

Curriculum Learning for Language-guided, Multi-modal Detection of Various Pathologies

Laurenz Adrian Heidrich^{*1,2}

LAURENZ.HEIDRICH@TUM.DE

Aditya Rastogi^{*2}

ADITYA.RASTOGI@UKBONN.DE

Priyank Upadhyay²

PRIYANK.UPADHYA@UKBONN.DE

Gianluca Brugnara^{2,3}

GIANLUCA.BRUGNARA@UKBONN.DE

Martha Foltyn-Dumitru²

MARTHA.FOLTYN-DUMITRU@UKBONN.DE

Benedikt Wiestler^{1,4}

B.WIESTLER@TUM.DE

Philipp Vollmuth^{2,3}

PHILIPP.VOLLMUTH@UKBONN.DE

¹ *AI for Image-Guided Diagnosis and Therapy (AI-IDT), School of Medicine and Health, Technical University of Munich, Munich, Germany*

² *Division for Computational Radiology & Clinical AI (CCIBonn.ai), University Hospital Bonn, Bonn, Germany*

³ *Division for Medical Image Computing (MIC), German Cancer Research Center (DFKZ), Heidelberg, Germany*

⁴ *Munich Center for Machine Learning, Munich, Germany*

Editors: Under Review for MIDL 2025

Abstract

Pathology detection in medical imaging is crucial for radiologists, yet current approaches that train specialized models for each region of interest often lack efficiency and robustness. Furthermore, the scarcity of annotated medical data, particularly for diverse phenotypes, poses significant challenges in achieving generalizability. To address these challenges, we present a novel language-guided object detection pipeline for medical imaging that leverages curriculum learning strategies, chosen for their ability to progressively train models on increasingly complex samples, thereby improving generalization across pathologies, phenotypes, and modalities. We developed a unified pipeline to convert segmentation datasets into bounding box annotations, and applied two curriculum learning approaches - teacher curriculum and bounding box size curriculum - to train a Grounding DINO model. Our method was evaluated on different tumor types in MRI and CT scans and showed significant improvements in detection accuracy. The teacher and bounding box size curriculum learning approaches yielded a 4.9% AP and 5.2% AP increase over baseline, respectively. The results highlight the potential of curriculum learning to optimize medical image analysis and clinical workflow by providing a versatile and efficient detection algorithm.

Keywords: Medical Image Analysis, Deep Learning, Tumor Detection, Curriculum Learning

1. Introduction and Related Work

Pathology detection on medical imaging is a cornerstone of modern radiology practice and plays a crucial role in diagnosis and treatment planning. Accurate pathology detection is

* Contributed equally

critical for determining the presence, location, and extent of abnormalities, guiding diagnostic accuracy, enabling targeted treatments, monitoring disease progression, and evaluating therapy efficacy. Simplifying associated workflows and reducing prediction complexity is essential to increase efficiency, remove barriers to clinical adoption, and improve prediction quality. This can be achieved by unifying specialized models into foundation models, and by reducing algorithmic and task complexity. In this context, segmentation algorithms can be substituted by detection-only algorithms for many clinical tasks, such as identifying new metastases or counting the number of existing lesions.

Recent advances in medical imaging have led to the development of foundation models capable of handling multiple modalities and interactive tasks (Ma et al., 2024a,b). These models show superior flexibility compared to specialized segmentation networks such as nnU-Net (Isensee et al., 2021), operating on different imaging modalities and accepting visual input prompts. In addition, recent developments have also demonstrated the benefits of language guidance in medical image analysis, enabling more efficient medical image interpretation, with several studies focusing on language-driven segmentation (Koleilat et al., 2024b; Li et al., 2024; Liu et al., 2023; Zhao et al., 2024a) and improving zero-shot and few-shot performance (Koleilat et al., 2024a). Notable work has been done in the area of medical object detection, particularly for brain tumors (Mercaldo et al., 2023; Chen et al., 2024a; Abdusalomov et al., 2023; He et al., 2023). However, there appears to be a lack of multi-modal and multi-pathology detection frameworks. The success of Grounding DINO (G-DINO) (Liu et al., 2025), an open-set language-guided object detector, has generated interest in its application to medical imaging, allowing the integration of different pathologies, modalities and text prompts into a single network. So far, only a few algorithms (Biswas, 2023; Xie et al., 2024; Ramesh et al., 2023) have taken advantage of this additional guidance for object detection in medical imaging. In this study we focus on developing a language-guided network to detect pathologies - specifically tumors - of various organs.

Unlike natural images, tumor detection poses unique challenges due to the significant variability in tumor phenotypes across patients, which demands large datasets to achieve generalization. However, the scarcity of annotated medical imaging data makes generalization challenging, especially for detecting smaller tumors (Abdusalomov et al., 2023; He et al., 2023). To address these shortcomings, we investigated different Curriculum Learning (CL) strategies (Bengio et al., 2009) to increase detection accuracy. CL, introduced by Bengio et al. (Bengio et al., 2009), has found several applications in the medical domain (Jiménez-Sánchez et al., 2019; Wei et al., 2021; Oksuz et al., 2019; Fischer et al., 2024) with the goal of improving performance by gradually increasing training complexity. The strategy used in this study, called data-level CL, gradually increases the complexity of the training samples: First, the model is trained on large, well-contrasted tumors to establish robust feature representations. Next, the training data is expanded to include smaller, less conspicuous tumors with increasing anatomical and modality variability. This progressive learning approach helps the model develop better generalization capabilities and improves its sensitivity to subtle pathological findings.

