The Missing Piece: A Case for Pre-Training in 3D Medical Object Detection

Katharina Eckstein[⋆,1,2,3], Constantin Ulrich[⋆,1,2], Michael Baumgartner[⋆⋆,1,4,5], Jessica Kächele[1,2,3], Dimitrios Bounias[1,2], Tassilo Wald[1,4,5], Ralf Floca[1,6], and Klaus H. Maier-Hein[1,2,3,4,5,7,8]

[1] German Cancer Research Center (DKFZ), Division of Medical Image Computing, Heidelberg, Germany
[2] Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany
[3] German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Core Center Heidelberg, Germany
[4] Helmholtz Imaging, DKFZ, Heidelberg, Germany
[5] Faculty of Mathematics and Computer Science, Heidelberg University, Germany
[6] Heidelberg Institute of Radiation Oncology (HIRO), National Center for Radiation Research in Oncology (NCRO), Heidelberg, Germany
[7] Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany
[8] National Center for Tumor Diseases (NCT) Heidelberg, Heidelberg, Germany
{katharina.eckstein,constantin.ulrich}@dkfz-heidelberg.de

Abstract. Large-scale pre-training holds the promise to advance 3D medical object detection, a crucial component of accurate computer-aided diagnosis. Yet, it remains underexplored compared to segmentation, where pre-training has already demonstrated significant benefits. Existing pre-training approaches for 3D object detection rely on 2D medical data or natural image pre-training, failing to fully leverage 3D volumetric information. In this work, we present the first systematic study of how existing pre-training methods can be integrated into state-of-the-art detection architectures, covering both CNNs and Transformers. Our results show that pre-training consistently improves detection performance across various tasks and datasets. Notably, reconstruction-based self-supervised pre-training outperforms supervised pre-training, while contrastive pre-training provides no clear benefit for 3D medical object detection. Our code is publicly available at: https://github.com/MIC-DKFZ/nnDetection-finetuning.

Keywords: 3D Object Detection· Self-Supervised Learning· Pre-training

1 Introduction

Accurate detection of anatomical structures and abnormalities in 3D medical imaging is crucial for reliable diagnosis and clinical decision-making. Unlike segmentation, which provides detailed structural delineation, detection focuses on

⋆ Equal contribution; co-first author order may be adjusted for individual use.
⋆⋆ Work done while at DKFZ, now at Siemens Healthineers.

localizing clinically relevant objects. Critically, detection excels in clinically relevant metrics, especially in high-stakes scenarios where completely missing an object can have far more severe consequences than minor inaccuracies in pixel-wise delineation [24]. Despite its clinical importance, research on 3D object detection has received significantly less attention than segmentation, as evidenced by medical image analysis challenges predominantly emphasizing segmentation tasks [23].

Recent advancements in 3D medical image segmentation have spurred interest in large-scale pre-training. For instance, Ulrich et al. introduced Multitalent [31], a framework that enables supervised training across multiple segmentation datasets. Moreover, self-supervised pre-training strategies [32,39,34,30,13] have demonstrated promising results for segmentation applications. Likewise, detection models might particularly benefit from pre-training due to the typically small size of annotated datasets and their tendency to over focus on local image features, rather than leveraging broader contextual information. However, despite the advancements in pre-training for segmentation, the impact of purely 3D large-scale pre-training remains unexplored for 3D object detection.

This gap was also acknowledged in one of the most recent and comprehensive studies on 3D medical object detection by Baumgartner et al., who extensively revised the nnDetection framework [5,4]. While their work made significant contributions to the field, it did not address the potential role of large-scale pre-training. Yet, beyond their work, research on pre-training strategies for medical object detection is virtually nonexistent, with only a handful of studies even touching upon this direction. Existing pre-training strategies for medical object detection have predominantly focused on 2D data, utilizing either natural image pre-training [35] or 2D medical data [20,10,26,6,21]. This is largely due to the scarcity of publicly available 3D object detection datasets with sufficient cases for effective pre-training. To partially capture 3D context, some methods extend pre-trained 2D models by integrating adjacent slices. This includes using ImageNet-pretrained backbones with 3D context slices added at the downstream stage [33,36] or pseudo-3D approaches that treat image channels (e.g., RGB) as separate slices during pre-training [38]. Another strategy relies on video-based pre-training, where adjacent frames are used to simulate the sequential nature of medical slices [1]. Notably, no prior study has systematically explored large-scale 3D pre-training for 3D medical object detection.

