LOCAL ATTENTION LEARNS FROM REGISTRATION FOR IN-CONTEXT MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In-Context Segmentation (ICS) uses a small set of annotated samples as a prompt and has proven to be a powerful paradigm for both domain and task generalization in medical image segmentation. However, current ICS models often assume that the query image and the provided context are spatially aligned—a condition frequently unmet in real-world clinical scenarios. This paper introduces OmniSeg, a novel ICS model that dynamically aligns context and query representations, with two core contributions: SpotBlock, a new module combining spatial and contextual dense attention, and a deep registration scheme that supervises this attention with a registration loss. Our method better leverages the context set, which is crucial for out-of-distribution scenarios, and thus achieves strong generalization. OmniSeg significantly outperforms state-of-the-art ICS methods, yielding an average gain of 9.8 Dice points against the best baseline across 10 held-out datasets. Our approach paves the way for a new generation of spatially-aware ICS models. Source code will be released upon publication.

1 INTRODUCTION

Medical image segmentation is a cornerstone of deep learning in healthcare. However, clinical deployment is often hampered by poor generalization across acquisition centers and the scarcity of expert annotations. To address these hurdles, *universal* segmentation models have recently emerged. By training on diverse, aggregated datasets, these models generalize to a wide array of target classes and modalities, even in low-shot settings. They thus provide robust results in data-scarce scenarios and can significantly accelerate medical image annotation within semi-automatic workflows.

Because they are not tied to a fixed label set, they must be instructed on what to segment in a given image. Therefore, different categories of universal models can be defined based on the modality of the prompt conditioning the segmentation task: spatial prompts (scribbles, clicks, etc.) around the target object for interactive models (Ma et al., 2024; Wong et al., 2024; Isensee et al., 2025), a textual description of the task for text-guided models (Zhao et al., 2025; Rokuss et al., 2026), or examples of the expected segmentation results on a subset of images for in-context segmentation (ICS) models.

UniverSeg (Butoi* et al., 2023), MultiverSeg (Wong et al., 2025) or Tyche (Rakic et al., 2024) are successful two-dimensional ICS models able to generalize to unseen and diverse segmentation tasks. They are based on a convolutional mechanism, the *CrossBlock*, which fuses the features of the query image and context set at different scales of a U-shaped network. These ICS approaches assume that the query image and the context set of segmentation samples are spatially aligned and fuse them in a pixel-wise fashion. However, this assumption is commonly violated in real conditions, as illustrated in Fig. 1 for hip segmentation in X-rays: the misalignment between the three context masks and the query makes MultiverSeg inaccurate. Moreover, the context features are aggregated through simple averaging, disregarding the observation substantiated by Gao et al. (2025) that a *good* context is better than a *large* context.

We propose OmniSeg, an ICS framework that registers context representations onto the query while spotting the most relevant information across the context. As shown in Fig. 1*(a)*, OmniSeg infers displacement fields registering the context masks onto the query, enabling the fusion of similar semantic structures and the aggregation of relevant information from the context set. As shown in Fig. 1*(b)*, OmniSeg's advantage over MultiverSeg increases with spatial misalignment. We make the following contributions:

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.

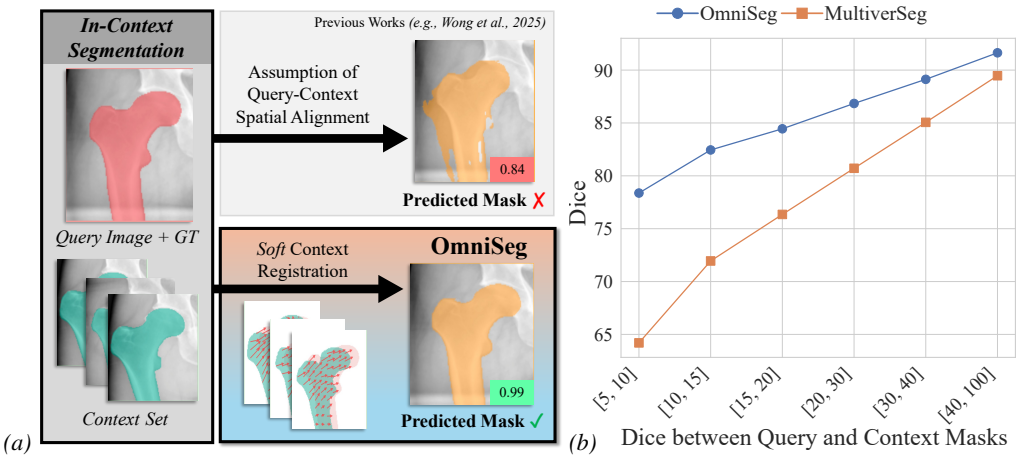

Figure 1: **In-context segmentation robust to query-context spatial misalignment.** *(a)* We propose a new in-context segmentation model, using a soft registration mechanism supervised in a deep manner, yielding few-shot segmentation that is more robust to misalignment between context and query. *(b)* Evaluated on a large and diverse test set, the performance gain of our method over MultiverSeg (Wong et al., 2025) increases as the spatial misalignment between the context set and the query grows.

- We introduce SpotBlock, a new module fusing query and context features (Section 3.1). This module combines two dense attention mechanisms, first across local neighborhoods, then across the context examples. On the one hand, the local attention relaxes the constraints of spatial alignment between query image and context examples. On the other hand, dense context pooling enables the model to better leverage relevant signals relative to the query from the context set.
- SpotBlock induces a Displacement Vector Field (DVF) approximating the spatial mapping between query and context feature maps, see red arrows in Fig. 1. To explicitly drive registration during training, we design a proxy for the unknown target DVF using an alignment loss based on segmentation masks (Section 3.2). We provide an analytical derivation demonstrating that this supervision has a dense contrastive effect on the query's representation learning, supporting a collaboration between alignment and segmentation objectives.
- We assemble a large and diverse dataset for universal medical ICS with a total of 1.8 million image-mask pairs covering 8 medical imaging modalities. Extensive experiments reported in Section 4 show that OmniSeg clearly surpasses current state-of-the-art models. We validate our architectural and supervision contributions through a comprehensive ablation study. OmniSeg remains effective with small context sizes and transfers particularly well to unseen tasks and domains.

2 Related Works

Medical image segmentation has evolved from U-Net (Ronneberger et al., 2015) and hybrid ConvNet-Transformer architectures (Chen et al., 2024; Petit et al., 2021; Cao et al., 2022) to foundation models such as SAM (Kirillov et al., 2023) and its medical adaptations (Ma et al., 2024; 2025), which enable broad anatomical coverage but remain dependent on ambiguity-prone human annotations or interactions. These models also fail to generalize across diverse tasks.

In-context segmentation (ICS) for medical images can be organized into three families. *Prototype-based methods* (Wang et al., 2019; Ouyang et al., 2022; Cuttano et al., 2026; Sun et al., 2026) match class prototypes against query features but collapse spatial structure into compact representations. *Atlas-based methods* (Ding et al., 2020; Zhang et al., 2025; He et al., 2022) register an annotated reference onto the query, preserving spatial structure but depending on a separate registration step. Recently, UniverSeg (Butoi* et al., 2023) established a foundational framework: a single in-context model can generalize to unseen medical segmentation tasks without task-specific fine-tuning, extending few-shot segmentation to universal segmentation. It initiated the *fusion-based methods* category, where models densely combine context and query features at multiple scales: UniverSeg introduced

the CrossBlock for this purpose within a U-shaped network, extended by MultiverSeg (Wong et al., 2025) and Tyche (Rakic et al., 2024). Other works augment ICS with generative objectives (Wang et al., 2023; Czolbe & Dalca, 2023; Hu et al., 2025), retrieval (Gao et al., 2025), prompt conditioning (Wu & Xu, 2024), or mixture-of-experts (Guo et al., 2026). These methods typically assume spatial alignment between query and context, moreover, they consider context examples as equally relevant, two assumptions commonly violated in practice.

Local attention for registration. Several works explore displacement-aware or spatially-restricted attention without directly supervising alignment for segmentation. Atlas-based methods (Ding et al., 2020; Zhang et al., 2025; He et al., 2022) align a reference to the query through a separate registration module, decoupled from the segmentation objective. COVER (He & Li, 2025) learns features invariant to spatial transformations by supervising the estimation of closed-form displacement fields computed between two augmented views of the *same* image from their feature maps. Our method takes inspiration from COVER, with key adaptations to the encoder, decoder, and skip connections that enable proper registration in ICS, plus a training loss acting as a proxy for the unknown DVF.

3 METHOD

Problem setup An in-context segmentation (ICS) model f_θ is trained on diverse datasets, i.e., image-mask pairs $\Omega = \{(x, y)\}$ across multiple medical segmentation tasks $\mathcal{T} = \{(\text{modality}, \text{axis}, \text{target})\}$, to achieve task-agnostic generalization. At inference, the segmentation task $t \in \mathcal{T}$ to perform is specified through a context set $\mathcal{C}^t = \{(x_i^{\mathcal{C}}, y_i^{\mathcal{C}})\}_{i\in[1,S]} \subset \Omega^t$, and the segmentation of a query image x is given by $\hat{y} = f_\theta(x, \mathcal{C}^t)$.

However, context and query are often spatially misaligned, making it difficult for state-of-the-art ICS methods to transfer spatial information specified by the context masks (see Fig. 1). To overcome this, we introduce a new in-context architecture (Section 3.1) and supervision scheme (Section 3.2) that register context and query in feature space.

OmniSeg leverages *soft* registration Fig. 2 gives an overview of OmniSeg. We adopt a U-shaped layout with a two-branch convolutional encoder which separately computes context and query dense representations. The bottleneck and decoder fuse context and query features through a succession of *SpotBlocks*, a new module presented in Section 3.1. At each decoder level, the SpotBlock modules perform dense local attention between query and context features, both spatially and across the context set. This dual mechanism provides robustness to spatial misalignment via soft registration while selecting relevant information from the context.

Each SpotBlock computes registration DVFs between the query and the context, which are used to warp the context features passed through the skip connections. During training, we explicitly supervise the DVFs. Since the target DVF is unknown, we use a sparse registration proxy loss, $\mathcal{L}_{\text{reg}}$, based on the alignment of segmentation masks. Importantly, we propose a deep supervision scheme that applies $\mathcal{L}_{\text{reg}}$ at each level of the decoder. A complete description of OmniSeg's architecture is provided in Appendix A.1.

3.1 SPOTBLOCK

ICS models must handle spatial and contextual information effectively. The way the context is aggregated and combined with query features is key to segmentation performance. We therefore propose a *Soft Registration Mechanism* (SRM) that matches features from the same anatomical region and avoids mixing unrelated information. To efficiently aggregate context samples, we introduce a *Dense Context Pooling* (DCP) mechanism that fetches, per pixel, the best features across the context set. SRM and DCP enable the model to better benefit from the context information, which is crucial in out-of-distribution scenarios where the model cannot solely rely on knowledge of seen tasks. As shown in Fig. 3, the combination of spatial and contextual re-embedding via SRM followed by DCP constitutes our proposed module: SpotBlock. SpotBlock takes as input a query feature map q and a set of S context feature maps $C = \{c_i\}_{i\in[1,S]}$ and outputs fused versions of these feature maps (q', C').

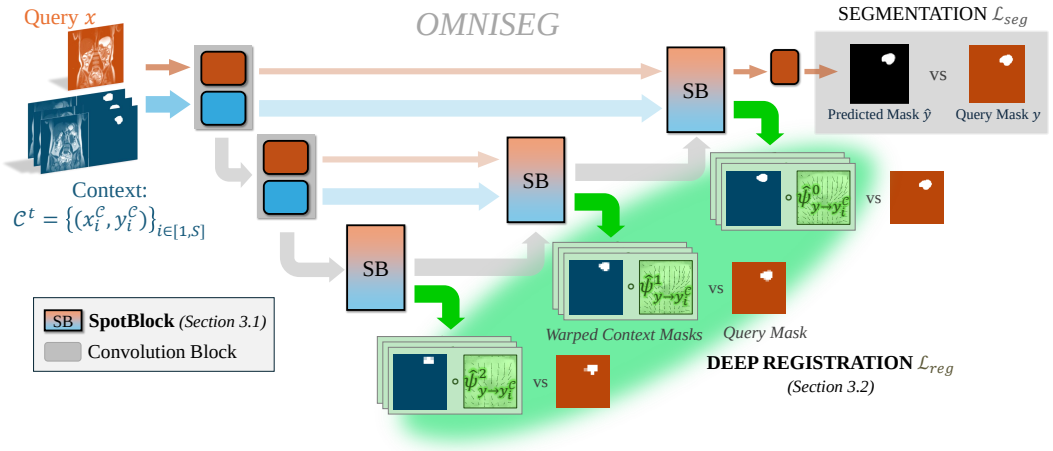

Figure 2: **OmniSeg overview**. Our U-shaped model takes as input a query image and a context set $\mathcal{C}^t = \{(x_i^{\mathcal{C}}, y_i^{\mathcal{C}})\}_{i\in[1,S]}$ of image/mask pairs prompting the segmentation task. After feature extraction via a two-branch convolutional encoder, the decoder fuses the context and query representations using SpotBlocks, which apply successive neighbor-wise and context-wise attentions (Section 3.1). Each SpotBlock induces a DVF that registers context features onto the query grid. We employ deep supervision to explicitly support registration via $\mathcal{L}_{\text{reg}}$. We show in Section 3.2 that $\mathcal{L}_{\text{reg}}$ complements the segmentation loss $\mathcal{L}_{\text{seg}}$ for ICS.

Soft Registration Mechanism (SRM) SRM pulls local context features toward corresponding pixel locations in the query by applying neighbor-wise cross-attention between query and context features. This can be interpreted as *softly* registering the query and context representations. First, convolution layers are applied to extract the query $Q \in \mathbb{R}^{H\times W\times F}$ feature maps from q along with the keys $K_i \in \mathbb{R}^{H\times W\times F}$ and values $V_i \in \mathbb{R}^{H\times W\times F_\nu}$ from each c_i. Q and K_i are then ℓ_2-normalized. We define $\Delta_{\mathcal{N}} = \{\Delta_j, j \in [1, N]\} \in \mathbb{R}^{N\times 2}$ as the set of offsets from the center of the neighborhood $\mathcal{N}$. We construct a key matrix $\mathbf{K}_i \in \mathbb{R}^{N\times H\times W\times F}$ and a value matrix $\mathbf{V}_i \in \mathbb{R}^{N\times H\times W\times F_\nu}$ that replicate K_i and V_i, respectively, with an offset of Δ_j, for all $j \in [1, N]$. We define the dense neighbor attention map $A_i^{\mathcal{N}}$ as:

$$A_i^{\mathcal{N}} = \underset{N}{\text{Softmax}}\left(\frac{Q \times_F \mathbf{K}_i}{T}\right) \quad \in \mathbb{R}^{N\times H\times W} \tag{1}$$

where $\times_F$ denotes tensor contraction along the feature dimension, and $\times_N$ and Softmax_N denote contraction and softmax along the neighbor dimension (see Appendix A.2). T is a fixed temperature hyperparameter. Finally, the registered context feature maps ν_i are obtained as the output of the dense local cross-attention:

$$\nu_i = A_i^{\mathcal{N}} \times_N \mathbf{V}_i \quad \in \mathbb{R}^{H\times W\times F_\nu} \tag{2}$$

Soft registration emerges from local attention The soft registration in Eq. (2) can be reinterpreted as computing a registration DVF from the query to each context sample i. The resulting DVF from the query feature map q to the context feature map c_i is given by:

$$\tilde{\psi}_{q\to c_i} = A_i^{\mathcal{N}} \times_N \Delta_{\mathcal{N}} \quad \in \mathbb{R}^{H\times W\times 2}. \tag{3}$$

Note that this DVF, when applied to the context feature maps, approximates the effect of the SRM, i.e., $\nu_i[p] \approx V_i[p + \tilde{\psi}_{q\to c_i}[p]]$. Chaining the SpotBlocks across levels incrementally registers the context features onto the query (see Fig. 6 in Section 4.3). The DVF computed from a given SpotBlock is thus the residual displacement at scale l. The total displacement $\hat{\psi}^l_{y\to y_i^{\mathcal{C}}}$ is obtained via composition.

Dense Context Pooling (DCP) While SRM pulls features within spatial neighborhoods, DCP pulls features across the softly registered context pool to extract information relevant to the query. We reuse the neighbor-wise query–context attention from Eq. (1) to pool the fused query–context features. Formally, the dense contextual attention $A^{\mathcal{C}}$ is computed as:

$$A^{\mathcal{C}} = \underset{S}{\text{Softmax}}\left(A^{\mathcal{N}} \times_N \frac{Q \times_F \mathbf{K}}{T}\right) \quad \in \mathbb{R}^{S\times H\times W} \tag{4}$$

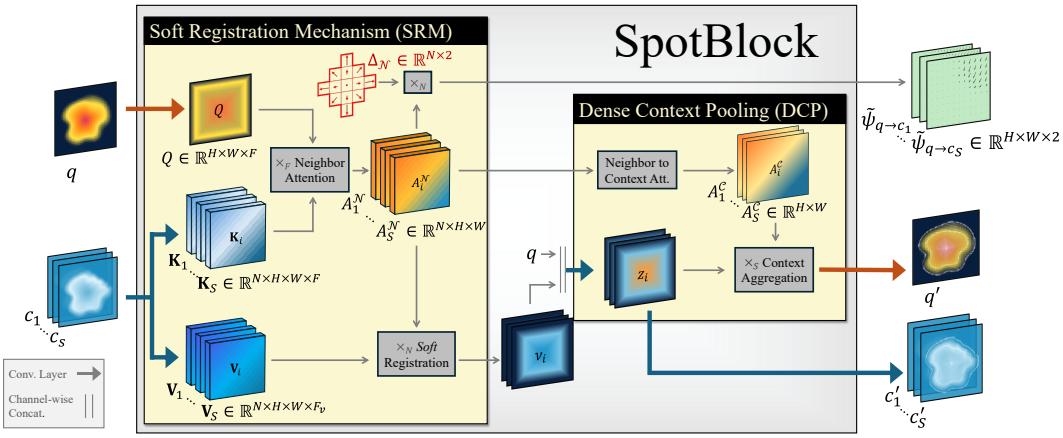

Figure 3: **Overview of the SpotBlock**. First, the Soft Registration Mechanism (SRM) registers each context feature map c_i for $i = 1, \dots, S$ onto the query feature map q through neighbor-wise cross-attention, giving ν_i (Eq. (2)). This translates into applying a DVF $\tilde{\psi}_{q \to c_i}$ (Eq. (3)), which we supervise at each SpotBlock with $\mathcal{L}_{\text{reg}}$ (Section 3.2). To select relevant registered context information, we perform Dense Context Pooling on fused query-context features Z_i using context attention maps $A^{\mathcal{C}}$ (Eq. (5)).

This dense attention map $A_i^{\mathcal{C}}$ indicates at a pixel level how much and where context sample i is similar to the query after registration. The registered context feature maps are fused with the query via concatenation and convolution, defining $Z = \{\text{ConvLayer}(q \parallel \nu_i)\}_{i \in [1,S]} \in \mathbb{R}^{S \times H \times W \times F_z}$. The aggregated feature map ζ across the context set is then obtained as:

$$\zeta = A^{\mathcal{C}} \times_S Z \quad \in \mathbb{R}^{H \times W \times F_z}. \tag{5}$$

The final outputs of a SpotBlock are the fused query q' and context feature maps $(c'_1, \dots, c'_S)$, obtained after applying a convolution layer to ζ and to $(Z_1, \dots, Z_S)$ separately. Further details regarding SpotBlock are provided in Appendix A.2.

3.2 Supervision

OmniSeg loss The loss supervising the training of OmniSeg consists of two objectives, segmentation and registration:

$$\mathcal{L}(\theta, \mathcal{T}) = \mathbb{E}_{t \sim \mathcal{T}} \mathbb{E}_{\{(x,y), \mathcal{C}^t\} \sim \Omega^t} \left[\mathcal{L}_{\text{seg}}(\hat{y}\ ;\ y) + \mathcal{L}_{\text{reg}}(\hat{\psi}\ ;\ y, \mathcal{C}^t) \right], \tag{6}$$

where $\hat{y}$ and $\hat{\psi}$ are the predicted mask and the DVFs output by OmniSeg.

