eISSN 2783-6851

VILNIUS TECH
Vilnius Gediminas Technical University

NEW TRENDS in COMPUTER SCIENCES

2026
Volume 4
Issue 2
Pages 208–226
https://doi.org/10.3846/ntcs.2026.26765

RESOLUTION PRESERVATION OVER ARCHITECTURAL COMPLEXITY: HIGH-RESOLUTION PATCH-BASED SEMI-SUPERVISED LEARNING FOR PANCREAS SEGMENTATION FROM CT IMAGES

Istiaque AHMED [1]✉, Shinthiya Nowsain PROMI [2], Kursat KOMURCU [3]

[1]*Institute of Computer Science, Faculty of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania*
[2]*Faculty of Fundamental Sciences, Vilnius Gediminas Technical University, Vilnius, Lithuania*
[3]*Artificial Intelligence Methods Laboratory, Institute of Computer Science, Faculty of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania*

Article History:
- received 21 April 2026
- accepted 18 September 2026

Abstract. Precise segmentation of the pancreas from computed tomography (CT) scans is critical for diagnosis and surgical planning but remains challenging due to the organ's small volume (<0.5% of the abdomen), high anatomical variability, and low contrast with surrounding tissues. Conventional deep learning methods typically downsample inputs to fit GPU memory, destroying fine-grained boundary information essential for small-organ delineation. This study proposes a high-resolution patch-based framework that preserves native CT resolution (512 × 512) through 256 × 256 patch extraction with domain-specific Hounsfield unit (HU) windowing. A lightweight 2D U-Net (7.8M parameters) achieves 0.809±0.015 Dice (peak 0.849) on the Medical Segmentation Decathlon (MSD) Task07 Pancreas dataset, competitive with nnU-Net (0.823 Dice, 88M parameters) and far above SAM (0.705) and MedSAM (0.439). Five-fold cross-validation across all 281 volumes yields a population-level Dice of 0.630±0.025. Integrating uncertainty-aware mean teacher (UA-MT) semi-supervised learning recovers 99.3% of supervised performance using only 50% of labelled data, with reproducible results across three seeds (σ = 0.012). On 80 external TCIA volumes, UA-MT (0.603 Dice) outperforms the supervised model (0.544), demonstrating superior cross-domain generalisation.

Keywords: pancreas segmentation, computed tomography, deep learning, resolution preservation, semi-supervised learning, uncertainty-aware mean teacher, patch-based segmentation, medical image segmentation, annotation efficiency.

✉Corresponding author. E-mail: *iah.utsho@gmail.com*

Notations

Variables and functions

λ – consistency weight of the semi-supervised loss;
α – exponential moving average decay of the teacher network;
T – number of Monte Carlo dropout forward passes;
τ – confidence threshold for FixMatch pseudo-labelling.

Flowchart symbols (Figure 1)

The pipeline in Figure 1 is drawn in the standard flowchart notation of ISO 5807 (equivalently, the ANSI flowchart symbol set): a rounded rectangle marks the start and the end of the procedure; a parallelogram marks input or output data; a plain rectangle marks a processing step; a rectangle with double vertical bars marks a predefined process, that is, a self-contained

Copyright © 2026 The Author(s). Published by Vilnius Gediminas Technical University

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

sub-procedure specified elsewhere in the text; a diamond marks a decision, with its outgoing branches labelled Yes and No; and a directed arrow marks the direction of control flow.

Abbreviations

ASD – average symmetric surface distance;
BCE – binary cross-entropy;
CNN – convolutional neural network;
CPS – cross-pseudo supervision;
CT – computed tomography;
DSC – Dice similarity coefficient;
EMA – exponential moving average;
HD95 – 95th-percentile Hausdorff distance;
HU – Hounsfield unit;
IoU – intersection over union;
MC – Monte Carlo;
MSD – Medical Segmentation Decathlon;
MT – mean teacher;
SAM – Segment Anything Model;
SSL – semi-supervised learning;
TCIA – The Cancer Imaging Archive;
UA-MT – uncertainty-aware mean teacher;
ViT – vision transformer.

1. Introduction

Pancreatic cancer is the third leading cause of cancer-related mortality in the United States, with a five-year survival rate of approximately 12% (Siegel et al., 2025). This dismal prognosis is largely attributable to late-stage diagnosis, as early pancreatic lesions are often clinically silent. Automated segmentation of the pancreas from abdominal CT scans is therefore a necessary step towards computer-aided diagnosis, surgical planning, and longitudinal disease monitoring. However, the pancreas is difficult to segment, it occupies a minute fraction (<0.5%) of the total abdominal volume, exhibits high inter-patient shape variability, and shares similar Hounsfield unit (HU) intensities with neighbouring organs such as the stomach, duodenum, and surrounding adipose tissue.

To address these challenges, deep learning approaches have been extensively deployed. The U-Net architecture (Ronneberger et al., 2015) and its volumetric extensions, V-Net (Milletari et al., 2016) and 3D U-Net (Çiçek et al., 2016), established the encoder-decoder paradigm with skip connections as the de facto standard for medical image segmentation. Subsequent work introduced attention mechanisms (Oktay et al., 2018) and recurrent saliency transformations (Q. Yu et al., 2018) to guide the network towards the small pancreatic region. More recently, vision transformers (Dosovitskiy et al., 2021), adapted for medical imaging through architectures such as TransUNet (J. Chen et al., 2024) and Swin-UNet (Cao et al., 2023), have demonstrated strong performance on large-organ segmentation tasks by capturing

long-range global dependencies. However, these architectures impose severe GPU memory demands, and a common practice is to aggressively downsample input volumes to fit entire 3D volumes or complex multi-layer transformer blocks into memory (Luo, Wang, et al., 2022).

In this paper, we question this common practice. We hypothesise that for small, low-contrast anatomical structures like the pancreas, the loss of fine-grained local texture and boundary sharpness caused by downsampling outweighs the benefits of advanced architectural complexity or global context. Despite the growing reliance on resolution-reducing pipelines, few studies have systematically isolated the effect of spatial resolution from architectural complexity for small-organ segmentation. Furthermore, consistency-based semi-supervised methods have shown promise in mitigating annotation scarcity (Yu et al., 2019), yet their performance remains tethered to the underlying spatial resolution of the feature maps.