To bridge this gap, we present a comprehensive study evaluating the impact of different large-scale pre-training strategies on 3D medical object detection. Specifically, our key contributions include:

1. **The First Comprehensive Study on Pre-Training Paradigms for 3D Object Detection** to analyze the impact of both supervised and self-supervised large-scale pre-training for 3D medical object detection across eight diverse downstream detection datasets.
2. **Evaluation Across Detection Architectures:** Model performance varies widely depending on dataset characteristics and annotation types, such as bounding boxes or segmentation masks, as demonstrated by Baumgartner

et al. [4]. To assess the generalization of pre-training strategies, we examine their transferability to two state-of-the-art detection models, Retina U-Net [17,4] and Deformable DETR [40,4], covering both CNN and Transformer.
3. **Comparison of Pre-Training Architectures for Pre-training:** ResEncL, a state-of-the-art model for semantic segmentation [16] that has shown improved downstream performance with self-supervised pre-training [32], and an adapted version of Retina U-Net, allowing both segmentation pre-training as well as downstream 3D object detection fine-tuning.

2 Methods

In this study, we evaluate the impact of large-scale pre-training on 3D medical object detection using two state-of-the-art architectures: Retina U-Net and Deformable DETR. Notably, both architectures are specifically designed for detection and cannot be directly applied to other tasks without modifications. Therefore, for pre-training, we adapt Retina U-Net for supervised segmentation and employ the state-of-the-art ResEncL model for both supervised and self-supervised learning [32,16]. We then transfer only the pre-trained backbone from these models to the detection networks for downstream fine-tuning, as visualized in fig. 1. Our experimental setup involves five development datasets and three independent testing datasets. The development datasets are employed to systematically investigate various fine-tuning strategies, enabling us to identify optimal approaches for adapting pre-trained models to 3D medical object detection.

2.1 3D Object Detection

Retina U-Net [17] is a single-stage, anchor-based object detector enhanced with semantic segmentation supervision. Its architecture extends the Feature Pyramid Network (FPN) of RetinaNet with additional high-resolution levels in the FPN's top-down pathway to support an auxiliary segmentation task, creating a U-Net-like symmetric structure (U-FPN), as visualized in fig. 1. The detection head, applied to the final four or five resolution levels, consists of a classification and a regression branch. The regression branch uses smooth L1 loss, while the classification branch employs binary focal loss. Segmentation is supervised with a combined cross-entropy and batch Dice loss function.

Deformable DETR [40] is a two-stage transformer-based detection architecture. In contrast to traditional DEtection TRansformer (DETR), Deformable DETR replaces global self-attention with a sparse deformable attention mechanism, significantly reducing computational complexity and enhancing efficiency by focusing on a small set of queries per attention operation. Additionally, Deformable DETR introduces iterative bounding box refinement, progressively updating the bounding boxes instead of predicting them from scratch. As visualized in fig. 1, Deformable DETR contains an encoder network as a first component to

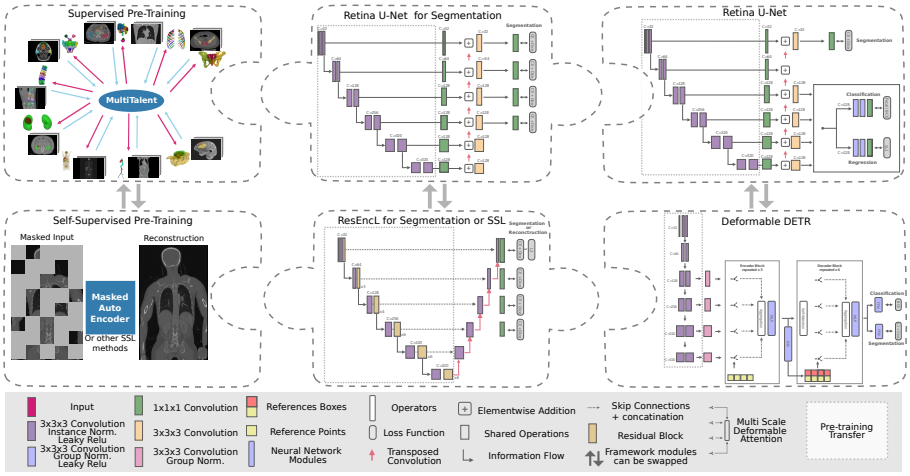

Fig. 1. A cross-framework bridge between nnDetection and its pre-training counterparts: Different pre-training paradigms (supervised and self-supervised), pre-training architectures (Retina U-Net and ResEncL), and detection-specific models (Retina U-Net and Deformable DETR) can be combined like puzzle pieces, offering a flexible and integrative approach to optimizing detection performance.