Segmentation supervision As the segmentation loss $\mathcal{L}_{\text{seg}}$, we use the sum of the standard soft Dice (Milletari et al., 2016) and focal (Lin et al., 2017) losses to supervise the predicted mask.

Registration supervision In the ICS setting, ground truth for the DVF is not directly available. Therefore, supervision is obtained by measuring how well the computed DVFs register the context masks onto the query masks, essentially evaluating the overlap between the warped context masks and the query mask through the soft Dice loss. We use masks rather than images to focus registration on the target shapes, since segmentation is the primary objective. We denote the associated loss as $\mathcal{L}_{\text{align}}$, defined for each context example $i \in [1, S]$ as:

$$\mathcal{L}_{\text{align}}(\hat{\psi}_{y \to y_i^{\mathcal{C}}}\ ;\ y, y_i^{\mathcal{C}}) = \mathcal{L}_{\text{Dice}}(y_i^{\mathcal{C}} \circ \hat{\psi}_{y \to y_i^{\mathcal{C}}}\ ;\ y) \tag{7}$$

Although $\hat{\psi}_{y \to y_i^{\mathcal{C}}}$ is computed from the query and context feature maps (Eq. (3)), we reuse the same field here to warp the corresponding context mask $y_i^{\mathcal{C}}$ onto the query mask y.

The registration supervision is performed at each depth level l of OmniSeg's decoder and for each context example i with $\mathcal{L}_{\text{align}}$. At each decoder level, we back-propagate only the alignment loss for

the best-aligned context example to limit the impact of outliers in the context set:

$$\mathcal{L}_{\text{reg}}(\hat{\psi}\ ;\ y, \mathcal{C}^t) = \frac{1}{\omega}\sum_{l=0}^{L} \alpha_l \min_{i\in[1,S]} \mathcal{L}_{\text{align}}(\hat{\psi}^l_{y\to y_i^{\mathcal{C}}}\ ;\ y, y_i^{\mathcal{C}}). \tag{8}$$

where $\alpha_l = \frac{1}{2^{L-l}}$ are the factors corresponding to resolution upscaling along the bottleneck and decoder, and $\omega = \sum_l \alpha_l$ is the normalization factor.

Registration supervision complements the segmentation objective The DVFs output by SpotBlocks are calculated from the query and context feature maps rather than explicitly predicted. Minimizing the registration loss therefore has a direct impact on the dense representations of query and context. We derive the gradient of $\mathcal{L}_{\text{align}}$ in Appendix B, showing that it decomposes into pixel-wise terms:

$$\nabla_\theta \mathcal{L}_{\text{align}}(\hat{\psi}_{y\to y_i^{\mathcal{C}}}\ ;\ y, y_i^{\mathcal{C}}) \propto -\sum_{p\in\text{pixels}} \lambda[p] \sum_{j=1}^{N} (\Delta_j \cdot n^*[p]) \nabla_\theta A^{\mathcal{N}}_{i,j}[p] \tag{9}$$

where $n^*[p]$ is the inward normal to the boundary of the registered context mask i and $\lambda[p]$ is positive if p is in the query foreground and negative otherwise. As further detailed in Appendix B, Eq. (9) reveals that $\mathcal{L}_{\text{align}}$ encourages foreground query features to be similar to the foreground context features and dissimilar to the background ones, with the opposite effect on background query features. Therefore, the loss induces a dense contrastive effect on query feature maps, pushing the features of the foreground and the background apart. This effect intuitively supports the segmentation objective, consistent with our assumption that registration supervision benefits segmentation.

4 Experiments

4.1 Experimental settings

Datasets There is no widely adopted, publicly available train-test split for evaluating foundation ICS methods. To create our 2D in-context dataset, we combined several open-source, task-specific segmentation datasets. Combining images from native 2D sources with slices extracted from 3D volumes, we assembled roughly 1.8 million image-mask pairs spanning 8 imaging modalities (MRI, CT, PET, ultrasound, X-ray, microscopy, fundus imaging, and OCT) and a wide array of targets (e.g., organs, tumors, and blood vessels). Detailed construction steps are described in Appendix C to ensure reproducibility of our results. Our dataset is organized into 80 subsets mirroring the original data sources and modalities. Following Butoi* et al. (2023); Wong et al. (2025), we reserved a selection of these subsets as held-out validation and evaluation splits to assess generalization performance. Notably, the held-out evaluation set contains both unseen target classes and an unseen modality, namely PET. For information about the degree of generalization of the datasets, see Table 8 of Appendix C.

Setup Following Wong et al. (2025), the final version of OmniSeg is trained with a synthetic task augmentation scheme described in Appendix D. During training, we uniformly sample a task $t = (\text{modality}, \text{axis}, \text{target})$, then the query and the context set are sampled across all the related subsets to favor the model's generalization capability. Following previous works, we perform model selection and hyperparameter tuning using seen and unseen validation subsets, and we evaluate OmniSeg on subsets left completely held out during training, accounting for a total of more than 200,000 image samples. We set the temperature to $T = 0.1$ (see Section 4.3) and the local neighborhood $\mathcal{N}$ to a discrete disk with a diameter of 5 pixels (see Appendix E.5). Our final model is trained with the Adam optimizer and a constant learning rate of $\eta = 10^{-4}$ and a batch size of 32 for 5 epochs, on 4 H100 GPUs for $\sim$48 hours.

Baselines We compare our model against pretrained medical ICS baselines, including UniverSeg (Butoi* et al., 2023), Tyche-TS (Rakic et al., 2024), and MultiverSeg (Wong et al., 2025). These U-shaped convolutional networks leverage CrossBlock layers to fuse query and context data via feature concatenation and averaging. In addition, we compare our method against generic, transformer-based ICS models, specifically SegGPT (Wang et al., 2023) along with INSID3 (Cuttano et al., 2026), built upon the natural-image foundation model DINOv3 (Siméoni et al., 2025).

Table 1: **Dice and HD95 scores on held-out evaluation subsets.** All models are evaluated with a context set size $S = 16$. The best performance is reported in **bold**, the second best is underlined. $*$ indicates that metrics have been computed on 10,000 samples.

Model	IBD (MRI)		ACDC (MRI)		TS Muscles (CT)		AMOS (CT)		autoPET (PET)	
	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓
SegGPT	35.3 ± 0.6	31.1 ± 0.7	51.8 ± 0.8	25.6 ± 0.8	31.1 ± 0.7*	48.0 ± 0.8*	36.2 ± 0.3	41.1 ± 0.4	2.4 ± 0.1	83.8 ± 0.2
INSID3	3.2 ± 0.1	90.0 ± 0.3	2.9 ± 0.2	84.7 ± 0.5	1.9 ± 0.1	93.4 ± 0.1	2.4 ± 0.1	89.2 ± 0.2	0.0 ± 0.0	99.3 ± 0.0
UniverSeg	53.8 ± 0.5	5.8 ± 0.2	71.6 ± 0.6	3.7 ± 0.1	59.9 ± 0.2	9.9 ± 0.1	58.3 ± 0.2	5.8 ± 0.1	9.2 ± 0.1	23.0 ± 0.1
Tyche-TS	47.3 ± 0.6	11.9 ± 0.5	68.7 ± 0.7	4.1 ± 0.2	54.1 ± 0.2	15.5 ± 0.1	50.4 ± 0.3	15.8 ± 0.2	19.1 ± 0.2	20.8 ± 0.2
MultiverSeg	58.6 ± 0.6	4.9 ± 0.1	76.8 ± 0.5	2.9 ± 0.1	71.1 ± 0.2	6.6 ± 0.1	66.6 ± 0.2	3.9 ± 0.0	17.5 ± 0.2	**17.7** ± 0.1
Ours	**63.1** ± 0.6	**4.3** ± 0.1	**83.1** ± 0.4	**1.9** ± 0.1	**83.0** ± 0.1	**3.1** ± 0.0	**79.3** ± 0.2	**2.4** ± 0.0	**28.9** ± 0.3	17.8 ± 0.1

Model	BUID (US)		HipXRay (XR)		PanDental (XR)		SCR (XR)		WBC (Micro.)		All Tasks Avg.	
	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓
SegGPT	34.5 ± 3.8	53.5 ± 4.5	94.3 ± 0.6	1.5 ± 0.4	93.4 ± 0.5	0.7 ± 0.1	81.6 ± 0.8	5.4 ± 0.6	84.3 ± 1.9	6.5 ± 1.4	37.1 ± 0.3	40.0 ± 0.4
INSID3	13.9 ± 1.5	60.9 ± 2.3	38.9 ± 1.2	33.6 ± 0.7	46.6 ± 1.5	17.9 ± 0.5	17.1 ± 0.3	55.3 ± 0.6	25.1 ± 0.9	26.1 ± 0.8	3.9 ± 0.0	87.8 ± 0.1
UniverSeg	58.7 ± 2.5	16.8 ± 1.1	91.6 ± 0.8	1.6 ± 0.3	93.2 ± 0.4	0.8 ± 0.1	84.4 ± 0.6	1.7 ± 0.1	87.0 ± 0.9	2.9 ± 0.3	59.7 ± 0.1	7.6 ± 0.1
Tyche-TS	60.7 ± 2.9	13.0 ± 1.3	93.1 ± 0.7	1.4 ± 0.3	93.4 ± 0.4	0.6 ± 0.1	84.3 ± 0.8	1.5 ± 0.1	89.0 ± 1.0	2.4 ± 0.3	54.1 ± 0.2	13.6 ± 0.1
MultiverSeg	64.8 ± 2.6	11.2 ± 1.0	96.4 ± 0.3	0.4 ± 0.2	94.4 ± 0.3	0.5 ± 0.1	89.3 ± 0.4	0.9 ± 0.1	**90.5** ± 1.0	2.2 ± 0.4	68.1 ± 0.1	5.3 ± 0.0
Ours	**67.4** ± 2.7	**10.9** ± 1.0	**97.0** ± 0.2	**0.4** ± 0.2	**94.9** ± 0.4	**0.5** ± 0.1	**91.8** ± 0.3	**0.7** ± 0.1	90.3 ± 0.8	**1.9** ± 0.3	**77.9** ± 0.1	**3.3** ± 0.0

Metrics We evaluate segmentation performance using the Dice score and the 95th percentile Hausdorff Distance (HD95), with the latter normalized by the image diagonal. Additionally, we employ these metrics to quantify the average alignment between the query mask and the context masks warped by the DVFs estimated by OmniSeg. For 3D imaging modalities, evaluation is performed on slices extracted along each of the three canonical axes as explained in Appendix C. To account for performance variability induced by the context set selection, we report the average performance of each sample across five independent, randomly sampled context sets, following Butoi* et al. (2023); Wong et al. (2025); Rakic et al. (2024). We compute the overall average by aggregating scores at the task level. Finally, we report 99% confidence intervals (CIs) for each metric across the held-out subsets computed as explained in Appendix E.1.

4.2 State-of-the-art comparison

Task generalization Results presented in Table 1 demonstrate that OmniSeg outperforms all baselines by a large margin. Specifically, it achieves a 9.8 Dice-point and 2.0 HD95-point average improvement over MultiverSeg, the second-best method on these held-out subsets. This gap widens to more than 18 Dice points when compared to UniverSeg and Tyche-TS. SegGPT and INSID3 are generic ICS models whose pre-training datasets do not focus on medical images. They therefore fail to effectively generalize to medical ICS tasks outside their seen modalities, despite the inclusion of limited medical data during training.

OmniSeg yields a performance gain of over 11 Dice points on the CT subsets, AMOS and TS Muscles, CT being the modality most frequently seen during training. Notably, this improvement is achieved despite AMOS being included in the training set of MultiverSeg. Furthermore, autoPET and BUID are tumor segmentation benchmarks that inherently exhibit significant spatial variance across samples due to heterogeneous tumor locations. OmniSeg performs well on these challenging subsets, achieving substantial performance gains despite the PET modality being entirely unseen during training. As detailed in Table 11 of Appendix E.3, the advantage of OmniSeg over MultiverSeg is consistent across tasks. This experimentally validates that our contributions allow better out-of-distribution generalization. Finally, OmniSeg performs on par with the baselines on the natively 2D HipXRay, PanDental, and WBC subsets, where the ICS task is less complex due to stronger spatial consistency and high visual contrast between the target and the background. We show qualitative segmentation results in Appendix E.9.

An evaluation of the computational cost of our approach compared to the baselines can be found in Appendix E.8. OmniSeg induces a moderate overhead compared to convolutional ICS methods such as MultiverSeg.

4.3 Ablations and model analysis

In this section, we assess the relevance of our design choices through ablation experiments on our architecture and supervision scheme as well as hyperparameters. We present the most salient

Table 2: **Ablation studies.** All models are trained on our dataset for 5 epochs. The Alignment Dice score for MultiverSeg corresponds to the average original alignment between query and context masks for reference.

Model configuration	Dice ↑	Align. Dice ↑
MultiverSeg (retrained)	72.0 ± 0.1	21.8 ± 0.1
SRM	73.3 ± 0.1	41.2 ± 0.1
SRM+DCP	77.1 ± 0.1	43.2 ± 0.1
SRM+DCP+*Deep Reg. Loss*	$\mathbf{77.9} \pm 0.1$	$\mathbf{50.7} \pm 0.1$

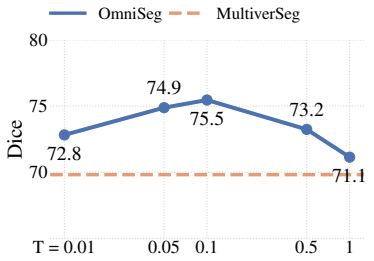

Figure 4: **Impact of temperature.** Task average segmentation performance w.r.t. the temperature at 2 epochs.

results of this study in the following paragraphs and report further analysis and per-dataset results in Appendix E.

Robustness to misalignment We assess the segmentation performance of our model for different levels of context-query misalignment by collecting the Dice score and Alignment Dice score for each query of our test datasets with one randomly selected context set. We visualize the binned results in Fig. 1*(b)*. OmniSeg's performance is more stable than MultiverSeg with respect to the context-query alignment, showing our model's robustness to misalignment.

Architecture and supervision design We study the effect of our core architecture and supervision design choices, in particular SpotBlock's DCP and our Deep Registration Loss (Deep Reg. Loss), against a retrained MultiverSeg baseline with our training setup. The results are presented in Table 2. Other relevant variants are studied in Appendix E.2 and confirm the relevance of our final model. Compared to MultiverSeg, using our two-branch convolutional encoder with SpotBlock's SRM in the decoder alone yields higher segmentation accuracy. Even without registration supervision, the DVFs output by SRM align context masks to query masks, improving the Alignment Dice score by 19.4 points over the initial alignment. Adding DCP increases the Dice score by a further 3.8 points, highlighting the complementarity of SRM and DCP within SpotBlocks. Finally, deep registration supervision provides additional segmentation gains while improving the Alignment Dice score by 7.5 points.

Impact of temperature The temperature is an important parameter of SpotBlocks, as it directly affects their dense contrastive behavior through the $\mathrm{Softmax}$ activations. We evaluate five values (0.01, 0.05, 0.1, 0.15, and 1) with models trained for 2 epochs and report the segmentation results on the test datasets in Fig. 4. Results on our held-out validation splits motivated our choice of $T = 0.1$. Test performance stays stable in a large range around this value. Nevertheless, detailed results in Appendix E.4 show that it is suboptimal for some of the test subsets.

Impact of the context size We report in Fig. 5*(a)* the Dice score with 99% CI on test datasets with a varying context size from 1 to 64 for MultiverSeg and OmniSeg, both trained with our training setup for 5 epochs. To limit the compute budget, the evaluation is capped at 1000 query samples per dataset. OmniSeg performs better for all context sizes, including small ones. With only 8 context samples, OmniSeg matches the segmentation accuracy of MultiverSeg with 64 context samples.

ICS vs. supervised methods We simulate the realistic scenario in which only a handful of annotated medical images are available. To assess few-shot performance, we split each task subset of the test datasets into 80%/20% train/test splits, train nnU-Net on nested subsets of sizes $S \in \{16, 32, 64, 128\}$, and use the same subsets as context sets for MultiverSeg and OmniSeg. Fig. 5*(b)* reports mean Dice with 99% CI over three test datasets, alongside a supervised upper-bound nnU-Net trained on the 80% train split. Without any task-specific retraining, OmniSeg outperforms nnU-Net in the few-shot regime on these subsets. On AMOS, OmniSeg at $S = 16$ already matches nnU-Net at $S = 128$ (79.1 ± 0.5 vs. 80.1 ± 0.6 Dice). This highlights OmniSeg's practical value in data-scarce clinical and research settings. For details on the experimental settings and further discussion, see Appendix E.7.

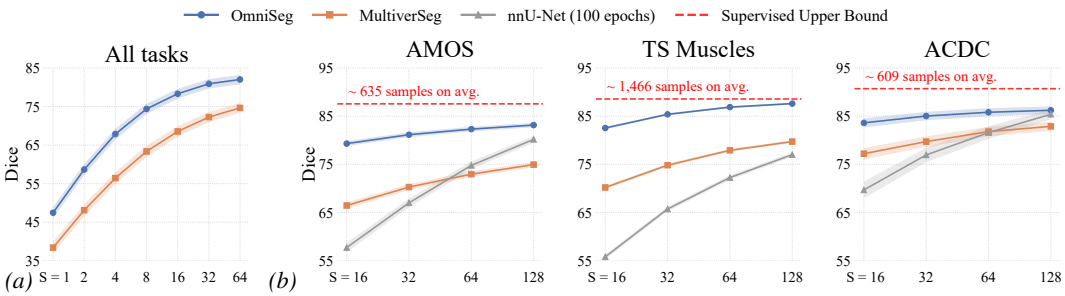

Figure 5: **Context size analysis.** *(a)* Average Dice w.r.t. the context set size. *(b)* Segmentation performance of nnU-Net vs. OmniSeg in a few-shot regime. The average training set size of the supervised upper bound across tasks is indicated on the figure. Out of the box, OmniSeg outperforms in-distribution supervised methods in the low-annotation regime.

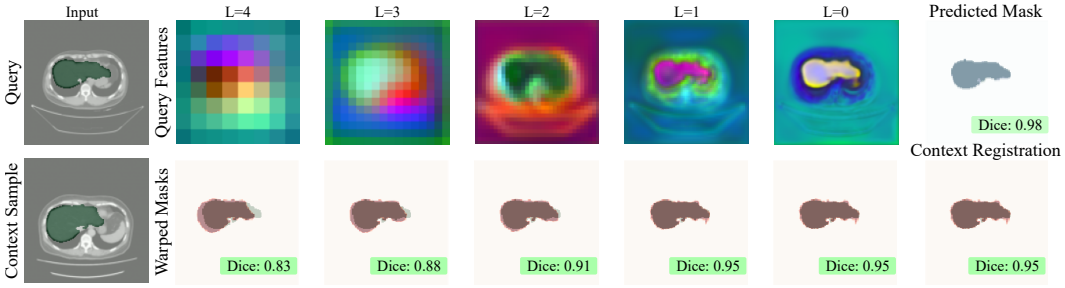

Figure 6: **Intermediate outputs of SpotBlocks across the decoder** ($S = 16$). The first row shows an RGB rendering of the query feature maps projected onto their first three principal components. The second row shows the context mask gradually registered onto the query mask by the DVFs computed by the SpotBlocks.

OmniSeg's internal representations Fig. 6 shows the intermediate outputs of the SpotBlocks across the decoder. We can see in the PCA plots of the query and context representations that the foreground, the background, and their boundaries are well separated in the feature space, even in the deeper blocks of the decoder. This illustrates the dense contrastive behavior of the SpotBlocks across the decoder. The bottom row shows the context mask gradually registered onto the query mask by the DVFs computed by the SpotBlocks from the local attention, illustrating the complementarity of the segmentation and registration objectives. Visualizations of dense attention maps can be found in Appendix E.6.

