QGA-TTT: QUERY-GUIDED AGGREGATION VIA TEST-TIME TRAINING FOR MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical image segmentation asks for a label at every position of a scene that holds many structures at once, and which class a region belongs to is decided in large part by what else the image contains. Every position of one image therefore has to be labelled against one and the same account of that image. Convolution reaches only as far as its receptive field, and pairwise attention lets positions compare with one another without producing a single shared description, so neither forms such an account. Test-time training layers fit a small set of fast weights inside the forward pass, so the map they apply is recomputed for every input. We use this property to build one shared account per image. At each of the deepest levels of a segmentation network, the level's positions together with a small set of learnable queries shared across every image are fitted, in a single forward pass, into one affine map, and every position reads the global information back through that same map. The shared queries enter every fitting, so decoder weights shared across images can read what the map returns. The queries are kept across two passes so that the second reads an account already placed in the coordinates of the current image, and the map is estimated twice by fittings that share no parameters. Across nine public datasets spanning endoscopy, ultrasound, optical coherence tomography and magnetic resonance imaging, the method reaches the best Dice on every one, by up to 2.81 points over the strongest comparison method. Ablations show each of the three constructions to be necessary, and both the number of fittings and the capacity of the account are interior optima.

1 INTRODUCTION

Segmentation of medical images supports measurement, planning and follow-up across essentially every imaging modality, and encoder decoder networks built from residual convolutions remain the arrangement that clinical pipelines are organised around (Ronneberger et al., 2015; Çiçek et al., 2016; Isensee et al., 2021). Much of the recent progress has come from changing how positions of a feature map exchange information, first through self-attention (Vaswani et al., 2017; Dosovitskiy et al., 2021; Liu et al., 2021; Chen et al., 2021; Hatamizadeh et al., 2022a; Zhou et al., 2023) and then through state space layers of linear cost (Gu & Dao, 2024; Zhu et al., 2024; Ma et al., 2024; Liu et al., 2024a).

Scenes that hold many structures at once put a demand on this exchange that none of these primitives meets. A region of tissue that looks like one structure in isolation is identified by the configuration of the structures around it, and the same appearance supports different labels in different images. All positions of one image consequently have to be labelled against one and the same account of what that image contains. A convolutional decoder cannot form such an account, because each position sees its receptive field together with weights that are shared across the whole dataset. Attention lets positions compare with one another, and each position forms its own comparison, so no single reading of the scene is shared across the level. Its cost also grows with the square of the number of positions, which confines global mixing to the deepest levels, where spatial detail has already been removed (Wang et al., 2018; Katharopoulos et al., 2020). Recurrent and state space layers do compress a level into a state, but that state is accumulated along an ordering, so positions early and late in it read different things.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

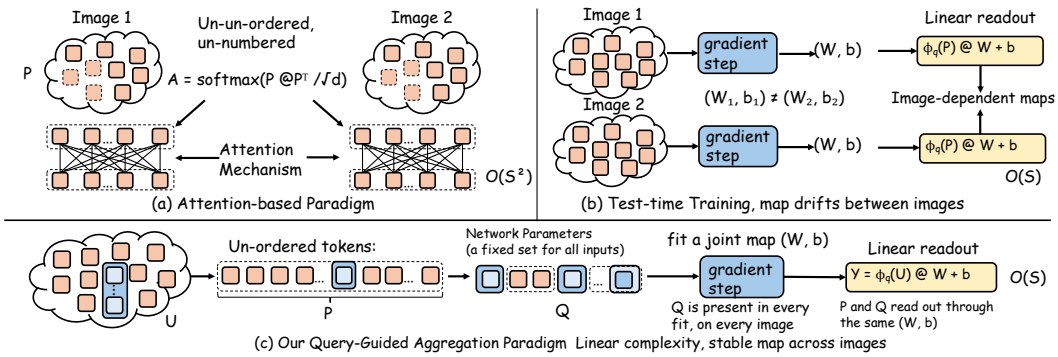


Figure 1: Three ways of mixing tokens. (a) Attention, standing also for state space layers, keeps its weights fixed once training ends and compares positions pairwise. (b) A test-time training layer fits a small map inside the forward pass, so two images leave it with different fitted weights. (c) QGA-TTT adds queries shared across every image to the working set the map is fitted from.

The obstacle is that forming an account of the current image calls for a computation that depends on that image, while a trained network applies the same weights to every input. A recent family of sequence layers removes this obstacle. A test-time training layer fits a small set of fast weights inside its own forward pass, so the map it applies is recomputed for each input (Sun et al., 2025; Behrouz et al., 2025), in the tradition of fast weight programmers (Schmidhuber, 1992; Schlag et al., 2021). The fitting runs identically during training and at inference and uses no objective beyond the one already present in the forward computation.

Our insight is that a single such fit, applied to a whole level at once, provides the shared account the task asks for. We fit one affine map from the level's own positions together with a small set of learnable queries that are parameters of the network and therefore identical for every image, and we have every position read the global information back through that same fitted map. The queries matter because a per-image fit alone would express each image in coordinates that image chooses for itself, which decoder weights shared across images cannot read. Two further constructions follow from the account being shared. Keeping the queries across two passes lets the second pass read an account already placed in the coordinates of the current image. Estimating the map twice with fittings that share no parameters keeps any single estimate from deciding the labelling of an entire level. We call the result Query-Guided Aggregation with test-time training, or QGA-TTT.

We identify the absence of a shared per-image account as the obstacle medical image segmentation runs into, and show that a layer which fits weights inside its forward pass supplies one. We build QGA-TTT from three constructions that each follow from that requirement. On nine public datasets across four modalities QGA-TTT improves on the strongest comparison method of every dataset, and ablations establish all three constructions as necessary and show the number of fittings and the capacity of the account to be interior optima.

2 Related Work

Backbones for medical segmentation. Encoder decoder networks with skip connections have organised medical segmentation since the U-Net (Ronneberger et al., 2015; Long et al., 2015), and the arrangement carried over to volumetric data (Çiçek et al., 2016; Milletari et al., 2016). Later work added nested skip pathways (Zhou et al., 2018; 2020) and attention on them (Oktay et al., 2018), while self-configuring pipelines showed a carefully tuned convolutional network to remain a strong competitor (Isensee et al., 2021; 2024). We leave this backbone in place and change what the deepest skip connections carry.