extract a feature representation from the input image. A point-wise convolution is applied to this feature representation to reduce the channel dimensionality. The extracted feature maps are flattened into a sequence of spatial tokens, with positional encodings added to retain spatial information. These tokens are then processed by a Transformer encoder-decoder architecture (3 encoder and 6 decoder blocks). The Deformable DETR detection head comprises one branch for classification (linear layer) and one for bounding box prediction (multi-layer perceptron). Focal loss is employed to account for dataset imbalances.

2.2 Pre-training Paradigms

Supervised Pre-training For large-scale supervised pretraining, we adopted the MultiTalent (MT) approach – a multi-dataset training paradigm introduced by Ulrich et al. [31]. To support this, we compiled a large-scale dataset collection of publicly available, pixel-wise annotated 3D medical images, comprising over 20,000 3D volumes sourced from 65 datasets, with more than 300,000 image-mask pairs. The dataset includes CT, MRI, and PET modalities. Due to length limitation, detailed dataset descriptions are omitted here but can be found in the associated arXiv publication. Notably, several datasets feature re-annotated publicly available images—for instance, the Abdomen Atlas [19] and Abdomen1K [22] datasets include images from the Medical Decathlon, among others. All datasets and images used for downstream fine-tuning were excluded from pre-training to avoid data leakage.

Self-Supervised Pre-training Self-supervised pre-training was performed using two large-scale medical imaging datasets: CT-RATE [12] and the Adolescent Brain Cognitive Development (ABCD) Study [28], totaling 91,768 training images. CT-RATE includes 25,692 non-contrast 3D CT scans, expanded to 50,188 volumes through multiple reconstructions from 21,304 unique patients. The ABCD Study, the largest U.S. longitudinal brain development study, contributed 41,580 brain images from 11,875 participants aged 9–10 at baseline, including T1-weighted, T2-weighted, and fMRI scans. We evaluated four self-supervised pre-training paradigms:
Models Genesis (MG) aims to reconstruct original image patches from transformed versions using non-linear intensity shifts, in-painting, out-painting, and local shuffling techniques [39].
Masked Autoencoder (MAE) utilizes a masked autoencoding strategy to reconstruct images, applying a 75% mask ratio to learn contextual features [13].
SparkMAE (S3D) modifies MAEs for CNN architectures to better process sparse inputs. It introduces sparse convolutions and normalization, where masking is reapplied after each convolution and normalization is restricted to non-masked values. A learnable mask token is used to fill masked areas for the encoder, followed by a densification convolution layer applied to all but the highest resolution feature maps [30].
VoCo leverages anatomical consistency by contrasting random sub-volumes against base crops to predict contextual overlap within 3D medical images [34].

Implementation We trained two MultiTalent networks: the state-of-the-art segmentation model ResEncL U-Net [16], and Retina U-Net [17]. The ResEncL U-Net employed a patch size of cubic 192 with a batch size of 12, while Retina U-Net used a patch size of cubic 128 and a batch size of 48. These differences in training parameters reflect that ResEncL U-Net was optimized for segmentation tasks, while Retina U-Net was designed for object detection. All SSL methods used the ResEncL architecture with a patch size of 192, and a batch size of 12. All networks were trained for 4,000 epochs using four NVIDIA A100 GPUs and a decreasing 'poly' learning rate schedule starting at 0.01 [8]. All pre-training data was preprocessed with z-score normalization and resampled to an isotropic voxel spacing of 1 mm. All other training parameters follow the default implementations in the corresponding open-source code-bases: All fine-tuning experiments are implemented within the nnDetection framework [5,4] and follow the default training scheme, including all hyperparameters. Therefore, the computational requirements match those reported in [4] and are independent of the pre-training. Supervised pre-training is conducted using the MultiTalent framework [31], while self-supervised pre-training is performed using the nnSSL framework[32], both inspired by nnU-Net [15]. This work establishes, for the first time, a cross-framework bridge between nnDetection and its pre-training counterparts, facilitating seamless integration across detection, segmentation and SSL paradigms.