The main objective of this study is therefore to quantify, through controlled experiments, the relative contribution of spatial-resolution preservation versus architectural complexity in pancreas segmentation from CT, and to determine whether a resolution-preserving patch-based pipeline can simultaneously match the accuracy of much larger models and reduce the annotation burden when combined with semi-supervised learning. The findings are intended to inform the design of practical clinical decision-support tools: hospitals operating on commodity GPUs benefit from a lightweight model that runs on standard hardware, radiology departments with limited annotation budgets benefit from the semi-supervised configuration that recovers 99.3% of fully supervised performance from only half the labels, and screening programmes that must generalise across scanner protocols benefit from the cross-domain robustness we observe on the external TCIA cohort.

This work makes five principal contributions. First, we provide a controlled resolution ablation study demonstrating that downsampling CT slices causes a severe collapse in segmentation performance for the pancreas, showing that spatial resolution is the primary performance driver. Second, we quantify the impact of domain-specific HU windowing ([–125, 275]), showing a +5.0% absolute Dice gain over standard abdominal windowing. Third, we demonstrate that a simple patch-based CNN operating at native resolution achieves competitive performance with nnU-Net (Isensee et al., 2021) and substantially outperforms vision transformers and the foundation model SAM (Kirillov et al., 2023). Fourth, we integrate the resolution-preserving framework with UA-MT semi-supervised learning and show that it recovers 99.3% of fully supervised performance using only 50% labelled data, with reproducible results across three independent seeds. Fifth, we validate the framework using 5-fold cross-validation across all 281 MSD volumes and on 80 external TCIA volumes without retraining, demonstrating consistent cross-dataset generalisation.

The remainder of this paper is organised as follows. Section 2 reviews related work on pancreas segmentation, the resolution–complexity trade-off, vision transformers and foundation models, and semi-supervised learning. Section 3 describes the proposed high-resolution patch-based framework and its experimental setup. Section 4 reports the empirical results, covering preprocessing and resolution, comparison with established methods, annotation efficiency, and cross-dataset generalisation. Section 5 discusses the findings and their clinical implications, and the final section concludes the paper.

2. Related work

2.1. Pancreas segmentation from CT

The automated segmentation of the pancreas has evolved through several paradigms. Roth et al. (2015) pioneered CNN-based pancreas segmentation with DeepOrgan, employing holistically-nested convolutional networks to localise the organ. This was subsequently extended through spatial aggregation of hierarchical feature maps (Roth et al., 2018), achieving Dice scores of approximately 0.78 on the NIH-82 dataset. The recurrent saliency transformation network (Q. Yu et al., 2018) introduced multi-stage refinement with iterative attention, reaching approximately 0.84 Dice on the same benchmark. Oktay et al. (2018) proposed Attention U-Net, which learns spatial attention gates that suppress irrelevant regions while emphasising relevant features for the pancreas. More recently, large-scale multi-organ segmentation systems such as TotalSegmentator (Wasserthal et al., 2023) have demonstrated broad anatomical coverage but are not specialised for the unique challenges of small, low-contrast organs. Zheng and Luo (2023) proposed extension-contraction transformations to improve pancreas boundary delineation. A common thread across successful pancreas-specific methods is the implicit preservation of spatial resolution through coarse-to-fine strategies; however, no prior work has explicitly studied the resolution-versus-complexity trade-off.

2.2. Resolution and complexity in medical image segmentation

The tension between spatial resolution and model complexity has been a recurring theme in medical image segmentation. The fully convolutional network (Long et al., 2015) established the encoder-decoder paradigm, and U-Net (Ronneberger et al., 2015) refined it with skip connections for fine-grained localisation. V-Net (Milletari et al., 2016) and 3D U-Net (Çiçek et al., 2016) extended this to volumetric data, but the cubic growth of 3D convolutions necessitates aggressive input downsampling. For a typical abdominal CT volume of $512 \times 512 \times D$ voxels, processing the full resolution in 3D requires memory that scales as $O(L^3)$, making it impractical on standard clinical GPUs without resolution reduction.

DeepLab (L.-C. Chen et al., 2018) introduced atrous (dilated) convolutions to maintain spatial resolution without pooling, though its adoption in 3D medical imaging remains limited due to the computational overhead of dilated 3D convolutions. The nnU-Net framework (Isensee et al., 2021) demonstrated that self-configuring pipelines with careful preprocessing and training strategies consistently outperform hand-tuned architectures, achieving near-state-of-the-art results across diverse medical segmentation tasks. However, even nnU-Net's self-configuration often selects reduced input resolutions when GPU memory is constrained. Despite these advances, the fundamental trade-off between input resolution and model capacity remains underexplored for small-organ applications, where the loss of boundary information from downsampling may be severe.

2.3. Vision transformers and foundation models

The vision transformer (Dosovitskiy et al., 2021) demonstrated that pure self-attention mechanisms can rival CNNs on natural image recognition, but typically require massive pre-training datasets. TransUNet (J. Chen et al., 2024) hybridised a CNN encoder with transformer layers for medical images, while Swin-UNet (Cao et al., 2023) proposed a pure transformer U-shaped architecture with shifted-window attention. UNETR (Hatamizadeh et al., 2022) applied vision transformers directly to 3D medical image segmentation. Most recently, foundation models such as the Segment Anything Model (SAM) (Kirillov et al., 2023) and its medical adaptation MedSAM (Ma et al., 2024) have demonstrated broad segmentation capabilities, though their effectiveness on small, low-contrast anatomical structures remains underexplored. Despite their massive scale, these models operate at fixed internal resolutions and their aggressive patch embeddings (typically 16 × 16) discard fine-grained spatial information.

2.4. Semi-supervised learning for medical image segmentation

The scarcity of voxel-level annotations is a known bottleneck in medical imaging (Yang et al., 2023; Jiao et al., 2024). Consistency regularisation has emerged as a dominant paradigm: the mean teacher framework (Tarvainen & Valpola, 2017) enforces prediction consistency between a student and its exponential moving average (the teacher) under input perturbations. L. Yu et al. (2019) introduced uncertainty-aware self-ensembling (UA-MT), filtering unreliable teacher predictions via Monte Carlo dropout uncertainty estimates. On the pseudo-labelling front, FixMatch (Sohn et al., 2020) combined consistency regularisation with confidence-based thresholding, while cross-pseudo supervision (CPS) (X. Chen et al., 2021) employs dual networks that generate pseudo-labels to supervise each other. Luo, Hu, et al. (2022) proposed cross-teaching between CNN and transformer networks for semi-supervised medical segmentation. However, no prior work has combined resolution-preserving input strategies with SSL for small-organ segmentation, which we address in this study.

3. Methodology

The proposed framework, summarised in Figure 1, comprises five stages: domain-specific HU windowing of native 512 × 512 CT slices, 256 × 256 patch extraction that preserves the original voxel spacing, training of a lightweight 2D U-Net (with optional UA-MT semi-supervised consistency loss on unlabelled patches), sliding-window inference with Gaussian blending, and 3D volume reconstruction.

