COMPARATIVE EVALUATION OF DEEP-LEARNING ARCHITECTURES FOR 3D THORACIC ORGAN-AT-RISK SEGMENTATION ON PRIVATE CLINICAL CT DATA

Denisa-Daniela FRIMU-PASCU[1], Ciprian DOBRE[2]

Abstract. *Automatic segmentation of thoracic organs at risk (OARs) can reduce contouring burden in radiotherapy, but performance depends on architecture, training recipe, and annotation availability. In this, we compared nnU-Net, STU-Net-S, MedNeXt-S, and TotalSegmentator-initialized STU-Net-B on a fixed private thoracic CT cohort of 159 evaluable cases and 15 OARs. The mean Dice was 0.653, 0.725, 0.734, and 0.748, respectively, with STU-Net-B being ranked highest across all organs, whereas MedNeXt-S led on better-supported structures. Thus, we demonstrate that model ranking is anatomy- and evidence-support-dependent, supporting organ-aware rather than single-score model selection.*

Keywords: medical image segmentation; thoracic CT; organs at risk; nnU-Net; STU-Net; MedNeXt; radiotherapy; partial annotation

DOI 10.56082/annalsarsciinfo.2026.2.54

1. Introduction

Accurate delineation of organs at risk is an important component of radiotherapy planning. Manual contouring requires substantial clinical effort and is affected by anatomical complexity, image quality, observer variability, treatment-specific requirements, and the number of structures that must be reviewed. Deep-learning-based segmentation, on the other hand, offers the possibility of reducing repetitive contouring work while provâiding consistent initial anatomical delineations for expert review.

U-shaped convolutional neural networks established the dominant architectural paradigm for biomedical image segmentation [1], while volumetric extensions demonstrated the benefit of using three-dimensional spatial context when processing medical images [2]. More recent systems have moved beyond a single fixed U-Net design. nnU-Net introduced a self-configuring approach in which preprocessing, patch geometry, architecture and training parameters are adapted to dataset characteristics [3]. STU-Net subsequently investigated systematic scaling and transfer learning for three-dimensional medical segmentation [4], while MedNeXt modernised volumetric convolutional networks using ConvNeXt-inspired residual blocks and large-kernel

[1]PhD Student, Doctoral School of Automatic Control and Computers, National University of Science and Technology POLITEHNICA Bucharest, Bucharest, Romania.
[2]Prof., Faculty of Automatic Control and Computers, National University of Science and Technology POLITEHNICA Bucharest, Bucharest, Romania, (ciprian.dobre@upb.ro).

processing [5]. Large multi-organ resources such as TotalSegmentator have additionally created opportunities for supervised pretraining across broad anatomical label spaces [6].

These developments raise an important practical question: does a more recent or larger architecture consistently provide superior performance when transferred to heterogeneous clinical radiotherapy data?

Answering this question is less straightforward than comparing one global Dice score. Real retrospective clinical contour archives differ substantially from curated research benchmarks. Not every structure is delineated in every patient, because contours are created according to treatment purpose, anatomical coverage and clinical need. Missing annotation therefore cannot automatically be interpreted as absence of the anatomical structure. Partial-label learning research has shown that treating unlabelled structures as conventional background may introduce incorrect supervision [7].

Evaluation is similarly affected. A global average can combine structures with very different anatomical characteristics and very different numbers of available reference contours. Metric-selection frameworks emphasise that segmentation performance should be interpreted according to the target anatomy, reference data, error characteristics and intended use rather than through a single universal metric [8].

The present study reports a comparative evaluation performed during doctoral research on private clinical radiotherapy CT data. Four three-dimensional segmentation configurations were evaluated:

1. nnU-Net v2 3D full-resolution as the reference configuration;
2. STU-Net-S trained from random initialisation;
3. MedNeXt-S trained from random initialisation;
4. STU-Net-B initialised from TotalSegmentator-pretrained weights.

The primary objective was to compare these configurations on the same fixed thoracic evaluation cohort using structure-specific ground-truth denominators. A second objective was to determine whether the apparent ranking of the models remained stable when the analysis was restricted to structures supported by larger numbers of annotated cases.

The central hypothesis was that model performance would be anatomy dependent and evaluation dependent, such that no architecture could be considered universally superior from a single aggregate score.

2. Material and method

2.1. Clinical imaging data

The experimental data originated from a private retrospective radiation-oncology imaging archive. The three-dimensional experimental programme contained 156 prepared model-development cases according to the executable dataset metadata.