In this work, we explore the potential of two different CL strategies on G-DINO’s detection performance by pre-training the network on the TotalSegmentator dataset (Wasserthal et al., 2023), followed by CL-based fine-tuning on tumor datasets spanning different imaging modalities (Magnetic Resonance Imaging (MRI) and Computed Tomography (CT))

and anatomical sites (brain metastasis, glioma, liver & kidney tumor). In addition, we have developed a pipeline to convert ground truth segmentations into bounding boxes by using morphological operations to consolidate them. This goes beyond the naive approach of drawing tight bounding boxes around segmentations. An extensive evaluation of the G-DINO baseline model was conducted, comparing its performance with models trained using two CL approaches: teacher CL (Weinshall et al., 2018) and bounding box CL (Shi and Ferrari, 2016). The results show a 4.9% improvement in Average Precision (AP) with teacher CL and a 5.2% increase in AP with bounding box CL compared to the baseline model. Based on the reviewed literature, this paper is among the initial efforts to:

1. Apply two different CL strategies to a language-guided detection network (G-DINO).
2. Train G-DINO jointly on different pathologies from various body regions and modalities, demonstrating the model’s versatility with limited datasets.
3. Develop and formalize a novel preprocessing pipeline to convert medical segmentation datasets into object detection datasets.

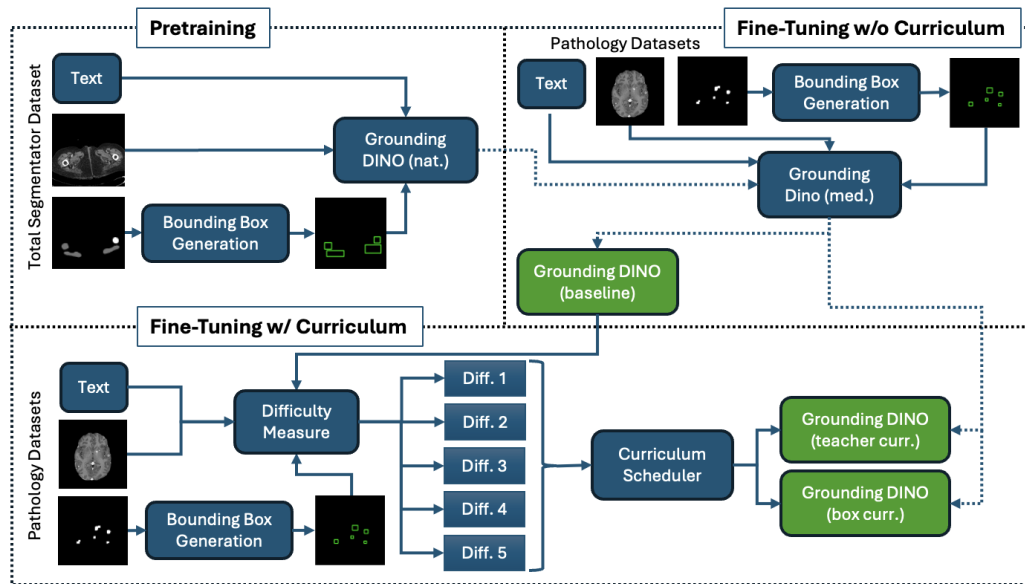


Figure 1: Overview of our method: As a first step, the natural image G-DINO model is pre-trained on the Total Segmentator dataset (top left). In a second step, the baseline is finetuned without CL on all pathology datasets (top right). Finally, two CL models are trained: for teacher CL, the baseline is used to guide the difficulty sorting, while for bounding box CL, the size of bounding boxes are used for difficulty sorting (bottom).

2. Method

Our methodological pipeline consists of three parts, as shown in Figure 1. The first part consists of pre-training the language-guided detection network G-DINO on a large multi-modal, multi-organ dataset to detect 163 different regions of interest - 104 for CT and 59 for MRI. The second part consists of fine-tuning the pre-trained G-DINO on the pathological datasets producing the baseline. Finally, the third part consists of fine-tuning the pre-trained model on the pathological datasets using two different CL strategies.

2.1. Ground Truth Bounding Box Generation

Since our experimental datasets are primarily designed for segmentation tasks, we developed a method to generate ground truth bounding boxes. To generate bounding boxes for the pathology datasets, we first removed all segmentation masks not related to pathologies, such as liver or liver vessel masks. Then, we merged masks if certain tumors had compartments (e.g., “contrast enhancing” and “necrotic” parts). Finally, we performed dilation on the binary segmentation mask to remove discontinuities in the tumor mask. This provides a more accurate bounding box for a tumor instead of separate bounding boxes for discontinuous regions of the same tumor and ensures that the bounding boxes are not overly tight, providing a more realistic representation similar to human annotation. We perform the dilation in 2 iterations with a 3×3 kernel. After performing the dilation, we then drew tight bounding boxes around the resulting segmentations (see Figure 2).

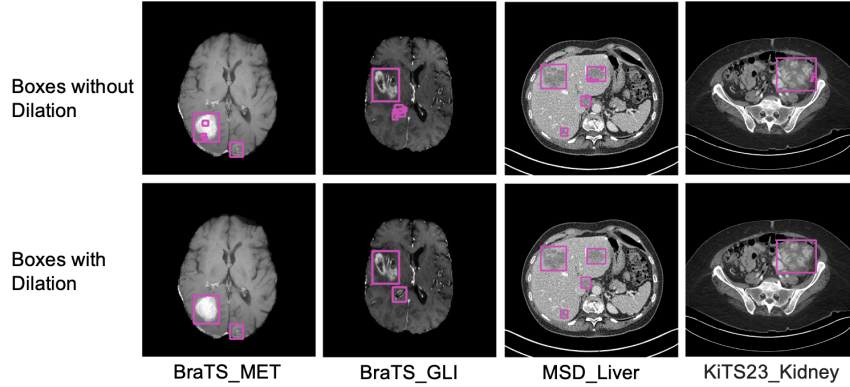


Figure 2: Depiction of the regularization effects of our bounding box pipeline using dilation.