Table 1. Development and Test Pool Datasets, including the numbers of images for training, validation and testing, the number of objects and the median spacing.

Dataset	Target	Modality	Split	Objects	Spacing [mm]
Dev D01 MSD Pancreas [2]	Pancreatic Tumor	CT	156/40/85	283	2.50x0.80x0.80
Dev D02 RibFrac [37]	Rib Fracture	CT	336/84/80	4422	1.25x0.74x0.74
Dev D03 KiTS21 [14]	Kidney Cyst, Tumor	CT	204/51/45	826	0.78x0.78x0.78
Dev D04 LIDC [3]	Lung Nodule (benign vs. malign.)	CT	690/173/155	1884	1.38x0.70x0.70
Dev D05 DUKE Breast [27]	Primary Breast Tumor	MRI	509/128/274	911	1.00x0.70x0.70
Test D06 LUNA16 [29]	Lung Nodule	CT	711/88/89	1186	1.25x0.70x0.70
Test D07 PN9 [25]	Lung Nodule	CT	6037/670/2091	40436	1.00x1.00x1.00
Test D08 CTA-A [7]	Brain Aneurysm	CT	948/238/152	1590	0.40x0.46x0.46

2.3 Downstream Datasets

We utilized a total of 8 datasets, comprising CT and MRI images with varying object types, to develop and evaluate all methods. The datasets were split into two pools: a development pool, which was used to determine the optimal parameters and make design decisions, and a test pool to evaluate the impact of different pre-trainings (table 1). From all datasets without an official split we separated hold-out test sets, comprising 15-30% of all images. The remaining images were split 80/20 into training and validation sets. Our experimental design builds upon the principles and processing steps established by Baumgartner et al. [4], ensuring consistency with their methodology. For PN9 (D07) experiments, we trained a single model on the training set, selected the post-processing parameters on the official validation set, and used the provided test set for our final evaluation. For CTA-A (D08) we split the data 80/20 into train and validation sets and utilized the internal test set (containing data from the same hospitals as the training data) for the final evaluation. To ensure a unified evaluation across the datasets, we employed the nnDetection metric calculations. For all datasets with official evaluation scripts, we will additionally provide the official evaluation in an arXiv version of this paper.

2.4 Metrics and Statistical Analysis

Detection performance was evaluated using the mean Average Precision (mAP) [17,24] at an IoU threshold of 0.1, emphasizing the diagnostic performance of the method and its ability to coarsely localize target objects. As an additional metric, the Free-response Receiver Operating Characteristic (FROC) [18,29] was employed with False Positive Per Image (FPPI) thresholds at [1/8, 1/4, 1/2, 1, 2, 4, 8]. To account for variations in object counts and task difficulty, rankings were computed via bootstrapping with 1000 iterations on the image level.

3 Experiments and Results

We explore large-scale pre-training for 3D object detection by evaluating four configurations: (**i**) Retina U-Net optimized for nnDetection (RetUNet), (**ii**) Deformable DETR with the Retina U-Net encoder (DefDETR), (**iii**) Retina U-Net

with an encoder from the ResEncL architecture (ResEnc-RetUNet), (**iv**) Deformable DETR with the ResEncL encoder (ResEnc-DefDETR).

Finding the Best Fine-Tuning Configuration: We identify the optimal fine-tuning configuration for each architecture based on MultiTalent (MT) pre-training, using an 80/20 train-validation split within the training set. As shown in table 2, using a fixed 1mm target spacing outperformed nnDetection's dataset-dependent spacing on these datasets. A learning rate of 0.1 was more effective than lower values, and transferring only encoder weights performed better than full model transfer. For ResEnc, fine-tuning with the pre-training patch size (192) showed no benefit for RetUNet and caused out-of-memory issues for DefDETR on a single A100 GPU node. Additionally, we explored strategies for handling multi-sequence datasets using D05 with four input channels. During MT pre-training, we assigned a unique stem per dataset to adjust the number of input channels, mapping them to a uniform 32-channel representation. For downstream fine-tuning, we tested three approaches: (**i**) Random initialization, (**ii**) Replicating a single-channel MRI stem [2], (**iii**) Using a stem from another four-sequence MRI dataset [11]. The third approach performed best. For SSL pre-training, we used the second-best random initialization instead.