Global context in vision backbones. Self-attention made pairwise exchange between all positions available to vision models (Vaswani et al., 2017; Dosovitskiy et al., 2021), and windowed and hierarchical variants brought the cost down enough for dense prediction (Liu et al., 2021; Xie et al., 2021a; Wang et al., 2018). Medical segmentation adopted this through hybrids that insert attention

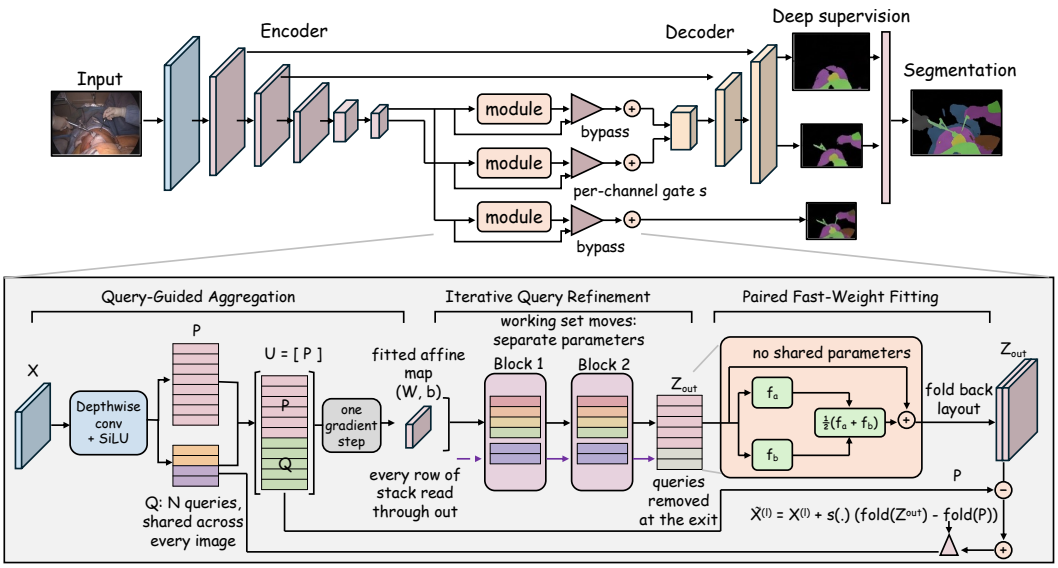


Figure 2: QGA-TTT in a segmentation network. The encoder runs in full and the three deepest skip connections each carry a correction that is formed from that level alone, scaled by a per-channel gate and added back before the decoder reads it. The lower panel opens one level. Local features and the shared queries form a working set, one affine map is fitted from it, every row reads out through that map, the queries are kept across both passes, and each pass averages two fittings that share no parameters.

into an encoder decoder skeleton (Chen et al., 2021; 2024b; Hatamizadeh et al., 2022b;a; Zhou et al., 2023; Xie et al., 2021b; Gao et al., 2021). Linear attention and related kernel approximations reduce the quadratic cost while keeping the pairwise form (Katharopoulos et al., 2020; Choromanski et al., 2021; Wang et al., 2020; Shen et al., 2021). State space layers replace pairwise comparison with a recurrence of linear cost (Gu et al., 2022; Gu & Dao, 2024; Dao & Gu, 2024) and have been carried into segmentation backbones (Zhu et al., 2024; Liu et al., 2024b; Ma et al., 2024; Liu et al., 2024a; Xing et al., 2024). All of these compute how one position relates to others, and none produces a single description of the level that every position then reads.

Layers that fit weights inside the forward pass. A recent class of sequence layer makes the hidden state itself a small model, fitted by gradient steps taken during the forward computation (Sun et al., 2025; Behrouz et al., 2025; Dalal et al., 2025). The idea descends from fast weight programmers (Schmidhuber, 1992; Ba et al., 2016), which were later connected to linear attention (Schlag et al., 2021; Von Oswald et al., 2023) and given parallel formulations that make them practical at scale (Yang et al., 2024; 2025). These layers sit alongside attention and state space layers as a third way of mixing tokens. The fitting inside them runs identically during training and at inference, carries no objective of its own and leaves every network parameter untouched. We build on a layer of this class and use the per-input map it produces as the account a level needs.

Fixed sets of non-spatial tokens. Carrying a small set of learnable rows alongside spatial features is an established device. Detection transformers decode objects from a fixed set of learnable queries (Carion et al., 2020; Zhu et al., 2021), and the pattern was extended to segmentation as mask queries (Cheng et al., 2021; 2022). Latent arrays compress large inputs into a fixed size representation (Jaegle et al., 2021; 2022; Lee et al., 2019), and registers added to vision transformers absorb artefacts that otherwise appear in the feature map (Darcet et al., 2024). Slot representations bind image content into a small number of learned slots (Locatello et al., 2020). In all of these the rows are read out as the model's output or as an intermediate representation. Ours are never read out. They enter a fit whose product is a map that the spatial positions use, and they hold that map in a form that is stable across images.

3 METHOD

3.1 OVERVIEW

We work with an encoder and decoder segmentation network built from residual convolutional blocks, the family that current medical segmentation pipelines are organised around (Ronneberger et al., 2015; Isensee et al., 2021). The encoder produces one feature map per resolution level and passes each of them to the decoder through a skip connection. Writing ℓ for a level, its feature map is $\mathbf{X}^{(\ell)} \in \mathbb{R}^{C_\ell \times S_\ell}$, where C_ℓ is the channel width and S_ℓ is the number of spatial positions, equal to the product of the spatial extents at that level. Flattening the spatial axes gives a set of S_ℓ tokens $\mathbf{Z}^{(\ell)} \in \mathbb{R}^{S_\ell \times C_\ell}$, with $\mathbf{z}_i$ the token at position i.

Our contribution acts on the three deepest levels, after the encoder has run in full and before the decoder reads them. At each of these levels we compute a correction of the same shape as $\mathbf{X}^{(\ell)}$ and add it to the skip connection. The encoder path itself is left untouched, so the features that flow from one encoder level to the next are the ones the encoder produced. Figure 2 places QGA-TTT in the network and opens up one level of it.

3.2 QUERY-GUIDED AGGREGATION

The account we want is a property of the set of positions, independent of their order, so it should be produced by an operation over that set. It also has to be readable by decoder weights that are shared across images, which rules out a description expressed in coordinates that each image chooses for itself. We obtain both by fitting, in a single forward pass, one affine map from the level's own tokens together with a small set of learnable rows that are shared across every image, and by having each position read the global information back through that same fitted map. Because the shared rows are identical for every input, part of what every fitting has to account for is the same whichever image is present. We call them queries, and we call the construction Query-Guided Aggregation. With the queries, the maps fitted for two different images end up closer to each other than without them (Section 5).

Let $\mathbf{Q} \in \mathbb{R}^{N \times C_\ell}$ hold the N queries, which are parameters of the network and therefore identical for every input. A depthwise convolution followed by a SiLU activation turns $\mathbf{X}^{(\ell)}$ into local features $\mathbf{P} \in \mathbb{R}^{S_\ell \times C_\ell}$, which supply the material the fit reads. Concatenating gives the working set

$$\mathbf{U} = [\,\mathbf{P}\,;\,\mathbf{Q}\,] \in \mathbb{R}^{(S_\ell + N) \times C_\ell}. \tag{1}$$

The test-time training layer forms three linear views of this set, written $\phi_q(\mathbf{U})$, $\phi_k(\mathbf{U})$ and $\phi_v(\mathbf{U})$, and splits the channels into H heads of width $d = C_\ell / H$. Starting from a learned initialisation $(\mathbf{W}_0, \mathbf{b}_0)$ it takes one gradient step on the objective of predicting the value view from the key view,

$$\mathcal{L}(\mathbf{W}, \mathbf{b}) = \big\| \,\mathrm{LN}\big(\phi_k(\mathbf{U})\mathbf{W} + \mathbf{b}\big) - \big(\phi_v(\mathbf{U}) - \phi_k(\mathbf{U})\big) \,\big\|^2, \tag{2}$$

which yields fitted weights $(\mathbf{W}, \mathbf{b})$ with $\mathbf{W} \in \mathbb{R}^{H \times d \times d}$. Every row of the working set then reads out through those same fitted weights,

$$\mathbf{y}_i = \phi_q(\mathbf{u}_i)\,\mathbf{W} + \mathbf{b}, \tag{3}$$

so the only channel through which one position influences another is the fitted map. Nothing in this computation ties a row's contribution to where that row sits. The layer carries no positional encoding, and it weights all rows of the working set equally, so a query acts only through the content it holds.