2.2. Grounding DINO

G-DINO (Liu et al., 2025) is an open object detector capable of identifying any object based on textual input, such as referring expressions or categories. Given a (Image, Text) pair input, G-DINO predicts multiple (Bounding Box, Noun Phrase) pairs with confidence scores for each detected entity. The noun phrase is the predicted semantic entity of the box and is derived from the input prompt in an open-set fashion. The model employs a dual-encoder-single-decoder architecture consisting of an image encoder for visual feature extraction, a text encoder for textual information processing, a feature enhancer for fusion of extracted features, a language-guided query selection module for query initialization, and a cross-modality decoder for bounding box refinement (Liu et al., 2025). We used the G-DINO implementation from the mmdetection framework (Zhao et al., 2024b) and adopted the focal loss (Lin et al., 2020) ($\gamma = 2.0$ and $\alpha = 0.25$) and the weighted L1 loss ($w = 5$) as loss functions. (Zhao et al., 2024b) For image and text encoders, we used Swin-Tiny (Liu et al., 2021) and bert-base-uncased (Devlin et al., 2018), respectively.

2.3. Curriculum Learning

Our CL approach includes two methods: teacher CL and bounding box CL. For the teacher CL, we used the baseline model to perform inference on the entire training dataset of

pathological images and computed the AP score for each image to sort them into five difficulty levels based on their evaluation scores (ranging from 1 for the easiest to 5 for the most difficult). This approach is common in data-level CL and is based on the premise that simpler training samples will receive high precision scores, while more complex and ambiguous samples will receive low scores. For false positives, we manually set the AP score to 0.0 if the network predicts a bounding box with a confidence level greater than 0.3. The bounding box curriculum classifies each training sample into one of five difficulty levels based on the size of the smallest bounding box present in the image. We found that bounding box size is a primary indicator of prediction difficulty, as shown by the performance of the baseline model (Table 1). This approach has the advantage of not requiring a trained baseline model a priori. In both CL approaches, we randomly assigned training samples without bounding boxes (i.e., no object of interest) to difficulty categories to maintain an equal distribution. To fine-tune the model on the pathological dataset using CL, we start with the easiest category and progressively introduce more difficult categories at each epoch. After 5 epochs, the network is fine-tuned on the complete data until convergence.

3. Experiments and Results

3.1. Pretraining & Datasets

In our experiments, we initialized the G-DINO architecture with weights published by (Zhao et al., 2024b), which were trained on several natural image datasets (Objects365 (Shao et al., 2019), GRIT (Peng et al., 2023), V3Det (Wang et al., 2023), Golden-G dataset (Kamath et al., 2021)). We first pretrained this network on the TotalSegmentator dataset, which consists of CT (Wasserthal et al., 2023) and MRI (D’Antonoli et al., 2024) scans of the entire human body, to adapt the image encoder weights to medical imaging modalities and to improve medical semantic understanding. We then fine-tuned the network on heterogeneous datasets spanning multiple modalities, pathologies, hospitals, and scanner manufacturers. This aggregated dataset includes both MRI and CT scans with four different detection targets: brain metastasis & glioma (MRI, T1-weighted contrast-enhanced) as well as liver & kidney tumor (CT, contrast-enhanced). For brain metastases, we used a dataset by Ramakrishnan et al. (Ramakrishnan et al., 2024) (abbreviated as Yale_BM in this paper) and the BraTS 2023 metastasis challenge (Moawad et al., 2023) (BraTS_MET). For glioma, we used the BraTS 2023 glioma challenge (Baid et al., 2021; Menze et al., 2015; Bakas et al., 2017) (BraTS_GLI). For liver tumor, we selected the Liver Tumor (MSD_Liver) and Hepatic Vessel (MSD_Hep_Vessel) challenges from the Medical Segmentation Decathlon (Antonelli et al., 2022). For kidney tumor, we used the KiTS23 dataset (KiTS23_Kidney) (Heller et al., 2023).

3.2. Data Preprocessing & Training Detail

Both CT and MRI datasets, originally in 3D NIfTI format, require preprocessing for compatibility with G-DINO, a 2D object detector. Following Ma et al. (Ma et al., 2024a), we clipped MRI images to their [0.05, 99.5] percentile and normalized them to [0, 255], while CT images were windowed (level = 40 and width = 400) before normalization. For the pathological datasets, we used organ segmentation masks to only retain slices containing the organ

of the associated pathology. Patients from all datasets were split into train/validation/test (0.7, 0.15, 0.15) sets. The resulting 2D training dataset consisted of 199,672 slices, of which 66,990 slices had bounding box annotations (i.e., tumors were present). During training, the BraTS_Glioma dataset was undersampled by a factor of 3 to ensure a balanced class distribution. We trained the models on two NVIDIA RTX A6000 GPUs with a batch size of 10 until convergence. The baseline model was trained with a learning rate of $1e^{-5}$ for the first 5 epochs, followed by a learning rate of $1e^{-6}$, while the curricula models were trained with $1e^{-5}$ for the first 7 epochs and $1e^{-6}$ for the remaining epochs to ensure equal data exposure and to compensate for shorter curricula epochs due to data exclusion. Figure A1 in the appendix illustrates the training loss and validation curves.