Each stage is designed to preserve native spatial detail until the final reconstruction, so that boundary information is never discarded through early downsampling. The remaining subsections describe the dataset, the resolution-preservation strategy, the network, the semi-supervised extension, the training protocol, and the evaluation metrics in turn.

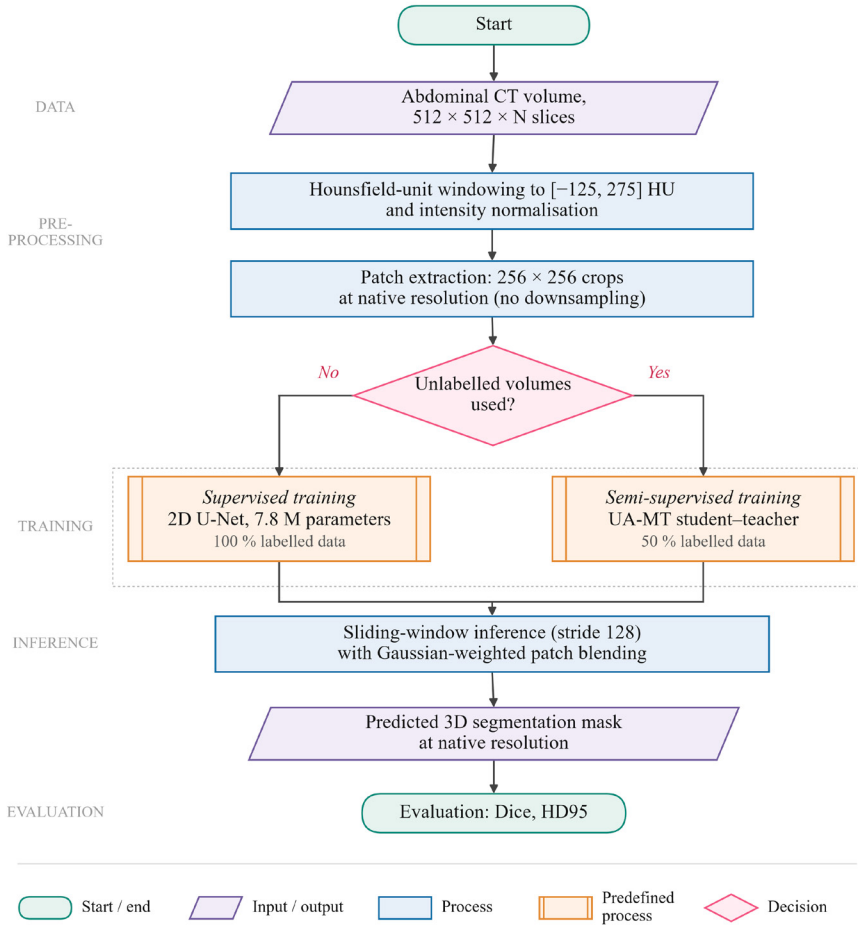


Figure 1. Flowchart of the proposed high-resolution patch-based pancreas segmentation pipeline, drawn in standard flowchart notation (ISO 5807 / ANSI); symbol definitions are provided in the Notations section

3.1. Data and preprocessing

We used the public Medical Segmentation Decathlon (MSD) Task07 Pancreas dataset (Antonelli et al., 2022), consisting of 281 contrast-enhanced portal venous phase abdominal CT volumes with voxel-level annotations for both pancreatic parenchyma and tumour. Following standard practice for pancreas segmentation benchmarks, we merged the parenchyma and tumour labels into a single binary foreground class (pancreas vs. background). The native axial resolution is 512 × 512 pixels with variable slice thickness ranging from 1.0 to 2.5 mm. We partitioned the dataset into 271 volumes for training, 10 for validation, and 4 held-out test

volumes for final 3D volumetric evaluation. After extracting 2D axial slices containing pancreatic tissue, the training set yielded 9,798 image-mask pairs. For cross-dataset evaluation, we additionally acquired the external TCIA Pancreas-CT dataset (N = 80) (Clark et al., 2013), collected at a different clinical site using different scanner protocols.

3.2. Resolution preservation strategy

Standard abdominal windowing clips HU values to [–100, 240]. We empirically identified and applied an optimised window of [–125, 275], which better captures the low-contrast pancreatic boundaries, yielding a +5.0% absolute Dice improvement. Instead of downsampling the native 512 × 512 axial slices, we employed a patch-based extraction strategy. During training, 256 × 256 patches were randomly cropped from the native-resolution slices, serving simultaneously as a data augmentation mechanism through spatial jittering. This approach preserves the original voxel spacing (≈0.7 mm × 0.7 mm) and allows the network to learn from uncompromised pixel data while maintaining manageable GPU memory footprints. During inference, a sliding-window approach with a stride of 128 pixels was used to reconstruct the full 512 × 512 probability map, with Gaussian weighting applied to smooth overlapping predictions at patch boundaries.

3.3. Network architecture

The framework employs a standard 2D U-Net (Ronneberger et al., 2015) adapted for high-resolution patch-based segmentation. The encoder consists of four downsampling levels, each featuring two consecutive 3 × 3 convolutional layers with ReLU activation, followed by 2 × 2 max-pooling. Feature channels progress from 32 to 64, 128, 256, and 512 at the bottleneck. The decoder mirrors the encoder, using 2 × 2 transposed convolutions for upsampling. Skip connections concatenate corresponding encoder features with decoder maps to recover fine spatial details. The final layer is a 1 × 1 convolution with sigmoid activation, producing a pixel-wise probability map. The total network comprises approximately 7.8M trainable parameters, deliberately lightweight to isolate the effect of input resolution from architectural complexity. The choice of a standard U-Net is intentional: more complex architectures would confound the analysis of whether performance gains arise from resolution preservation or architectural innovation. The moderate model capacity is well-matched to the 256 × 256 input patches, avoiding both underfitting and the overfitting risks associated with overparameterised models on limited medical datasets.