Two properties follow. The fitted map has size $H(d^2+d)$, which is set by the channel width alone, so the amount of information that crosses between positions at a level does not grow as that level carries more positions. And the number of queries follows the label space, independent of the spatial grid, so the same N serves a level holding a few dozen positions and a level holding several thousand.

3.3 ITERATIVE QUERY REFINEMENT

A single fit leaves the queries in the state they were trained to start from. They enter the working set carrying what is common across the training distribution, and the fit reads them in that state, so the account they help form has not yet been placed in the coordinates of the image at hand.

Running the aggregation a second time repairs this, provided the queries survive between the two passes. The first pass writes into them what this image contains, and the second pass reads the image again with queries that have already been placed. The tokens carry the level's content. Their values are written back to spatial positions and are held there by the segmentation objective, which leaves them unable to hold a summary of the level. The queries occupy no spatial position and are discarded at the exit, so they are free to carry state, and they form the only route by which the account built in the first pass reaches the second.

Concretely the queries are inserted once, the working set passes through two blocks with separate parameters, and the query rows are removed only at the exit,

$$\mathbf{U}^{(1)} = \mathrm{Block}_1(\mathbf{U}), \qquad \mathbf{U}^{(2)} = \mathrm{Block}_2(\mathbf{U}^{(1)}), \qquad \mathbf{Z}^{\text{out}} = \mathbf{U}^{(2)}_{1:S_\ell}. \tag{4}$$

The fast weights of each block start from that block's own learned initialisation and are discarded when the block returns, which is what makes the query rows the carrier. Keeping the queries across the two passes costs no parameters and no additional fitting.

3.4 Paired Fast-Weight Fitting

Because every position reads through one and the same fitted map, an error in that map is not confined to a region. It shifts the decision at every position of the level at once, which makes the reliability of a single fit worth spending capacity on.

We therefore estimate the map twice within a block, with two fittings that share no parameters, and average their read-outs. The two fittings see the same working set and differ only in the parameters that form their views and their initialisations, so averaging them reduces the influence of any one estimate without changing what is being estimated. Writing f_a and f_b for the two fittings and LN for layer normalisation, a block computes

$$\mathrm{Block}(\mathbf{U}) = \mathbf{U} + \tfrac{1}{2}\Big(f_a\big(\mathrm{LN}(\mathbf{U})\big) + f_b\big(\mathrm{LN}(\mathbf{U})\big)\Big). \tag{5}$$

Table 4 reports segmentation quality with one fitting and with three.

3.5 Gated injection

The three constructions above produce $\mathbf{Z}^{\text{out}}$ from $\mathbf{P}$. Folding both back to the spatial layout and writing $\mathbf{s} \in \mathbb{R}^{C_\ell}$ for a per-channel gate, the level's skip connection becomes

$$\widetilde{\mathbf{X}}^{(\ell)} = \mathbf{X}^{(\ell)} + \mathbf{s} \odot \big(\mathrm{fold}(\mathbf{Z}^{\text{out}}) - \mathrm{fold}(\mathbf{P})\big). \tag{6}$$

The subtraction in Equation 6 leaves exactly the updates the two blocks accumulated. The depthwise convolution supplies the material the fit reads and never enters the output as a feature of its own, so what reaches the decoder is the original skip connection plus a purely global correction.

The gate is initialised near zero, so the network starts at the identity map and training sets how much correction each channel receives. The correction reaches every position of a level at once, and an untrained global path would therefore perturb the whole level during the first epochs. Starting at the identity map removes that effect and lets the correction grow as the gate is trained. Because $\mathbf{s}$ is a parameter, the amount of correction the trained network uses can be read from it directly.

The account describes which structures the image contains, which is information about identity, so the correction is added on the side of the network that assigns labels. Parameters and compute are reported in Section 4.3.

4 Experiments

4.1 Setup

We evaluate on nine public medical segmentation datasets covering endoscopy, ultrasound, optical coherence tomography and magnetic resonance imaging. Six are two dimensional, OCT5k (Arikan et al., 2025), m2caiSeg (Maqbool et al., 2020), Endoscapes2023 (Murali et al., 2023; Mascagni

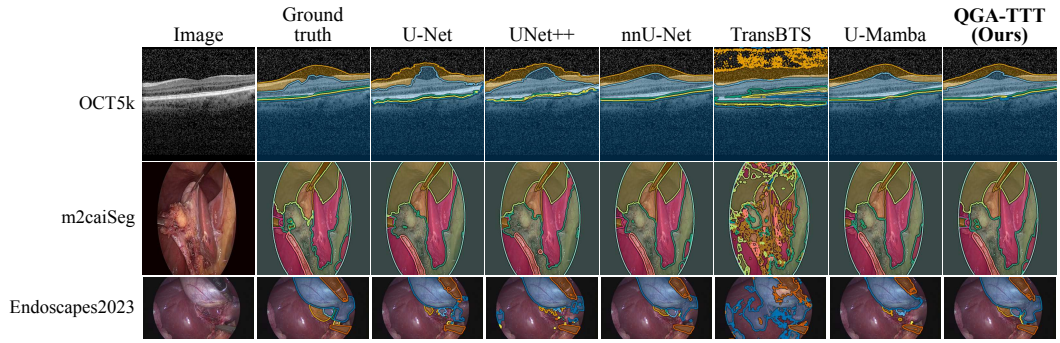


Figure 3: Segmentation on the three datasets whose label spaces carry several classes, one case per dataset. Each row shows the case whose Dice under QGA-TTT is closest to the median of that dataset, and the five comparison methods shown run on all nine datasets. Appendix D shows all nine datasets with every comparison method.

et al., 2025), Kvasir-SEG (Jha et al., 2020), PW-BALFC (Shi et al., 2025) and the pubic symphysis and fetal head set PSFHS (Chen et al., 2024a), and three are volumetric, UT-EndoMRI (Liang et al., 2025b), GCM2025 (Liang et al., 2025a) and CirrMRI600+ (Jha et al., 2025). Their label spaces range from one foreground class to eighteen.