3.3. Curriculum Learning - Difficulty categories

We categorized the training data into five difficulty levels based on two heuristics to obtain two CL strategies as explained in Section 2.3. After training the baseline model, we observed that the size of the bounding box correlated with the precision score and thus could be an indicator of the difficulty of the samples (Table 1). To standardize bounding box sizes, we calculated their areas relative to the image area. We then sorted the samples based on the smallest bounding box present in each slice, ensuring an even distribution across categories. Slices without bounding boxes were randomly assigned to maintain an equal distribution of annotated and unannotated samples. For teacher CL, we additionally included slices without ground-truth bounding boxes but with false positive predictions from the baseline model and assigned them an AP score of 0.0. We then created intervals to maintain a roughly equal distribution across categories. Dataset distributions across categories are shown in the appendix for both heuristics in Table A1 and Table A2, respectively.

3.4. Experimental Setup

After pre-training on the TotalSegmentator dataset, we simultaneously fine-tuned G-DINO on the full pathological training dataset to obtain three models: a baseline model and two CL-based models using the bounding box and teacher principles introduced in Section 2.3. During training, model weights were evaluated on a merge of the individual validation sets, and the best performing (mean-AP over all detection targets) weights were selected. Subsequent testing was performed on all datasets individually. We evaluated the object detection results using the COCO metrics (Lin et al., 2014). We used AP values at different IoU thresholds: AP@0.5 and AP@0.75 with IoU thresholds of 0.5 and 0.75, respectively. The unspecified AP represents the average metric across IoU thresholds between 0.5 and 0.95 in 0.05 increments. Additionally, we evaluated the predictions separately for different bounding box sizes, where AP small, AP medium, AP large refer to the AP of ground truth bounding boxes with areas of $[0, 32^2]$, $[32^2, 96^2]$, and $[96^2, \infty)$ pixels, respectively.

3.5. Results

Table 1 shows the quantitative test results for all models. Figure 3 shows the predictions of our three models alongside the ground truth for one case from each dataset. Both CL approaches improved performance on average AP metrics, with the bounding box CL model achieving the highest gains (+5.2% AP, +6.1% AP@0.75, +6.0% AP@0.5 over baseline).

Dataset	Curr.	AP (%) @0.5	AP (%) @0.75	AP (%)	AP (%) large	AP (%) medium	AP (%) small
Overall	Without	69.7	50.6	46.5	69.3	62.0	30.9
	Box	75.7	56.7	51.7	70.0	65.8	35.6
	Teacher	75.5	56.0	51.4	72.7	65.5	35.6
Yale_BM	Without	61.5	48.2	42.2	-	69.9	37.3
	Box	79.5	64.9	56.9	-	81.5	52.4
	Teacher	76.0	61.8	54.3	-	77.5	50.8
BraTS_MET	Without	66.7	54.1	45.9	-	76.9	42.2
	Box	82.7	65.9	57.4	-	83.8	54.2
	Teacher	81.5	64.5	56.6	-	83.6	53.0
BraTS_GLI	Without	84.8	72.2	65.8	-	81.1	48.1
	Box	85.2	72.5	66.0	-	81.3	48.4
	Teacher	84.5	71.8	65.4	-	80.7	47.5
MSD_Liver	Without	61.4	32.1	33.2	66.6	46.8	19.9
	Box	60.4	33.2	32.7	63.1	44.9	21.9
	Teacher	61.3	30.2	31.9	65.8	46.0	19.0
MSD_Hep_Vessel	Without	69.4	37.7	39.2	62.8	43.7	15.0
	Box	70.5	43.5	43.1	67.8	47.5	17.6
	Teacher	72.4	42.9	42.9	68.5	46.4	17.8
KiTS23_Kidney	Without	74.6	59.5	52.9	78.4	53.8	22.6
	Box	75.7	60.4	53.9	79.0	56.0	19.0
	Teacher	77.0	64.5	57.4	83.8	58.6	25.2

Table 1: Overview of the AP scores on individual datasets and averaged to an overall score.

Both CL models outperformed the baseline in all size-constrained AP scores, with the largest gains in the most difficult categories (AP small: +4.7% for both models, AP medium: +3.5% & +3.8% for bounding box CL and teacher CL, respectively). The results thus support our hypothesis that CL improves performance especially for the most difficult samples with the smallest tumors. Looking at individual datasets, the CL models performed best in 5 out of 6 datasets for all metrics, with the MSD_Liver dataset showing slight underperformance in this context (-0.5% AP for bounding box CL, -1.3% AP for teacher CL). Overall, the results indicate that CL generally improves model performance, especially for challenging detection tasks with small to medium bounding boxes.

3.6. Ablation Studies

To show the effect of CL, and also to follow recent work (Wu et al., 2023; Chen et al., 2024b; Braun et al., 2017) that proposes a hard-to-easy methodology (anti-CL), we also trained our models in such a setting. The results are shown in the appendix in Table A3. Anti-box CL achieves an overall AP score of 49.5%, which is 3.0% better than the baseline and 2.2% less than regular bounding box CL. Anti-teacher CL, on the other hand, achieves an overall AP score of 51.2%, which is 3.7% better than the baseline and only slightly (0.2%) worse than regular teacher CL. Both CL and anti-CL proved to be effective for both

difficulty sorting approaches and outperformed the baseline. Our results show that when training samples were sorted by difficulty based on the performance of the baseline model, the sorting order had little effect on final accuracy. However, when the difficulty sorting was based on bounding box size, the sample order during training had a more pronounced effect, leading to greater variation in accuracy. This suggests that difficulty sorting based on manual heuristics interacts more with learning dynamics than teacher-based difficulty sorting.