Table 2. Finding the best fine-tuning configuration for each architecture. Validation results on five development datasets, reporting mean Average Precision (mAP) with MultiTalent pre-training.

					mAP@IoU 0.1						Stem ablation D05		
Model	**Transfer**	**Spacing**	**Patch**	**LR**	**D01**	**D02**	**D03**	**D04**	**Mean**	**Rank**	**RND**	**1Ch**[2]	**4Ch**[11]
RetUNet	Backbone	Default	128^3	1e-2	80.32	74.71	80.81	63.00	74.71	3.50	-	-	-
	Backbone	1x1x1	128^3	1e-3	86.96	73.59	79.50	66.86	76.73	3.25	-	-	-
	All	1x1x1	128^3	1e-2	87.51	**76.40**	84.04	66.61	78.64	2.00	-	-	-
	Backbone	1x1x1	128^3	1e-2	**88.25**	75.74	**85.35**	**67.32**	**79.16**	**1.25**	87.17	84.83	**88.16**
DefDETR	Backbone	Default	128^3	3e-4	73.32	76.86	**85.57**	61.96	74.43	1.75	-	-	-
	Backbone	1x1x1	128^3	3e-4	**90.06**	**77.82**	84.60	**63.87**	**79.09**	**1.25**	85.96	85.70	**87.89**
ResEnc-RetUNet	Backbone	1x1x1	192^3	1e-2	**92.06**	73.09	**84.23**	63.51	78.22	**1.50**	-	-	-
	Backbone	1x1x1	128^3	1e-2	90.38	**75.84**	83.61	**65.96**	**78.95**	**1.50**	84.88	**86.12**	85.11
ResEnc-DefDETR	Backbone	1x1x1	192^3	3e-4	OOM	OOM	OOM	OOM	-	2.00	-	-	-
	Backbone	1x1x1	128^3	3e-4	**90.53**	**77.65**	**85.49**	**66.70**	**80.09**	**1.00**	**88.28**	85.02	87.28

Impact of Pre-training To evaluate the impact of pre-training, we trained two baseline models from scratch for comparison: one following the architecture and configuration (e.g. median target spacing) recommended by nnDetection ("default") and another with a fixed architecture and target spacing cubic 1mm to match the pre-trained models ("fixed"). Overall, pre-trained models consistently outperform their non-pretrained counterparts across all architectures, as demonstrated in table 3 and fig. 2. Among the pre-training strategies, self-supervised reconstruction-based approaches (MAE, MG, S3D) yield the best results across all datasets. In contrast, contrastive pre-training (VoCo) underperforms relative

to training from scratch. Supervised pre-training (MT) also leads to notable performance gains. Overall, pre-training provides a more substantial performance boost for Deformable DETR than for Retina U-Net. Furthermore, the ResEnc backbone surpasses its Retina U-Net counterpart in performance but requires more VRAM and has a higher parameter count. Notably, a fixed architecture with a target spacing of 1 mm, when trained from scratch, achieves better rankings across datasets and models than the nnDetection configuration.

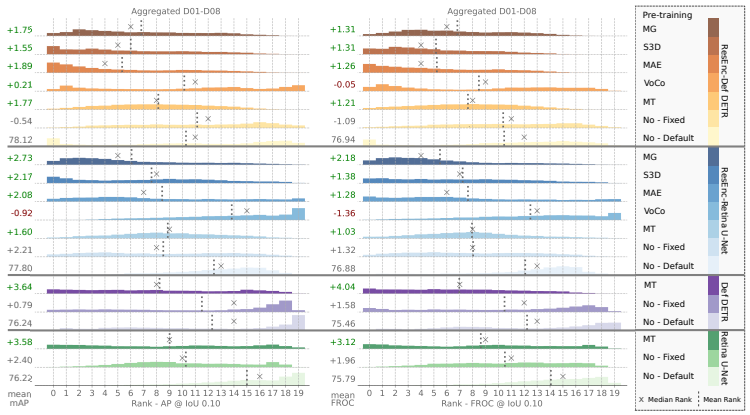

Fig. 2. Reconstruction-based SSL pre-training enhances detection the most. Aggregated ranking distributions for the test splits, that were derived from bootstrapping with 1000 iterations for each model and aggregated across all datasets D01-D08. Next to the ranking distribution, we report the difference in mAP and FROC of each method compared to the default nnDetection baseline for each architecture.