Comparison methods are a self-configuring convolutional pipeline (Isensee et al., 2021), twenty segmentation architectures spanning convolutional, attention based and state space designs, and three networks already built on test-time training layers, two of them for medical segmentation (Zhou et al., 2026; Xu, 2024) and one a general vision backbone that reports semantic segmentation among its downstream tasks (Han et al., 2026) (Ronneberger et al., 2015; Zhou et al., 2018; Chen et al., 2021; Hatamizadeh et al., 2022b;a; Zhou et al., 2023; Xie et al., 2021b; Roy et al., 2023; Huang et al., 2023; Ma et al., 2024; Liu et al., 2024a; Ruan et al., 2025; Wu et al., 2025; Wang et al., 2024; Zhang et al., 2021; Wang et al., 2021; Li et al., 2025; Gao et al., 2021; Lee et al., 2023). Each is trained under the same budget on every dataset it supports, and we report the strongest of them per dataset by Dice. Architectures with a two dimensional implementation only are excluded from the volumetric comparison, so no method is credited with a score obtained under a configuration the others did not use.

All models train for 100 epochs of 250 iterations on fold 0, with deep supervision (Lee et al., 2015) and AdamW (Loshchilov & Hutter, 2019). Our model uses a fixed learning rate of 3×10^{-4}, eight queries and head width sixteen. We report Dice (Dice, 1945) averaged over foreground classes and the 95th percentile Hausdorff distance (HD95) (Huttenlocher et al., 1993), following standard practice (Maier-Hein et al., 2024; Antonelli et al., 2022). On Endoscapes2023 every configuration, ours and all comparison methods, uses bfloat16 mixed precision with float32 sliding window accumulation, because the large frames of that dataset overflow the range of float16 activations.

4.2 Main results

Table 1 reports every method on every dataset it runs on. QGA-TTT reaches the best Dice on all nine datasets, with the largest gains on the datasets whose label spaces carry several classes. The last family collects networks that already build on test-time training layers, which makes it the closest point of comparison. Placing the fit on a whole level and anchoring it with queries shared across images improves on all three wherever they run.

Appendix A reports HD95 for the same methods and datasets.

Figure 3 shows the recurring difference, the assignment of whole regions to the right class.

Table 1: Dice on nine medical segmentation datasets, averaged over foreground classes. Best per dataset in bold, second best underlined. Within each family methods are ordered as those running on both kinds of data, then those whose public implementation is two dimensional only, then those that are volumetric only. A dash marks a dataset a method does not run on, because its public implementation covers only the other kind of data. All methods share the training budget, split and evaluation code.

	two dimensional						volumetric		
Method	OCT5k	m2caiSeg	Endoscapes	Kvasir	PW-BALFC	PSFHS	UT-Endo	GCM2025	CirrMRI
Convolutional									
U-Net	51.98	42.94	49.20	84.50	71.99	96.47	73.25	36.31	89.72
UNet++	50.53	45.86	44.67	84.39	73.07	96.40	76.03	38.59	89.66
nnU-Net	60.34	50.66	53.54	85.32	<u>74.17</u>	97.26	77.26	44.30	94.23
U-KAN	57.53	46.07	48.41	83.68	69.24	97.02	–	–	–
MedNeXt	–	–	–	–	–	–	74.22	35.57	89.76
STU-Net	–	–	–	–	–	–	<u>79.15</u>	44.91	93.50
Attention based									
TransBTS	33.90	18.78	14.59	36.96	30.59	91.58	65.04	21.24	84.17
TransUNet	56.23	50.04	<u>53.97</u>	<u>86.54</u>	69.87	96.67	–	–	–
TransFuse	44.48	45.12	49.26	77.39	67.44	94.78	–	–	–
UTNet	53.17	48.94	48.16	84.35	72.01	96.73	–	–	–
Swin UNETR	52.19	46.12	47.59	74.52	65.12	96.82	–	–	–
UNETR	–	–	–	–	–	–	68.20	26.68	85.43
nnFormer	–	–	–	–	–	–	62.14	29.70	85.10
CoTr	–	–	–	–	–	–	76.50	11.74	89.83
3D UX-Net	–	–	–	–	–	–	70.98	29.89	88.06
State space									
U-Mamba	60.81	49.75	48.65	84.58	73.97	96.81	78.60	<u>45.08</u>	<u>94.53</u>
Mamba-UNet	46.33	43.11	47.50	82.70	66.76	96.37	–	–	–
VM-UNet	39.02	41.01	49.07	80.56	64.09	95.59	–	–	–
VM-UNet-V2	42.09	43.96	45.23	78.30	56.38	95.49	–	–	–
H-vmunet	50.21	42.63	39.71	80.82	65.65	95.37	–	–	–
Swin-UMamba	50.84	49.08	48.33	84.01	71.98	97.07	–	–	–
Test-time training based									
TTT-UNet	<u>61.23</u>	49.70	51.12	86.06	72.07	<u>97.30</u>	78.96	42.36	93.38
ViT^3	55.80	<u>51.50</u>	50.12	84.45	72.20	96.98	–	–	–
Med-TTT	50.13	45.35	43.16	75.84	64.52	95.47	–	–	–
QGA-TTT	**64.04**	**53.25**	**56.34**	**87.11**	**74.37**	**97.46**	**79.93**	**45.65**	**94.83**

4.3 Cost

Measured on one case against the self-configuring convolutional pipeline at the same patch size (Table 2), the relative overhead of QGA-TTT shrinks as the patch grows, from 7.2 times the time and 53% more memory on the smallest patch to 4.3 times and 23% on the volumetric one, because the cost of QGA-TTT is set by the channel width of the three deepest levels while the cost of the convolutional stack grows with the resolution it processes. A patch large enough to be clinically useful is therefore the regime where QGA-TTT is cheapest relative to the network it sits in. On the volumetric dataset its peak memory is within 3% of U-Mamba, the strongest comparison method there, and below that of UNet++ (Appendix B).

Table 6 in Appendix B extends Table 2 to every comparison method under the same measurement. Parameter count does not decide the ranking of Table 1. TransUNet carries 105.9 million parameters and nnFormer 149.1 million, both more than QGA-TTT, and both score below it on every dataset they run on.

Table 2: Forward pass on one case at the patch size each dataset's plan specifies, measured on one GeForce RTX 4090 with half precision autocast, a batch of one and the deep supervision heads off as at inference, averaged over twenty repetitions after five warm up passes. Memory is the peak memory of the pass, weights included.

		Parameters (M)		Time (ms)		Memory (MiB)	
Dataset	Patch	nnU-Net	QGA-TTT	nnU-Net	QGA-TTT	nnU-Net	QGA-TTT
OCT5k	512×512	46.3	87.3	4.6	33.3	401	614
Endoscapes2023	512×896	46.3	87.3	5.4	35.2	507	732
GCM2025	$20 \times 256 \times 320$	43.5	71.1	20.0	85.8	1095	1343

Table 3: Each construction in turn. A tick marks a construction the model keeps. Values are Dice; Δ is the difference from the complete model.

Construction kept			m2caiSeg		Endoscapes	
Queries	Carried over	Paired fitting	Dice	Δ	Dice	Δ
			51.99	-1.26	54.93	-1.41
✓			52.78	-0.47	55.43	-0.91
	✓	✓	53.02	-0.22	54.82	-1.51
✓	✓		51.73	-1.52	53.71	-2.62
✓	✓	✓	**53.25**	—	**56.34**	—

4.4 Ablations

Ablations run on m2caiSeg and Endoscapes2023, whose label spaces carry eighteen and six foreground classes. Each changes one factor and keeps the split, preprocessing, augmentation, budget, optimiser, schedule, checkpoint rule and evaluation code of the complete model.