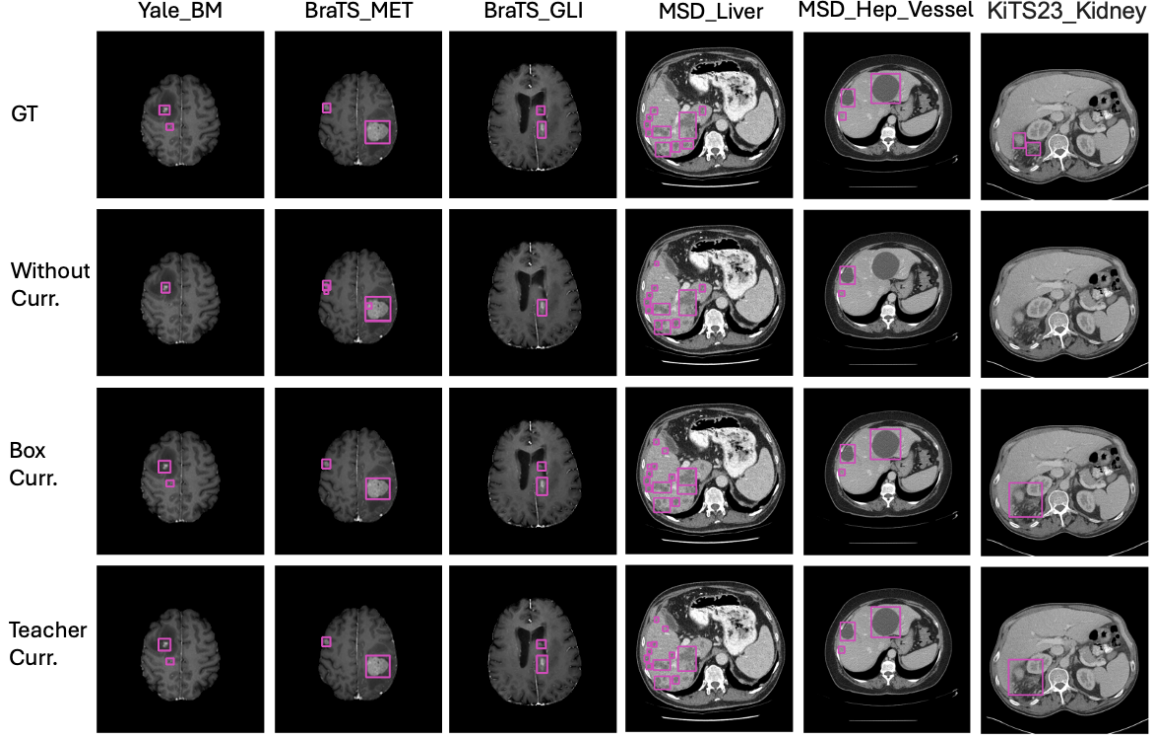


Figure 3: Visualization of the predicted results with an illustrative case from each dataset.

4. Conclusion

Our study demonstrates the effectiveness of CL strategies in multimodal medical image object detection. Implementation of bounding box size-based and teacher-guided curricula improved overall detection accuracy, particularly for small and medium-sized objects. However, the lack of improvement on the MSD_Liver dataset necessitates further investigation into dataset-specific factors affecting the effectiveness of CL. We observed that both sorting heuristics can be effectively applied in an anti-curriculum fashion, with only the teacher heuristic being able to match the regular sorting approach. Thus, studying the necessary conditions for successful anti-CL needs further investigation. Despite these promising results, challenges in CL implementation remain. Data-level CL still requires hand-crafted difficulty categories and predefined scheduling. In addition, the computational overhead of CL, especially in the teacher-guided approach, warrants consideration in balancing improved performance with increased training time.

Acknowledgments

P.V. is funded through the Else Kröner Clinician Scientist Endowed Professorship (reference number: 2022_EKCS.17). A.R. is funded through the Bonfor Startup Postdoc Fellowship (reference number: 2024-1B-10)

References

- Akmalbek Bobomirzaevich Abdusalomov, Mukhriddin Mukhiddinov, and Taeg Keun Whangbo. Brain tumor detection based on deep learning approaches and magnetic resonance imaging. *Cancers*, 15(16), 2023. ISSN 2072-6694. doi: 10.3390/cancers15164172. URL <https://www.mdpi.com/2072-6694/15/16/4172>.
- Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Ildoo Kim, Klaus Maier-Hein, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30695-9. URL <https://www.nature.com/articles/s41467-022-30695-9>.
- Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. doi: 10.48550/arXiv.2107.02314. URL <https://arxiv.org/abs/2107.02314>.
- Pyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data*, 4, 9 2017. ISSN 2052-4463. doi: 10.1038/sdata.2017.117. URL <https://www.nature.com/articles/sdata2017117>.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML ’09, pages 41–48. Association for Computing Machinery, 2009. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553380. URL <https://dl.acm.org/doi/10.1145/1553374.1553380>.