5 Conclusion

In this work, we introduced OmniSeg, a new ICS model that substantially outperforms current methods in task generalization. This improved performance is driven by two main contributions: first, the SpotBlock module, which performs a *soft* registration of the context onto the query via dense local attention followed by dense context pooling, second, a deep registration supervision scheme that supports segmentation by inducing a dense contrastive effect on the feature maps. Additionally, OmniSeg scales effectively with context size and shows robustness to query/context spatial misalignment. We believe that OmniSeg can serve as a foundation for medical image segmentation, enabling high-quality results with minimal annotation effort. While OmniSeg is highly versatile, its registration-aware design may perform suboptimally on anatomical structures with significant inter-subject topological variation, such as blood vessels. Finally, recent works (Wu & Xu, 2024; Wong et al., 2025; Guo et al., 2026) designed frameworks combining multiple universal segmentation paradigms: in-context, interactive or prototype-guided. Similarly, extending OmniSeg beyond ICS to broader segmentation paradigms and applications is a promising direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing the mathematical analysis, partially implement methods, provide feedback on research methodology via mock reviews, and generate synthetic data (see Appendix D). Additionally, we used these tools to format quantitative results into figures and tables, review the paper for grammatical or orthographical errors, improve overall readability, and identify existing literature. We also used LLMs to generate figure-rendering scripts and LaTeX boilerplate. In both cases, manual review and adjustment steps were performed. Generative AI tools were not used for formulating mathematical claims, proposing hypotheses, thematic data analysis, dataset cleaning, translation, or interpreting results. We have reviewed all AI-assisted work and ensured that it did not alter any quantitative result or change any finding or interpretation. Any code related to the core of our method was thoroughly reviewed manually and furthermore verified via unit testing and qualitative toy examples. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with AI assistance.

ETHICS STATEMENT

All medical images and annotations in this study were obtained from publicly available, anonymized datasets. We made no attempt to re-identify patients or recover any personal data, and no sensitive health information was accessed or used.

REPRODUCIBILITY STATEMENT

Thorough architectural descriptions of OmniSeg and SpotBlocks are provided in Appendices A.1 and A.2, respectively. Source code will be made publicly available upon publication. The composition of our dataset and its preparation, including preprocessing and splitting strategies, are presented in detail in Appendix C. Furthermore, our training pipeline and its hyperparameters are described in Appendix D.

REFERENCES

AA Ahmed, MM Elmohr, D Fuentes, MA Habra, SB Fisher, ND Perrier, M Zhang, and KM Elsayes. Radiomic mapping model for prediction of ki-67 expression in adrenocortical carcinoma. *Clinical Radiology*, 75(6):479–e17, 2020.

Zeeshan Ahmed, Munawar Ahmed, Attiya Baqai, and Fahim Aziz Umrani. Intraretinal cystoid fluid, 2022. URL https://www.kaggle.com/ds/2277068.

Tugba Akinci D'Antonoli, Lucas K. Berger, Ashraya K. Indrakanti, Nathan Vishwanathan, Jakob Weiss, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, Alexandra Walter, Elmar M. Merkle, Daniel T. Boll, Hanns-Christian Breit, Andrew Phillip Nicoli, Martin Segeroth, Joshy Cyriac, Shan Yang, and Jakob Wasserthal. Totalsegmentator mri: Robust sequence-independent segmentation of multiple anatomic structures in mri. *Radiology*, 314(2):e241613, 2025. doi: 10.1148/radiol.241613. URL https://doi.org/10.1148/radiol.241613. PMID: 39964271.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Jalen Anthony Benson, Weiruo Zhang, Ann NC Leung, et al. A radiogenomic dataset of non-small cell lung cancer. *Scientific Data*, 5(1):1–9, 2018.

Pedro RAS Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Yannick Kirchhoff, Maximilian Rokuss, Ziyan Huang, Jin Ye, Junjun He, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus H. Maier-Hein, Paul Jaeger, Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, Yong Xia, Zhaohu Xing, Lei Zhu, Yousef Sadegheih, Afshin Bozorgpour, Pratibha Kumari, Reza Azad, Dorit Merhof, Pengcheng Shi, Ting Ma, Yuxin Du, Fan Bai, Tiejun Huang, Bo Zhao, Haonan Wang, Xiaomeng Li, Hanxue Gu, Haoyu Dong, Jichen Yang, Maciej A. Mazurowski, Saumya Gupta, Linshan Wu, Jiaxin Zhuang, Hao Chen, Holger Roth, Daguang Xu, Matthew B. Blaschko, Sergio Decherchi, Andrea Cavalli, Alan L. Yuille, and Zongwei Zhou. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? *Neural Information Processing Systems*, 2024.

Bahareh Behboodi, Francois-Xavier Carton, Matthieu Chabanas, Sandrine de Ribaupierre, Ole Solheim, Bodil KR Munkvold, Hassan Rivaz, Yiming Xiao, and Ingerid Reinertsen. Open access segmentations of intraoperative brain tumor ultrasound images. *Medical Physics*, 51(9):6525–6532, 2024.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

Harsharan Singh Bhatia, Andreas-David Brunner, Furkan Öztürk, Saketh Kapoor, Zhouyi Rong, Hongcheng Mai, Marvin Thielert, Mayar Ali, Rami Al-Maskari, Johannes Christian Paetzold, et al. Spatial proteomics in three-dimensional intact specimens. *Cell*, 185(26):5040–5058, 2022.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, Fabian Lohöfer, Julian Walter Holch, Wieland Sommer, Felix Hofmann, Alexandre Hostettler, Naama Lev-Cohain, Michal Drozdzal, Michal Marianne Amitai, Refael Vivanti, Jacob Sosna, Ivan Ezhov, Anjany Sekuboyina, Fernando Navarro, Florian Kofler, Johannes C. Paetzold, Suprosanna Shit, Xiaobin Hu, Jana Lipková, Markus Rempfler, Marie Piraud, Jan Kirschke, Benedikt Wiestler, Zhiheng Zhang, Christian Hülsemeyer, Marcel Beetz, Florian Ettlinger, Michela Antonelli, Woong Bae, Míriam Bellver, Lei Bi, Hao Chen, Grzegorz Chlebus, Erik B. Dam, Qi Dou, Chi-Wing Fu, Bogdan Georgescu, Xavier Giró i Nieto, Felix Gruen, Xu Han, Pheng-Ann Heng, Jürgen Hesser, Jan Hendrik Moltz, Christian Igel, Fabian Isensee, Paul Jäger, Fucang Jia, Krishna Chaitanya Kaluva, Mahendra Khened, Ildoo Kim, Jae-Hun Kim, Sungwoong Kim, Simon Kohl, Tomasz Konopczynski, Avinash Kori, Ganapathy Krishnamurthi, Fan Li, Hongchao Li, Junbo Li, Xiaomeng Li, John Lowengrub, Jun Ma, Klaus Maier-Hein, Kevis-Kokitsi Maninis, Hans Meine, Dorit Merhof, Akshay Pai, Mathias Perslev, Jens Petersen, Jordi Pont-Tuset, Jin Qi, Xiaojuan Qi, Oliver Rippel, Karsten Roth, Ignacio Sarasua, Andrea Schenk, Zengming Shen, Jordi Torres, Christian Wachinger, Chunliang Wang, Leon Weninger, Jianrong Wu, Daguang Xu, Xiaoping Yang, Simon Chun-Ho Yu, Yading Yuan, Miao Yue, Liping Zhang, Jorge Cardoso, Spyridon Bakas, Rickmer Braren, Volker Heinemann, Christopher Pal, An Tang, Samuel Kadoury, Luc Soler, Bram van Ginneken, Hayit Greenspan, Leo Joskowicz, and Bjoern Menze. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023.

Nathan Bloch, Anant Madabhushi, Henkjan Huisman, John Freymann, Justin Kirby, Michael Grauer, Andinet Enquobahrie, Carl Jaffe, Laurence Clarke, and Keyvan Farahani. Nci-isbi 2013 challenge: Automated segmentation of prostate structures. *The Cancer Imaging Archive*, 2015.

Stephen R Bowen, William TC Yuh, Daniel S Hippe, Wei Wu, Savannah C Partridge, Saba Elias, Guang Jia, Zhibin Huang, George A Sandison, Dennis Nelson, et al. Tumor radiomic heterogeneity: multiparametric functional imaging to characterize variability and predict response following cervical cancer radiation therapy. *Journal of Magnetic Resonance Imaging*, 47(5):1388–1396, 2018.

W Brahmi, I Jdey, and F Drira. Panoramic dental xray dataset. *Kaggle*, 2025.

Walid Brahmi and Imen Jdey. Automatic tooth instance segmentation and identification from panoramic x-ray images using deep cnn. *Multimedia Tools and Applications*, 83(18):55565–55585, 2024.

Walid Brahmi, Imen Jdey, and Fadoua Drira. Exploring the role of convolutional neural networks (cnn) in dental radiography segmentation: A comprehensive systematic literature review. *Engineering Applications of Artificial Intelligence*, 133:108510, 2024.

Victor Ion Butoi*, Jose Javier Gonzalez Ortiz*, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Universeg: Universal medical image segmentation. *International Conference on Computer Vision*, 2023.

Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2):577–590, 2013.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the European Conference on Computer Vision Workshops(ECCVW)*, 2022.

Adrian Carballal, Francisco J Novoa, Carlos Fernandez-Lozano, Marcos García-Guimaraes, Guillermo Aldama-López, Ramón Calviño-Santos, José Manuel Vazquez-Rodriguez, and Alejandro Pazos. Automatic multiscale vascular image segmentation algorithm for coronary angiography. *Biomedical Signal Processing and Control*, 46:1–9, 2018.

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, 2024. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2024.103280. URL https://www.sciencedirect.com/science/article/pii/S1361841524002056.

Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8: 132665–132676, 2020.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6): 1045–1057, 2013.

CREMI. Miccai challenge on circuit reconstruction from electron microscopy images, 2016. URL https://cremi.org/. Accessed: 2020-10-27.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. INSID3: Training-free in-context segmentation with DINOv3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Steffen Czolbe and Adrian V Dalca. Neuralizer: General neuroimage analysis without re-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6217–6230, 2023.

Wangbin Ding, Lei Li, Xiahai Zhuang, and Liqin Huang. Cross-modality multi-atlas segmentation using deep neural networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12263 of *Lecture Notes in Computer Science*, pp. 243–253. Springer, 2020. doi: 10.1007/978-3-030-59716-0_23. arXiv:2008.08946.

Reuben Dorent, Aaron Kujawa, Marina Ivory, Spyridon Bakas, Nicola Rieke, Samuel Joutard, Ben Glocker, Jorge Cardoso, Marc Modat, Kayhan Batmanghelich, et al. Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. *Medical Image Analysis*, 83:102628, 2023.

Vanessa Gonzalez Duque, Alexandra Marquardt, Yordanka Velikova, Lilian Lacourpaille, Antoine Nordez, Marion Crouzier, Hong Joo Lee, Diana Mateus, and Nassir Navab. Ultrasound segmentation analysis via distinct and completed anatomical borders. *International Journal of Computer Assisted Radiology and Surgery*, 19(7):1419–1427, 2024.

Sofija Engelson, Jan Ehrhardt, Timo Kepp, Joshua Niemeijer, and Heinz Handels. Lnq challenge 2023: Learning mediastinal lymph node segmentation with a probabilistic lymph node atlas. *arXiv preprint arXiv:2406.03984*, 2024.

Andriy Fedorov, Michael Schwier, David Clunie, Christian Herz, Steve Pieper, Ron Kikinis, Clare Tempany, and Fiona Fennessy. An annotated test-retest collection of prostate multiparametric mri. *Scientific Data*, 5(1):1–13, 2018.

Yunhe Gao, Di Liu, Zhuowei Li, Yunsheng Li, Dongdong Chen, Mu Zhou, and Dimitris N. Metaxas. Show and segment: Universal medical image segmentation via in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20830–20840, June 2025.

Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601, 2022.

Sergios Gatidis, Marcel Früh, Matthias P. Fabritius, Sijing Gu, Konstantin Nikolaou, Christian La Fougère, Jin Ye, Junjun He, Yige Peng, Lei Bi, Jun Ma, Bo Wang, Jia Zhang, Yukun Huang, Lars Heiliger, Zdravko Marinov, Rainer Stiefelhagen, Jan Egger, Jens Kleesiek, Ludovic Sibille, Lei Xiang, Simone Bendazzoli, Mehdi Astaraki, Michael Ingrisch, Clemens C. Cyran, and Thomas Küstner. Results from the autopet challenge on fully automated lesion segmentation in oncologic pet/ct imaging. *Nature Machine Intelligence*, 6(11):1396–1405, 2024.

Bangwei Guo, Yunhe Gao, Meng Ye, Difei Gu, Yang Zhou, Leon Axel, and Dimitris N. Metaxas. K-prism: A knowledge-guided and prompt integrated universal medical image segmentation model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gvRf95K4im.

Daniel Gut. X-ray images of the hip joints. *Mendeley Data*, 1, 2021.

Yuting He and Shuo Li. Vector contrastive learning for pixel-wise pretraining in medical vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19827–19837, October 2025.

Yuting He, Rongjun Ge, Xiaoming Qi, Yang Chen, Jiasong Wu, Jean-Louis Coatrieux, Guanyu Yang, and Shuo Li. Learning better registration to learn better few-shot medical image segmentation: Authenticity, diversity, and robustness. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022. doi: 10.1109/TNNLS.2022.3190452.

Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, 67:101821, 2021.

Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific Data*, 9(1): 762, 2022.

Jiesi Hu, Hanyang Peng, Yanwu Yang, Xutao Guo, Yang Shang, Pengcheng Shi, Chenfei Ye, and Ting Ma. Neuroverse3d: Developing in-context learning universal model for neuroimaging in 3d. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21721–21731, 2025.

Fabian Isensee, Maximilian Rokuss, Lars Krämer, Stefan Dinkelacker, Ashis Ravindran, Florian Stritzke, Benjamin Hamm, Tassilo Wald, Moritz Langenberg, Constantin Ulrich, Jonathan Deissler, Ralf Floca, and Klaus Maier-Hein. nninteractive: Redefining 3d promptable segmentation, 2025. URL https://arxiv.org/abs/2503.08373.

Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging*, 33(2):233–245, 2013.

Yuanfeng Ji, Haotian Bai, Chongjian GE, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, and Ping Luo. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Doris Kaltenecker, Rami Al-Maskari, Moritz Negwer, Luciano Hoeher, Florian Kofler, Shan Zhao, Mihail Todorov, Zhouyi Rong, Johannes Christian Paetzold, Benedikt Wiestler, et al. Virtual reality-empowered deep-learning analysis of brain cells. *Nature Methods*, 21(7):1306–1315, 2024.

A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Kendall J Kiser, Arko Barman, Sonja Stieb, Clifton D Fuller, and Luca Giancardo. Novel autosegmentation spatial similarity metrics capture the time required to correct segmentations better than traditional metrics in a thoracic cavity segmentation workflow. *Journal of Digital Imaging*, 34: 541–553, 2021.

Markus Krönke, Christine Eilers, Desislava Dimova, Melanie Köhler, Gabriel Buschner, Lilit Schweiger, Lemonia Konstantinidou, Marcus Makowski, James Nagarajah, Nassir Navab, et al. Tracked 3d ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. *Plos one*, 17(7):e0268550, 2022.

Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE Transactions on Medical Imaging*, 38(11):2556–2568, 2019.

Zeki Kuş and Musa Aydin. Medsegbench: A comprehensive benchmark for medical image segmentation in diverse data modalities. *Scientific Data*, 11(1), November 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-04159-2. URL http://dx.doi.org/10.1038/s41597-024-04159-2.

Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, pp. 103285, 2024a.

Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised models transfer to 3d image segmentation? In *The Twelfth International Conference on Learning Representations*, 2024b.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Zudi Lin, Donglai Wei, Mariela D Petkova, Yuelong Wu, Zergham Ahmed, Silin Zou, Nils Wendt, Jonathan Boulanger-Weill, Xueying Wang, Nagaraju Dhanyasi, et al. Nucmm dataset: 3d neuronal nuclei instance segmentation at sub-cubic millimeter scale. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 164–174. Springer, 2021.

Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.

Xiangde Luo, Jia Fu, Yunxin Zhong, Shuolin Liu, Bing Han, Mehdi Astaraki, Simone Bendazzoli, Iuliana Toma-Dasu, Yiwen Ye, Ziyang Chen, Yong Xia, Yanzhou Su, Jin Ye, Junjun He, Zhaohu Xing, Hongqiu Wang, Lei Zhu, Kaixiang Yang, Xin Fang, Zhiwei Wang, Chan Woong Lee, Sang Joon Park, Jaehee Chun, Constantin Ulrich, Klaus H. Maier-Hein, Nchongmaje Ndipenoch, Alina Miron, Yongmin Li, Yimeng Zhang, Yu Chen, Lu Bai, Jinlong Huang, Chengyang An, Lisheng Wang, Kaiwen Huang, Yunqi Gu, Tao Zhou, Mu Zhou, Shichuan Zhang, Wenjun Liao, Guotai Wang, and Shaoting Zhang. Segrap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma. *Medical Image Analysis*, 101:103447, 2025.

Jun Ma, Yixin Wang, Xingle An, Cheng Ge, Ziqi Yu, Jianan Chen, Qiongjie Zhu, Guoqiang Dong, Jian He, Zhiqiang He, Tianjia Cao, Yuntao Zhu, Ziwei Nie, and Xiaoping Yang. Towards data-efficient learning: A benchmark for covid-19 ct lung and infection segmentation. *Medical Physics*, 48(3):1197–1210, 2021. doi: https://doi.org/10.1002/mp.14676.

Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, Shuiping Gou, Franz Thaler, Christian Payer, Darko Štern, Edward G.A. Henderson, Dónal M. McSweeney, Andrew Green, Price Jackson, Lachlan McIntosh, Quoc-Cuong Nguyen, Abdul Qayyum, Pierre-Henri Conze, Ziyan Huang, Ziqi Zhou, Deng-Ping Fan, Huan Xiong, Guoqiang Dong, Qiongjie Zhu, Jian He, and Xiaoping Yang. Fast and low-gpu-memory abdomen ct organ segmentation: The flare challenge. *Medical Image Analysis*, 82:102616, 2022a.

Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2022b.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, January 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z.

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos, 2025. URL https://arxiv.org/abs/2504.03600.

Jun Ma, Sumin Kim, Yuyin Zhou, and Bo Wang (eds.). *Foundation Models for 3D Biomedical Image Segmentation*, volume 1 of *Lecture Notes in Computer Science*, 2026. Springer Cham. ISBN 978-3-032-23495-7. doi: 10.1007/978-3-032-23496-4. eBook ISBN: 978-3-032-23496-4.

Francesco Marzola, Nens Van Alfen, Jonne Doorduin, and Kristen M Meiburger. Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment. *Computers in biology and medicine*, 135:104623, 2021.

Manca Žerovnik Mekuč, Ciril Bohak, Samo Hudoklin, Byeong Hak Kim, Rok Romih, Min Young Kim, and Matija Marolt. Automatic segmentation of mitochondria and endolysosomes in volumetric electron microscopy data. *Computers in Biology and Medicine*, 119:103693, 2020. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2020.103693. URL https://www.sciencedirect.com/science/article/pii/S0010482520300792.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10): 1993–2024, 2014.

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. Ieee, 2016.

Ahmed Morshid, Khaled M. Elsayes, Ahmed M. Khalaf, Mohamed M. Elmohr, Jian Yu, Ahmed O. Kaseb, Mohamed Hassan, Armeen Mahvash, Zeshan Wang, John D. Hazle, and David Fuentes. A machine learning model to predict hepatocellular carcinoma response to transcatheter arterial chemoembolization. *Radiology: Artificial Intelligence*, 1(5):e180021, 2019.

Andriy Myronenko, Dong Yang, Yufan He, and Daguang Xu. Aorta segmentation from 3d ct in miccai seg. a. 2023 challenge. In *MICCAI Challenge on Segmentation of the Aorta*, pp. 13–18. Springer, 2023.

Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervised learning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(7):1837–1848, 2022. doi: 10.1109/TMI.2022.3150682.

Danielle F Pace, Hannah TM Contreras, Jennifer Romanowicz, Shruti Ghelani, Imon Rahaman, Yue Zhang, Patricia Gao, Mohammad Imrul Jubair, Tom Yeh, Polina Golland, et al. Hvsmr-2.0: A 3d cardiovascular mr dataset for whole-heart segmentation in congenital heart disease. *Scientific Data*, 11(1):721, 2024.

Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-net transformer: Self and cross attention for medical image segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pp. 267–276. Springer, 2021.

Gašper Podobnik, Primož Strojan, Primož Peterlin, Bulat Ibragimov, and Tomaž Vrtovec. Han-seg: The head and neck organ-at-risk ct and mr segmentation dataset. *Medical physics*, 50(3): 1917–1927, 2023.

Yulei Qin, Mingjian Chen, Hao Zheng, Yun Gu, Mali Shen, Jie Yang, Xiaolin Huang, Yue-Min Zhu, and Guang-Zhong Yang. Airwaynet: a voxel-connectivity aware approach for accurate airway segmentation using convolutional neural networks. In *International conference on medical image computing and computer-assisted intervention*, pp. 212–220. Springer, 2019.

Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132:104319, 2021.

Marianne Rakic, Hallee E. Wong, Jose Javier Gonzalez Ortiz, Beth A. Cimini, John V. Guttag, and Adrian V. Dalca. Tyche: Stochastic in-context learning for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11159–11173, June 2024.

Maximilian Rokuss, Moritz Langenberg, Yannick Kirchhoff, Fabian Isensee, Benjamin Hamm, Constantin Ulrich, Sebastian Regnery, Lukas Bauer, Efthimios Katsigiannopulos, Tobias Norajitra, and Klaus Maier-Hein. Voxtell: Free-text promptable universal 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37538–37557, 2026.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Holger R Roth, Le Lu, Ari Seff, Kevin M Cherry, Joanne Hoffman, Shijun Wang, Jiamin Liu, Evrim Turkbey, and Ronald M Summers. A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part I 17*, pp. 520–527, 2014.

Holger R Roth, Ziyue Xu, Carlos Tor-Díez, Ramon Sanchez Jacob, Jonathan Zember, Jose Molto, Wenqi Li, Sheng Xu, Baris Turkbey, Evrim Turkbey, et al. Rapid artificial intelligence solutions in a pandemic - the covid-19-20 lung ct lesion segmentation challenge. *Medical Image Analysis*, 82: 102605, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2010.

Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004.

Karen-Helene Støverud, David Bouget, Andre Pedersen, Håkon Olav Leira, Tore Amundsen, Thomas Langø, and Erlend Fagertun Hofstad. Aeropath: An airway segmentation benchmark dataset with challenging pathology and baseline method. *Plos one*, 19(10):e0311416, 2024.

Sujun Sun, Mingwu Ren, and Haofeng Zhang. Training-free cross-domain few-shot segmentation via robust semantic representation and matching. *arXiv preprint arXiv:2606.24297*, 2026.

David Svoboda, Michal Kozubek, and Stanislav Stejskal. Generation of digital phantoms of cell nuclei and simulation of image formation in 3d image cytometry. *Cytometry Part A: The Journal of the International Society for Advancement of Cytometry*, 75(6):494–509, 2009.

Mihail Ivilinov Todorov, Johannes Christian Paetzold, Oliver Schoppe, Giles Tetteh, Suprosanna Shit, Velizar Efremov, Katalin Todorov-V¨olgyi, Marco D¨uring, Martin Dichgans, Marie Piraud, et al. Machine learning analysis of whole mouse brain vasculature. *Nature methods*, 17(4):442–449, 2020.

Jasper W van der Graaf, Miranda L van Hooff, Constantinus FM Buckens, Matthieu Rutten, Job LC van Susante, Robert Jan Kroeze, Marinus de Kleuver, Bram van Ginneken, and Nikolas Lessmann. Lumbar spine segmentation in mr images: a dataset and a public benchmark. *Scientific Data*, 11 (1):264, 2024.

Bram Van Ginneken, Mikkel B Stegmann, and Marco Loog. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical image analysis*, 10(1):19–40, 2006.

Santiago Vitale, José Ignacio Orlando, Emmanuel Iarussi, and Ignacio Larrabide. Improving realism in patient-specific abdominal ultrasound simulation using cyclegans. *International journal of computer assisted radiology and surgery*, 15(2):183–192, 2020.

Kareem Wahid, Cem Dede, Mohamed Naser, and Clifton Fuller. Training dataset for hntsmrg 2024 challenge, 2024. URL https://doi.org/10.5281/zenodo.11199559.

Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1130–1140. IEEE, 2023.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5 (5), 2023.

Donglai Wei, Kisuk Lee, Hanyu Li, Ran Lu, J Alexander Bae, Zequan Liu, Lifu Zhang, Márcia dos Santos, Zudi Lin, Thomas Uram, et al. Axonem dataset: 3d axon instance segmentation of brain cortical regions. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pp. 175–185. Springer, 2021.

Hallee E. Wong, Marianne Rakic, John Guttag, and Adrian V. Dalca. Scribbleprompt: Fast and flexible interactive segmentation for any biomedical image. *European Conference on Computer Vision (ECCV)*, 2024.

Hallee E. Wong, Jose Javier Gonzalez Ortiz, John Guttag, and Adrian V. Dalca. Multiverseg: Scalable interactive segmentation of biomedical imaging datasets with in-context guidance. *International Conference on Computer Vision*, 2025.

Junde Wu and Min Xu. One-Prompt to Segment All Medical Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11302–11312, 2024.

Jinzhong Yang, Harini Veeraraghavan, Samuel G Armato III, Keyvan Farahani, Justin S Kirby, Jayashree Kalpathy-Kramer, Wouter van Elmpt, Andre Dekker, Xiao Han, Xue Feng, et al. Autosegmentation for thoracic radiation treatment planning: a grand challenge at aapm 2017. *Medical physics*, 45(10):4568–4581, 2018.

Kaiyuan Yang, Fabio Musio, Yihui Ma, Norman Juchler, Johannes C Paetzold, Rami Al-Maskari, Luciano Höher, Hongwei Bran Li, Ibrahim Ethem Hamamci, Anjany Sekuboyina, et al. Benchmarking the cow with the topcow challenge: Topology-aware anatomical segmentation of the circle of willis for cta and mra. *ArXiv*, pp. arXiv–2312, 2025.

Weihao Yu, Hao Zheng, Minghui Zhang, Hanxiao Zhang, Jiayuan Sun, and Jie Yang. Break: Bronchi reconstruction by geodesic transformation and skeleton embedding. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2022.

Minghui Zhang, Yangqian Wu, Hanxiao Zhang, Yulei Qin, Hao Zheng, Wen Tang, Corey Arnold, Chenhao Pei, Pengxin Yu, Yang Nan, et al. Multi-site, multi-domain airway tree modeling. *Medical Image Analysis*, 90:102957, 2023.

Ziyu Zhang, Yi Yu, Simeng Zhu, Ahmed Aly, Yunhe Gao, Ning Gu, and Yuan Xue. Atlas is your perfect context: One-shot customization for generalizable foundational medical image segmentation. *arXiv preprint arXiv:2512.18176*, 2025.

Shan Zhao, Mihail Ivilinov Todorov, Ruiyao Cai, AI-Maskari Rami, Hanno Steinke, Elisabeth Kemter, Hongcheng Mai, Zhouyi Rong, Martin Warmer, Karen Stanic, et al. Cellular and molecular probing of intact human organs. *Cell*, 180(4):796–812, 2020.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Xiao Zhou, Ya Zhang, Yanfeng Wang, and Weidi Xie. Large-vocabulary segmentation for medical images with text prompts. *npj Digital Medicine*, 8(1):566, 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-01964-w. URL https://www.nature.com/articles/s41746-025-01964-w.

Hao Zheng, Yulei Qin, Yun Gu, Fangfang Xie, Jie Yang, Jiayuan Sun, and Guang-Zhong Yang. Alleviating class-wise gradient imbalance for pulmonary airway segmentation. *IEEE transactions on medical imaging*, 40(9):2452–2462, 2021.

Wei Zheng, Cheng Peng, Zeyuan Hou, Boyu Lyu, Mengfan Wang, Xuelong Mi, Shuoxuan Qiao, Yinan Wan, and Guoqiang Yu. Nis3d: A completely annotated benchmark for dense 3d nuclei image segmentation. *Advances in Neural Information Processing Systems*, 36:4741–4752, 2023.

Xin Zheng, Yong Wang, Guoyou Wang, and Jianguo Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71, 2018. doi: https://doi.org/10.1016/j.micron.2018.01.010. URL https://www.sciencedirect.com/science/article/pii/S0968432817303037.

Zhangnan Zhong, Li Huang, Shi-Ting Feng, Haiwei Lin, Xinyue Wang, Baolan Lu, Kangyang Cao, Xuehua Li, and Bingsheng Huang. A comprehensive dataset of magnetic resonance enterography images with intestinal segment annotations. *Scientific Data*, 12(1):425, 2025.

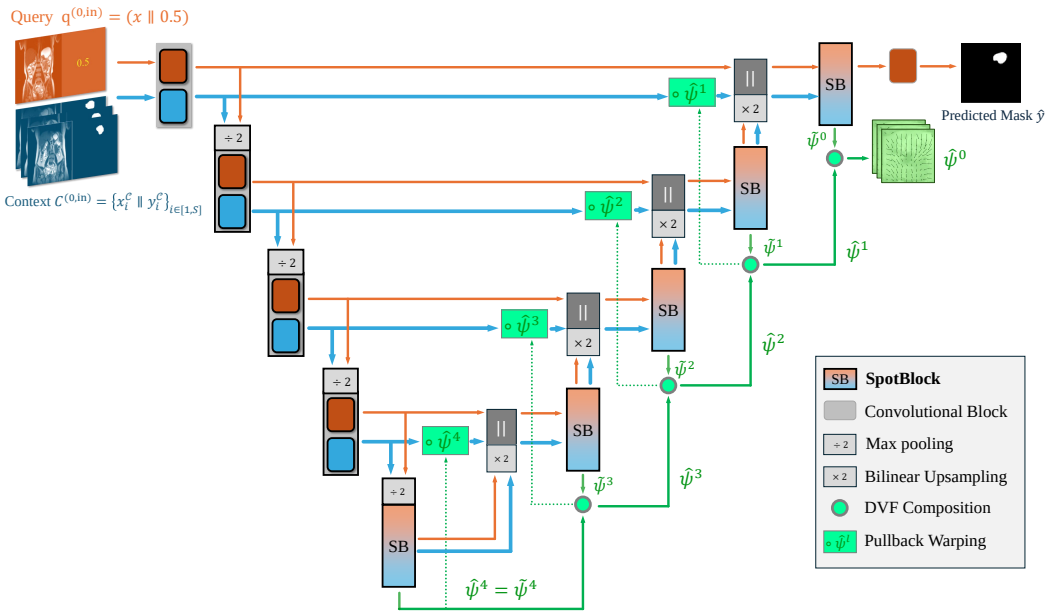

Figure 7: **OmniSeg architecture**. The two-branch convolutional encoder separately processes the query and context images. At the bottleneck and at each decoder level, a SpotBlock module fuses the query and context feature maps via SRM followed by DCP. The induced DVF is used to warp the context skip features before they are concatenated with the decoder output.

A Model

To ensure complete reproducibility, this section details the extended architectural design of OmniSeg (Appendix A.1) and SpotBlock (Appendix A.2).

A.1 Detailed architecture

The overall architecture of OmniSeg is shown in Fig. 7. OmniSeg takes as input the query image concatenated with a dummy channel and the concatenated context image and mask pairs $C^{(0,\text{in})} = \{x_i^{\mathcal{C}} \parallel y_i^{\mathcal{C}}\}_{i\in[1,S]}$ along the channel dimension. Let L be the number of levels of our U-shaped model, $L = 4$ for our final model. The dummy channel appended to the query input, $q^{(0,\text{in})} = (x \parallel 0.5)$, matches the number of input channels of the context branch, whose second channel carries the context mask. Each encoder block consists of one convolution block per branch, separately computing the query and context feature maps. For $l \in \{0, \dots, L-1\}$:

$$q^{(l,\text{enc})} = \text{ConvBlock}(q^{(l,\text{in})}) \tag{10}$$

$$c_i^{(l,\text{enc})} = \text{ConvBlock}(c_i^{(l,\text{in})}) \quad \text{for } i = 1, 2, \dots, S \tag{11}$$

$$(q^{(l+1,\text{in})}, C^{(l+1,\text{in})}) = (\text{Down}(q^{(l,\text{enc})}),\ \text{Down}(C^{(l,\text{enc})})) \tag{12}$$

At the bottleneck $l = L$, a SpotBlock module applies a first soft registration step with a large field of view:

$$(q^{(L,\text{dec})}, C^{(L,\text{dec})}, \tilde{\psi}^L_{q\to c_i}) = \text{SpotBlock}(q^{(L,\text{in})}, C^{(L,\text{in})}) \tag{13}$$

The decoder SpotBlocks then fuse the context and query features. For $l \in \{L-1, \dots, 0\}$:

$$q^{(l,\text{skip})} = q^{(l,\text{enc})} \parallel \text{Up}(q^{(l+1,\text{dec})}) \tag{14}$$

$$c_i^{(l,\text{skip})} = c_i^{(l,\text{enc})} \circ \hat{\psi}^{l+1}_{y\to y_i^{\mathcal{C}}} \parallel \text{Up}(c_i^{(l+1,\text{dec})}) \tag{15}$$

$$(q^{(l,\text{dec})}, C^{(l,\text{dec})}, \tilde{\psi}^l_{q\to c_i}) = \text{SpotBlock}(q^{(l,\text{skip})}, C^{(l,\text{skip})}) \tag{16}$$

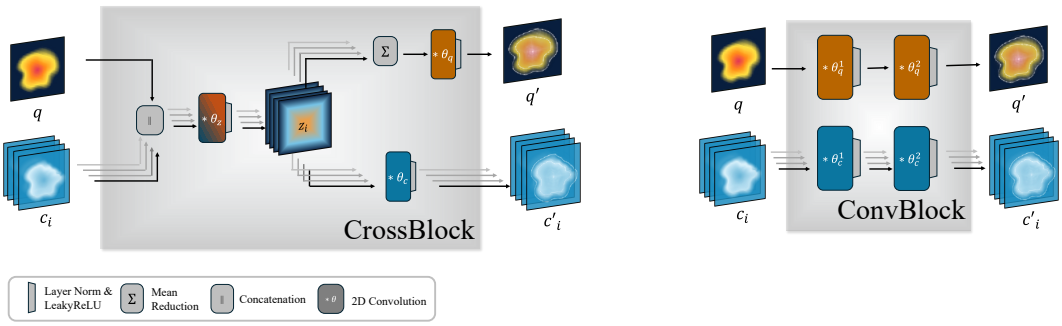

Figure 8: **Convolutional block and CrossBlock architectures.** The CrossBlock module was introduced by Butoi* et al. (2023). We employ an encoder that applies separate convolutions to the query and the context.

where $f \circ \psi$ denotes applying the DVF ψ to a 2D field f by pullback transformation. Down and Up are max-pooling and bilinear upsampling, respectively. Skip connections are performed via concatenation. Furthermore, the skip context features are registered using the DVF computed by the previous SpotBlock, as specified in Eq. (15). This ensures that skip features and the output of the previous decoder block have been spatially deformed in a consistent way. The segmentation logits are finally obtained by applying a single 1×1 convolution layer:

$$\hat{y} = \mathrm{Conv}(q^{(0,\mathrm{dec})}, \theta_{\hat{y}}) \tag{17}$$

As each SpotBlock of the decoder only outputs the incremental registration DVFs $\tilde{\psi}^l_{q \to c_i}$, we iteratively compose these DVFs to obtain the total displacement $\hat{\psi}^l_{y \to y_i^c}$ registering the context i on the query. At the bottleneck, $\hat{\psi}^L_{y \to y_i^c} = \tilde{\psi}^L_{q \to c_i}$, and for every depth level $l \in \{L-1, \ldots, 0\}$, we compose the residual field $\tilde{\psi}^l_{q \to c_i}$ with the upsampled total field from the level below, denoting this composition operator by $\oplus$:

$$\hat{\psi}^l_{y \to y_i^c} = \tilde{\psi}^l_{q \to c_i} \oplus \hat{\psi}^{l+1}_{y \to y_i^c} = \tilde{\psi}^l_{q \to c_i} + \overbrace{\hat{\psi}^{(l+1,\mathrm{up})}_{y \to y_i^c} \circ \tilde{\psi}^l_{q \to c_i}}^{\text{Vector fusion}}. \tag{18}$$

Scale alignment

Space alignment

At the final decoder stage, we obtain the predicted global, full-resolution registration DVF $\hat{\psi}^0_{y \to y_i^c}$ as the aggregation of all residual DVFs. Note that all predicted DVFs represent pullback registration flows: although the vectors originate at query coordinates and terminate at corresponding context locations, they dictate the inverse grid sampling required to register context feature maps to the query space (see Fig. 9).

A.2 SpotBlock in depth

Soft Registration Mechanism (SRM) As explained in Section 3.1, the dense local attention between the query and the context pixels is computed based on three tensors $Q \in \mathbb{R}^{H \times W \times F}$, $\mathbf{K} \in \mathbb{R}^{S \times N \times H \times W \times F}$ and $\mathbf{V} \in \mathbb{R}^{S \times N \times H \times W \times F_\nu}$. We detail how these tensors are obtained below. First, convolution layers are applied separately to the input query features q and the context features C:

$$Q = \mathrm{LN}(\sigma(\mathrm{Conv}(q, \theta^Q))) \tag{19a}$$

$$K_i = \mathrm{LN}(\sigma(\mathrm{Conv}(c_i, \theta^K))) \text{ for } i = 1, 2, \ldots, S \tag{19b}$$

$$V_i = \mathrm{LN}(\sigma(\mathrm{Conv}(c_i, \theta^V))) \text{ for } i = 1, 2, \ldots, S \tag{19c}$$

where LN denotes layer normalization and σ is the LeakyReLU activation function. Note that the same convolution kernels θ^K and θ^V are applied to all context samples. Next, K_i and V_i are repeated

Table 3: Architectural details of ConvBlock, CrossBlock, and SpotBlock across the different levels l of the U-shaped architecture. All convolutions have a stride of 1 and are padded to preserve the dimensions of the inputs. θ^Q, θ^K, and θ^V have no bias.