A separate fixed evaluation registry contained 160 thoracic patient records. One record could not be reconstructed because the corresponding raw DICOM CT data were unavailable and was permanently excluded before final scoring. The resulting evaluation cohort contained 159 evaluable thoracic cases.

The 15-structure segmentation ontology consisted of: heart, left lung, right lung, spinal cord, esophagus, liver, left kidney, right kidney, trachea, thyroid, stomach, pancreas, spleen, aorta, and duodenum. The presence of upper-abdominal structures reflects the anatomical coverage of thoracic radiotherapy CT acquisitions and the contours available in the clinical archive. The clinical data were de-identified for research analysis and were used under the confidentiality and research-data-use arrangements established for the doctoral research.

2.2. Annotation availability

The clinical archive was partially annotated. Consequently, all 15 structures were not available as reference contours in every CT study. The numbers of cases containing valid ground-truth contours ranged from 159 for the spinal cord to only 8 for the duodenum. Heart and lungs were represented in most cases, whereas several upper-abdominal and less routinely delineated structures had substantially smaller denominators. In our case, a missing contour was not treated as evidence that the anatomical structure was absent. Standard overlap metrics were therefore calculated only for cases in which a valid reference contour for the corresponding structure was available.

The study documentation describes a multidisciplinary annotation workflow involving radiotherapy physicians, radiation-oncology residents, a clinical oncologist and a medical physicist, including review of initial contours and resolution of disagreements through discussion. However, a complete quantitative inter-observer analysis was not available and is therefore not claimed in this study.

2.3. DICOM conversion and preprocessing

Source CT series and RTSTRUCT contours were converted to volumetric image-label representations for three-dimensional processing. In particular, DICOM CT slices were ordered according to physical slice-position metadata rather than file names. During technical quality assurance, a z-orientation problem affecting an earlier conversion was identified. The conversion procedure was corrected on 23 June 2026, the evaluation volumes were reconstructed, predictions were regenerated, and all orientation-affected earlier evaluation results were superseded.

The executable nnU-Net planning metadata specified a target spacing of 3.00 × 1.07421875 × 1.07421875 mm (z × y × x).

For the nnU-Net reference configuration, the patch size was 160 × 128 × 80 voxels and the batch size recorded in the executable planning file was 2.

CT normalisation followed the nnU-Net preprocessing configuration. The executable dataset fingerprint recorded foreground bounds of 0 and 170 on the stored model-ready intensity scale, with a mean of 71.95 and standard deviation of 67.63. Because the preserved conversion metadata do not establish that this scale corresponds directly to unmodified native Hounsfield units, these values are treated as model-ready intensity statistics rather than native-HU thresholds.

Spatial and intensity augmentation in the reference workflow included the standard nnU-Net operations, such as rotation, scaling, Gaussian noise and blur, brightness and contrast modification, simulated low resolution, gamma transformation and mirroring.

2.4. Compared segmentation configurations

The benchmark compared four contemporary three-dimensional segmentation configurations. Their principal experimental characteristics are summarised in Table 1.

Table 1. Experimental configuration of the four segmentation arms.

Model	*Initialisation*	*Patch / batch*	*Sampling*	*Initial learning rate*	*Training characteristic*
nnU-Net v2	Standard nnU-Net	160×128×80 / 2	nnU-Net default	Framework configuration	Long-budget reference; best validation EMA pseudo-Dice checkpoint near epoch 377
STU-Net-S	Random	160×128×80 / 2	Foreground oversampling = 1.0	0.01	12-hour training cap
MedNeXt-S	Random	160×128×64 / 1	OAR-centred crops	0.01, SGD	12-hour training cap; sliding-window inference
STU-Net-B	TotalSegmentator pretrained	160×128×80 / 1	Foreground oversampling = 1.0	0.001	12-hour fine-tuning cap

The nnU-Net reference used a six-stage three-dimensional PlainConvUNet with feature widths of 32, 64, 128, 256, 320 and 320, instance normalisation and leaky-ReLU activation. STU-Net-S represents the smaller scalable U-Net configuration and was trained without TotalSegmentator pretraining. MedNeXt-S represents a modern convolutional alternative employing ConvNeXt-inspired volumetric blocks and larger convolutional kernels. The local implementation used a smaller patch depth and organ-centred cropping because of computational and memory

requirements. Finally, STU-Net-B combines a larger STU-Net configuration with supervised initialisation from TotalSegmentator-derived weights.