3.4. Semi-supervised learning framework

To address the clinical bottleneck of voxel-level annotation, we investigated the annotation efficiency of three SSL paradigms. The mean teacher (MT) (Tarvainen & Valpola, 2017) enforces MSE consistency between a student model and the exponential moving average (EMA) of its own weights on unlabelled data. The uncertainty-aware mean teacher (UA-MT) (L. Yu et al., 2019) extends MT by using Monte Carlo dropout (T = 8 stochastic passes) to estimate the teacher's epistemic uncertainty, with the student learning only from regions

where the teacher is highly confident. Cross-pseudo supervision (CPS) (X. Chen et al., 2021) trains two independently initialised networks that generate pseudo-labels to supervise each other. We additionally evaluated FixMatch (Sohn et al., 2020), which combines weak-strong augmentation consistency with hard pseudo-labelling at a confidence threshold of $\tau = 0.95$, to compare consistency-based and pseudo-labelling paradigms. For semi-supervised experiments, we fixed the validation and test sets and partitioned the 271-volume training pool into labelled and unlabelled subsets at 10%, 25%, and 50% ratios, with nested splits ensuring fair comparison.

3.5. Training details

All experiments were conducted on an NVIDIA Tesla V100 GPU (32 GB) on a university HPC cluster, using TensorFlow 2.x. Models were optimised with Adam (learning rate = 10^{-4}) for 100 epochs with batch size 32. The training set of 9,798 patch-mask pairs was shuffled at each epoch. Binary cross-entropy (BCE) served as the primary supervised loss function. We additionally evaluated a combined Dice+BCE loss to assess whether directly optimising for volumetric overlap provides benefits for the extreme class imbalance inherent to pancreas segmentation (<0.5% foreground). The nnU-Net baseline used its self-configured settings (SGD optimiser, polynomial learning rate schedule, Dice+cross-entropy loss) for 1,000 epochs in its 3d_fullres configuration. For semi-supervised experiments, the EMA decay coefficient was set to $\alpha = 0.999$; UA-MT employed dropout rate 0.1 in all convolutional blocks for Monte Carlo uncertainty estimation with $T = 8$ stochastic forward passes. The consistency weight λ was ramped sigmoidally to a maximum of 10.0 over the first 40 epochs, allowing the student to first learn basic representations from labelled data before incorporating the teacher's consistency signal. No additional data augmentation beyond random patch cropping was applied, isolating the effect of resolution preservation from augmentation strategies.

3.6. Evaluation metrics

We report the Dice similarity coefficient (DSC) for volumetric overlap, intersection over union (IoU) for training monitoring, the 95th-percentile Hausdorff distance (HD95) for boundary delineation quality, and the average symmetric surface distance (ASD). HD95 measures the 95th percentile of symmetric surface distances between predicted and ground-truth boundaries, providing a reliable estimate of worst-case boundary error while being less sensitive to outlier voxels than the standard Hausdorff distance.

4. Results

4.1. Preprocessing and resolution

Our initial experiments isolated the effect of HU windowing on segmentation accuracy. The standard abdominal window [–100, 240] achieved an average Dice of 0.765, while the optimised pancreas-specific window [–125, 275] improved the average to 0.815, an absolute gain of +5.0%. The improvement was most pronounced on the challenging Pancreas 005 case (+15.46% Dice), where the pancreas boundary is particularly ambiguous due to low contrast

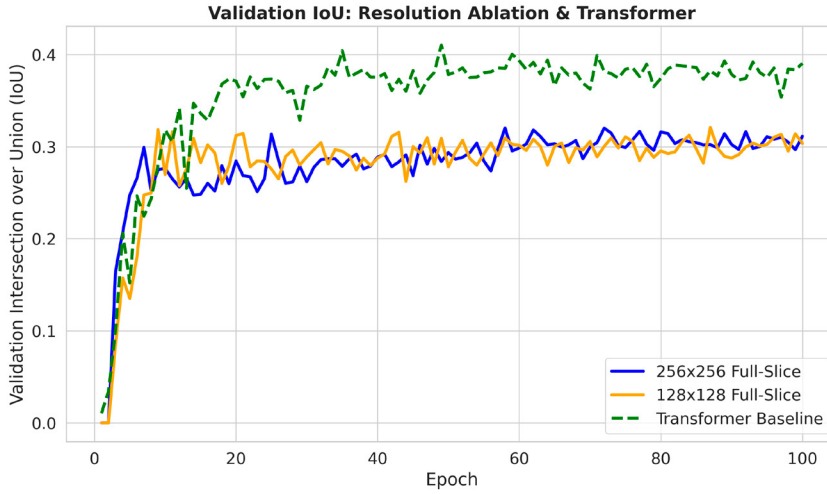


Figure 2. Validation IoU curves for the resolution ablation study

with surrounding fat and vascular structures. This confirms the sensitivity of CNNs to input intensity distributions and supports the value of domain-specific preprocessing as a first-order optimisation, before any architectural modifications.

To isolate the impact of input resolution from architectural complexity, we conducted an ablation study using the identical 7.8M U-Net architecture across three input configurations. The 128 × 128 full-slice model converged rapidly but stagnated at a validation IoU of 0.304. The 256 × 256 full-slice model showed only marginal improvement, reaching 0.311 IoU. Both substantially underperformed the patch-based model (≈0.69 IoU), a 55% relative degradation. This indicates that the pancreas is effectively "lost" to the network when native resolution is compromised. The global context provided by seeing the entire abdomen in the downsampled slices could not compensate for the loss of local boundary texture at the sub-millimetre level.

To provide theoretical grounding for this empirical result, we analysed the spatial frequency content of CT slices using 2D Fourier transform analysis. Downsampling acts as a low-pass filter that irreversibly destroys high-frequency components encoding fine structural details such as organ boundaries and tissue texture. The mean boundary edge strength (measured via Sobel analysis at the pancreas boundary) decreases from 0.436 at native resolution to 0.345 at 256 × 256 (–21%) and 0.269 at 128 × 128 (–38%). This progressive degradation of boundary contrast directly explains the IoU collapse: the pancreas boundary becomes indistinguishable from surrounding tissue at reduced resolutions. Figure 2 shows the corresponding validation IoU curves.

Taken together, the ablation and the frequency-domain analysis indicate that resolution loss, rather than limited model capacity, is the dominant cause of the performance gap. This motivates the patch-based design that we next compare against established segmentation methods.

4.2. Comparison with established methods and computational cost

Table 1 and Figure 3 summarise the comparison across all evaluated methods. The patch-based CNN achieved 0.809±0.015 Dice across four independent runs (peak 0.849), competitive with nnU-Net (0.823 Dice, 88M parameters, 1,000 epochs) while using 11× fewer parameters and 10× fewer training epochs. The vision transformer baseline (4.7M parameters) trained from scratch on identical patches achieved only 0.390 IoU; this comparison primarily demonstrates that transformers require large-scale pretraining to be effective on small medical datasets ($N < 300$), rather than an inherent architectural inferiority. SAM ViT-B achieved 0.705 Dice with ground-truth bounding-box prompts (best-case scenario), and MedSAM scored only 0.439 despite being fine-tuned on 1.5M medical image-mask pairs. Without prompts, SAM collapsed to 0.097. These results indicate that foundation models, despite their massive scale (91M parameters), do not overcome the resolution limitation for small-organ segmentation.