Table 3. High variance across different pre-training paradigms and architectures on the test splits of the dev. and test pool datasets. The overall best metric is underlined, while the best for each architecture is highlighted in bold.

		mAP@IoU 0.10										FROC@IoU 0.10									
Model	**Pre-Training**	**D01**	**D02**	**D03**	**D04**	**D05**	**D06**	**D07**	**D08**	**Mean**	**Rank**	**D01**	**D02**	**D03**	**D04**	**D05**	**D06**	**D07**	**D08**	**Mean**	**Rank**
RetUNet	No - Default	73.16	78.00	78.38	66.65	78.95	82.70	67.12	84.81	76.22	16.78	79.50	65.09	73.62	62.89	85.40	84.69	65.00	90.14	75.79	14.56
	No - Fixed	79.15	**80.35**	81.40	67.12	75.41	**83.33**	**68.34**	**93.89**	78.62	9.33	85.04	**67.49**	76.06	64.57	81.49	**84.95**	**66.38**	**96.03**	77.75	10.11
	MT [31]	**83.89**	79.48	**81.83**	**68.87**	**82.44**	82.15	67.29	92.48	**79.80**	**8.89**	**89.41**	66.40	**77.48**	**65.83**	**87.90**	84.44	65.14	94.67	**78.91**	**8.78**
DefDETR	No - Default	67.65	79.93	**83.43**	62.07	71.06	82.04	**70.18**	**93.59**	76.24	11.89	72.61	67.62	77.48	59.06	78.47	84.44	67.59	**96.37**	75.45	11.00
	No - Fixed	68.72	80.29	79.49	**69.30**	74.64	**84.31**	69.98	89.55	77.03	10.89	73.45	68.74	**78.25**	**68.39**	81.18	**86.10**	**67.73**	92.40	77.03	9.00
	MT [31]	**83.96**	**80.93**	80.61	66.99	**82.77**	83.55	69.58	90.70	**79.89**	**8.11**	**87.90**	**69.43**	77.99	65.87	**88.11**	85.84	67.36	93.42	**79.49**	**7.33**
ResEnc-RetUNet	No - Default	73.84	79.25	79.57	65.75	78.64	83.13	68.52	93.68	77.80	13.44	78.66	67.00	74.39	62.70	84.52	85.33	66.20	96.26	76.88	11.11
	No - Fixed	**83.75**	79.35	81.25	68.33	79.80	83.69	68.97	94.95	80.01	8.00	88.40	67.26	76.19	65.50	85.56	86.48	66.97	96.60	79.12	7.67
	MT [31]	78.84	79.70	82.39	67.91	80.00	83.45	69.52	93.35	79.40	8.11	84.71	66.60	78.25	65.87	86.39	85.97	**67.69**	95.12	78.83	7.67
	VoCo [34]	74.47	**80.35**	76.00	66.87	78.65	82.51	65.86	90.34	76.88	14.78	79.66	67.39	72.07	62.98	86.44	86.48	63.77	92.74	76.44	12.00
	MAE [9]	83.73	78.18	80.26	69.07	81.25	83.03	69.26	94.27	79.88	8.67	**90.25**	63.05	75.80	66.29	86.70	**87.12**	67.27	96.15	79.08	6.67
	S3D [30]	79.87	78.58	81.10	**70.45**	81.24	**83.82**	69.20	**95.44**	79.96	8.33	86.39	65.85	76.83	**68.30**	85.97	86.22	67.25	96.60	79.18	7.89
	MG [39]	82.64	79.36	**83.37**	68.85	**82.06**	83.13	**69.72**	95.05	**80.52**	**6.11**	87.39	**67.65**	**79.67**	66.39	**87.38**	86.99	67.60	**96.71**	**79.97**	**3.89**
ResEnc-DefDETR	No - Default	67.87	78.86	**85.33**	64.30	80.83	82.89	69.32	**95.59**	78.12	10.63	73.11	67.26	78.89	61.76	85.19	84.57	67.01	**97.73**	76.94	10.75
	No - Fixed	67.20	80.20	80.86	66.77	81.60	84.16	69.42	90.44	77.58	11.44	73.11	68.21	77.61	65.08	86.50	85.84	67.50	92.40	77.03	10.22
	MT [31]	**82.30**	81.12	82.19	67.09	81.34	83.39	69.28	92.47	79.90	7.89	**86.89**	70.31	78.51	65.83	86.18	85.71	67.58	93.65	79.33	7.44
	VoCo [34]	74.98	80.22	81.70	63.72	**82.72**	80.55	70.98	91.81	78.34	10.67	79.50	70.51	78.51	62.42	**87.80**	83.04	**68.89**	93.88	78.07	8.33
	MAE [9]	74.50	82.09	82.21	67.16	81.82	85.68	**71.56**	95.07	**80.01**	**3.78**	80.00	69.92	78.89	66.15	86.39	87.24	69.51	96.94	79.38	3.67
	S3D [30]	73.29	**82.44**	82.71	**67.40**	82.14	85.58	70.39	93.44	79.67	4.89	80.50	**70.61**	**79.15**	**66.85**	87.07	**87.50**	68.46	95.35	**79.44**	**3.56**
	MG [39]	78.62	81.47	82.71	66.76	81.41	**85.73**	70.44	91.81	79.87	6.44	86.05	70.15	78.64	65.08	86.44	86.86	68.28	93.99	**79.44**	6.00