The experiment should therefore not be interpreted as a pure architecture ablation. The models were evaluated on a common task and fixed evaluation cohort, but training duration, initialisation, patch geometry, sampling and batch size were not identical. The results consequently describe the performance of the tested architecture-and-training recipes rather than proving intrinsic superiority of one architecture independent of training conditions.

2.5. Model selection and evaluation governance

Checkpoint selection for each arm was based on validation evidence rather than evaluation-cohort performance. No within-arm hyperparameter or checkpoint selection was performed using the final evaluation cohort.

The complete four-model research programme developed sequentially during the doctoral work and was not a prospectively blinded architecture trial. The fixed-cohort designation therefore refers to stable final evaluation membership and validation-based model selection rather than to prospective preregistration of all four experiments.

One experimental fold and one recorded seed were available for each final model configuration.

2.6. Quantitative evaluation

The Dice similarity coefficient was used as the principal overlap metric:

$$Dice = 2TP / (2TP + FP + FN)$$

where TP, FP and FN denote true-positive, false-positive and false-negative segmented voxels, respectively.

Additional evaluation outputs included intersection over union, precision, recall, F1 and a common benchmark HD95 output. Because the exact implementation and physical-unit handling of the preserved aggregate HD95 output could not be completely reconstructed, HD95 was retained only as secondary within-benchmark evidence and was not used for external numerical comparison.

For each organ, performance was calculated only among patients for whom a valid ground-truth contour was available.

When valid ground truth was present but no corresponding overlap was obtained, Dice contributed zero to the overlap analysis. Cases without a clinical contour were excluded from conventional organ-level Dice aggregation.

The benchmark also recorded how often a model generated a structure for a case in which no clinical annotation for that structure was available. This quantity is

referred to here as prediction-without-annotation rate, rather than conventional false-positive rate, because an unlabelled organ may still be anatomically present.

The principal aggregate endpoint was the unweighted mean Dice across the 15 structures. Each anatomical structure therefore contributed equally regardless of its number of annotated patients.

2.7. Annotation-support sensitivity analysis

Because annotation availability differed substantially between structures, an additional post hoc descriptive analysis evaluated whether architecture ranking was stable when poorly supported structures were progressively excluded.

Mean Dice was recalculated for all 15 structures, then for structures with ground-truth contours in at least 30 cases. Following this, we increased the structures with ground-truth contours to include at least 50 cases, and finally used the four structures with the highest annotation availability: heart, left lung, right lung and spinal cord.

The N≥30 and N≥50 values are evidence-support filters only. They were not defined as clinical thresholds or prospective statistical endpoints.

The statistical interpretation of this study is descriptive. No claims of statistical superiority are made, because repeated-seed experiments, paired significance testing and bootstrap confidence intervals were not part of the completed benchmark.

3. Results and discussions

3.1. Overall segmentation performance

The four configurations produced different aggregate performance profiles. Table 2 summarises the principal benchmark results.

Table 2. Overall performance of the four 3D segmentation configurations.

Model	*Mean Dice, 15 structures*	*High-availability Dice*	*Reported aggregate HD95**	*Prediction-without-annotation rate*
nnU-Net	0.653	0.746	186.6	0.595
STU-Net-S	0.725	0.760	164.1	0.592
MedNeXt-S	0.734	0.822	239.9	0.604
STU-Net-B	0.748	0.812	141.3	0.588

**HD95 is used only as a descriptive common-output measure within this benchmark because complete implementation and physical-unit metadata were not available for retrospective verification.*

Using the unweighted mean across all 15 structures, STU-Net-B achieved the highest score, with a mean Dice of 0.748, followed by MedNeXt-S at 0.734, STU-Net-S at 0.725, and the long-budget nnU-Net reference at 0.653.

The improvement of the later configurations over the nnU-Net reference was not uniform across all organs. Importantly, the model with the highest overall mean was not the same model that produced the highest performance on the four structures with the greatest annotation availability. MedNeXt-S obtained 0.822 on this subset compared with 0.812 for STU-Net-B.

This already indicates that an overall benchmark average is insufficient to characterise the relative strengths of the models.

3.2. Organ-specific performance

Table 3 presents the mean Dice for each anatomical structure and explicitly reports the corresponding number of available ground-truth cases.