Table 1. Quantitative comparison on MSD Task07 Pancreas (*patches from native 512 × 512 slices; †IoU metric)

Method	Input res.	Labelled	Params (M)	Dice ↑	Dataset
nnU-Net	Auto	100%	88.0	0.823	MSD
SAM ViT-B (bbox)	Fixed	100%	91.0	0.705	MSD
MedSAM (bbox)	Fixed	100%	91.0	0.439	MSD
ViT baseline	256^{2}*	100%	4.7	0.390†	MSD
Ablation 256^{2}	256^{2}	100%	7.8	0.311†	MSD
Ablation 128^{2}	128^{2}	100%	7.8	0.304†	MSD
Ours (multi-seed)	256^{2}*	100%	7.8	0.809±0.015	MSD
Ours (best run)	256^{2}*	100%	7.8	0.849	MSD
UA-MT 50%	256^{2}*	50%	7.8	0.803	MSD

Note: SAM/MedSAM use ground-truth bounding-box prompts (best-case). Resolution-ablation rows use the same 7.8M U-Net on downsampled full slices, isolating the effect of resolution from architecture.

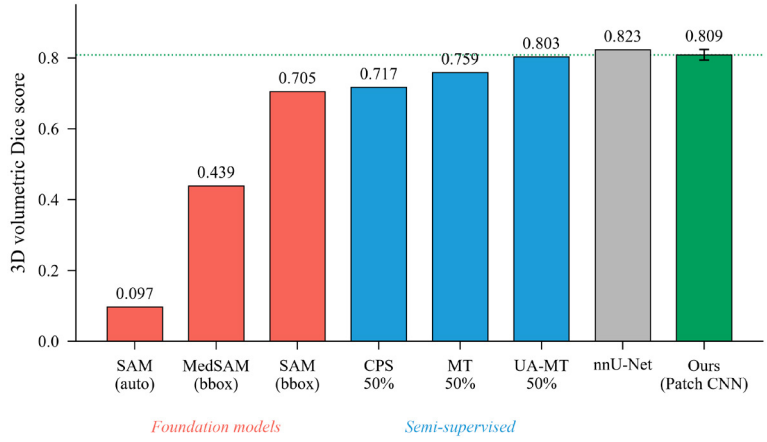


Figure 3. Model comparison on MSD Task07 Pancreas (error bars: standard deviation across four runs)

Beyond accuracy, the practical cost of each model determines its clinical applicability. Table 2 reports the parameter count, training budget, and inference time per volume.

Table 2. Computational efficiency (parameters, training epochs, inference time per volume)

Method	Params (M)	Epochs	Inf. time / vol.
nnU-Net	88.0	1,000	N/A
SAM ViT-B	91.0	Pre-trained	≈15 s
Ours (patch U-Net)	7.8	100	≈7 s

The lightweight footprint enables practical deployment: the model fits comfortably within the memory of a single commodity GPU and completes a full volume in roughly seven seconds, roughly twice as fast as the foundation-model baseline. Competitive accuracy is therefore obtained without the hardware burden typically associated with high-performing pancreas segmentation.

4.3. Annotation efficiency via semi-supervised learning

Table 3 and Figure 4 present the SSL annotation efficiency results. UA-MT consistently dominates across all data regimes. At just 10% labelled data (27 volumes), UA-MT (0.397 IoU) approaches the performance of mean teacher at 25% (0.396 IoU), effectively halving the required annotation effort. At the 25% ratio, UA-MT (0.484 IoU) outperforms mean teacher at 50% (0.457 IoU). CPS requires a larger labelled set to stabilise; at 10% and 25%, mutual supervision between two weakly-initialised networks leads to representation collapse, and only at 50% did CPS reach 0.400 IoU. For interpretability, Table 3 and Figure 4 report validation-set metrics (best IoU and the corresponding approximate Dice, respectively) used for model selection across label budgets; these differ from the held-out 3D volumetric test Dice of the champion models reported in Table 4 (for example, 0.803 for UA-MT at 50%).

To verify that UA-MT's superior performance is not an artefact of a single favourable initialisation, we repeated the UA-MT 50% experiment across three independent seeds (42, 123, 456). The resulting validation IoU values were 0.468, 0.469, and 0.447, yielding a mean of 0.462±0.012. The low variance confirms that UA-MT's dominance is reproducible and not attributable to initialisation sensitivity.

Table 4 presents the per-case 3D Dice scores for the champion models, and Figure 5 shows the corresponding qualitative segmentation overlays on the most challenging case (Pancreas 005). The supervised model achieved 0.809±0.015 across four independent runs. The UA-MT model trained on only 50% of labels reached 0.803±0.091, recovering 99.3% of

Table 3. SSL annotation efficiency: best validation IoU at each label budget

Method	10% (27 vols)	25% (67 vols)	50% (135 vols)
Mean teacher	0.194	0.396	0.457
CPS	0.246	0.333	0.400
UA-MT	0.397	0.484	0.479

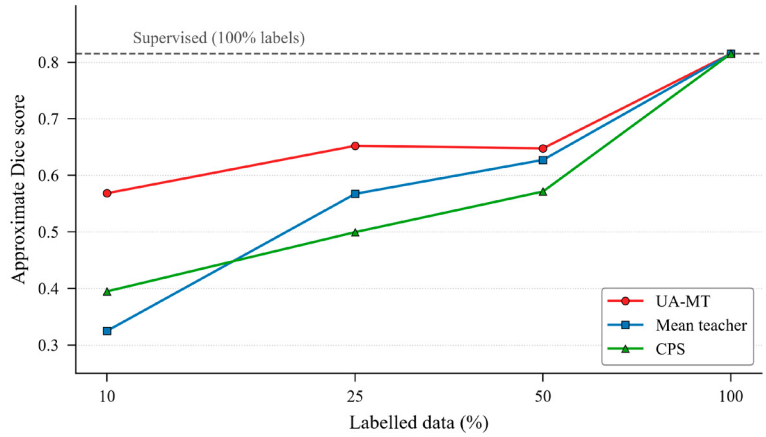


Figure 4. Annotation efficiency: approximate Dice as a function of labelled-data budget for the three SSL paradigms

the supervised mean with half the annotation effort. On the most challenging case (Pancreas 005), UA-MT (0.716) actually outperforms the fully supervised model (0.662), suggesting that uncertainty filtering provides a stronger learning signal for ambiguous medical boundaries than labelled data alone. In contrast, the pseudo-labelling baseline FixMatch reached only 0.687 Dice at the 50% budget, well below all three consistency-based methods; we examine this gap in Section 4.