	Module		
Level	**ConvBlock**	**CrossBlock**	**SpotBlock**
$l = 0$	$\theta^{(q,1)}, \theta^{(c,1)}$: Conv(out $= 256$, ker $= 3\times3$) $\theta^{(q,2)}, \theta^{(c,2)}$: Conv(out $= 256$, ker $= 3\times3$)	θ^z : Conv(out $= 256$, ker $= 3\times3$) θ^q, θ^c : Conv(out $= 256$, ker $= 3\times3$)	θ^Q, θ^K : Conv(out $= 16$, ker $= 3\times3$) θ^V : Conv(out $= 64$, ker $= 3\times3$) θ^Z : Conv(out $= 64$, ker $= 3\times3$) θ^q, θ^c : Conv(out $= 64$, ker $= 3\times3$)
N_{params}	1.2M	1.2M	0.6M
$l = 1$	$\theta^{(q,1)}, \theta^{(c,1)}$: Conv(out $= 256$, ker $= 3\times3$) $\theta^{(q,2)}, \theta^{(c,2)}$: Conv(out $= 256$, ker $= 3\times3$)	θ^z : Conv(out $= 256$, ker $= 3\times3$) θ^q, θ^c : Conv(out $= 256$, ker $= 3\times3$)	θ^Q, θ^K : Conv(out $= 32$, ker $= 3\times3$) θ^V : Conv(out $= 128$, ker $= 3\times3$) θ^Z : Conv(out $= 128$, ker $= 3\times3$) θ^q, θ^c : Conv(out $= 128$, ker $= 3\times3$)
N_{params}	2.4M	2.4M	1.9M
$l = 2, 3$	$\theta^{(q,1)}, \theta^{(c,1)}$: Conv(out $= 256$, ker $= 3\times3$) $\theta^{(q,2)}, \theta^{(c,2)}$: Conv(out $= 256$, ker $= 3\times3$)	θ^z : Conv(out $= 256$, ker $= 3\times3$) θ^q, θ^c : Conv(out $= 256$, ker $= 3\times3$)	θ^Q, θ^K : Conv(out $= 64$, ker $= 3\times3$) θ^V : Conv(out $= 256$, ker $= 3\times3$) θ^Z : Conv(out $= 256$, ker $= 3\times3$) θ^q, θ^c : Conv(out $= 256$, ker $= 3\times3$)
N_{params}	2.4M	2.4M	4.7M
$l = 4$ Bottleneck		θ^z : Conv(out $= 256$, ker $= 3\times3$) θ^q, θ^c : Conv(out $= 256$, ker $= 3\times3$)	θ^Q, θ^K : Conv(out $= 16$, ker $= 3\times3$) θ^V : Conv(out $= 64$, ker $= 3\times3$) θ^Z : Conv(out $= 64$, ker $= 3\times3$) θ^q, θ^c : Conv(out $= 64$, ker $= 3\times3$)
N_{params}	-	2.4M	3.2M

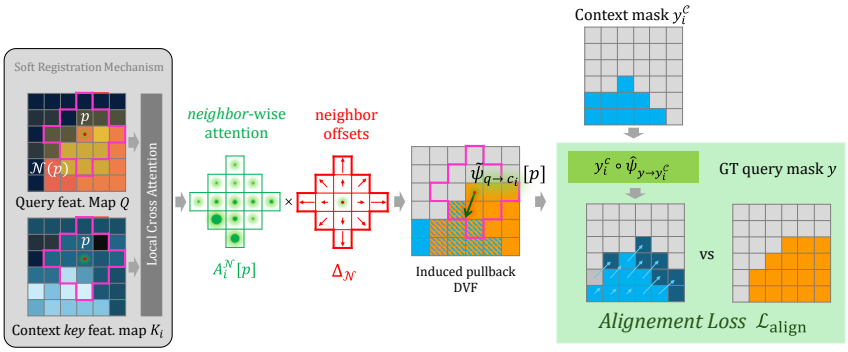

Figure 9: **Alignment loss**. For a pixel p and a context example i, the displacement vector $\tilde{\psi}_{q\to c_i}[p]$ is obtained by weighting the neighborhood offsets $\Delta_{\mathcal{N}}$ by the query–context neighbor-wise similarities. When p belongs to the query mask, the registration loss encourages $\tilde{\psi}_{q\to c_i}[p]$ to point toward the context mask $y_i^{\mathcal{C}}$. This encourages the query features within the query mask to be closer to the context features of pixels inside the context masks than to those outside.

and offset according to the neighborhood structure $\Delta_{\mathcal{N}}$ so as to align the query features with the features derived from neighboring pixels in the context at position p:

$$\mathbf{K}_{i,j}[p] = K_i[p + \Delta_j] \text{ for } j = 1, 2, \ldots, N$$
$$\mathbf{V}_{i,j}[p] = V_i[p + \Delta_j] \text{ for } j = 1, 2, \ldots, N$$

The dense local attention $A^{\mathcal{N}}$ is then computed following Eq. (1). The tensor contraction operator $\times_F$ used in this equation collapses the shared feature dimension of Q and $\mathbf{K}_i$ by performing an inner product along this axis. We break down the effect of this operator below:

$$\begin{aligned}(Q \times_F \mathbf{K}_i)[j,h,w] &= \sum_{m\in[1,F]} Q[h,w,m]\mathbf{K}_i[j,h,w,m] \\ &= \sum_{m\in[1,F]} Q[h,w,m]K_i[h+\Delta_j^0, w+\Delta_j^1, m]\end{aligned}$$

for all $(j,h,w) \in [1,N]\times[1,H]\times[1,W]$. The dense local attention $A^{\mathcal{N}}$ enables the computation of the softly registered context maps ν_i according to Eq. (2).

Dense Context Pooling (DCP) The DCP step selects relevant information across the context. We exploit the previously computed neighbor-wise query-to-context similarity $Q \times_F \mathbf{K}$ in Eq. (4) to compute a pixel-wise context attention $A^{\mathcal{C}}$. $Q \times_F \mathbf{K}$ is reweighted by the attention $A^{\mathcal{N}}$ and collapsed across the *neighbor* dimension. The motivation for this reweighting, instead of simple averaging, is that the dense context attention $A_i^{\mathcal{C}}[p]$ between the query and context i at pixel p should be high if there is at least one pixel within the neighborhood of p in K_i that matches $Q[p]$. As illustrated in Fig. 9, if we were to only average the similarities across $\mathcal{N}[p]$, the overall score would be low, discarding information from a relevant context example.

As in Butoi* et al. (2023), we merge the query and context features via concatenation, forming $Z \in \mathbb{R}^{S\times H\times W\times F_z}$. In our case, the spatial coherence of this concatenation is strengthened by the soft registration step: context maps and query maps are fused after being aligned. Precisely, Z is defined as:

$$Z_i = \text{LN}(\sigma(\text{Conv}(q \parallel \nu_i, \theta^Z))) \text{ for } i = 1, 2, \ldots, S \tag{20}$$

The dense context attention $A^{\mathcal{C}}$ is used to aggregate these merged query-context maps (Eq. (5)) to compute the updated query feature map.

The final outputs of the SpotBlock (q', C') are obtained after a last convolution step:

$$q' = \text{LN}(\sigma(\text{Conv}(\zeta, \theta^q))) \tag{21}$$
$$c_i' = \text{LN}(\sigma(\text{Conv}(Z_i, \theta^c))) \text{ for } i = 1, 2, \ldots, S \tag{22}$$

The dimensions of all the convolution kernels used in OmniSeg are detailed in Table 3.

B CONTRASTIVE EFFECT OF THE REGISTRATION LOSS

As explained in the main body of the article, at every level l, SpotBlock modules, and more specifically the Soft Registration Mechanism, induce DVFs $\tilde{\psi}_{q\to c_i}[p]$ that register context features onto the query grid (see Eq. (3)). We explicitly supervise these DVFs by evaluating how well they register the context masks $y_i^{\mathcal{C}}$ onto the query mask y, thereby defining a loss $\mathcal{L}_{\text{align}}$.

Our experiments show that this loss boosts in-context segmentation performance. We argue that the synergy between this deep registration loss and the segmentation objective arises because the gradient of $\mathcal{L}_{\text{align}}$ has a pixel-wise contrastive effect on the query feature maps, as we detail below. More specifically, it moves the pixel features in the query foreground toward the features in context foregrounds and away from the features in context backgrounds. The converse holds for query background features *mutatis mutandis*. Indeed, because the DVFs are computed in closed form from the attention weights, they depend on the learned query and context features (Eq. (19)). Optimizing $\mathcal{L}_{\text{align}}$ thus has a direct effect on these representations.

For simplicity, let us assume that $\tilde{\psi} = \hat{\psi}$, which holds at the bottleneck of OmniSeg (see Eq. (18)). The same reasoning extends to downstream decoder blocks. Adapting the definition of $\mathcal{L}_{\text{align}}$ for a given context i, we get:

$$\mathcal{L}_{\text{align}}(\hat{\psi}_{y\to y_i^{\mathcal{C}}}\ ;\ y, y_i^{\mathcal{C}}) = \mathcal{L}_{\text{Dice}}(y_i^{\mathcal{C}} \circ \tilde{\psi}_{q\to c_i}\ ;\ y) = 1 - \text{Dice}(\tilde{y}_i, y) \tag{23}$$

Minimizing this loss is equivalent to maximizing the Dice score between the registered context mask $y_i^{\mathcal{C}} \circ \tilde{\psi}_{q\to c_i}$ and the query mask y. To lighten notation, we write $\tilde{y}_i = y_i^{\mathcal{C}} \circ \tilde{\psi}_{q\to c_i}$. As y and $y_i^{\mathcal{C}}$ are binary masks, the soft Dice score can be expressed as:

$$\text{Dice}(\tilde{y}_i, y) = 2\frac{\sum\limits_{p\in\text{pixels}} y[p]\,\tilde{y}_i[p]}{\sum\limits_{p\in\text{pixels}} y[p]^2 + \tilde{y}_i[p]^2} = \frac{2\,\mathcal{I}}{\mathcal{U}} \tag{24}$$

The numerator $\mathcal{I}$ quantifies the overlap of the two masks, while the denominator $\mathcal{U}$ has a normalization and regularization role, encouraging them to have similar sizes. Let us derive the gradient of Eq. (24) with respect to the model parameters θ. First, we express the gradients of $\mathcal{I}$ and $\mathcal{U}$ separately:

$$\nabla_\theta \mathcal{I} = \sum_{p\in\text{pixels}} y[p]\,\nabla_\theta \tilde{y}_i[p] \qquad \text{and} \qquad \nabla_\theta \mathcal{U} = 2\sum_{p\in\text{pixels}} \tilde{y}_i[p]\,\nabla_\theta \tilde{y}_i[p]. \tag{25}$$

The quotient rule then gives:

$$\begin{aligned}
\nabla_\theta\,\text{Dice}(\tilde{y}_i, y) &= \frac{2}{\mathcal{U}^2}\left(\mathcal{U}\,\nabla_\theta\mathcal{I} - \mathcal{I}\,\nabla_\theta\mathcal{U}\right) \\
&= \frac{2}{\mathcal{U}}\left(\nabla_\theta\mathcal{I} - \frac{\mathcal{I}}{\mathcal{U}}\,\nabla_\theta\mathcal{U}\right) \\
&= \frac{2}{\mathcal{U}}\sum_{p\in\text{pixels}}\Big[y[p]\,\nabla_\theta\tilde{y}_i[p] - \text{Dice}(\tilde{y}_i, y)\,\tilde{y}_i[p]\,\nabla_\theta\tilde{y}_i[p]\Big] \\
&= \frac{2}{\mathcal{U}}\sum_{p\in\text{pixels}}\left(y[p] - \tilde{y}_i[p]\,\text{Dice}(\tilde{y}_i, y)\right)\nabla_\theta\tilde{y}_i[p]
\end{aligned} \tag{26}$$

Denoting $\lambda[p] = y[p] - \tilde{y}_i[p]\,\text{Dice}(\tilde{y}_i, y)$, we can write:

$$\nabla_\theta\,\text{Dice}(\tilde{y}_i, y) = \frac{2}{\mathcal{U}}\sum_{p\in\text{pixels}}\lambda[p]\,\nabla_\theta\tilde{y}_i[p] \tag{27}$$

The gradient of the loss can therefore be analyzed at a pixel level. Since $\text{Dice}(\tilde{y}_i, y) \in [0, 1]$, if p belongs to the query foreground (i.e., $y[p] = 1$) then $\lambda[p] \geq 0$. Otherwise, if p belongs to the query background then $\lambda[p] \leq 0$. It remains to compute the per-pixel factor $\nabla_\theta\tilde{y}_i[p]$ with respect to the pixel-wise local similarities (see Eq. (1)). The chain rule gives:

$$\nabla_\theta\tilde{y}_i[p] = \nabla_\theta\left(y_i^{\mathcal{C}}[p + \tilde{\psi}_{q\to c_i}[p]]\right) = \nabla_p y_i^{\mathcal{C}}[p + \tilde{\psi}_{q\to c_i}[p]] \cdot \nabla_\theta\tilde{\psi}_{q\to c_i}[p] \tag{28}$$

The first factor of this scalar product can be seen as the value at pixel p of the pullback of the spatial gradient of $y_i^{\mathcal{C}}$, i.e., $\nabla_p y_i^{\mathcal{C}} \circ \tilde{\psi}_{q \to c_i}$. This gradient map $\nabla_p y_i^{\mathcal{C}} \circ \tilde{\psi}_{q \to c_i}$ is non-zero only for pixels close to the border of the target. It is close rather than *exactly on* the border because we define $y_i^{\mathcal{C}}$ outside its discrete grid through bilinear interpolation. In that case, it corresponds to the inward normal of the target boundary $n^*[p] = \nabla_p y_i^{\mathcal{C}}[p + \tilde{\psi}_{q \to c_i}[p]]$. Based on Eq. (3), we get:

$$\nabla_\theta \tilde{y}_i[p] = \nabla_\theta \left(y_i^{\mathcal{C}}[p + \tilde{\psi}_{q \to c_i}[p]] \right) = \sum_{j=1}^{N} (\Delta_j \cdot n^*[p]) \nabla_\theta A_{i,j}^{\mathcal{N}}[p]. \tag{29}$$

Substituting this expression into Eq. (27), we obtain Eq. (9):

$$\nabla_\theta \mathcal{L}_{\text{align}}(\hat{\psi}_{y \to y_i^{\mathcal{C}}} \,;\, y, y_i^{\mathcal{C}}) = -\frac{2}{\mathcal{U}} \sum_{p \in \text{pixels}} \lambda[p] \sum_{j=1}^{N} (\Delta_j \cdot n^*[p]) \nabla_\theta A_{i,j}^{\mathcal{N}}[p]$$

The neighborhood of p, $\mathcal{N}[p]$, can be split into two halves depending on whether $\Delta_j \cdot n^*[p]$ is positive or negative. The positive half $\mathcal{N}[p]^+$ is oriented toward the interior of the mask $y_i^{\mathcal{C}}$, and the negative half $\mathcal{N}[p]^-$ in the opposite direction. As a reminder, only the pixels p such that $p + \tilde{\psi}_{q \to c_i}[p]$ lies near an edge of $y_i^{\mathcal{C}}$ contribute to the gradient. The fact that $\tilde{\psi}_{q \to c_i}[p]$ is a convex combination of the neighborhood offsets implies that $\mathcal{N}[p]$ intersects $y_i^{\mathcal{C}}$. Under the weak assumption that $y_i^{\mathcal{C}}$ has locally low curvature, $\mathcal{N}[p]^+$ will have a larger intersection with $y_i^{\mathcal{C}}$ than $\mathcal{N}[p]^-$ (see Fig. 10). As such, favoring similarity with neighbors in $\mathcal{N}[p]^+$ over $\mathcal{N}[p]^-$ translates to encouraging query features at p to be closer to context features inside the context masks than outside; the symmetric argument holds when p is in the query background, pushing query features away from the context foreground and toward the context background. Since $A_{i,j}^{\mathcal{N}}[p] = \text{Softmax}(Q[p] \cdot K_i[p + \Delta_j]/T)$, the sign of $\lambda[p]$ dictates which half is favored when minimizing $\mathcal{L}_{\text{align}}$.

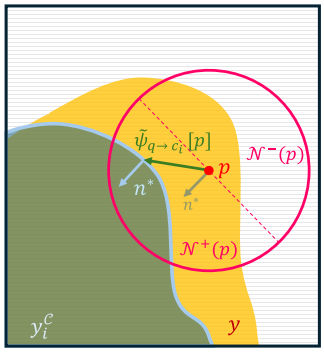

Figure 10: **Local gradient effect**. Illustration of the terms of Eq. (29).

Bringing it all together:

- If p belongs to the query foreground, i.e., $y[p] = 1$, then $\lambda[p] \geq 0$. Based on Eq. (9), minimizing $\mathcal{L}_{\text{align}}$ favors higher similarity between the query features $Q[p]$ and the context key features $K_i[p + \Delta_j]$ for neighbors $j \in \mathcal{N}[p]^+$ and lower similarity for neighbors in $\mathcal{N}[p]^-$. Therefore, the gradient reinforces the similarity between query and context pixel features belonging to the foreground and, conversely, pushes the query foreground features away from the context background features.
- Otherwise, if p belongs to the query background, i.e., $y[p] = 0$, then $\lambda[p] \leq 0$. In that case, the effect is inverted. With the same reasoning, the gradient brings query and context pixel features belonging to the background closer together and separates the query background features from the context foreground features.

Table 4 lays out these four sign combinations of $\lambda[p]$ and $\Delta_j \cdot n^*[p]$ reflecting the local configuration as a two-way table, together with their effect on the attention weight $A_{i,j}^{\mathcal{N}}[p]$ and on the underlying features. This analysis reveals a dense contrastive effect of the registration loss. Intuitively, this should help the segmentation objective by making the pixel features inside the query mask more distinguishable from those outside. Note that this contrastive property is most effective near the overlap between the boundaries of the registered context masks and the query masks. This further motivates our deep registration scheme, since these overlaps cover broader regions of the image at deeper levels of OmniSeg.

C Dataset

We build our dataset by assembling multiple public datasets. We relied in part on pre-existing multi-dataset collections, BiomedSegFM (Ma et al., 2026) and MedSegBench (Kuş & Aydin, 2024). Both

Table 4: **Local configurations of the registration loss gradient.** This table summarizes how the registration loss $\mathcal{L}_{\text{align}}$ gradient affects the local query and context features depending on the local configuration.

	p **is in the query foreground** $(y[p] = 1 \Rightarrow \lambda[p] \geq 0)$	p **is in the query background** $(y[p] = 0 \Rightarrow \lambda[p] \leq 0)$
$p + \Delta_j$ **likely in the foreground of context** i $(j \in \mathcal{N}[p]^+ \Leftrightarrow \Delta_j \cdot n^*[p] > 0)$	$\downarrow \mathcal{L}_{\text{align}} \sim \uparrow A^{\mathcal{N}}_{i,j}[p]$ • Query features $Q[p]$ **pulled toward** context features $K_i[p + \Delta_j]$.	$\downarrow \mathcal{L}_{\text{align}} \sim \downarrow A^{\mathcal{N}}_{i,j}[p]$ • Query features $Q[p]$ **pushed away** from context features $K_i[p + \Delta_j]$.
$p + \Delta_j$ **likely in the background of context** i $(j \in \mathcal{N}[p]^- \Leftrightarrow \Delta_j \cdot n^*[p] < 0)$	$\downarrow \mathcal{L}_{\text{align}} \sim \downarrow A^{\mathcal{N}}_{i,j}[p]$ • Query features $Q[p]$ **pushed away** from context features $K_i[p + \Delta_j]$.	$\downarrow \mathcal{L}_{\text{align}} \sim \uparrow A^{\mathcal{N}}_{i,j}[p]$ • Query features $Q[p]$ **pulled toward** context features $K_i[p + \Delta_j]$.

provide pre-processed medical images and segmentation masks from multiple sources. In our dataset, all multi-class segmentation masks are split into one binary mask per class. We further convert instance segmentation masks into semantic segmentation masks. Following previous works (Wong et al., 2025; Butoi* et al., 2023), every image/mask pair is resized to 128×128.

Our collection is organized into 80 subsets, each associated with a modality and a set of anatomical or pathological classes. These subsets either correspond to a single source dataset or a subdivision defined in BiomedSegFM which may combine several sources.

Furthermore, we group these 2D image/mask sample pairs per segmentation task. We define a segmentation task as the triplet $(\text{modality}, \text{axis}, \text{target})$. Tasks for which fewer than 32 samples are available are filtered out.

We randomly select 10% of the training split to serve as an in-distribution validation split. Additionally, we built held-out validation and evaluation sets containing segmentation tasks and modalities left unseen during training in order to evaluate our model for task generalization. Both splits cover at least two imaging modalities.

Tables 5 to 7 detail the composition of our dataset, its splits and subsets. The column "Exc/Tot" denotes the number of anatomical targets that only appear in the associated dataset.

2D slice extraction We defined a 2D slice extraction scheme to make use of 3D datasets and favor geometrical diversity across the slices. As shown in Fig. 11, for each volume and 3D multilabel segmentation mask pair (X, Y) and each anatomical target j segmented in Y, we extract 3 image slices and 2D binary masks per canonical axis (i.e., a total of 9 extracted 2D pairs per volume and per anatomical structure).

For each axis, we extract the slice i_{largest} displaying the largest section of the anatomical target j.

$$i_{\text{largest}} = \arg\max_{i \in \text{axis}} \sum_{p \in \text{slice}_i} \mathbf{1}(Y_i[p] = j) \tag{30}$$

Two additional slices are selected at intermediate positions between this slice and the extremities $(i_{\text{top}}, i_{\text{bot.}})$ of the segmented target along the considered axis.

$$i_{\text{inter-top}} = \frac{i_{\text{largest}} + i_{\text{top}}}{2}$$
$$i_{\text{inter-bot.}} = \frac{i_{\text{largest}} + i_{\text{bot.}}}{2}$$

The pairs of 2D images and binary segmentation masks are defined for each axis as $(X_i, \mathbf{1}(Y_i = j))$ for $i \in \{i_{\text{inter-bot.}}, i_{\text{largest}}, i_{\text{inter-top}}\}$.