Table 3 takes the three constructions in turn. Removing the queries, discarding them between the two passes, and reducing the two fittings to one each cost quality on both datasets, so all three carry weight. The first block of Table 4 separates the fit from the path it sits on. Replacing the fit with a plain linear aggregate keeps the shape of the mechanism, in that one global map is still formed from the image and read by every position, and it costs quality on both datasets. The gain therefore depends on the fit itself.

The remaining blocks vary how much of each quantity the model gets. The number of fittings and the capacity of the fitted map are interior optima, in that quality falls on both sides of the setting we use and does so on both datasets. Attaching QGA-TTT to the deepest level alone also costs quality, so the three levels are not redundant.

5 Analysis

The map the fitting produces can be read directly off a trained model. Its stable rank, the squared Frobenius norm divided by the squared largest singular value, is one for a map acting along a single direction and k for one spread evenly over k directions. Averaged over the validation images and over the heads of a level, every fitted map on m2caiSeg and Endoscapes2023 lands between 1.01 and 1.42, while the learned initialisation each fitting starts from lands between 2.43 and 3.58, and the two ranges do not meet. The leading direction takes half to four fifths of the total singular mass. Figure 4 shows the spectrum at the deepest level and the same measurement at every level and every fitting. One gradient step turns a map spread over many directions into one acting along close to a single direction, which is the form a level needs if all of its positions are to read one account. Halving the width of the map costs 1.13 and 1.11 Dice (Table 4), so the full width supplies the space from which each image and head chooses its direction.

Beyond the shape of a single fitted map, the queries also bring the maps of different images closer together. We measure the distance at the deepest enhanced level, where global mixing has always been placed in a segmentation network and where QGA-TTT on that level alone already recovers 1.52 and 1.31 Dice over nnU-Net in Table 4.

Table 4: Ablations of the source of the global map and of the amount of each quantity. The first block keeps the position of QGA-TTT and its gated injection and changes only the source of the map. Quality falls on both sides of the number of fittings and of the capacity of the fitted map.

		m2caiSeg		Endoscapes	
Factor	Setting	Dice	Δ	Dice	Δ
Source of the map	nothing, gate at zero	52.25	−1.00	55.55	−0.79
	a linear aggregate	52.96	−0.28	54.86	−1.48
	a fit	**53.25**	—	**56.34**	—
Fittings per pass	one	51.73	−1.52	53.71	−2.62
	two	**53.25**	—	**56.34**	—
	three	52.84	−0.41	53.84	−2.50
Capacity of the map	halved	52.12	−1.13	55.23	−1.11
	as proposed	**53.25**	—	**56.34**	—
	doubled	52.82	−0.43	53.53	−2.81
Queries	**eight**	**53.25**	—	**56.34**	—
	thirty two	51.83	−1.42	55.24	−1.09
Levels carrying QGA-TTT	deepest only	52.18	−1.07	54.85	−1.49
	deepest three	**53.25**	—	**56.34**	—

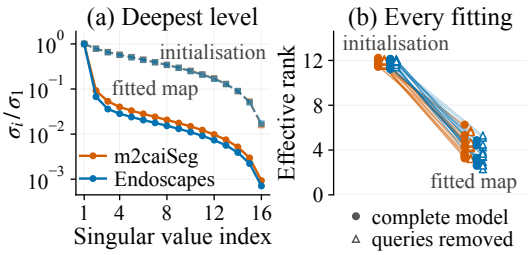


Figure 4: The map one gradient step produces acts along close to a single direction. (a) Singular values of the fitted map and of the learned initialisation it starts from at the deepest enhanced level, averaged over the validation images and heads and divided by the leading value of each curve. (b) Every level and every fitting on m2caiSeg and Endoscapes2023 against its initialisation, for the complete model and for the control trained without queries. Effective rank is the exponential of the entropy of the normalised singular values.

For every pair of validation images we take the fitted weights of the level, subtract them and divide by their magnitudes. The complete model and the control trained without queries see the same pairs of images, and we compare them pair by pair. Pairing matters on m2caiSeg, whose validation set holds consecutive frames of one procedure, so that near repeated images put mass near zero for both models. Across the four fittings of the level and both datasets the paired median distance of the complete model is 0.600 to 0.900 of the control, and the complete model is the closer of the two on 64 to 98 percent of the pairs. Figure 5 in Appendix C gives the pooled distributions and the comparison fitting by fitting.

The gate of Equation 6 starts at 10^{-3}, so any larger value is something training put there. Across all nine datasets the mean magnitude of the trained gate lands between 3.8×10^{-3} and 9.5×10^{-2}, and the shallowest of the three levels receives the largest gate on eight of the nine. The network places most of the contribution at the level that still carries the most positions.

6 Conclusion

Medical image segmentation asks for every position of an image to be labelled against one account of what that image contains, and the token mixing primitives segmentation networks are built from do not form one. QGA-TTT obtains such an account by fitting one affine map, inside a single forward pass, from a level's own positions together with a small set of queries shared across images. It improves on the strongest comparison method of nine public datasets in four modalities, and ablations show each of its three constructions to be necessary.

REFERENCES

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, 2022. doi: 10.1038/s41467-022-30695-9.

Mustafa Arikan, James Willoughby, Sevim Ongun, Ferenc Sallo, Andrea Montesel, Hend Ahmed, Ahmed Hagag, Marius Book, Henrik Faatz, Maria Vittoria Cicinelli, Amani A. Fawzi, Dominika Podkowinski, Marketa Cilkova, Diana Morais De Almeida, Moussa Zouache, Ganesham Ramsamy, Watjana Lilaonitkul, and Adam M. Dubis. OCT5k: A dataset of multi-disease and multi-graded annotations for retinal layers. *Scientific Data*, 12, 2025. doi: 10.1038/s41597-024-04259-z.

Jimmy Ba, Geoffrey E. Hinton, Volodymyr Mnih, Joel Z. Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/9f44e956e3a2b7b5598c625fcc802c36-Abstract.html.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In *Advances in Neural Information Processing Systems 38*, pp. 125925–125962, 2025. doi: 10.52202/085713-3786.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, volume 12346 of *Lecture Notes in Computer Science*, pp. 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Lecture Notes in Computer Science, pp. 424–432. Springer, 2016. doi: 10.1007/978-3-319-46723-8_49.

Gaowen Chen, Jieyun Bai, Zhanhong Ou, Yaosheng Lu, and Huijin Wang. PSFHS: Intrapartum ultrasound image dataset for AI-based segmentation of pubic symphysis and fetal head. *Scientific Data*, 11, 2024a. doi: 10.1038/s41597-024-03266-4.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. URL https://arxiv.org/abs/2102.04306.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024b. doi: 10.1016/j.media.2024.103280.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17864–17875. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/950a4152c2b4aa3ad78bdd6b366cc179-Abstract.html.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289, 2022. doi: 10.1109/CVPR52688.2022.00135.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with Performers. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2009.14794.

Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17702–17711, 2025. doi: 10.1109/CVPR52734.2025.01649.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071. PMLR, 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.16588.

Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945. doi: 10.2307/1932409.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.11929.

Yunhe Gao, Mu Zhou, and Dimitris N. Metaxas. UTNet: A hybrid transformer architecture for medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pp. 61–71. Springer, 2021. doi: 10.1007/978-3-030-87199-4_6.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2111.00396.

Dongchen Han, Yining Li, Tianyu Li, Zixuan Cao, Ziming Wang, Jun Song, Yu Cheng, Bo Zheng, and Gao Huang. ViT^3: Unlocking test-time training in vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 51–61, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Han_ViT3_Unlocking_Test-Time_Training_in_Vision_CVPR_2026_paper.html.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Lecture Notes in Computer Science, pp. 272–284. Springer, 2022a. doi: 10.1007/978-3-031-08999-2_22.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR: Transformers for 3D medical image segmentation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1748–1758, 2022b. doi: 10.1109/WACV51458.2022.00181.

Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, and Yu Qiao. STU-Net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv preprint arXiv:2304.06716*, 2023. URL https://arxiv.org/abs/2304.06716.

Daniel P. Huttenlocher, Gregory A. Klanderman, and William J. Rucklidge. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993. doi: 10.1109/34.232073.

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. doi: 10.1038/s41592-020-01008-z.

Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jäger. nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024*, Lecture Notes in Computer Science, pp. 488–498. Springer, 2024. doi: 10.1007/978-3-031-72114-4_47.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2107.14795.

Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-SEG: A segmented polyp dataset. In *MultiMedia Modeling*, Lecture Notes in Computer Science, pp. 451–462. Springer, 2020. doi: 10.1007/978-3-030-37734-2_37.

Debesh Jha, Onkar Kishor Susladkar, Vandan Gorade, Elif Keles, Matthew Antalek, Deniz Seyithanoglu, Timurhan Cebeci, Halil Ertugrul Aktas, Gulbiz Dagoglu Kartal, Sabahattin Kaymakoglu, Sukru Mehmet Erturk, Yuri Velichko, Daniela P. Ladner, Amir A. Borhani, Alpay Medetalibeyoglu, Gorkem Durak, and Ulas Bagci. Large scale MRI collection and segmentation of cirrhotic liver. *Scientific Data*, 12, 2025. doi: 10.1038/s41597-025-05201-7.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165. PMLR, 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 562–570. PMLR, 2015. URL https://proceedings.mlr.press/v38/lee15a.html.

Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A. Landman. 3D UX-Net: A large kernel volumetric ConvNet modernizing hierarchical transformer for medical image segmentation. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2209.15076.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Chenxin Li, Xinyu Liu, Wuyang Li, Cheng Wang, Hengyu Liu, Yifan Liu, Zhen Chen, and Yixuan Yuan. U-KAN makes strong backbone for medical image segmentation and generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5):4652–4660, 2025. doi: 10.1609/aaai.v39i5.32491.

Jiaming Liang, Lihuan Dai, Xiaoqi Sheng, Xiangguang Chen, Chun Yao, Guihua Tao, Qibin Leng, Hongmin Cai, and Xi Zhong. HWA-UNETR: Hierarchical window aggregate UNETR for 3D multimodal gastric lesion segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2025*, Lecture Notes in Computer Science, pp. 273–282. Springer, 2025a. doi: 10.1007/978-3-032-05141-7_27.

Xiaomin Liang, Linda A. Alpuing Radilla, Kamand Khalaj, Haaniya Dawoodally, Chinmay Mokashi, Xiaoming Guan, Kirk E. Roberts, Sunil A. Sheth, Varaha S. Tammisetti, and Luca Giancardo. A multi-modal pelvic MRI dataset for deep learning-based pelvic organ segmentation in endometriosis. *Scientific Data*, 12, 2025b. doi: 10.1038/s41597-025-05623-3.

Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, Hairong Zheng, and Shanshan Wang. Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024*, Lecture Notes in Computer Science, pp. 615–625. Springer, 2024a. doi: 10.1007/978-3-031-72114-4_59.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems 37*, pp. 103031–103063, 2024b. doi: 10.52202/079017-3273.

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)*, pp. 9992–10002, 2021. doi: 10.1109/ICCV48922.2021.00986.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11525–11538. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8511df98c02ab60aea1b2356c013bc0f-Abstract.html.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1711.05101.

Jun Ma, Feifei Li, and Bo Wang. U-Mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. URL https://arxiv.org/abs/2401.04722.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, A. Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew B. Blaschko, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Robert Haase, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, Brennan

Nichyporuk, Felix Nickel, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Maarten van Smeden, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, and Paul F. Jäger. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, 2024. doi: 10.1038/s41592-023-02151-z.

Salman Maqbool, Aqsa Riaz, Hasan Sajid, and Osman Hasan. m2caiSeg: Semantic segmentation of laparoscopic images using convolutional neural networks. *arXiv preprint arXiv:2008.10134*, 2020. URL https://arxiv.org/abs/2008.10134.

Pietro Mascagni, Deepak Alapatt, Aditya Murali, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Guido Costamagna, Didier Mutter, Jacques Marescaux, Bernard Dallemagne, and Nicolas Padoy. Endoscapes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. *Scientific Data*, 12, 2025. doi: 10.1038/s41597-025-04642-4.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016. doi: 10.1109/3DV.2016.79.

Aditya Murali, Deepak Alapatt, Pietro Mascagni, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Guido Costamagna, Didier Mutter, Jacques Marescaux, Bernard Dallemagne, and Nicolas Padoy. The Endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. *arXiv preprint arXiv:2312.12429*, 2023. URL https://arxiv.org/abs/2312.12429.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. URL https://arxiv.org/abs/1804.03999.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F. Jäger, and Klaus H. Maier-Hein. MedNeXt: Transformer-driven scaling of ConvNets for medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023*, Lecture Notes in Computer Science, pp. 405–415. Springer, 2023. doi: 10.1007/978-3-031-43901-8_39.

Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. VM-UNet: Vision Mamba UNet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2025. doi: 10.1145/3767748.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9355–9366. PMLR, 2021. URL https://proceedings.mlr.press/v139/schlag21a.html.

Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992. doi: 10.1162/neco.1992.4.1.131.

Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3530–3538, 2021. doi: 10.1109/WACV48630.2021.00357.

Xin Shi, Qing Huang, Teng Xu, Hongwen Mei, Tingwei Quan, Xiuli Wang, Yinghan Shi, Ye Hu, Zhimei Duan, Fei Xie, Sifan Li, Lixin Xie, and Kaifei Wang. PW-BALFC, a clinical dataset for detection and instance segmentation of bronchoalveolar lavage fluid cell. *Scientific Data*, 12, 2025. doi: 10.1038/s41597-025-05452-4.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (Learn at test time): RNNs with expressive hidden states. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57503–57522. PMLR, 2025. URL https://proceedings.mlr.press/v267/sun25h.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. URL https://arxiv.org/abs/2006.04768.

Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. TransBTS: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pp. 109–119. Springer, 2021. doi: 10.1007/978-3-030-87193-2_11.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7794–7803, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Non-Local_Neural_Networks_CVPR_2018_paper.html.

Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. Mamba-UNet: UNet-like pure visual Mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*, 2024. URL https://arxiv.org/abs/2402.05079.

Renkai Wu, Yinghao Liu, Pengchen Liang, and Qing Chang. H-vmunet: High-order vision Mamba UNet for medical image segmentation. *Neurocomputing*, 624:129447, 2025. doi: 10.1016/j.neucom.2025.129447.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html.

Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pp. 171–180. Springer, 2021b. doi: 10.1007/978-3-030-87199-4_16.

Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. SegMamba: Long-range sequential modeling Mamba for 3D medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024*, Lecture Notes in Computer Science, pp. 578–588. Springer, 2024. doi: 10.1007/978-3-031-72111-3_54.

Jiashu Xu. Med-TTT: Vision test-time training model for medical image segmentation. *arXiv preprint arXiv:2410.02523*, 2024. doi: 10.48550/arXiv.2410.02523. URL https://arxiv.org/abs/2410.02523.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems 37*, pp. 115491–115522, 2024. doi: 10.52202/079017-3668.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2412.06464.

Yundong Zhang, Huiye Liu, and Qiang Hu. TransFuse: Fusing transformers and CNNs for medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pp. 14–24. Springer, 2021. doi: 10.1007/978-3-030-87193-2_2.

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnFormer: Volumetric medical image segmentation via a 3D transformer. *IEEE Transactions on Image Processing*, 32:4036–4045, 2023. doi: 10.1109/TIP.2023.3293771.

Rong Zhou, Zhengqing Yuan, Zhiling Yan, Weixiang Sun, Kai Zhang, Yiwei Li, Yanfang Ye, Xiang Li, Lichao Sun, and Lifang He. TTT-UNet: Enhancing U-Net with test-time training layers for biomedical image segmentation. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 3679–3703. PMLR, 2026. URL https://proceedings.mlr.press/v315/zhou26a.html.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Lecture Notes in Computer Science, pp. 3–11. Springer, 2018. doi: 10.1007/978-3-030-00889-5_1.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2020. doi: 10.1109/TMI.2019.2959609.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62429–62442. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhu24f.html.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.04159.

A BOUNDARY DISTANCES

Table 5: 95th percentile Hausdorff distance (HD95) on the nine datasets, in voxels, lower is better. Rows follow Table 1. Best per dataset in bold, second best underlined.

	two dimensional						volumetric		
Method	OCT5k	m2caiSeg	Endoscapes	Kvasir	PW-BALFC	PSFHS	UT-Endo	GCM2025	CirrMRI
Convolutional									
U-Net	25.37	35.64	48.33	44.05	66.29	0.25	12.30	90.25	31.66
UNet++	26.07	37.45	52.35	47.28	60.40	0.28	12.58	85.70	40.74
nnU-Net	<u>13.72</u>	28.08	<u>31.18</u>	37.94	55.97	0.17	12.18	**74.12**	18.93
U-KAN	16.99	26.43	40.86	45.87	66.70	0.21	–	–	–
MedNeXt	–	–	–	–	–	–	13.57	93.42	36.92
STU-Net	–	–	–	–	–	–	<u>9.93</u>	76.77	20.95
Attention based									
TransBTS	73.52	107.15	103.77	224.09	234.74	2.85	16.62	123.15	47.52
TransUNet	15.44	**25.10**	31.74	**33.52**	64.29	0.23	–	–	–
TransFuse	61.86	28.68	43.04	40.29	74.05	0.56	–	–	–
UTNet	21.89	27.15	41.66	45.90	61.06	0.22	–	–	–
Swin UNETR	39.26	31.48	45.03	69.19	63.98	0.22	–	–	–
UNETR	–	–	–	–	–	–	23.89	117.05	52.68
nnFormer	–	–	–	–	–	–	17.76	95.76	58.23
CoTr	–	–	–	–	–	–	13.76	139.16	38.58
3D UX-Net	–	–	–	–	–	–	14.77	117.44	52.05
State space									
U-Mamba	13.79	28.97	34.03	39.25	55.75	0.21	19.99	79.65	**10.85**
Mamba-UNet	45.72	34.56	37.64	42.73	73.53	0.26	–	–	–
VM-UNet	36.20	29.66	36.31	51.78	83.91	0.40	–	–	–
VM-UNet-V2	65.24	30.61	55.31	61.19	100.72	0.43	–	–	–
H-vmunet	34.52	33.72	54.49	53.30	66.94	0.45	–	–	–
Swin-UMamba	22.61	31.70	45.71	44.16	<u>49.56</u>	0.18	–	–	–
Test-time training based									
TTT-UNet	13.77	27.21	32.89	43.29	56.26	<u>0.16</u>	11.55	<u>75.11</u>	26.79
ViT^3	18.38	<u>26.22</u>	38.20	44.58	**49.44**	0.20	–	–	–
Med-TTT	46.47	30.06	48.27	63.86	71.04	0.44	–	–	–
QGA-TTT	**12.26**	29.40	**30.19**	<u>34.45</u>	55.04	**0.16**	**9.34**	82.43	<u>17.21</u>

B COST OF EVERY METHOD

Table 6: Forward pass on one case for every method, measured as in Table 2. Time is in milliseconds, memory is the peak memory of the pass in MiB with weights included, and parameters are in millions. Rows follow Table 1. A dash marks data a method's public implementation does not cover. Entries marked † are measured at the patch their implementation accepts, 512×512 for TransUNet, TransFuse and VM-UNet-V2 on Endoscapes2023, 128^3 for TransBTS, and $16 \times 256 \times 320$ for MedNeXt, UNETR, nnFormer, CoTr and 3D UX-Net.

	OCT5k, 512×512			Endoscapes2023, 512×896			GCM2025, $20 \times 256 \times 320$		
Method	Time	Memory	Params	Time	Memory	Params	Time	Memory	Params
Convolutional									
U-Net	3.1	155	2.0	3.6	266	2.0	13.8	872	5.7
UNet++	6.1	375	2.4	9.7	650	2.4	48.1	2207	7.0
nnU-Net	4.6	401	46.3	5.4	507	46.3	20.0	1095	43.5
U-KAN	13.3	306	25.4	20.3	445	25.4	–	–	–
MedNeXt	–	–	–	–	–	–	53.7†	1053	6.0
STU-Net	–	–	–	–	–	–	31.5	1282	58.2
Attention based									
TransBTS	32.9†	1132	32.8	33.2†	1148	32.9	32.8†	1132	32.8
TransUNet	15.4	786	105.9	16.0†	788	105.9	–	–	–
TransFuse	10.5	204	26.4	10.3†	207	26.6	–	–	–
UTNet	9.6	306	10.0	12.6	489	10.0	–	–	–
Swin UNETR	19.0	390	25.1	20.5	555	25.1	–	–	–
UNETR	–	–	–	–	–	–	15.2†	858	90.5
nnFormer	–	–	–	–	–	–	28.9†	1242	149.1
CoTr	–	–	–	–	–	–	20.2†	855	41.9
3D UX-Net	–	–	–	–	–	–	52.6†	1268	53.0
State space									
U-Mamba	9.2	528	64.8	10.0	646	64.8	45.1	1309	62.3
Mamba-UNet	23.2	498	35.9	26.7	717	35.9	–	–	–
VM-UNet	13.8	427	27.5	17.7	627	27.5	–	–	–
VM-UNet-V2	14.1	391	22.8	14.4†	393	22.8	–	–	–
H-vmunet	71.2	116	9.0	69.4	164	9.0	–	–	–
Swin-UMamba	23.9	633	59.9	29.3	831	59.9	–	–	–
Test-time training based									
TTT-UNet	10.3	526	64.2	11.1	644	64.2	45.3	1308	62.0
ViT^3	38.1	246	29.5	37.6	342	29.5	–	–	–
Med-TTT	68.2	642	44.2	64.0	658	44.2	–	–	–
QGA-TTT	33.3	614	87.3	35.2	732	87.3	85.8	1343	71.1