4 Discussion

This work systematically studies the impact of large-scale pre-training on 3D medical object detection, showing that reconstruction-based self-supervised learning outperforms supervised pre-training. It also bridges nnDetection with pre-training frameworks, enabling a unified approach for medical image analysis. However, supervised pre-training was limited to segmentation tasks due to the scarcity of large 3D medical detection datasets, though segmentation annotations could be converted for detection. Whether organ detection pre-training truly enhances lesion detection remains questionable. Furthermore, similar to Baumgartner et al. [4], we observed high variability across tasks, which prevented us from identifying a single pre-training architecture combination that consistently outperformed all others. Finally, future work should investigate performance in low-data regimes and explore efficient fine-tuning strategies such as linear probing or LoRA, as these aspects were beyond the scope of this study.

Acknowledgments. This work was partially funded by the Helmholtz Foundation Model Initiative (HFMI) under the subproject The Human Radiome Project (THRP), and by Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science.

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

References

1. Amiriparian, S., Meiners, A., Rothenpieler, D., et al.: Universal lesion detection utilising cascading r-cnns and a novel video pretraining method. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (2023)
2. Antonelli, M., Reinke, A., Bakas, S., et al.: The medical segmentation decathlon. Nature Communications (2022)
3. Armato III, S.G., McLennan, G., Bidaut, L., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. Medical Physics (2011)
4. Baumgartner, M., Ickler, M.K., Jaeger, P.F., et al.: nndetection: A self-configuring method for volumetric 3d object detection. Under Review (2025)
5. Baumgartner, M., Jäger, P.F., Isensee, F., Maier-Hein, K.H.: nndetection: A self-configuring method for medical object detection. In: MICCAI (2021)
6. Benčević, M., Habijan, M., Galić, I., Pizurica, A.: Self-supervised learning as a means to reduce the need for labeled data in medical image analysis. In: 2022 30th European Signal Processing Conference (EUSIPCO) (2022)
7. Bo, Z.H., Qiao, H., Tian, C., et al.: Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network. Patterns (2021)
8. Chen, L.C., Papandreou, G., Kokkinos, I., et al.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)