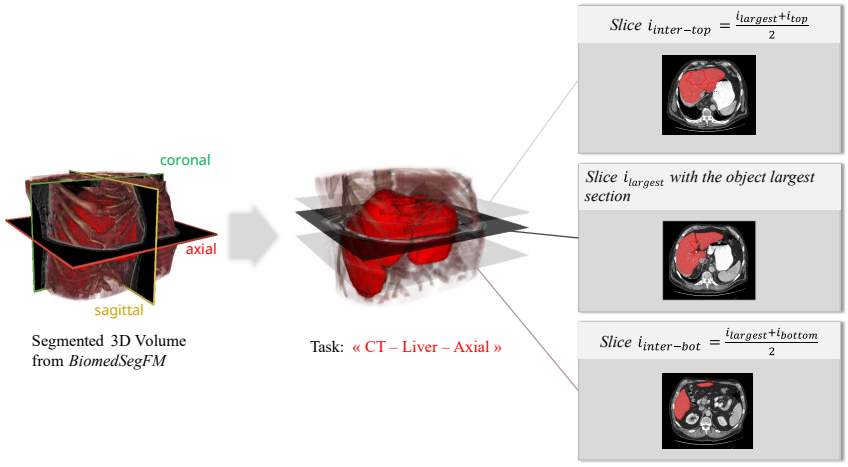

Figure 11: The 2D slice extraction process from 3D segmented volumes of **BiomedSegFM**.

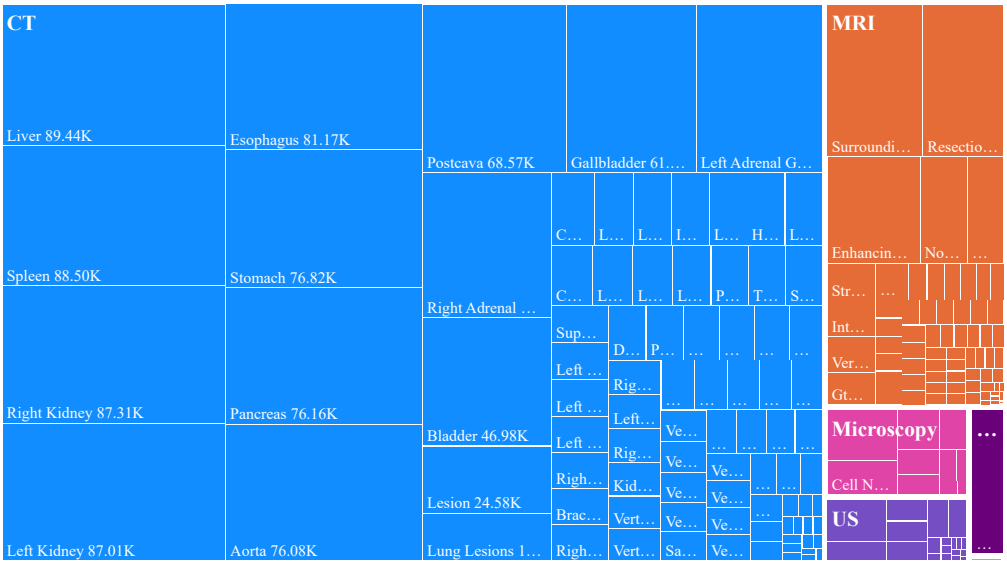

Figure 12: Distribution of images per modality and target class in the training split. To compensate for the imbalance between modalities, we employ a uniform task sampling strategy during the training of the final model.

All images and masks are padded to a square shape. We discard pairs whose segmented target covers fewer than 9 pixels, as well as pairs whose initial shape is too thin (e.g., when one axis of the original volume is small compared to the other), i.e., $\min(H, W)/\max(H, W) < 0.15$.

Remarks:

- RGB images of the fundus datasets are converted to grayscale.
- The training split of the dataset includes noisy samples due in particular to orientation mislabeling and image artifacts.

Generalization degree of the held-out test subsets. The 10 held-out test subsets used in Table 1 exhibit varying degrees of domain shift with respect to the training data. To provide a transparent

Table 5: Training datasets. Datasets are split into training and in-distribution validation splits.

Modality	Subset	Target Description	Targets (Exc/Tot)	Acquisitions	Samples
CT (1,292,815)	Abdomen1K (Ma et al., 2022b;a)	*Abdomen organs*	0/5	1,000	44,733
	AbdomenAtlas (Li et al., 2024a)	*Abdomen organs*	1/13	8,128	833,537
	AdrenalTumor (Ahmed et al., 2020)	*Adrenal gland lesions*	1/1	48	348
	Aeropath (Støverud et al., 2024)	*Airways*	1/1	27	243
	AirwayTree (Zhang et al., 2023)	*Airway tree*	1/1	280	2,520
	Aorta (Myronenko et al., 2023)	*Aortic Vessel Trees*	1/1	51	456
	COVID19-Infection (Roth et al., 2022)	*Covid-19 Infection in CTs*	1/1	198	8,808
	ColonTumor (Antonelli et al., 2022)	*Colon cancer primaries*	1/1	114	487
	KidneyTumor (Heller et al., 2021)	*Kidney lesions*	1/1	441	4,934
	LiverTumor (Antonelli et al., 2022; Morshid et al., 2019)	*Liver lesions*	1/1	656	6,534
	LungLesion (Antonelli et al., 2022; Bakr et al., 2018)	*Lung lesions*	1/1	6,969	17,400
	Lymph Node (Roth et al., 2014; Engelson et al., 2024)	*Lymph node*	1/1	981	7,715
	PancreasTumor (Antonelli et al., 2022)	*Pancreas tumors*	1/1	593	2,778
	SegRap HeadNeckTumor (Luo et al., 2025)	*Head-Neck Lesions*	1/1	216	1,854
	ThoracicOrgans (Yang et al., 2018)	*Chest organs*	0/5	54	2,398
	TopCoW24 (Yang et al., 2025)	*Circle of Willis*	0/1	125	1,023
	TotalSeg cardiac (Wasserthal et al., 2023)	*Cardiovascular structures*	9/18	1,057	108,532
	TotalSeg organs (Wasserthal et al., 2023)	*Abdomen organs*	9/24	1,059	141,166
	TotalSeg-vertebrae (Wasserthal et al., 2023)	*Vertebrae*	25/26	557	82,767
	WholeBodyTumor (Gatidis et al., 2022)	*Tumorous lesions*	0/1	408	24,582
MRI (205,635)	AMOS (Ji et al., 2022)	*Abdomen organs*	1/15	54	6,263
	BraTS-T1c (Menze et al., 2014)	*Brain lesions*	0/4	1,215	30,273
	BraTS-T1n (Menze et al., 2014)	*Brain lesions*	0/4	1,215	30,486
	BraTS-T2f (Menze et al., 2014)	*Brain lesions*	0/4	1,215	30,319
	BraTS-T2w (Menze et al., 2014)	*Brain lesions*	0/4	1,215	30,549
	CHAOS-T1 (Kavur et al., 2021)	*Abdomen organs*	0/4	18	646
	CHAOS-T2 (Kavur et al., 2021)	*Abdomen organs*	0/4	18	648
	CervicalCancer (Bowen et al., 2018)	*Cervical Cancer Tumor*	1/1	7	32
	HNTS-MRG (Wahid et al., 2024)	*Head and Neck Tumors*	1/1	261	4,514
	ISLES ADC (Hernandez Petzsche et al., 2022)	*Stroke Lesions*	0/1	212	2,607
	ISLES DWI (Hernandez Petzsche et al., 2022)	*Stroke Lesions*	0/1	212	2,660
	ProstateADC (Antonelli et al., 2022)	*Transition Zone*	1/1	257	1,975
	ProstateT2 (Antonelli et al., 2022; Bloch et al., 2015; Litjens et al., 2014)	*Prostate*	0/1	386	3,325
	QIN-PROSTATE-Lesion (Fedorov et al., 2018)	*Prostate lesions*	1/1	59	123
	RESECT (Pace et al., 2024)	*Brain Tumor*	0/1	23	129
	SegThy (Krönke et al., 2022)	*Thyroid and blood vessels*	2/5	14	617
	Spider IVD (van der Graaf et al., 2024)	*Intervertebral Discs*	0/1	289	3,779
	Spider Spine (van der Graaf et al., 2024)	*Spinal Canal*	1/1	293	879
	Spider Vertebrae (van der Graaf et al., 2024)	*Vertebrae*	0/1	253	3,586
	CrossMoDA (Dorent et al., 2023)	*Tumor and Cochlea*	2/4	205	2,234
	TopCoW24 (Yang et al., 2025)	*Brain blood vessels*	0/1	125	1,111
	TotalSeg (Akinci D'Antonoli et al., 2025)	*Abdomen organs*	0/46	227	37,570
	WMH FLAIR (Kuijf et al., 2019)	*White matter hyperintensities*	0/1	153	5,590
	WMH T1 (Kuijf et al., 2019)	*White matter hyperintensities*	0/1	153	5,720
US (26,617)	Cardiac (Leclerc et al., 2019)	*Heart chamber in US sequence*	0/3	900	7,510
	Low-limb-Leg (Duque et al., 2024)	*Leg muscles*	2/3	32	864
	RESECT (Behboodi et al., 2024)	*Brain Tumors*	0/1	158	8,370
	SegThy (Krönke et al., 2022)	*Thyroid and blood vessels*	2/3	22	585
	AbdomenUS (Vitale et al., 2020)	*Abdomen Organs*	4/8	566	1,711
	MuscleUS (Marzola et al., 2021)	*Limb Muscles*	2/3	1,688	8,169
Microscopy (36,314)	AxonEM (Wei et al., 2021)	*Axon*	1/1	17	767
	MUCIC-HL60 (Svoboda et al., 2009)	*Cell Nuclei*	1/1	120	7,186
	NIS3D (Zheng et al., 2023)	*Nuclei*	0/1	6	2,564
	SELMA3D AD plaques (Kaltenecker et al., 2024)	*Alzheimer's Disease Plaque*	1/1	30	3,331
	SELMA3D neural activity (Kaltenecker et al., 2024)	*Neural Activity*	1/1	18	3,237
	SELMA3D nucleus (Kaltenecker et al., 2024)	*Nucleus*	0/1	11	5,129
	SELMA3D vessel (Kaltenecker et al., 2024)	*Vessels*	1/1	21	1,009
	CREMI (CREMI, 2016)	*Synaptic Clefts*	1/1	3	2,594
	NUCMM (Lin et al., 2021)	*Nuclei*	0/1	49	8,416
	UroCell Endolysosomes (Mekuč et al., 2020)	*Endolysosomes*	1/1	5	403
	UroCell Mitochondria (Mekuč et al., 2020)	*Mitochondria*	1/1	5	1,678
X-Rays (15,380)	Covid19-Radio (Chowdhury et al., 2020; Rahman et al., 2021)	*Lungs*	0/1	14,814	14,814
	LungXR (Jaeger et al., 2013; Candemir et al., 2013)	*Lungs*	0/1	566	566
OCT (703)	Cysto Fluid (Ahmed et al., 2022)	*Cystoid Macular Edema*	1/1	703	703
Fundus (37)	ChaseDB1 (Carballal et al., 2018)	*Vessels*	0/1	19	19
	Drive (Staal et al., 2004)	*Vessels*	0/1	18	18
Total	**66 Subsets**		**155**	**50,812**	**1,577,501**

Table 6: **Out-of-domain validation datasets**. These subsets contain held-out target classes. We used this split to validate the generalization capabilities of the model during training.

Modality	Subset	Target Description	Targets (Exc/Tot)	Acquisitions	Samples
CT (16,789)	HaN-Seg (Podobnik et al., 2023)	*Head neck organs*	23/29	38	9,260
	Lung (Ma et al., 2021; Kiser et al., 2021; Antonelli et al., 2022)	*Lungs in chest CT*	0/2	419	7,529
MRI (8,486)	HVSMR (Pace et al., 2024)	*Heart anatomy*	3/8	120	8,324
	Left Atrium (Antonelli et al., 2022)	*Heart anatomy*	0/1	18	162
Total	**4 Subsets**		39	**595**	**25,275**

Table 7: **Out-of-domain test subsets**. These subsets are used to evaluate OmniSeg's few-shot capabilities. They cover tasks seen during training on other data sources as well as completely held-out ones.

Modality	Subset	Description	Targets (Exc/Tot)	Acquisitions	Samples
CT	AMOS (CT) (Ji et al., 2022)	*Abdominal organs*	0/15	270	35,755
	TotalSeg muscles (Wasserthal et al., 2023)	*Bones and muscles*	5/23	1,097	126,432
MRI	Heart ACDC (Bernard et al., 2018)	*Heart structures*	2/3	270	6,858
	IBD (Zhong et al., 2025)	*Digestive system*	8/10	114	9,416
US	BUID (Al-Dhabyani et al., 2020)	*Breast lesions*	2/2	647	647
Microscopy	WBC (Zheng et al., 2018)	*White blood cells*	1/2	400	800
XR	PanDental (Brahmi et al., 2025)	*Jaw in dental XR*	1/1	116	116
	HipXRay (Gut, 2021)	*Hip joint in XR*	2/2	140	280
	SCR (Van Ginneken et al., 2006)	*Lungs, Clavicles and Heart*	4/5	1,235	1,235
PET	autoPET (Gatidis et al., 2022)	*Tumor lesions in FDG PET*	0/1	623	55,966
Total	**10 Subsets**		**62**	**4,912**	**237,535**

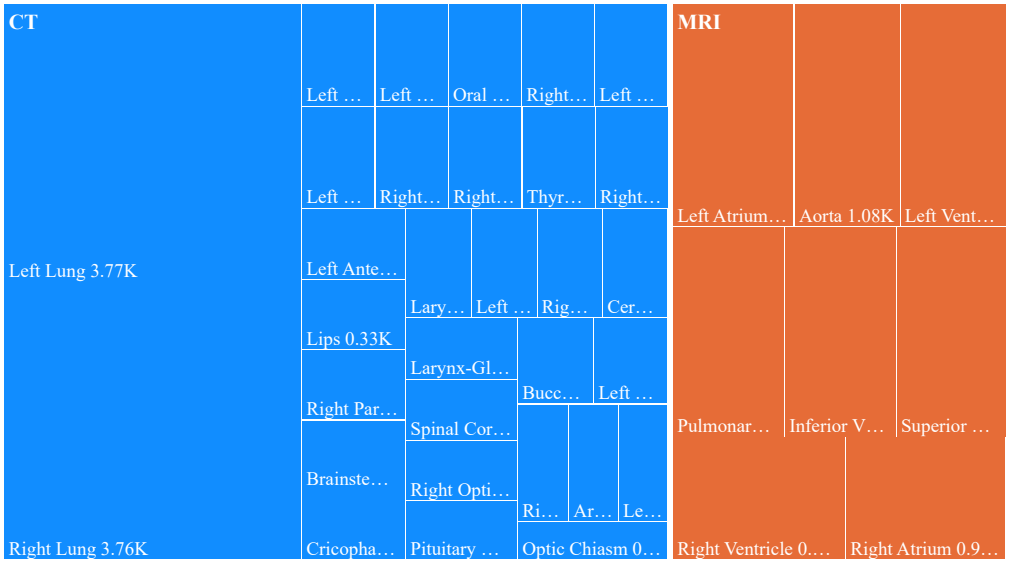

Figure 13: Distribution of images per modality and target class in the held-out validation split. The validation split includes both seen and unseen structures during training.



Figure 14: Distribution of images per modality and target class in the held-out test split. It contains both an unseen modality (PET) and unseen target classes. We evaluate the pretrained baselines and the variants of OmniSeg on a large quantity of images to assess the significance of our results.

analysis of OmniSeg's generalization, we categorize each subset according to its relationship with the training set in Table 8. Four categories emerge:

1. **Shared source, new targets.** TotalSeg Muscles shares the same acquisitions as the TotalSeg cardiac, organs, and vertebrae subsets used during training but for other targets (bones, organs). Therefore, the only domain shift lies in the targets. This subset evaluates the model's capacity to generalize to new targets for the same set of patients.
2. **Partially shared source, cross-modality.** autoPET PET originates from a dataset that also contributes to the training split under a different imaging modality. The CT component of the dataset appears in training as WholeBodyTumor (Gatidis et al., 2022) (408 acquisitions), while the PET component (623 acquisitions) is held out for evaluation. The patient cohorts are not disjoint between the two splits (303 acquisitions in common, i.e., about half of autoPET). However, the PET imaging modality is entirely absent from the training set. This dataset is particularly challenging since it covers a lesion segmentation task, which implies a higher chance of spatial misalignment between query and context masks.
3. **No source overlap, same targets.** All AMOS segmentation tasks (abdominal organs in CT) are covered by the training split. The domain shift separating this held-out dataset from the training is therefore restricted to the data source. As such, it allows us to evaluate the capacity of the model to generalize to different data sources at test time. Note that both Butoi* et al. (2023) and Wong et al. (2025) use this dataset as training data and that patients of the AMOS (MRI) subset are distinct from the CT subset.
4. **No source overlap, new targets.** The remaining seven subsets—IBD, ACDC, BUID, HipXRay, PanDental, SCR, and WBC—come from fully independent cohorts with no patient-level or dataset-level overlap with the training data. These subsets introduce partially or entirely unseen target classes.

D TRAINING SETUP

Data augmentation Following Butoi* et al. (2023) and Wong et al. (2025), we adopt in-task and task-level augmentation strategies (see Table 9). We also complement the training with synthetic

Table 8: **Generalization taxonomy of the held-out test subsets.** Each subset is categorized by its degree of overlap with the training data. "Exc/Tot" denotes the number of target classes exclusive to the test subset over the total, across all modalities. **Modality**, **Patients**, and **Target class** indicate whether each factor was seen in training (**Seen**), entirely absent (**Unseen**), or only partly overlapping (**Partially unseen**).

Category	Subset	Exc/Tot	Modality	Patients	Target class
Shared source, new targets	TS Muscles (CT)	5/23	Seen	Seen	Partially unseen
Partially shared source, cross-modality	autoPET (PET)	0/1	Unseen	Partially unseen	Seen
No source overlap, same targets	AMOS (CT)	0/15	Seen	Unseen	Seen
No source overlap, new targets	IBD (MRI)	8/10	Seen	Unseen	Partially unseen
	ACDC (MRI)	2/3	Seen	Unseen	Partially unseen
	BUID (US)	2/2	Seen	Unseen	Unseen
	HipXRay (XR)	2/2	Seen	Unseen	Unseen
	PanDental (XR)	1/1	Seen	Unseen	Unseen
	SCR (XR)	4/5	Seen	Unseen	Partially unseen
	WBC (Micro.)	1/2	Seen	Unseen	Partially unseen

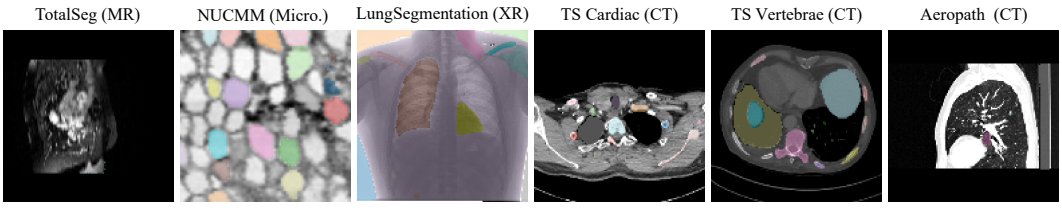

Figure 15: **Pseudo-masks generated with SAM for some training datasets.**

query/context sets. For all the images of the training split, pseudo-masks are obtained with the panoptic segmentation mode of SAM (Kirillov et al., 2023), creating new image/mask pairs. We use the ViT-B version of SAM with a uniform click prompt grid of 16×16 points. See examples of pseudo-masks in Fig. 15. These pairs are sampled as queries with a probability of 0.25 during training. In that case, a synthetic context set is created by applying various transformations to this query image (Table 9) as visualized in Fig. 16.

Uniform task sampling For the final model, we randomly sample a task $(\text{modality}, \text{axis}, \text{target})$ among all tasks at each training step (see Algorithm 1). This ensures more balanced frequencies of tasks seen during training.