Table 3. Organ-specific mean Dice and structure-specific annotation support.

Structure	*N with GT*	*nnU-Net*	*STU-Net-S*	*MedNeXt-S*	*STU-Net-B*
Heart	150	0.845	0.874	0.873	0.844
Left lung	143	0.652	0.694	0.710	0.778
Right lung	142	0.711	0.688	0.910	0.830
Spinal cord	159	0.776	0.784	0.796	0.794
Esophagus	64	0.606	0.627	0.643	0.646
Liver	61	0.902	0.917	0.918	0.911
Left kidney	57	0.814	0.851	0.842	0.823
Right kidney	54	0.829	0.829	0.830	0.834
Trachea	19	0.747	0.648	0.670	0.725
Thyroid	30	0.628	0.592	0.424	0.649
Stomach	17	0.796	0.744	0.760	0.758
Pancreas	9	0.592	0.578	0.607	0.557
Spleen	11	0.900	0.905	0.864	0.919
Aorta	9	0.000	0.730	0.745	0.754
Duodenum	8	0.000	0.420	0.416	0.400

No model dominated all structures.

STU-Net-B was numerically strongest for the left lung, esophagus, right kidney, thyroid, spleen and aorta. MedNeXt-S led for the right lung, spinal cord, liver and pancreas. STU-Net-S achieved the highest values for the heart and left kidney, while the nnU-Net reference retained the highest numerical Dice for the trachea and stomach.

The differences for several well-supported organs were very small. Heart Dice, for example, was 0.874 for STU-Net-S and 0.873 for MedNeXt-S. Spinal-cord

performance ranged only from 0.776 to 0.796. Liver Dice ranged from 0.902 to 0.918. These differences should therefore be regarded as descriptive numerical differences rather than evidence of clinically or statistically meaningful superiority.

The results instead support the broader conclusion that segmentation performance is anatomy dependent.

3.3. Large and well-supported structures

Large anatomical structures generally achieved relatively strong segmentation performance, although architecture-specific variation remained evident.

The liver produced Dice values above 0.90 for every architecture. Heart performance was similarly consistent, ranging from 0.844 to 0.874.

Lung segmentation showed greater between-model variability. STU-Net-B achieved the highest left-lung Dice at 0.778, whereas MedNeXt-S achieved a markedly higher right-lung Dice of 0.910.

The MedNeXt-S left-right lung results were asymmetric, with Dice values of 0.710 for the left lung and 0.910 for the right lung. The available experimental evidence does not establish whether this difference arose from cohort characteristics, annotations, model behaviour, crop configuration or another pipeline component. A causal interpretation of the asymmetry would therefore be unsupported.

3.4. Smaller and less consistently annotated structures

Performance interpretation becomes substantially less certain for structures represented by few reference annotations. The aorta, duodenum, pancreas and spleen had only 9, 8, 9 and 11 annotated cases, respectively. Although STU-Net-S, MedNeXt-S and STU-Net-B generated considerably higher aortic Dice values than the nnU-Net reference, these results are based on a very small denominator. Similarly, duodenal Dice ranged from 0.400 to 0.420 for the later comparative arms, but only eight cases contained a valid ground-truth contour. These results remain useful as exploratory evidence of model behaviour but should not be interpreted as stable population-level rankings.

The difficulty is not merely statistical. In retrospective radiotherapy data, low annotation availability can reflect clinical contouring practice, anatomical coverage and treatment relevance. Consequently, performance on sparsely represented structures captures interactions among model architecture, sampling, supervision and data availability.

3.5. Model-ranking sensitivity to annotation support

The most important secondary finding was that the identity of the highest-scoring model changed when the aggregate score was restricted to better-supported anatomical structures.

Table 4. Sensitivity of mean Dice ranking to minimum structure-specific annotation support.

Model	*All 15 structures*	*N≥30, 9 structures*	*N≥50, 8 structures*	*Four highest-availability structures*
nnU-Net	0.653	0.751	0.767	0.746
STU-Net-S	0.725	0.762	0.783	0.760
MedNeXt-S	0.734	0.772	0.815	0.822
STU-Net-B	0.748	0.790	0.808	0.812

STU-Net-B remains the numerical leader when all 15 structures were included and when analysis was restricted to the nine structures represented in at least 30 cases. However, when only structures with at least 50 available reference contours were included, MedNeXt-S became the highest-scoring configuration, with a mean Dice of 0.815 compared with 0.808 for STU-Net-B. MedNeXt-S also led the four highest-availability structures, with mean Dice of 0.822 compared with 0.812 for STU-Net-B.