To characterise boundary delineation quality beyond volumetric overlap, Table 5 reports the HD95 and ASD metrics. The supervised model with Dice+BCE loss achieved the best boundary accuracy (HD95: 7.40 mm, ASD: 2.51 mm). The UA-MT model maintained competitive boundary quality (HD95: 9.54 mm), consistent with its strong Dice performance. The mean teacher exhibited a substantially elevated HD95 of 18.31 mm, driven by a large boundary error on Pancreas 006 (HD95: 45.40 mm).

Across all label budgets, UA-MT thus preserves both volumetric overlap and boundary fidelity, whereas the alternative semi-supervised schemes trade one for the other. These results establish UA-MT as the most reliable configuration before we examine its behaviour across the full population and on external data in the next subsection.

Table 4. Volumetric 3D Dice scores per held-out test case (*mean across four runs). For the supervised configuration, Mean ± Std reports the mean and standard deviation across the four independent runs; for the semi-supervised configurations, it reports the mean and sample standard deviation across the four test cases

Configuration	001	004	005	006	Mean ± Std
Supervised*	0.907	0.829	0.662	0.836	0.809±0.015
UA-MT 50%	0.892	0.881	0.716	0.724	0.803±0.096
MT 50%	0.900	0.850	0.648	0.636	0.759±0.136
CPS 50%	0.899	0.791	0.596	0.582	0.717±0.154

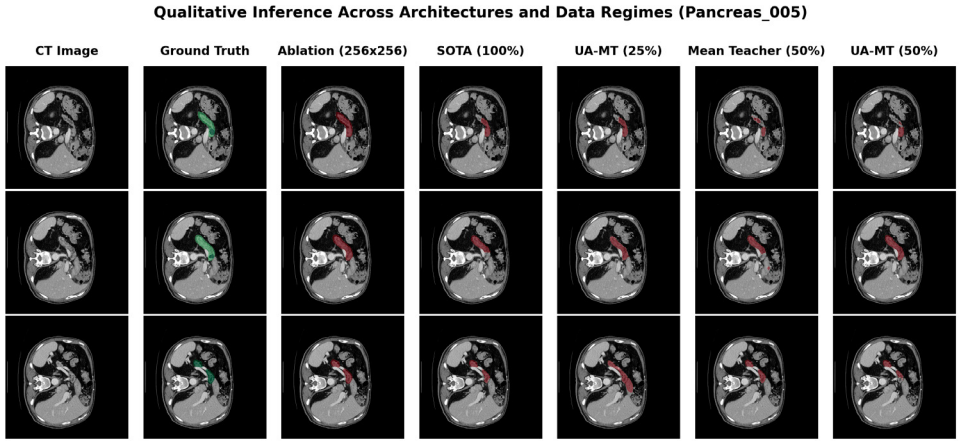


Figure 5. Qualitative comparison on the challenging Pancreas 005 case. Columns (left to right): CT image, ground truth, ablation (256×256), supervised (SOTA, 100% labels), UA-MT (25%), mean teacher (50%), UA-MT (50%)

Table 5. Boundary delineation quality (HD95 and ASD) on the four test cases. The Supervised (BCE) row is the single checkpoint carried into this boundary analysis; its four-case mean Dice (0.815) therefore differs from the four-run mean of 0.809±0.015 reported in Tables 1 and 4. All ± values are SD across the four cases

Method	Mean Dice	HD95 (mm)	ASD (mm)
Supervised (BCE)	0.815	9.35±2.21	2.97±0.73
Supervised (Dice+BCE)	0.824	7.40±3.65	2.51±1.35
UA-MT 50%	0.803	9.54±3.88	3.08±1.40
Mean teacher 50%	0.759	18.31±15.88	4.20±2.13
CPS 50%	0.717	11.29±3.78	3.39±1.21

4.4. Cross-validation, generalisation, and robustness

To provide a statistically rigorous, population-level performance estimate, we conducted patient-level 5-fold cross-validation across all 281 MSD volumes. Each fold trained on approximately 224 patients and evaluated on the remaining 56–57 using full 3D sliding-window inference. Table 6 presents the per-fold results.

The population-level mean Dice of 0.630±0.025 is lower than the 0.809±0.015 reported on the four held-out test cases. This gap is expected and warrants careful interpretation: the 5-fold evaluation encompasses the full spectrum of anatomical difficulty present in the MSD dataset, including extreme cases with atypical pancreatic morphology, very small organ volumes, or unusual positioning that yield per-volume Dice as low as 0.06. In contrast, the four held-out test cases represent moderate-to-easy anatomy. The low cross-fold standard deviation (0.027) confirms that the training procedure is stable and that performance variation is driven by patient difficulty rather than training stochasticity. This population-level estimate is more representative of expected real-world deployment performance and aligns with the

Table 6. Five-fold cross-validation results (3D volumetric Dice across all 281 MSD volumes). Overall is the population mean over all 281 out-of-fold predictions, with the sample SD across the five fold means; per-fold Std is the within-fold SD of per-volume Dice (population)

Fold	N (validation)	Mean Dice	Std
Fold 0	57	0.633	0.150
Fold 1	56	0.590	0.175
Fold 2	56	0.660	0.150
Fold 3	56	0.637	0.163
Fold 4	56	0.628	0.173
Overall	281	0.630±0.025	–

well-documented observation that pancreas segmentation remains inherently challenging across the full patient spectrum.

We additionally evaluated a combined Dice+BCE loss against the standard BCE-only formulation. Contrary to expectations, the Dice+BCE model achieved an average 3D Dice of 0.824, comparable to the BCE-only multi-seed mean (0.809±0.015) but lower than the best BCE-only run (0.849). The per-case analysis reveals that Dice+BCE degrades performance on the hardest case, Pancreas 005 (0.632 vs. 0.662 for BCE-only), while showing marginal improvements on easy cases. We attribute this to the interaction between the Dice loss gradient and the extreme class imbalance: on slices where the pancreas is very small, the Dice loss exhibits high gradient variance, introducing instability. This result supports the broader thesis that for small-organ segmentation, standard losses with appropriate resolution and preprocessing are more effective than sophisticated loss engineering.