Training settings We train all the variants of our model with a context size of $S = 16$ and temperature $T = 0.1$. We use the Adam optimizer with a constant learning rate $\eta = 10^{-4}$. The final model is trained with a batch size of 32 split on four NVIDIA H100 GPUs for a total of 5 epochs over 2 days.

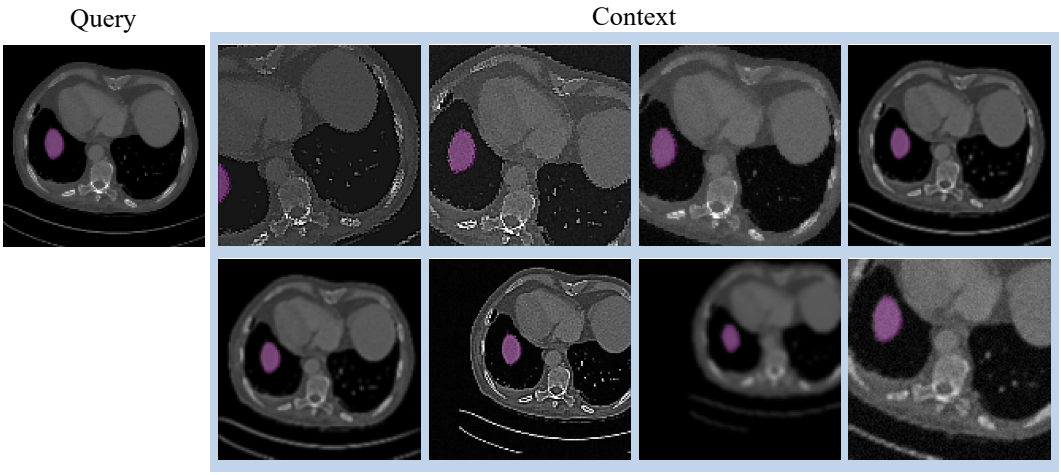

Figure 16: **Synthetic context set examples.** The top-left corner shows the query image with its label map sampled from the pseudo-mask. The other 8 image-mask pairs are the context set obtained from the synthetically augmented query image-mask pair.

Algorithm 1 OmniSeg training loop using Adam with learning rate $\eta = 10^{-4}$ over tasks $\mathcal{T}$, main architecture f_θ, in-task augmentations $\mathrm{Aug}_{\mathrm{img}}$ and task augmentations $\mathrm{Aug}_{\mathrm{task}}$

for $k = 1, \ldots,$ NumTrainSteps **do**
  $t = (\mathrm{modality}, \mathrm{axis}, \mathrm{target}) \sim \mathcal{T}$ ▷ Sample Task
  if $p \sim \mathcal{U} >$ threshold **then**
    $(x^t, y^t) \sim \Omega^t$ ▷ Sample Query
    $\mathcal{C}^t = \{(x_i^t, y_i^t)\}_{i \in [1,S]} \sim \Omega^t \setminus \{(x^t, y^t)\}$ ▷ Sample Context Set
  else
    Generate synthetic context set
    $(x^t, y^t) \sim \Omega^t$ ▷ Sample Query
    $y^t_{\mathrm{pseudo}} \leftarrow \mathrm{SAM}(x^t)$ ▷ Replace mask with pseudo-mask
    $y^t_{\mathrm{label}} \sim \mathcal{U}(y^t_{\mathrm{pseudo}})$ ▷ Sample label map from pseudo-mask
    $\mathcal{C}^t \leftarrow \{\mathrm{Aug}_{\mathrm{synth}}(x^t, y^t_{\mathrm{label}})\}_{i \in [1,S]}$ ▷ Synthetically generate context set
  end if
  $x^t, y^t \leftarrow \mathrm{Aug}_{\mathrm{img}}(x^t, y^t)$ ▷ Augment Query
  $\mathcal{C}^t \leftarrow \{\mathrm{Aug}_{\mathrm{img}}(x_i^t, y_i^t)\}_{i \in [1,S]}$ ▷ Augment Context Set
  $x^t, y^t, \mathcal{C}^t \leftarrow \mathrm{Aug}_{\mathrm{task}}(x^t, y^t, \mathcal{C}^t)$ ▷ Task Aug
  $\hat{y}, \hat{\psi} \leftarrow f_\theta(x^t, \mathcal{C}^t)$ ▷ Predict label map
  $\ell \leftarrow \mathcal{L}_{\mathrm{seg}}(\hat{y}; y^t) + \mathcal{L}_{\mathrm{reg}}(\hat{\psi}; y^t, \mathcal{C}^t)$ ▷ Compute loss
  $\theta \leftarrow \mathrm{Adam}(\theta, \nabla_\theta \ell, \eta)$ ▷ Gradient step
end for

Table 9: **Augmentations used to train OmniSeg. (a)** Within-task data augmentations are randomly sampled for each example to increase diversity within a task. **(b)** Task augmentations are sampled per task and applied to all examples in that task to increase task-level diversity. **(c)** Synthetic task augmentations are used when generating synthetic context sets. Each augmentation is sampled with probability p.

Augmentation	p	Parameters
Random Affine	0.25	degrees $\in [-25, 25]$ translation $\in [0, 0.1]$ scale $\in [0.9, 1.1]$
Brightness Contrast	0.25	brightness $\in [-0.1, 0.1]$ contrast $\in [0.5, 1.5]$
Elastic Transform	0.8	$\alpha \in [1, 2.5]$ $\sigma \in [7, 9]$
Sharpness	0.25	sharpness $= 5$
Gaussian Blur	0.25	$\sigma \in [0.1, 1.0]$ $k = 5$
Gaussian Noise	0.25	$\mu \in [0, 0.05]$ $\sigma \in [0, 0.05]$

(a) Within-Task Augmentations

Augmentation	p	Parameters
Random Affine	0.5	degrees $\in [0, 360]$ translation $\in [0, 0.2]$ scale $\in [0.8, 1.1]$
Brightness Contrast	0.5	brightness $\in [-0.1, 0.1]$ contrast $\in [0.8, 1.2]$
Gaussian Blur	0.5	$\sigma \in [0.1, 1.1]$ $k = 5$
Gaussian Noise	0.5	$\mu \in [0, 0.05]$ $\sigma \in [0, 0.05]$
Elastic Transform	0.5	$\alpha \in [1, 2]$ $\sigma \in [6, 8]$
Sharpness	0.5	sharpness $= 5$
Horizontal Flip	0.5	None
Vertical Flip	0.5	None
Sobel Edges Label	0.5	None
Flip Intensities	0.5	None

(b) Task Augmentations

Augmentation	p	Parameters
Random Affine	0.8	degrees $\in [-25, 25]$ translation $\in [0, 0.2]$ scale $\in [0.9, 1.5]$
Brightness Contrast	0.5	brightness $\in [-0.1, 0.1]$ contrast $\in [0.5, 1.5]$
Elastic Transform	0.8	$\alpha \in [1, 10]$ $\sigma \in [8, 15]$
Sharpness	0.5	sharpness $= 5$
Gaussian Blur	0.5	$\sigma \in [0.1, 1.5]$ $k = 5$
Gaussian Noise	0.5	$\mu \in [0, 0.05]$ $\sigma \in [0, 0.05]$

(c) Synthetic Task Augmentations

E Extended experiments

In this section, we provide detailed results per dataset, supplementary experiments and discussion.

E.1 Confidence Interval

For each task k (a modality $\times$ target $\times$ axis combination) we compute a mean m_k and standard error $\mathrm{SEM}_k = s_k/\sqrt{n_k}$ over its n_k per-case scores, with its degree of freedom $\mathrm{df}_k = n_k - 1$. The macro-average over K tasks weights each task equally:

$$m = \frac{1}{K}\sum_{k=1}^{K} m_k, \qquad \mathrm{SE} = \frac{1}{K}\sqrt{\sum_{k=1}^{K} \mathrm{SEM}_k^2}.$$

The confidence interval uses a Student-t quantile with a Welch–Satterthwaite effective degrees of freedom

$$\mathrm{df} = \frac{\left(\sum_k \mathrm{SEM}_k^2\right)^2}{\sum_k \mathrm{SEM}_k^4/\mathrm{df}_k}, \qquad \mathrm{CI} = m \pm t_{\mathrm{df},\,1-\alpha/2} \cdot \mathrm{SE},$$

with $\alpha = 0.01$ (corresponding to 99% confidence interval).

E.2 Architecture and supervision ablations

We detail in Table 10 the per-dataset performance of the variants of our ablation analysis (Section 4.3, highlighted rows) and of other relevant variants. For this ablation study, all models are trained for 5 epochs on our training set, with temperature set to 0.1, and model selection is done on our validation split. We use a fixed context set size of 16. We discuss below the per-dataset trends observed in Table 10 for each variant.

Architecture ablations (without registration loss). We assess the impact of different architectural choices for the encoder and decoder, and the benefit of Dense Context Pooling in SpotBlock.

- **MultiverSeg (retrained)**: We compare our model against MultiverSeg trained with our training setup. Our training setup improves MultiverSeg performance by 3.9 Dice points.
- **CrossBlock Enc. + SpotBlock Dec.**: We keep the OmniSeg configuration with $L = 4$, but replace the separate convolutional blocks of the encoder with CrossBlocks. This variant tests whether fusing query and context features in the encoder is beneficial. Separate streams for query and context allow computing context features once when running inference on

several images with a fixed context set. Our results confirm that keeping query and context features separate in the encoder improves segmentation performance.

- **Conv Enc. + CrossBlock Dec.**: In this variant, SpotBlocks are replaced by CrossBlocks in the decoder to verify that our performance gains are not solely due to using a separate encoder. The results support this hypothesis: this variant underperforms not only OmniSeg but also MultiverSeg on most datasets.
- ***w/o* Skip Co. Registration**: We assess the importance of warping encoder context features in the skip connections with the computed DVFs (cf. Eq. (15)). This highlights the importance of spatial coherence in the feature flow for the registration ability that emerges in SpotBlocks even without explicit supervision. Skip-connection features are concatenated with the previous decoder output, which has already been aligned to the query via the SRM. Without warping the context features, the skip and decoder activations live on mismatched spatial grids.
- ***w/o* DCP**: Starting from our two-branch convolutional encoder and SpotBlock decoder, we remove the DCP step from the SpotBlocks. Specifically, Eq. (5) is replaced by $\zeta = \mathrm{Avg}(Z_i)$. Removing DCP greatly degrades performance, confirming the benefit of dense context pooling especially where there is a large visual heterogeneity of the samples, like AMOS, IBD or BUID.
- **Full Architecture**: This variant combines all of our architectural contributions in a model trained with segmentation loss only. It outperforms the MultiverSeg baseline, including on subsets where the spatial alignment across samples is high, like HipXRay or PanDental.

These ablations show that SpotBlock performs better when query and context features are encoded separately and the skip context features are registered, which validates the spatial coherence of the feature flow. DCP is also crucial for the overall performance of SpotBlock, especially for data subsets with high visual heterogeneity.

Registration supervision ablations. We assess the importance of our registration supervision strategy, which consists of supervising the DVF computed for each decoder stage with decreasing weights (from bottleneck to output) and using the minimum alignment loss over the context set, i.e., supervising the best matching context example.

- ***Shallow* Reg. Loss**: In this variant, $\mathcal{L}_{\text{reg}}$ is computed from the last predicted DVF only. The results highlight the benefit of deep supervision at all decoder levels compared with standard shallow supervision, with a limited but consistent advantage for the deep scheme.
- ***Full-context* Reg. Loss**: We train this variant to assess our choice of back-propagating only the lowest alignment loss across the context set. In other words, the minimum in Eq. (8) is replaced by an average. This variant scores 0.1 Dice points above our model, within the confidence interval, although it performed worse on the validation set.
- ***Balanced* Reg. Loss**: In our default model, we assign decreasing weights to the alignment loss from the bottleneck to the output block. In this variant, every decoder level receives the same weight. Equal weighting slightly hurts overall performance. In particular, it seems to entail lower performance on the two tumor segmentation subsets BUID and autoPET for which the coarse alignment at the bottleneck level is more important for the final result as the tumor may be located in different parts of the query and context images.

These ablations show the benefit of our registration supervision strategy, and especially the use of deep supervision.

E.3 Per-task performance

To analyze OmniSeg's performance across anatomical structures, Table 11 reports the Dice score on AMOS (CT) and IBD (MRI) per task (target and axis) for the pretrained MultiverSeg (MVS) and our final version of OmniSeg. OmniSeg outperforms its closest baseline across nearly all these segmentation tasks. The gain is notably high for smaller anatomical structures like the adrenal glands or the esophagus. The gap with MultiverSeg remains high for targets left entirely held-out from the training set.

Table 10: **Detailed segmentation performance for ablation studies.** $*$ indicates that evaluation is done on 10,000 samples.

Model	IBD (MRI)	ACDC (MRI)	TS Muscles (CT)	AMOS (CT)	autoPET (PET)
MultiverSeg (retrained)	58.8 ± 0.6	78.5 ± 0.5	74.8 ± 0.2	74.7 ± 0.2	24.3 ± 0.3
Architecture (w/o Reg. Loss)					
CrossBlock Enc. + SpotBlock Dec.	58.7 ± 0.6	78.7 ± 0.5	75.8 ± 0.2	70.8 ± 0.2	23.8 ± 0.2
Conv Enc. + CrossBlock Dec.	56.8 ± 0.6	76.1 ± 0.6	69.4 ± 0.2	75.7 ± 0.3	25.9 ± 0.3
w/o Skip Co. Registration	58.7 ± 0.6	78.6 ± 0.5	74.9 ± 0.2	73.9 ± 0.2	26.8 ± 0.3
w/o DCP	59.1 ± 0.6	80.1 ± 0.4	78.1 ± 0.2	73.7 ± 0.2	23.5 ± 0.3
Full Architecture	63.2 ± 0.6	82.9 ± 0.4	81.9 ± 0.1	78.4 ± 0.2	24.1 ± 0.3
Registration Supervision					
Shallow Reg. Loss	61.7 ± 0.6	82.5 ± 0.4	81.5 ± 0.1	77.5 ± 0.2	24.2 ± 0.2
Full-context Reg. Loss	**63.8** ± 0.6	82.2 ± 0.5	**83.0** ± 0.1	**79.3** ± 0.2	27.0 ± 0.3
Balanced Reg. Loss	63.2 ± 0.6	**83.3** ± 0.4	82.7 ± 0.1	79.1 ± 0.2	25.8 ± 0.3
Ours	63.1 ± 0.6	83.1 ± 0.4	**83.0** ± 0.1	**79.3** ± 0.2	**28.9** ± 0.3

Model	BUID (US)	HipXRay (XR)	PanDental (XR)	SCR (XR)	WBC (Micro.)	Avg.
MultiverSeg (retrained)	59.6 ± 2.6	95.8 ± 0.3	94.3 ± 0.4	89.0 ± 0.5	88.7 ± 0.9	72.0 ± 0.1
Architecture (w/o Reg. Loss)						
CrossBlock Enc. + SpotBlock Dec.	58.2 ± 2.5	95.6 ± 0.4	93.0 ± 0.5	88.2 ± 0.8	84.6 ± 1.2	71.2 ± 0.1
Conv Enc. + CrossBlock Dec.	60.1 ± 2.7	95.2 ± 0.5	93.9 ± 0.4	87.1 ± 0.7	86.5 ± 0.9	69.5 ± 0.2
w/o Skip Co. Registration	56.6 ± 2.8	94.6 ± 0.5	93.9 ± 0.5	83.0 ± 0.7	87.3 ± 1.0	71.6 ± 0.1
w/o DCP	59.0 ± 2.7	96.5 ± 0.3	94.4 ± 0.3	90.5 ± 0.4	89.9 ± 0.8	73.3 ± 0.1
Full Architecture	64.3 ± 2.7	**97.1** ± 0.3	**95.3** ± 0.3	91.5 ± 0.4	90.0 ± 0.8	77.1 ± 0.1
Registration Supervision						
Shallow Reg. Loss	63.2 ± 2.6	**97.1** ± 0.3	94.7 ± 0.4	91.3 ± 0.4	89.4 ± 0.8	76.3 ± 0.1
Full-context Reg. Loss	**68.2** ± 2.6	96.8 ± 0.3	94.9 ± 0.4	91.2 ± 0.4	**90.4** ± 0.8	**78.0** ± 0.1
Balanced Reg. Loss	66.7 ± 2.7	97.0 ± 0.3	94.8 ± 0.4	91.7 ± 0.4	89.9 ± 0.8	77.7 ± 0.1
Ours	67.4 ± 2.7	97.0 ± 0.2	94.9 ± 0.4	91.8 ± 0.3	90.3 ± 0.8	77.9 ± 0.1

Table 11: **Per-target, per-axis Dice score on AMOS and IBD.** Dice scores of OmniSeg (OMS) and MultiverSeg (MVS) per task. For each sub-column, the best model is highlighted in bold. $*$ indicates anatomical targets absent from the training split.

		Axial		Coronal		Sagittal		Avg.	
Subset	**Target**	MVS	OMS	MVS	OMS	MVS	OMS	MVS	OMS
AMOS (CT)	Aorta	76.3	**90.9**	60.6	**84.0**	66.5	**82.9**	67.8	**85.9**
	Bladder	83.4	**86.7**	73.6	**80.0**	72.7	**79.0**	76.6	**81.9**
	Duodenum	45.9	**61.0**	43.0	**57.6**	39.8	**61.3**	42.9	**60.0**
	Esophagus	56.9	**73.2**	43.6	**65.6**	48.6	**66.0**	49.7	**68.3**
	Gallbladder	72.3	**81.1**	57.0	**73.7**	63.0	**77.6**	64.1	**77.5**
	Inferior vena cava	69.4	**84.8**	58.8	**75.3**	60.1	**76.1**	62.8	**78.7**
	Left adrenal gland	54.9	**73.7**	41.0	**67.1**	37.0	**66.5**	44.3	**69.1**
	Right adrenal gland	59.4	**75.1**	53.3	**71.8**	51.5	**72.3**	54.7	**73.1**
	Left kidney	88.0	**95.3**	87.0	**93.8**	82.3	**92.9**	85.8	**94.0**
	Right kidney	87.9	**95.1**	88.6	**94.1**	85.9	**93.5**	87.5	**94.2**
	Liver	90.7	**92.9**	90.2	**93.9**	88.2	**93.4**	89.7	**93.4**
	Pancreas	61.9	**75.2**	51.8	**69.4**	46.7	**65.7**	53.5	**70.1**
	Prostate	79.6	**80.4**	72.3	**73.4**	69.4	**71.4**	73.8	**75.1**
	Spleen	86.5	**93.0**	80.8	**90.6**	79.5	**91.0**	82.3	**91.5**
	Stomach	70.5	**78.2**	59.9	**76.3**	58.7	**74.0**	63.0	**76.2**
	Average	72.2	**82.4**	64.1	**77.8**	63.3	**77.6**	66.6	**79.3**
IBD (MRI)	Stomach	78.6	**81.5**	78.5	**83.7**	75.4	**82.3**	77.5	**82.5**
	Duodenum	58.5	**64.5**	53.4	**64.1**	54.1	**62.5**	55.3	**63.7**
	Small Intestine*	68.2	**70.7**	66.8	**73.7**	68.9	**70.6**	68.0	**71.7**
	Cecum*	61.1	**61.9**	51.2	**55.6**	**43.6**	42.7	52.0	**53.4**
	Appendix*	43.3	**45.3**	45.1	**52.0**	36.5	**43.0**	41.6	**46.8**
	Ascending colon*	68.0	**72.3**	63.8	**68.1**	60.0	**64.6**	63.9	**68.3**
	Transverse colon*	52.2	**55.0**	50.5	**58.4**	48.9	**54.1**	50.5	**55.8**
	Descending colon*	58.7	**64.9**	56.2	**64.5**	51.3	**55.7**	55.4	**61.7**
	Sigmoid colon*	53.1	**53.9**	48.1	**51.5**	50.4	**52.1**	50.5	**52.5**
	Rectum*	74.6	**77.1**	68.7	**74.4**	70.1	**73.1**	71.1	**74.9**
	Average	61.6	**64.7**	58.2	**64.6**	55.9	**60.1**	58.6	**63.1**

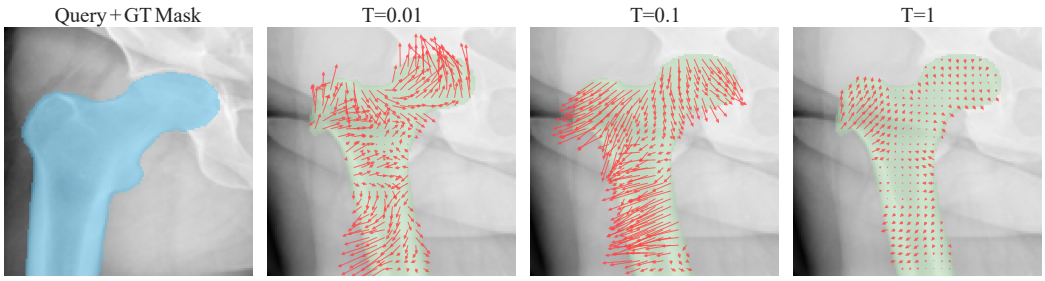

Figure 17: **Impact of temperature on the DVF.** The query and its mask are shown in the leftmost panel. The three other panels show the full-resolution DVFs computed for the same context example with different temperatures, cropped by the mask.