C DISTANCE BETWEEN THE MAPS TWO IMAGES FIT

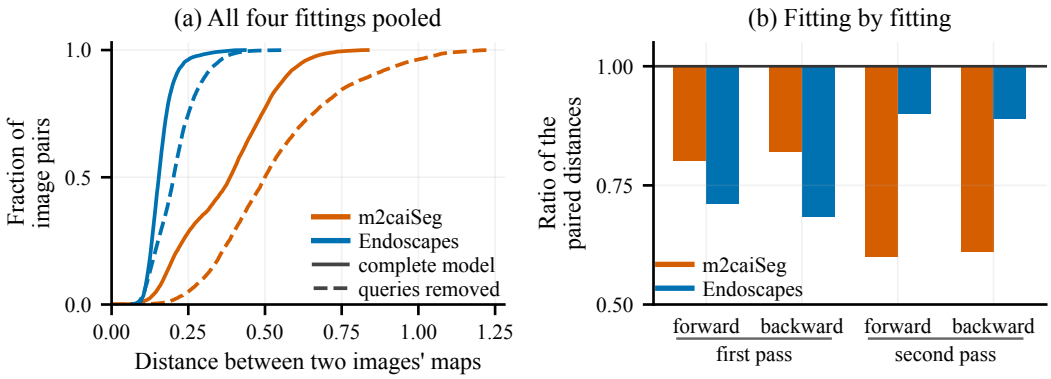


Figure 5: Distance between the maps two images fit, at the deepest enhanced level. Panel (a) pools the four fittings of the level and gives the fraction of image pairs whose maps lie within a given distance. Panel (b) compares the two models pair by pair and gives the median of the ratio, one fitting at a time, so a bar below one marks a fitting whose maps sit closer together in the complete model.

D COMPLETE QUALITATIVE COMPARISON

Figure 6 extends Figure 3 to all nine datasets and to every comparison method that runs on each, fifteen on the two dimensional datasets and eleven on the volumetric ones. Each row shows the case whose Dice under QGA-TTT is closest to the median of that dataset, and the rows for OCT5k, m2caiSeg and Endoscapes2023 show the same cases in the same windows as Figure 3. Volumetric cases are shown at the axial slice carrying the most annotated voxels, and every method in a row is shown at that same slice. A row is cropped only as far as the tightest square that still leaves at least 95% of every method's predicted foreground inside the panel, and the two rows for which no such square exists are shown at full frame, which is why their panels are wider than they are tall.

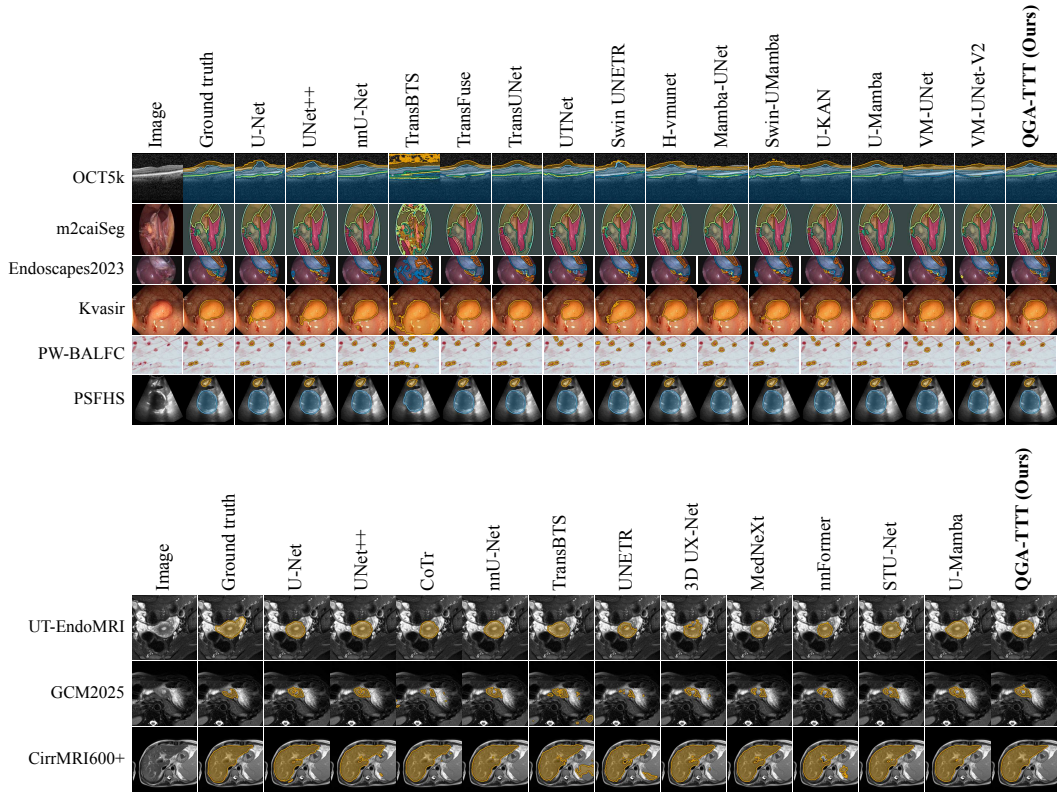


Figure 6: Segmentation on one case per dataset with every comparison method available on that dataset. Cases, slices and crop windows are chosen as described in the text. The two blocks carry different method sets, so each has its own header row. Across both blocks the crop hides at most 4.3% of any one method's prediction.

E REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of our results. Section 3 specifies every component of the method, and Section 4.1 gives the datasets, the splits, the training settings and the evaluation metrics. Section 4.3 and Appendix B report how the cost of every method is measured, and Appendix A gives the boundary distances on every dataset. All datasets used in this paper are publicly available. The code and the trained models will be released upon publication.

F LLM USAGE STATEMENT

We used large language models as writing aids to improve the readability of the manuscript and as search tools to find related work. We reviewed all suggested revisions, checked every cited paper against its original publication, and take responsibility for the content of this paper.