To evaluate domain robustness, we performed zero-shot inference on 80 external TCIA Pancreas-CT volumes (Clark et al., 2013) without any retraining or domain adaptation. The fully supervised model achieved 0.544±0.165 Dice, while UA-MT (50% labels) achieved 0.603±0.167 Dice, outperforming the supervised model by +5.9% absolute. This counter-intuitive result suggests that consistency regularisation on unlabelled data acts as an implicit domain regulariser, preventing overfitting to the source domain and learning more transferable anatomical representations. The single failure case (Dice = 0.000) was traced to an extreme anatomical variant with atypical pancreatic positioning.

Finally, a common failure mode for 2D medical imaging models is slice-wise instability, where the network produces inconsistent predictions on adjacent axial slices, leading to jagged 3D reconstructions. We quantified spatial stability by computing the 2D Dice score for every axial slice within representative test volumes. For the in-domain MSD Case 001, the model achieved a mean slice-wise Dice of 0.913 with a standard deviation of only 0.076, confirming highly stable predictions across the axial span. For the external TCIA Case 052, we observed a mean slice-wise Dice of 0.722 with a standard deviation of 0.246, reflecting the increased difficulty of cross-domain generalisation. The absence of complete failures (Dice ≈ 0) in the middle of either organ volume indicates that the high-resolution patch approach effectively captures the local continuity of the pancreas across the Z-axis, despite processing each slice independently.

5. Discussion

The results of this study provide consistent evidence that spatial resolution preservation is the single most impactful factor for pancreas segmentation from CT images. The IoU collapse from 0.69 to approximately 0.31 when downsampling, despite the use of an identical network architecture, shows clearly that the information destroyed by resolution reduction is not recovered by architectural sophistication alone. The Fourier analysis provides a principled explanation: downsampling acts as a low-pass filter that disproportionately attenuates the high-frequency boundary signals (38% edge strength loss at 128 × 128) that are essential for delineating the pancreas from adjacent soft tissues of similar radiodensity.

The 11.6% Dice gap between UA-MT (0.803) and FixMatch (0.687) illuminates a fundamental mismatch between hard pseudo-labelling and the nature of pancreatic boundaries. FixMatch's confidence threshold ($\tau = 0.95$) is designed for tasks with clear class boundaries, but pancreas-duodenum interfaces exhibit gradual intensity transitions where model confidence rarely exceeds this threshold. Consequently, FixMatch discards most of the learning signal from unlabelled boundary regions. UA-MT addresses this through two complementary mechanisms: the EMA teacher provides temporally-smoothed predictions that are inherently more stable than single-pass outputs, and the uncertainty weighting (via $T = 8$ MC dropout passes) down-weights genuinely noisy voxels while still propagating gradients from uncertain but informative boundary regions.

The finding that UA-MT generalises better than full supervision on the external TCIA dataset (0.603 vs. 0.544 Dice) has important clinical implications. It suggests that hospitals deploying semi-supervised models trained with partial annotations can expect more reliable performance across different scanner protocols and patient populations than with fully supervised alternatives. The consistency regularisation effectively acts as a domain generaliser: by enforcing prediction stability under perturbations on unlabelled data from the source domain, the model learns to extract features that are invariant to intensity shifts, noise characteristics, and resolution differences that distinguish one scanner from another.

The 5-fold cross-validation results (0.630±0.025 population-level Dice) provide a more conservative assessment of expected real-world performance than the 0.809 reported on four held-out test cases. The population-level estimate includes the full difficulty spectrum of the MSD dataset, with per-volume Dice ranging from 0.06 to 0.89. The low cross-fold standard deviation confirms that the training procedure is stable, and the performance variance is inherent to the anatomical difficulty of individual cases rather than to training instability.

The HD95 analysis reveals complementary information to the Dice metric. While the Dice+BCE loss achieved the best boundary accuracy (HD95: 7.40 mm), the mean teacher exhibited a pronounced boundary error on Pancreas 006 (HD95: 45.40 mm), producing a large false-positive island distant from the true organ. Multi-metric evaluation is therefore necessary in medical image segmentation: a model can achieve acceptable volumetric overlap while producing clinically unacceptable boundary errors. For surgical planning applications, where precise boundary delineation determines resection margins, HD95 may be more clinically relevant than Dice.

From a practical standpoint, the framework offers several advantages for clinical deployment. The lightweight architecture (7.8M parameters) enables inference on modest GPU

hardware, the patch-based strategy naturally handles variable input sizes without resizing, and the semi-supervised capability addresses the fundamental scalability bottleneck of expert annotation. The finding that UA-MT generalises better than full supervision is particularly relevant for federated learning scenarios, where models must perform consistently across heterogeneous scanner configurations and patient populations.

6. Conclusions

This study examined the common practice of employing increasingly complex, memory-intensive architectures for medical image segmentation at the expense of native spatial resolution. Through controlled ablation studies on pancreas segmentation from CT images, we demonstrated three key findings. First, downsampling CT volumes from native 512 × 512 resolution to 256 × 256 or 128 × 128 causes a marked collapse in segmentation fidelity (IoU dropping from 0.69 to approximately 0.31), with Fourier analysis confirming 38% boundary edge strength loss. Second, a simple patch-based 2D CNN (0.809±0.015 Dice, 7.8M parameters) achieves competitive performance with nnU-Net (0.823 Dice, 88M parameters) while substantially outperforming vision transformers and foundation models, and 5-fold cross-validation across all 281 volumes confirmed stable training with a population-level Dice of 0.630±0.025. Third, combining resolution preservation with UA-MT semi-supervised learning recovers 99.3% of supervised performance using only 50% of labelled data (HD95: 9.54 vs. 9.35 mm), with reproducible results across three seeds (σ = 0.012), while providing superior domain generalisation on 80 external TCIA volumes (0.603 vs. 0.544 Dice).

Several limitations should be acknowledged. The framework operates on 2D axial slices and ignores inter-slice continuity. Preliminary extension to multi-class segmentation (background, pancreas parenchyma, and tumour) revealed that while pancreas Dice remained comparable (0.673), the tumour class achieved a Dice of only 0.035, indicating that dedicated architectures with class-aware loss functions and volumetric context are needed for tumour detection. Future work should explore 2.5D approaches incorporating multi-slice context while maintaining native in-plane resolution.

These findings collectively advocate for a data-centric paradigm in small-organ segmentation, where spatial fidelity and domain-specific preprocessing take precedence over architectural complexity. While the current study focuses on the pancreas, the resolution-preservation hypothesis is likely generalisable to other small, low-contrast anatomical structures such as the adrenal glands, gallbladder, and pancreatic duct. Combining the resolution-preserving framework with active learning, so that annotators are guided to the volumes most informative for the model under a fixed budget, is a natural next step for clinical deployment.

