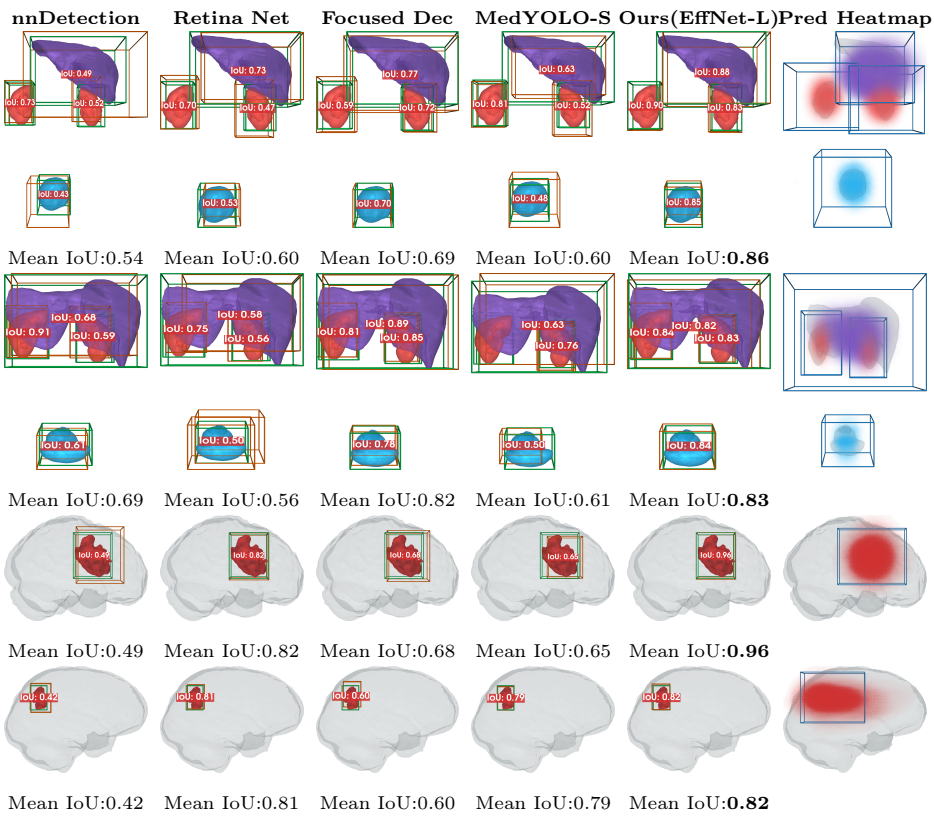

Fig. 2: Qualitative comparison of different detection methods on the Total Segmentator [22] (top two rows) and BraTS 2021 [2] (bottom two rows) datasets. Ground truth and predicted bounding boxes are denoted in green and red, respectively. The last column visualizes our first-stage heatmap predictions alongside the subsequently decoded RoIs (blue boxes). These multi-channel heatmaps explicitly indicate the predicted centers for **kidney**, **bladder**, and **liver** for the former dataset, and **brain tumors** for the latter.

Table 2 reveals MedCenterDet's distinct superiority in detecting relatively smaller structures. On the Total Segmentator dataset, our models demonstrate a substantial margin of improvement over baseline methods in both precision and average recall for relatively smaller objects. This fine-grained precision is structurally driven by the proposed CAL module. As evidenced by the ablation study in Table 1, integrating CAL yields dramatic performance leaps, particularly in enhancing the discriminative power of our lightweight configurations. For our highly compact EffNet-L, CAL provides crucial explicit guidance that prevents the limited-capacity network from being overwhelmed by complex 3D background noise, driving a massive performance surge. Conversely, the over-parameterized ResNet-50 inherently possesses sufficient representational capacity to implicitly

learn robust, noise-resistant features. As a result, the explicit regularization provided by CAL yields diminishing returns on the heavier backbone.

Visual comparisons in Fig. 2 further demonstrate the localization precision of our approach. Across both datasets, MedCenterDet yields predicted bounding boxes that consistently achieve the highest mean IoU against the ground truth; the last column illustrates that the heatmap effectively localizes the target centers, from which RoIs are directly decoded to constrain and guide the final high-precision bounding box regression.