Importantly, none of the underlying per-organ predictions changed during this analysis. Only the rule determining which structures contributed to the aggregate score changed.

This ranking reversal demonstrates that a segmentation leaderboard is partly a property of the aggregation policy and evidence distribution, rather than solely a property of the network architecture. For heterogeneous clinical datasets, reporting the number of reference cases available for every organ is therefore necessary for meaningful model comparison.

3.6. Interpretation of architectural differences

The tested architectures represent different strategies for processing three-dimensional anatomical information.

nnU-Net provides a highly engineered self-configuring baseline. Its strength lies not only in the U-Net topology but also in dataset fingerprinting, preprocessing, patch planning, augmentation, optimisation and inference decisions [3]. Consequently, it remains an important reference against which newer architectures should be evaluated.

STU-Net extends the U-Net family through systematic depth and width scaling and was designed to benefit from large-scale supervised pretraining [4]. In the present

experiment, the small STU-Net-S configuration was trained from random initialisation, whereas STU-Net-B was initialised from TotalSegmentator-derived weights.

The higher all-structure Dice obtained by STU-Net-B is compatible with the hypothesis that increased model capacity and transferred anatomical representations can be beneficial for heterogeneous multi-organ segmentation. However, the experiment cannot isolate these effects. STU-Net-B differed simultaneously in architecture size, initialisation, batch size and optimisation settings. The result therefore does not demonstrate that model size or pretraining alone caused the improvement.

MedNeXt follows a different strategy. Its ConvNeXt-inspired volumetric design uses modern residual blocks and large convolutional kernels to increase contextual processing while retaining convolutional inductive biases [5]. Its strong performance on the heart, lungs and spinal cord suggests that this processing strategy can be competitive for well-represented anatomical structures.

However, the MedNeXt experiment also used organ-centred crops and a different patch depth, meaning that the observed performance reflects the complete architecture-and-training recipe.

This distinction is important for the interpretation of applied deep-learning research. An architecture name is rarely the only independent variable. Preprocessing, sampling, supervision, initialisation, optimisation, inference strategy and training duration are part of the computational system that produces the final segmentation.

3.7. Partial annotation as an evaluation problem

Partial annotation was a major property of the clinical cohort.

A structure that is not represented by an RTSTRUCT contour cannot automatically be treated as anatomically absent. A model prediction in such a case may represent correct but unannotated anatomy. Calling every such prediction a false positive would therefore overstate segmentation error.

The prediction-without-annotation rates were similar across the four configurations, ranging from 0.588 to 0.604. These values should be interpreted primarily as indicators of interaction between the model output and incomplete clinical annotation, rather than as conventional anatomical false-positive rates.

The same issue potentially affects model training. No explicit missing-label mask, marginalised loss or equivalent partial-label-aware objective could be confirmed in the available final training configurations. Future experiments should therefore investigate losses that distinguish unlabelled anatomy from true background, such

as marginal and exclusion approaches developed specifically for partially supervised multi-organ segmentation [7].

3.8. Implications for model selection

The implications are as follows. If model selection had been based exclusively on the complete 15-organ mean Dice, STU-Net-B would have been selected. If the priority had instead been performance on the best-supported structures, MedNeXt-S would have been selected.

For an individual anatomical target, the preferred architecture could change again.

The practical implication is that there is no universally optimal model in the present benchmark. Model selection should instead depend on:

- the target anatomical structures;
- the amount and quality of reference evidence available for those structures;
- the clinical or engineering endpoint;
- the consequences of missed or over-segmented anatomy;
- the computational constraints of deployment;
- the training and transfer-learning strategy;
- patient-level failure behaviour in addition to aggregate performance.

A deployment-oriented segmentation system may ultimately benefit from architecture selection or model routing according to anatomy rather than enforcing one network as the optimal solution for every structure.

3.9. Limitations

First, this is a single private clinical dataset and no independent external clinical validation cohort was evaluated.

Second, one fold and one recorded seed were available for each architecture. The experiment therefore does not quantify variability across random initialisations or data partitions.

Third, the models were not trained with identical computational budgets. The nnU-Net reference had a substantially longer training trajectory, while the three later comparative arms were constrained to approximately 12 hours.