Acknowledgements

The computational experiments were performed on the High-Performance Computing (HPC) cluster of the Faculty of Mathematics and Informatics, Vilnius University. All persons named in this section have agreed to be named.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

AI acknowledgement

The authors acknowledge the use of a generative artificial intelligence assistant (Claude Sonnet 5, Anthropic) for language editing and for drafting parts of the manuscript text. The tool was not used to generate, select, or interpret any data, result, analysis, figure, or reference. All AI-assisted text was reviewed, verified, and revised by the authors, who take full responsibility for the content, accuracy, and integrity of the work. Artificial intelligence is not listed as an author.

Data availability statement

The MSD Task07 Pancreas dataset is publicly available at http://medicaldecathlon.com/. The TCIA Pancreas-CT dataset is available at https://wiki.cancerimagingarchive.net/. Code and trained models are available at https://github.com/Utshox/pancreas-segmentation-resolution-study.

Note on prior presentation

A preliminary version of this work will be presented as a poster at the eStream 2026 conference. The present manuscript is a substantially revised and expanded version, adding five-fold cross-validation, HD95 boundary metrics, multi-seed reproducibility analysis, external validation on TCIA, and an extended discussion.

Author contributions

IA conceived the study and was responsible for the design and development of the data analysis, the software implementation, the experiments, and the visualisations, and wrote the first draft of the article. SNP contributed to data curation, validation of the experimental results, and revision of the article. KK supervised the work, provided the computational resources, and was responsible for reviewing and editing the article. All authors read and approved the final version.

Disclosure statement

The authors declare no competing financial, professional, or personal interests from other parties.

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., van Ginneken, B., Bilello, M., Bilic, P., Christ, P. F., Do, R. K. G., Gollub, M. J., Heckers, S. H., Huisman, H., Jarnagin, W. R., … Cardoso, M. J. (2022). The medical segmentation decathlon. *Nature Communications*, *13*, Article 4128. https://doi.org/10.1038/s41467-022-30695-9

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2023). Swin-UNet: UNet-like pure transformer for medical image segmentation. In *Computer Vision – ECCV 2022 Workshops* (pp. 205–218). Springer. https://doi.org/10.1007/978-3-031-25066-8_9

Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M. P., Zhang, S., Xing, L., Lu, L., Yuille, A., & Zhou, Y. (2024). TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, *97*, Article 103280. https://doi.org/10.1016/j.media.2024.103280

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision* (pp. 833–851). https://doi.org/10.1007/978-3-030-01234-2_49

Chen, X., Yuan, Y., Zeng, G., & Wang, J. (2021). Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2613–2622). IEEE. https://doi.org/10.1109/cvpr46437.2021.00264

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *MICCAI 2016* (pp. 424–432). Springer. https://doi.org/10.1007/978-3-319-46723-8_49

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., & Prior, F. (2013). The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, *26*(6), 1045–1057. https://doi.org/10.1007/s10278-013-9622-7

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An image is worth 16x16 words*: *Transformers for image recognition at scale*. ArXiv. https://doi.org/10.48550/arXiv.2010.11929

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., & Xu, D. (2022). UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1748–1758). IEEE. https://doi.org/10.1109/WACV51458.2022.00181

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, *18*(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z

Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., & Jin, C. (2024). Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, *169*, Article 107840. https://doi.org/10.1016/j.compbiomed.2023.107840

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3992–4003). IEEE. https://doi.org/10.1109/ICCV51070.2023.00371

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3431–3440). IEEE. https://doi.org/10.1109/cvpr.2015.7298965

Luo, X., Hu, M., Song, T., Wang, G., & Zhang, S. (2022). *Semi-supervised medical image segmentation via cross-teaching between CNN and Transformer*. arXiv. https://doi.org/10.48550/arXiv.2112.04894

Luo, X., Wang, G., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Metaxas, D. N., & Zhang, S. (2022). Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, *80*, Article 102517. https://doi.org/10.1016/j.media.2022.102517

Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). MedSAM: Segment anything in medical images. *Nature Communications*, *15*(1), Article 654. https://doi.org/10.1038/s41467-024-44824-z

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision* (pp. 565–571). IEEE. https://doi.org/10.1109/3dv.2016.79

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). *Attention U-Net*: *Learning where to look for the pancreas*. arXiv. https://doi.org/10.48550/arXiv.1804.03999

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI 2015* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Roth, H. R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E. B., & Summers, R. M. (2015). DeepOrgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *MICCAI 2015* (pp. 556–564). Springer. https://doi.org/10.1007/978-3-319-24553-9_68

Roth, H. R., Lu, L., Lay, N., Harrison, A. P., Farag, A., Sohn, A., & Summers, R. M. (2018). Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. *Medical Image Analysis*, *45*, 94–107. https://doi.org/10.1016/j.media.2018.01.006

Siegel, R. L., Kratzer, T. B., Giaquinto, A. N., Sung, H., & Jemal, A. (2025). Cancer statistics, 2025. *CA*: *A Cancer Journal for Clinicians*, *75*(1), 10–45. https://doi.org/10.3322/caac.21871

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., & Raffel, C. (2020). *FixMatch*: *Simplifying semi-supervised learning with consistency and confidence*. ArXiv. https://doi.org/10.48550/arXiv.2001.07685

Tarvainen, A., & Valpola, H. (2017). *Mean teachers are better role models*: *Weight-averaged consistency targets improve semi-supervised deep learning results*. arXiv. https://doi.org/10.48550/arXiv.1703.01780

Wasserthal, J., Breit, H.-C., Meyer, M. T., Pradella, M., Hinck, D., Sauter, A. W., Heye, T., Boll, D. T., Cyriac, J., Yang, S., Bach, M., & Segeroth, M. (2023). TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiology*: *Artificial Intelligence*, *5*(5), Article e230024. https://doi.org/10.1148/ryai.230024

Yang, X., Song, Z., King, I., & Xu, Z. (2023). A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, *35*(9), 8934–8954. https://doi.org/10.1109/TKDE.2022.3220219

Yu, Q., Xie, L., Wang, Y., Zhou, Y., Fishman, E. K., & Yuille, A. L. (2018). Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8280–8289). IEEE. https://doi.org/10.1109/cvpr.2018.00864

Yu, L., Wang, S., Li, X., Fu, C.-W., & Heng, P.-A. (2019). Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *MICCAI 2019* (pp. 605–613). Springer. https://doi.org/10.1007/978-3-030-32245-8_67

Zheng, Y., & Luo, J. (2023). Extension–contraction transformation network for pancreas segmentation in abdominal CT scans. *Computers in Biology and Medicine*, *152*, Article 106410. https://doi.org/10.1016/j.compbiomed.2022.106410