4 Conclusion

In this paper, we propose MedCenterDet, an anchor-free 3D medical object detection framework designed to overcome the computational overhead and rigid hyperparameter dependency of traditional anchor-based methods. By directly predicting object centers and dimensions via anisotropic Gaussian heatmap regression, and introducing a CAL module to suppress background noise, our method effectively enhances the discriminative power of target-structure features for precise, multi-scale two-stage bounding box calibration. Future work will explore the framework's effectiveness in resolving highly dense or overlapping targets, alongside its generalizability across broader clinical imaging modalities.

Acknowledgments This work was supported by the Shenzhen University–Lingnan University Joint Research Programme (105026), the Lam Woo Research Fund (185611) at Lingnan University, the National Natural Science Foundation of China (62471313) and Guangdong Natural Science Foundation (2024A1515030143) at Shenzhen University.

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

References

1. Almahasneh, M., Xie, X., Paiement, A.: AttentNet: Fully convolutional 3D attention for lung nodule detection. arXiv preprint arXiv:2407.14464 (2024)
2. 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)
3. Baumgartner, M., Jäger, P.F., Isensee, F., Maier-Hein, K.H.: nnDetection: A self-configuring method for medical object detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 530–539. Springer (2021)
4. 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)

5. Chen, Y., Zhang, C., Chen, B., Huang, Y., Sun, Y., Wang, C., Fu, X., Dai, Y., Qin, F., Peng, Y., et al.: Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases. Computers in Biology and Medicine **170**, 107917 (2024)
6. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: CenterNet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6569–6578 (2019)
7. Frazier, J., Mathai, T.S., Liu, J., Paul, A., Summers, R.M.: 3D universal lesion detection and tagging in CT with self-training. arXiv preprint arXiv:2504.05201 (2025)
8. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1440–1448 (2015)
9. Hesamian, M.H., Jia, W., He, X., Kennedy, P.: Deep learning techniques for medical image segmentation: Achievements and challenges. Journal of Digital Imaging **32**(4), 582–596 (2019)
10. Jaeger, P.F., Kohl, S.A., Bickelhaupt, S., Isensee, F., Kuder, T.A., Schlemmer, H.P., Maier-Hein, K.H.: Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. In: Machine Learning for Health Workshop. pp. 171–183. PMLR (2020)
11. Joskowicz, L., Cohen, D., Caplan, N., Sosna, J.: Inter-observer variability of manual contour delineation of structures in CT. European Radiology **29**(3), 1391–1399 (2019)
12. Law, H., Deng, J.: CornerNet: Detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 734–750 (2018)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2980–2988 (2017)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision. pp. 740–755. Springer (2014)
15. Luo, X., Song, T., Wang, G., Chen, J., Chen, Y., Li, K., Metaxas, D.N., Zhang, S.: SCPM-Net: An anchor-free 3D lung nodule detection network using sphere representation and center points matching. Medical Image Analysis **75**, 102287 (2022)
16. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BraTS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2014)
17. Rao, Y., Chen, G., Lu, J., Zhou, J.: Counterfactual attention learning for fine-grained visual categorization and re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1025–1034 (2021)
18. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 658–666 (2019)
19. Sobek, J., Medina Inojosa, J.R., Medina Inojosa, B.J., Rassoulinejad-Mousavi, S., Conte, G.M., Lopez-Jimenez, F., Erickson, B.J.: MedYOLO: A medical image object detection framework. Journal of Imaging Informatics in Medicine **37**(6), 3208–3216 (2024)

20. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9627–9636 (2019)
21. Wang, Y., Yu, Q., Yan, K., Li, H., Guo, D., Zhang, L., Lu, L., Shen, N., Wang, Q., Ding, X., Ye, X., Jin, D.: LN-DETR: Effective lymph nodes detection in CT scans using location debiased query selection and contrastive query representation in transformer. arXiv preprint arXiv:2404.03819 (2024)
22. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
23. Wittmann, B., Navarro, F., Shit, S., Menze, B.: Focused decoding enables 3D anatomical detection by transformers. arXiv preprint arXiv:2207.10774 (2022)
24. Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C.: Distribution-aware coordinate representation for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7093–7102 (2020)
25. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)