- Risab Biswas. Polyp-SAM++: Can a text guided SAM perform better for polyp segmentation? *arXiv preprint arXiv:2308.06623*, 2023. doi: 10.48550/arXiv.2308.06623. URL <http://arxiv.org/abs/2308.06623>.
- Stefan Braun, Daniel Neil, and Shih-Chii Liu. A curriculum learning method for improved noise robustness in automatic speech recognition. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 548–552. IEEE, 2017. doi: 10.23919/EUSIPCO.2017.8081267. URL <https://ieeexplore.ieee.org/document/8081267>.
- Aruna Chen, Da Lin, and Qiqi Gao. Enhancing brain tumor detection in MRI images using YOLO-NeuroBoost model. *Frontiers in Neurology*, 15, 2024a. ISSN 1664-2295. doi: 10.3389/fneur.2024.1445882. URL <https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2024.1445882/full>.
- Junfan Chen, Jun Yang, Anfei Fan, Jinyin Jia, Chiyu Zhang, and Wei Li. Apan: Anti-curriculum pseudo-labelling and adversarial noises training for semi-supervised medical image classification. In *Pattern Recognition and Computer Vision: 7th Chinese Conference, PRCV 2024, Urumqi, China, October 18–20, 2024, Proceedings, Part XIV*, page 163–177, Berlin, Heidelberg, 2024b. Springer-Verlag. ISBN 978-981-97-8495-0. doi: 10.1007/978-981-97-8496-7_12. URL https://link.springer.com/chapter/10.1007/978-981-97-8496-7_12.
- Tugba Akinci D’Antonoli, Lucas K Berger, Ashraya K Indrakanti, Nathan Vishwanathan, Jakob Weiß, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, et al. Totalsegmentator mri: Sequence-independent segmentation of 59 anatomical structures in mr images. *arXiv preprint arXiv:2405.19492*, 2024. doi: 10.48550/arXiv.2405.19492. URL <https://arxiv.org/abs/2405.19492>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2018. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423/>.
- Stefan M. Fischer, Lina Felsner, Richard Osuala, Johannes Kiechle, Daniel M. Lang, Jan C. Peeken, and Julia A. Schnabel. Progressive growing of patch size: Resource-efficient curriculum learning for dense prediction tasks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference, Marrakesh, Morocco, October 6–10, 2024, Proceedings, Part IX*, pages 510–520. Springer-Verlag, 2024. ISBN 978-3-031-72113-7. doi: 10.1007/978-3-031-72114-4_49. URL https://link.springer.com/chapter/10.1007/978-3-031-72114-4_49.
- Jingzhen He, Jing Wang, Zeyu Han, Baojun Li, Mei Lv, and Yunfeng Shi. Cancer detection for small-size and ambiguous tumors based on semantic fpn and transformer. *PLOS ONE*, 18(2):1–17, 02 2023. doi: 10.1371/journal.pone.0275194. URL <https://doi.org/10.1371/journal.pone.0275194>.

- Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, Yoel Shoshan, Flora Gilboa-Solomon, Yasmeen George, Xi Yang, Jianpeng Zhang, Jing Zhang, Yong Xia, Mengran Wu, Zhiyang Liu, Ed Walczak, Sean McSweeney, Ranveer Vasdev, Chris Hornung, Rafat Solaiman, Jamee Schoephoerster, Bailey Abernathy, David Wu, Safa Abdulkadir, Ben Byun, Justice Spriggs, Griffin Struyk, Alexandra Austin, Ben Simpson, Michael Hagstrom, Sierra Virnig, John French, Nitin Venkatesh, Sarah Chan, Keenan Moore, Anna Jacobsen, Susan Austin, Mark Austin, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. *arXiv preprint arXiv:2307.01984*, 2023. doi: 10.48550/arXiv.2307.01984. URL <https://arxiv.org/abs/2307.01984>.
- Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01008-z. URL <https://www.nature.com/articles/s41592-020-01008-z>.
- Amelia Jiménez-Sánchez, Diana Mateus, Sonja Kirchhoff, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Miguel A González Ballester, and Gemma Piella. Medical-based deep curriculum learning for improved fracture classification. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI* 22, pages 694–702. Springer, 2019. doi: 10.1007/978-3-030-32226-7_77. URL https://link.springer.com/chapter/10.1007/978-3-030-32226-7_77.
- Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021. doi: 10.1109/ICCV48922.2021.00180. URL <https://ieeexplore.ieee.org/document/9710994>.
- Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-samv2: Towards universal text-driven medical image segmentation. *arXiv preprint arXiv:2409.19483*, 2024a. doi: 10.48550/arXiv.2409.19483. URL <https://arxiv.org/abs/2409.19483>.
- Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. MedCLIP-SAM: Bridging Text and Image Towards Universal Medical Image Segmentation. In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15012. Springer Nature Switzerland, October 2024b. doi: 10.1007/978-3-030-32226-7_77. URL https://link.springer.com/chapter/10.1007/978-3-031-72390-2_60.
- Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: Language meets vision transformer in medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(1):96–107, 2024. doi: 10.1109/TMI.2023.3291719. URL <https://ieeexplore.ieee.org/document/10172039>.

- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. doi: 10.1109/TPAMI.2018.2858826. URL <https://ieeexplore.ieee.org/document/8417976>.
- Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A. Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21095–21107, 2023. doi: 10.1109/ICCV51070.2023.01934. URL <https://ieeexplore.ieee.org/document/10376801>.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 38–55. Springer Nature Switzerland, 2025. ISBN 978-3-031-72970-6. doi: 10.1007/978-3-031-72970-6_3. URL https://link.springer.com/chapter/10.1007/978-3-031-72970-6_3.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Los Alamitos, CA, USA, October 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00986. URL <https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00986>.
- Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024a. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z. URL <https://www.nature.com/articles/s41467-024-44824-z>.
- Jun Ma, Sumin Kim, Feifei Li, Mohammed Baharoon, Reza Asakereh, Hongwei Lyu, and Bo Wang. Segment anything in medical images and videos: Benchmark and deployment. *arXiv preprint arXiv:2408.03322*, 2024b. doi: 10.48550/arXiv.2408.03322. URL <http://arxiv.org/abs/2408.03322>.
- Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R.

- Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftexharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. doi: 10.1109/TMI.2014.2377694. URL <https://ieeexplore.ieee.org/document/6975210>.
- Francesco Mercaldo, Luca Brunese, Fabio Martinelli, Antonella Santone, and Mario Cesarelli. Object detection for brain cancer detection and localization. *Applied Sciences*, 13(16):9158, 2023. ISSN 2076-3417. doi: 10.3390/app13169158. URL <https://www.mdpi.com/2076-3417/13/16/9158>.
- Ahmed W Moawad, Anastasia Janas, Ujjwal Baid, Divya Ramakrishnan, Rachit Saluja, Nader Ashraf, Leon Jekel, Raisa Amiruddin, Maruf Adewole, Jake Albrecht, et al. The brain tumor segmentation (brats-mets) challenge 2023: Brain metastasis segmentation on pre-treatment mri. *arXiv preprint arXiv:2306.00838*, 2023. URL <https://arxiv.org/abs/2306.00838>.
- Ilkay Oksuz, Bram Ruijsink, Esther Puyol-Antón, James R. Clough, Gastao Cruz, Aurelien Bustin, Claudia Prieto, Rene Botnar, Daniel Rueckert, Julia A. Schnabel, and Andrew P. King. Automatic cnn-based detection of cardiac mr motion artefacts using k-space data augmentation and curriculum learning. *Medical Image Analysis*, 55:136–147, 2019. ISSN 1361-8415. doi: <https://doi.org/10.1016/j.media.2019.04.009>. URL <https://www.sciencedirect.com/science/article/pii/S1361841518306765>.
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. doi: 10.48550/arXiv.2306.14824. URL <https://arxiv.org/abs/2306.14824>.
- Divya Ramakrishnan, Leon Jekel, Saahil Chadha, Anastasia Janas, Harrison Moy, Nazanin Maleki, Matthew Sala, Manpreet Kaur, Gabriel Cassinelli Petersen, Sara Merkaj, Marc von Reppert, Ujjwal Baid, Spyridon Bakas, Claudia Kirsch, Melissa Davis, Khaled Bousabarah, Wolfgang Holler, MingDe Lin, Malte Westerhoff, Sanjay Aneja, Fatima Memon, and Mariam S. Aboian. A large open access dataset of brain metastasis 3d segmentations on MRI with clinical and imaging information. *Scientific Data*, 11(1):254, 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03021-9. URL <https://www.nature.com/articles/s41597-024-03021-9>.
- Dhanush Babu Ramesh, Rishika Iytha Sridhar, Pulakesh Upadhyaya, and Rishikesan Kamaleswaran. Lung grounded-SAM (LuGSAM): A novel framework for integrating text prompts to segment anything model (SAM) for segmentation tasks of ICU chest x-rays.

- TechRxiv*, 2023. doi: 10.36227/techrxiv.24224761.v1. URL <http://dx.doi.org/10.36227/techrxiv.24224761.v1>.
- Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438, 2019. doi: 10.1109/ICCV.2019.00852. URL <https://ieeexplore.ieee.org/document/9009553>.
- Miaojing Shi and Vittorio Ferrari. Weakly supervised object localization using size estimates. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 105–121. Springer, 2016. doi: 10.1007/978-3-319-46454-1_7. URL https://link.springer.com/chapter/10.1007/978-3-319-46454-1_7.
- Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19844–19854, 2023. doi: 10.1109/ICCV51070.2023.01817. URL <https://ieeexplore.ieee.org/document/10376855>.
- Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: robust segmentation of 104 anatomical structures in CT images. *arXiv preprint arXiv:2208.05868*, 2023. doi: 10.48550/arXiv.2208.05868. URL <http://arxiv.org/abs/2208.05868>.
- Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Mustafa Nasir-Moin, Naofumi Tomita, Lorenzo Torresani, Jason Wei, and Saeed Hassanpour. Learn like a pathologist: Curriculum learning by annotator agreement for histopathology image classification. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2472–2482, 2021. doi: 10.1109/WACV48630.2021.00252. URL <https://ieeexplore.ieee.org/document/9423073>.
- Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5238–5246. PMLR, 10–15 Jul 2018. URL <https://proceedings.mlr.press/v80/weinshall18a.html>.
- Hao Wu, Jun Sun, and Qi You. Semi-supervised learning for medical image classification based on anti-curriculum learning. *Mathematics*, 11(6):1306, 2023. ISSN 2227-7390. doi: 10.3390/math11061306. URL <https://www.mdpi.com/2227-7390/11/6/1306>.
- Yuxin Xie, Tao Zhou, Yi Zhou, and Geng Chen. Simtxtseg: Weakly-supervised medical image segmentation with simple text cues. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15008. Springer Nature Switzerland, October 2024. ISBN 978-3-031-72111-3. doi: 10.1007/978-3-031-72111-3_60. URL https://link.springer.com/chapter/10.1007/978-3-031-72111-3_60.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moungh-Wen, Brian Piening, Carlo Bifulco, Mu Wei, Hoifung Poon, and Sheng Wang. BiomedParse: a biomedical foundation model for image parsing of everything everywhere all at once. *arXiv preprint arXiv:2405.12971*, 2024a. doi: 10.48550/arXiv.2405.12971. URL <http://arxiv.org/abs/2405.12971>.

Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024b. doi: 10.48550/arXiv.2401.02361. URL <http://arxiv.org/abs/2401.02361>.