9. Chen, Z., Agarwal, D., Aggarwal, K., et al.: Masked image modeling advances 3d medical image analysis. In: IEEE/CVF WACV (2023)
10. Cheng, P., Lin, L., Lyu, J., et al.: Prior: Prototype representation joint learning from medical images and reports. In: ICCV (2023)
11. Grøvik, E., Yi, D., Iv, M., et al.: Deep learning enables automatic detection and segmentation of brain metastases on multisequence mri. Journal of Magnetic Resonance Imaging (2020)
12. Hamamci, I.E., Er, S., Almas, F., et al.: Developing generalist foundation models from a multimodal dataset for 3d computed tomography. arXiv:2403.17834 (2024)
13. He, K., Chen, X., Xie, S., et al.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
14. Heller, N., Isensee, F., Trofimova, D., et al.: The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. arXiv:2307.01984 (2023)
15. Isensee, F., Jaeger, P.F., Kohl, S.A., et al.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods (2021)
16. Isensee, F., Wald, T., Ulrich, C., et al.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. arXiv preprint arXiv:2404.09556 (2024)
17. Jaeger, P.F., Kohl, S.A., Bickelhaupt, S., et al.: Retina u-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. In: Machine Learning for Health Workshop (2020)
18. Jin, L., Yang, J., Kuang, K., et al.: Deep-learning-assisted detection and segmentation of rib fractures from ct scans: Development and validation of fracnet. eBioMedicine (2020)
19. Li, W., Qu, C., Chen, X., et al.: Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. Medical Image Analysis (2024)
20. Liu, C., Cheng, S., Chen, C., et al.: M-flag: Medical vision-language pre-training with frozen language models and latent space geometry optimization. In: MICCAI (2023)
21. Liu, C., Shah, A., Bai, W., Arcucci, R.: Utilizing synthetic data for medical vision-language pre-training: Bypassing the need for real images. arXiv:2310.07027 (2023)
22. Ma, J., Zhang, Y., Gu, S., et al.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
23. Maier-Hein, L., Eisenmann, M., Reinke, A., et al.: Why rankings of biomedical image analysis competitions should be interpreted with care. Nature communications (2018)
24. Maier-Hein, L., Reinke, A., Godau, P., et al.: Metrics reloaded: recommendations for image analysis validation. Nature methods (2024)
25. Mei, J., Cheng, M.M., Xu, G., et al.: Sanet: A slice-aware network for pulmonary nodule detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
26. Müller, P., Kaissis, G., Zou, C., Rueckert, D.: Joint learning of localized representations from medical images and reports. In: ECCV (2022)
27. Saha, A., Harowicz, M.R., Grimm, L.J., et al.: A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. British Journal of Cancer (2018)
28. Saragosa-Harris, N.M., Chaku, N., MacSweeney, N., et al.: A practical guide for researchers and reviewers using the abcd study and other large longitudinal datasets. Developmental Cognitive Neuroscience (2022)

29. Setio, A.A.A., Traverso, A., de Bel, T., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge. Medical Image Analysis (2017)
30. Tian, K., Jiang, Y., Diao, Q., et al.: Designing bert for convolutional networks: Sparse and hierarchical masked modeling. arXiv preprint arXiv:2301.03580 (2023)
31. Ulrich, C., Isensee, F., Wald, T., et al.: Multitalent: A multi-dataset approach to medical image segmentation. In: MICCAI (2023)
32. Wald, T., Ulrich, C., Lukyanenko, S., et al.: Revisiting mae pre-training for 3d medical image segmentation. arXiv preprint arXiv:2410.23132 (2024)
33. Wang, X., Han, S., Chen, Y., Gao, D., Vasconcelos, N.: Volumetric attention for 3d medical image segmentation and detection. In: MICCAI (2019)
34. Wu, L., Zhuang, J., Chen, H.: Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In: CVPR (2024)
35. Wu, Y., Zhou, Y., Saiyin, J., et al.: Zero-shot nuclei detection via visual-language pre-trained models. In: MICCAI (2023)
36. Yan, K., Cai, J., Zheng, Y., et al.: Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in ct. IEEE Transactions on Medical Imaging (2020)
37. Yang, J., Shi, R., Jin, L., et al.: Deep rib fracture instance segmentation and classification from ct on the ribfrac challenge. arXiv:2402.09372 (2024)
38. Zhang, S., Xu, J., Chen, Y.C., et al.: Revisiting 3d context modeling with supervised pre-training for universal lesion detection in ct slices. In: MICCAI (2020)
39. Zhou, Z., Sodha, V., Pang, J., et al.: Models genesis. Medical image analysis (2020)
40. Zhu, X., Su, W., Lu, L., et al.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2021)