Fourth, STU-Net-B used supervised pretraining whereas STU-Net-S and MedNeXt-S were randomly initialised. The STU-Net-B result therefore reflects both architecture and transfer learning.

Fifth, MedNeXt-S used a different patch depth and organ-centred sampling strategy.

Sixth, annotation availability varied substantially by organ. Low-denominator results, particularly for aorta, duodenum, pancreas and spleen, should therefore be regarded as exploratory.

Seventh, explicit partial-label-aware optimisation was not documented for the final model-training configurations.

Eighth, the preserved aggregate HD95 implementation and physical-unit conventions were not sufficiently complete for the metric to support external quantitative comparison.

Finally, the benchmark provides technical segmentation evidence and does not by itself establish clinical acceptability, treatment-planning equivalence, reduction in clinician workload or improved patient outcomes. Those questions require prospective clinical evaluation.

4. Conclusions

This study compared nnU-Net, STU-Net-S, MedNeXt-S and TotalSegmentator-initialised STU-Net-B for three-dimensional multi-organ segmentation on a fixed private thoracic radiotherapy CT cohort containing 159 evaluable cases.

STU-Net-B achieved the highest unweighted mean Dice across the complete 15-structure ontology, with 0.748, compared with 0.734 for MedNeXt-S, 0.725 for STU-Net-S and 0.653 for nnU-Net.

However, the ranking was not invariant. MedNeXt-S became the highest-scoring configuration when the evaluation was restricted to structures supported by at least 50 reference contours and also led the four highest-availability structures.

The results therefore do not support the conclusion that one architecture is universally superior. Instead, they demonstrate an interaction among architecture, training recipe, anatomy, annotation availability and aggregation policy.

For heterogeneous real-world clinical data, architecture evaluation should consequently move beyond a single global Dice score. Structure-specific denominators, organ-level results and sensitivity of aggregate rankings to evidence support should be reported explicitly.

Future work should combine repeated-seed evaluation, external validation, partial-label-aware optimisation and systematic patient-level failure analysis. These additions would permit a stronger distinction between genuine architectural effects and performance differences produced by training configuration or data composition.

The principal engineering conclusion of this study is therefore that model selection for three-dimensional OAR segmentation should be anatomy-aware, evidence-aware and purpose-specific rather than architecture-name-driven.

R E F E R E N C E S

[1] O. Ronneberger, P. Fischer, T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, LNCS 9351, pp. 234–241, 2015. doi:10.1007/978-3-319-24574-4_28.

[2] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, *3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation*, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, LNCS 9901, pp. 424–432, 2016. doi:10.1007/978-3-319-46723-8_49.

[3] F. Isensee, P.F. Jaeger, S.A.A. Kohl, J. Petersen, K.H. Maier-Hein, *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation*, Nature Methods, vol. 18, pp. 203–211, 2021.

[4] Z. Huang, H. Wang, Z. Deng, J. Ye, Y. Su, H. Sun, J. He, Y. Gu, L. Gu, S. Zhang, Y. Qiao, *STU-Net: Scalable and Transferable Medical Image Segmentation Models Empowered by Large-Scale Supervised Pre-training*, arXiv:2304.06716, 2023.

[5] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P.F. Jäger, K.H. Maier-Hein, *MedNeXt: Transformer-driven Scaling of ConvNets for Medical Image Segmentation*, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023, pp. 405–415, 2023. doi:10.1007/978-3-031-43901-8_39.

[6] J. Wasserthal, H.C. Breit, M.T. Meyer, et al., *TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images*, Radiology: Artificial Intelligence, vol. 5, no. 5, e230024, 2023. doi:10.1148/ryai.230024.

[7] G. Shi, L. Xiao, Y. Chen, S.K. Zhou, *Marginal loss and exclusion loss for partially supervised multi-organ segmentation*, Medical Image Analysis, vol. 70, 101979, 2021. doi:10.1016/j.media.2021.101979.

[8] L. Maier-Hein et al., *Metrics Reloaded: Recommendations for image analysis validation*, Nature Methods, vol. 21, no. 2, pp. 195–212, 2024. doi:10.1038/s41592-023-02151-z.

[9] A.A. Taha, A. Hanbury, *Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool*, BMC Medical Imaging, vol. 15, article 29, 2015. doi:10.1186/s12880-015-0068-x.