Appendix A. Training Loss and Validation Scores for all Trained Models

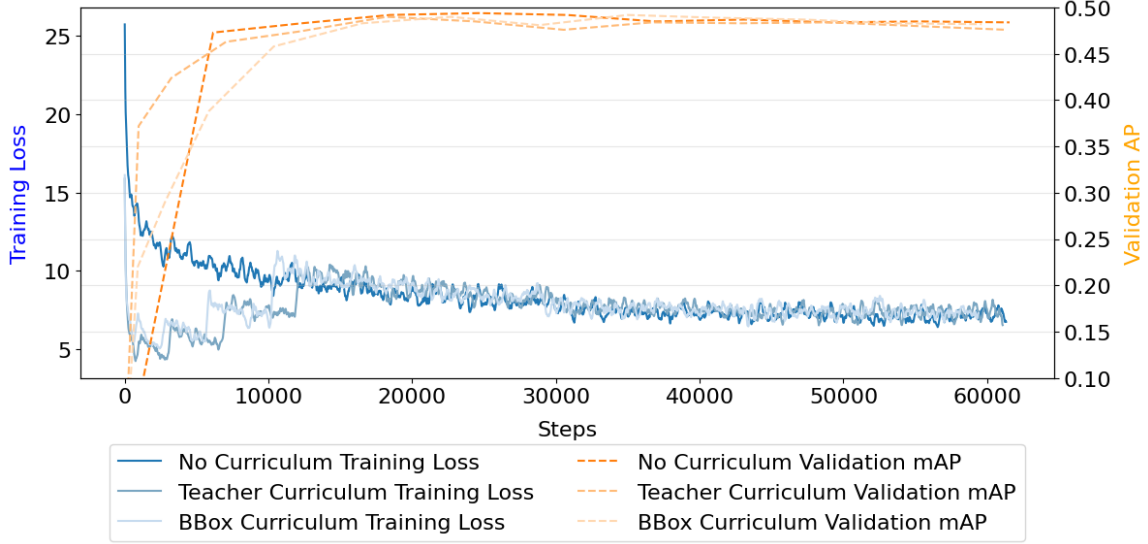


Figure A1: Training loss curves and validation AP plotted for the baseline and the two CL-models. The x-axis denotes training steps and the vertical lines at the top of the graph denote the start of new epochs for each model.

Appendix B. Data Distribution for Bounding Box Curriculum

Area Interval \ Datasets	[0, 0.21)	[0.21, 0.72)	[0.72, 1.75)	[1.75, 3.49)	[3.49, 100]	total
Yale_BM	2037	1596	881	509	167	5190
BraTS_GLI	3825	5186	7933	10346	10783	38073
BraTS_MET	2969	1768	1067	725	227	6756
MSD_Liver	2897	1294	510	196	225	5122
MSD_Hep_Vessel	717	917	758	383	504	3279
KiTS23_Kidney	1200	2619	1934	1340	1477	8570
total	13645	13380	13083	13499	13383	66990

Table A1: Distribution of image slices with ground truth annotations across datasets for the bounding box sorting approach: Based on the smallest bounding box present in a slice, the slice gets sorted into a particular difficulty interval. The intervals are defined by standardized area of the bounding box, with the smallest area intervals being the hardest category.

Appendix C. Data Distribution for Teacher Curriculum

AP Interval \ Datasets	[0, 0.30)	[0.30, 0.69)	[0.69, 0.87)	[0.87, 0.9)	[0.9, 1.0]	total
Yale_BM	878	1166	1875	1069	399	5387
BraTS_GLI	6245	5242	7198	10281	10437	39403
BraTS_MET	1232	1704	1993	1349	698	6976
MSD_Liver	1181	2102	1337	531	166	5317
MSD_Hep_Vessel	1024	863	894	509	235	3525
KiTS23_Kidney	1676	1101	2160	2288	1854	9079
total	12236	12178	15457	16027	13789	69687

Table A2: Distribution of image slices with ground truth annotations and false positive predictions across datasets for the teacher sorting approach: Based on the baseline inference performance, the slice gets sorted into a particular difficulty interval. The intervals are defined by AP scores, with the lowest AP score intervals being the hardest category.

Appendix D. Evaluation Scores for Anti-Curriculum Approaches

Dataset	Curr.	AP (%) @0.5	AP (%) @0.75	AP (%)	AP (%) large	AP (%) medium	AP (%) small
Overall	Without	69.7	50.6	46.5	69.3	62.0	30.9
	Anti Box	73.9	54.2	49.5	67.6	64.0	34.1
	Anti Teacher	75.5	56.2	51.2	70.1	65.1	35.8
Yale_BM	Without	61.5	48.2	42.2	-	69.9	37.3
	Anti Box	79.2	64.3	56.0	-	81.9	51.1
	Anti Teacher	79.2	63.5	55.7	-	80.4	51.8
BraTS_MET	Without	66.7	54.1	45.9	-	76.9	42.2
	Anti Box	82.4	65.9	57.0	-	81.9	53.9
	Anti Teacher	83.5	68.3	58.1	-	83.9	55.2
BraTS_GLI	Without	84.8	72.2	65.8	-	81.1	48.1
	Anti Box	85.1	72.2	65.9	-	80.9	48.7
	Anti Teacher	85.9	72.8	66.0	-	81.2	48.6
MSD_Liver	Without	61.4	32.1	33.2	66.6	46.8	19.9
	Anti Box	55.3	25.1	28.0	66.5	41.7	15.1
	Anti Teacher	57.2	28.8	30.2	65.0	42.2	18.1
MSD_Hep_Vessel	Without	69.4	37.7	39.2	62.8	43.7	15.0
	Anti Box	72.0	42.8	41.5	62.7	46.6	15.7
	Anti Teacher	70.4	42.3	41.6	64.8	45.9	17.5
KiTS23_Kidney	Without	74.6	59.5	52.9	78.4	53.8	22.6
	Anti Box	69.2	54.9	48.7	73.6	50.7	20.1
	Anti Teacher	76.9	61.7	55.3	80.4	57.0	23.6

Table A3: Overview of the detection accuracies for the anti-curriculum models on all datasets as well as an overall score (mean over the 6 datasets).