E.4 Impact of the temperature T

Fig. 17 illustrates how the temperature affects the computed DVF and, thus, the dense attention predicted by SpotBlock. At low temperature, $\mathrm{Softmax}$ acts as an $\arg\max$ operation leading to a noisy pairing between query and context pixels. This prevents an accurate registration of the context on the query and entails an irregular DVF as shown in the figure for $T = 0.01$. At higher temperatures, the neighbor-wise attention computed in Eq. (1) becomes uniform, the computed DVF therefore tends to zero (see Eq. (3)). Fig. 18 reports the dataset-specific impact of the temperature hyperparameter on OmniSeg. Overall, performance remains stable for temperatures below 0.15. Although the setting selected after held-out validation ($T = 0.1$) is slightly suboptimal on certain datasets (e.g., BUID and SCR), it yields the best average performance across all evaluated values.

E.5 Impact of the neighborhood $\mathcal{N}$

The local neighborhood $\mathcal{N}$ defines the spatial extent over which SpotBlock matches query and context pixels, and therefore the maximum displacement a single DVF can express at a given decoder stage. We ablate this hyperparameter by varying the disk diameter, training models for 5 epochs with all other settings at their default values ($T = 0.1$, context set size of 16), and evaluating on test subsets restricted to 1000 query samples (randomly subsampled when the test set is larger). This choice primarily affects behavior at the bottleneck, where residual DVFs are bounded by the upscaling factor of the decoder as they only correct the misalignment appearing at the higher resolution. Table 12 reports the average Dice score over our held-out test datasets for diameters 3, 5 and 7 pixels. A diameter of 3 is too restrictive: the attainable offsets are too small to compensate for the misalignment between query and context samples, which propagates to the segmentation output. A diameter of 7 does not help either and even scores slightly below the diameter of 5 retained for OmniSeg. This drop may be related to a suboptimal pairing with the temperature: T is kept fixed across neighborhood sizes, while it normalizes a $\mathrm{Softmax}$ over $N = |\mathcal{N}|$ neighbors, whose number grows quadratically with the diameter. A wider neighborhood therefore spreads the attention over more candidate matches, some of them far from the query pixel, which makes the resulting DVF noisier, an effect comparable to the one induced by the temperature alone in Fig. 17. A joint search over the diameter and the temperature would likely be needed to benefit from wider neighborhoods, which we leave for future work. Finally, the complexity of the model is quadratic in the diameter of the neighborhood, so increasing it comes at a significant computational cost.

Table 12: Average Dice for different diameters of $\mathcal{N}$.

Diameter	**Dice** ↑
3	74.2
5	**75.5**
7	74.8

E.6 Qualitative analysis of the dense context pooling

Fig. 19 illustrates that the pixel-wise similarity between query and context representations evolves across the SpotBlocks of the decoder. A region in the context maps may have a high similarity with the query at a given scale but lower in others. This figure also reveals the specific behavior of the last SpotBlock. It seems to play the role of selecting the overall most relevant context examples, the similarity being more balanced across the context examples in the deeper blocks.

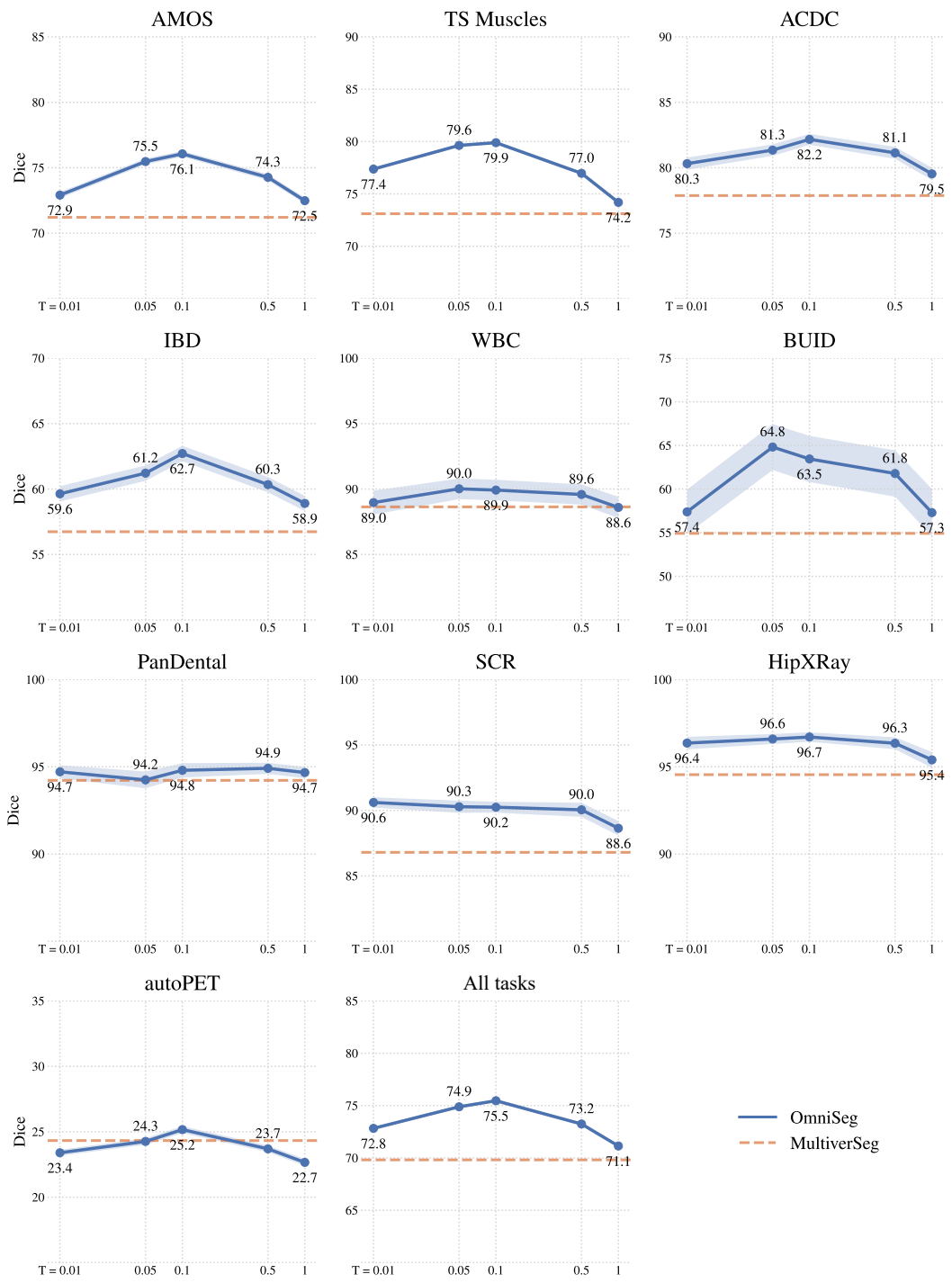

Figure 18: **Impact of temperature on ICS performance across datasets**. Models are trained for two epochs with varying temperatures, selecting $T = 0.1$ for the final model based on held-out validation.

E.7 In-context segmentation vs. supervised methods

In this section, we describe the experimental settings used to assess how many training samples a fully supervised model needs to outperform an in-context segmentation model. We first split each task

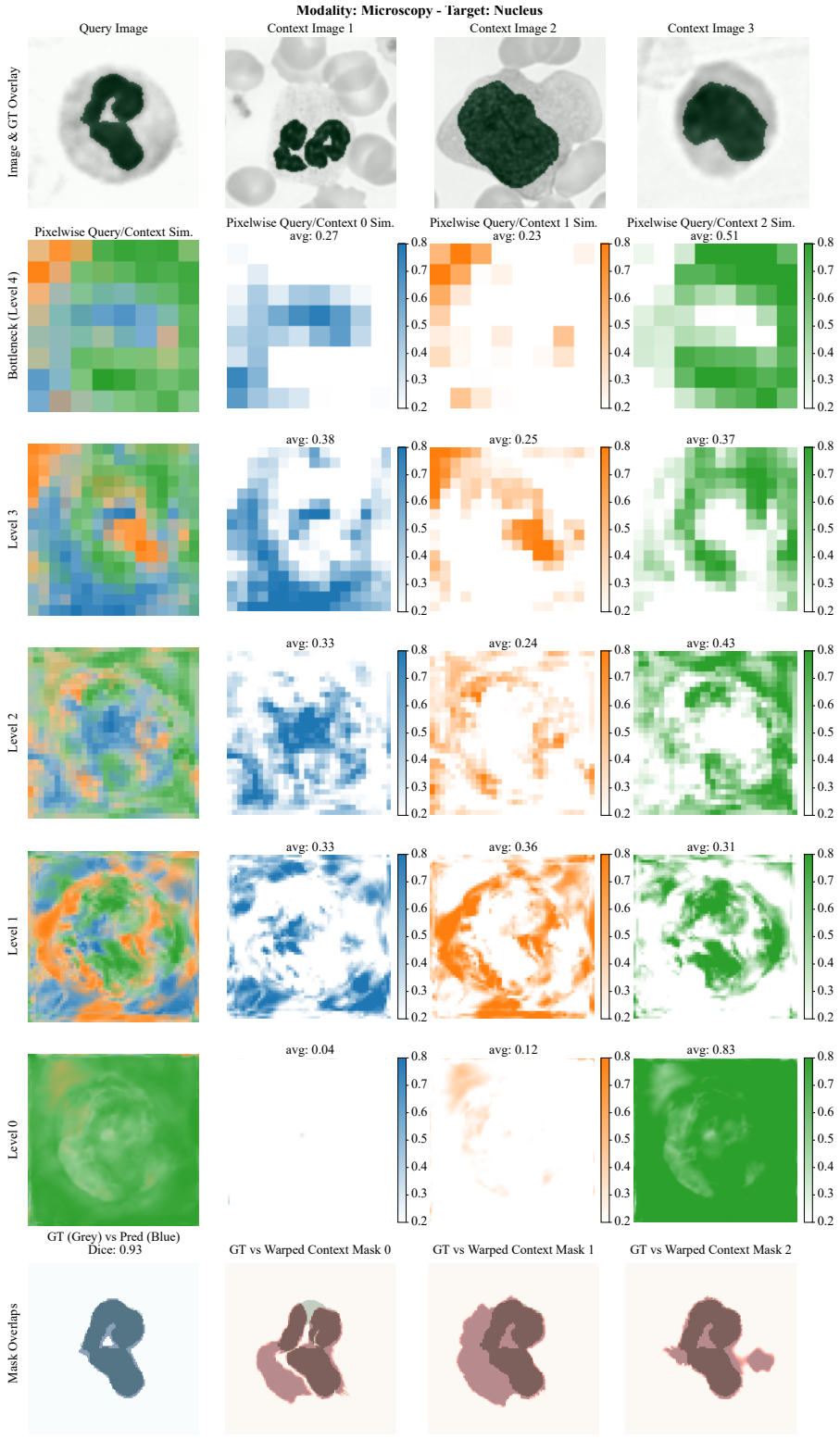

Figure 19: **Dense query/context attention maps.** The figure shows the dense attention maps $A^{\mathcal{C}}$ between query pixels and corresponding context pixels used by SpotBlocks to perform the Dense Context Pooling. Each context sample is arbitrarily associated with a color. The first column is the superposition of the similarity maps per context sample.

Table 13: Model complexity and inference efficiency for batch size 1, context size 16, and input size 128×128 on a single H100 GPU.

Model	# Params (M)	Inference time (ms)	FLOPs (G)
UniverSeg	1.18	3.89 ± 0.37	54
Tyche-TS	1.70	5.10 ± 0.03	91
MultiverSeg	18.90	29.81 ± 0.26	869
INSID3	303.23	18.75 ± 1.03	533
SegGPT	370.72	416.75 ± 1.63	20 980
Ours	23.50	44.66 ± 0.13	1230

subset of our test datasets into 80% train and 20% test splits. To estimate a supervised upper bound for each task, we train nnU-Net on the full 80% split and evaluate it on the remaining 20%. Then, to compare in-context segmentation models and a supervised baseline in a few-shot regime, we construct nested subsets of randomly selected samples from the 80% split, of sizes $S = 16, 32, 64, 128$. These subsets are used as training sets for nnU-Net and as context sets for OmniSeg and MultiverSeg, and the same 20% split is used for testing. We train nnU-Net for 250 epochs for the supervised upper bound and for 100 epochs in the few-shot regime to avoid overfitting. As discussed in the main body, OmniSeg offers usable segmentation results in a few-shot regime where supervised methods struggle, which can prove useful in clinical and research contexts. We report in Fig. 20 the results on our 10 held-out test datasets and the global average. nnU-Net also presents a larger 99% CI, indicative of failure cases, while OmniSeg shows more stable performance. Indeed, nnU-Net even struggles on in-distribution cases because of the long-tail distribution of medical datasets: each patient presents a different anatomy or pathology. OmniSeg, on the other hand, benefits from the diversity of its large pretraining dataset and learned to ground its predictions in the context examples thanks to its architecture and supervision. This experiment shows that in-context segmentation methods are effective out-of-the-box universal models for low-annotation regimes.

E.8 Computational cost and model size

We report in Table 13 parameter counts and inference costs for baselines and our model. Our model yields better performance than state-of-the-art (SOTA) methods at a moderate computational overhead.

E.9 Additional visualizations

We provide in Fig. 21 the visual predictions for our model and two baselines for one example from each held-out test dataset, with context set size $S = 16$. OmniSeg is less prone to smudges thanks to its stronger spatial awareness.

F General discussion

F.1 Societal impact and fairness

By requiring only a handful of context examples rather than large annotated cohorts, OmniSeg could lower the annotation burden for clinical deployment in settings where expert labels are scarce. Being a multi-task model, OmniSeg can be leveraged for many use cases while capitalizing on shared user experiences and integration processes. Realizing this benefit broadly depends on how well the model generalizes beyond its training distribution. As detailed in Appendix C, our dataset, though large and spanning 8 modalities, is compiled from publicly available sources acquired predominantly in North American, Western European, and East Asian hospitals, and largely from adult patients. While our held-out evaluation (Section 4.2) shows that OmniSeg generalizes to unseen sources, targets, and even an unseen modality, whether this generalization extends to other demographics, age groups, or acquisition protocols typical of lower-resource settings remains untested. Given that the reduced-annotation benefit of ICS would be most valuable precisely in such under-represented

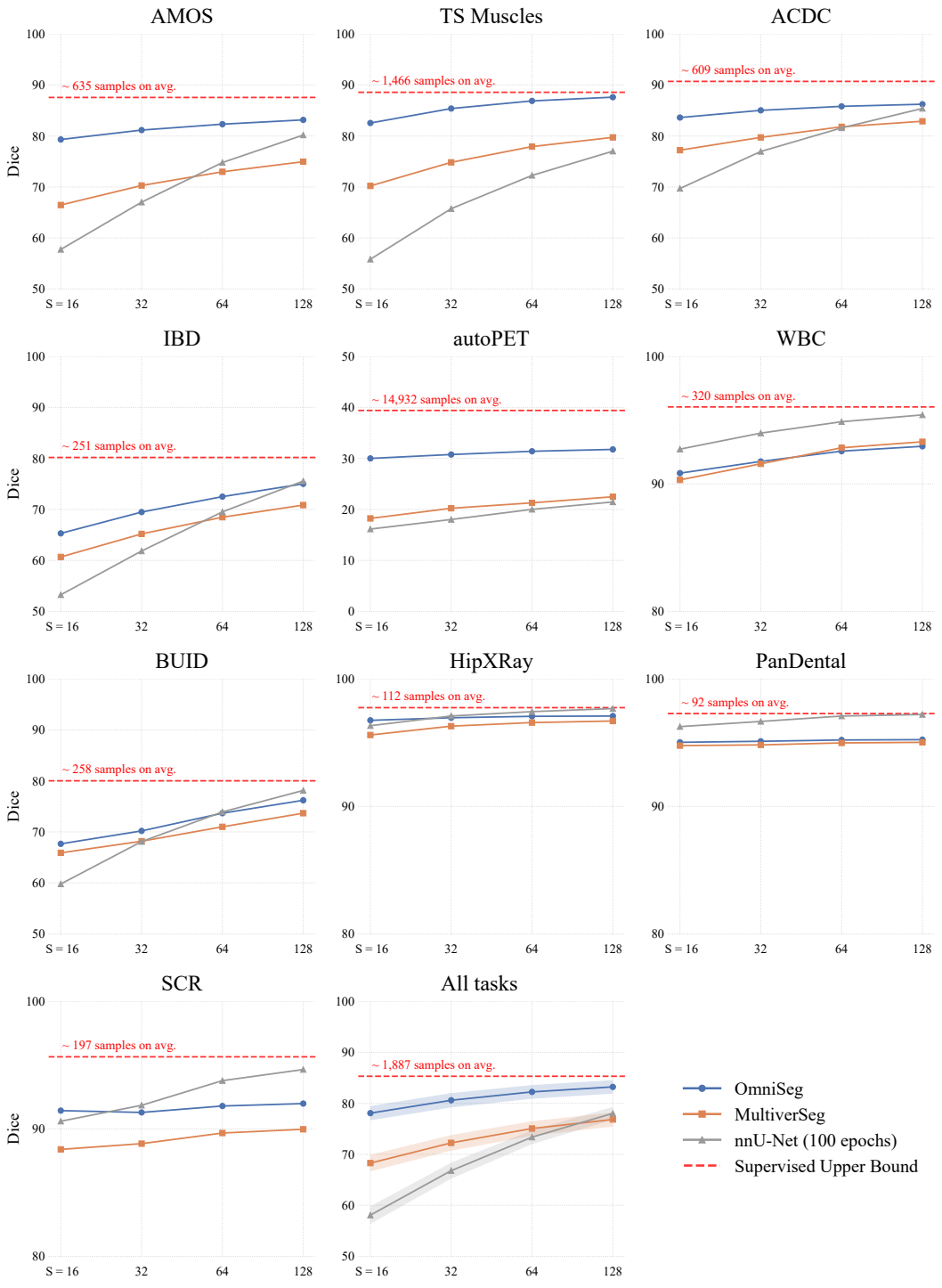

Figure 20: **Supervised vs. in-context segmentation.** Subsets of size $S = 16, 32, 64, 128$ are used as training sets for nnU-Net and as context sets for OmniSeg.

settings, closing this evaluation gap is an important direction for future work to ensure the technology benefits patients equitably.

F.2 Environmental impact

We coarsely estimate the total environmental footprint of the computations for this project at 3 t CO_2e, based on the energy performance of the cluster we used. This includes all training and experimentation related to the project, including preliminary runs not reported in the final paper. For comparison, this roughly equates to two round-trip intercontinental flights[1].

F.3 Limitations

This study is restricted to 2D binary image segmentation at a resolution of 128×128, following previous work. Adaptations are probably needed to port this method to 3D, given the heavy computational cost induced by the local attention along this extra dimension. Furthermore, OmniSeg adopts a registration paradigm to boost universal ICS performance. While this implicitly exploits the consistency of shapes across medical segmentation tasks, this approach may not be well-suited when the topologies of the anatomical targets vary substantially across subjects.

[1] https://icec.icao.int/calculator

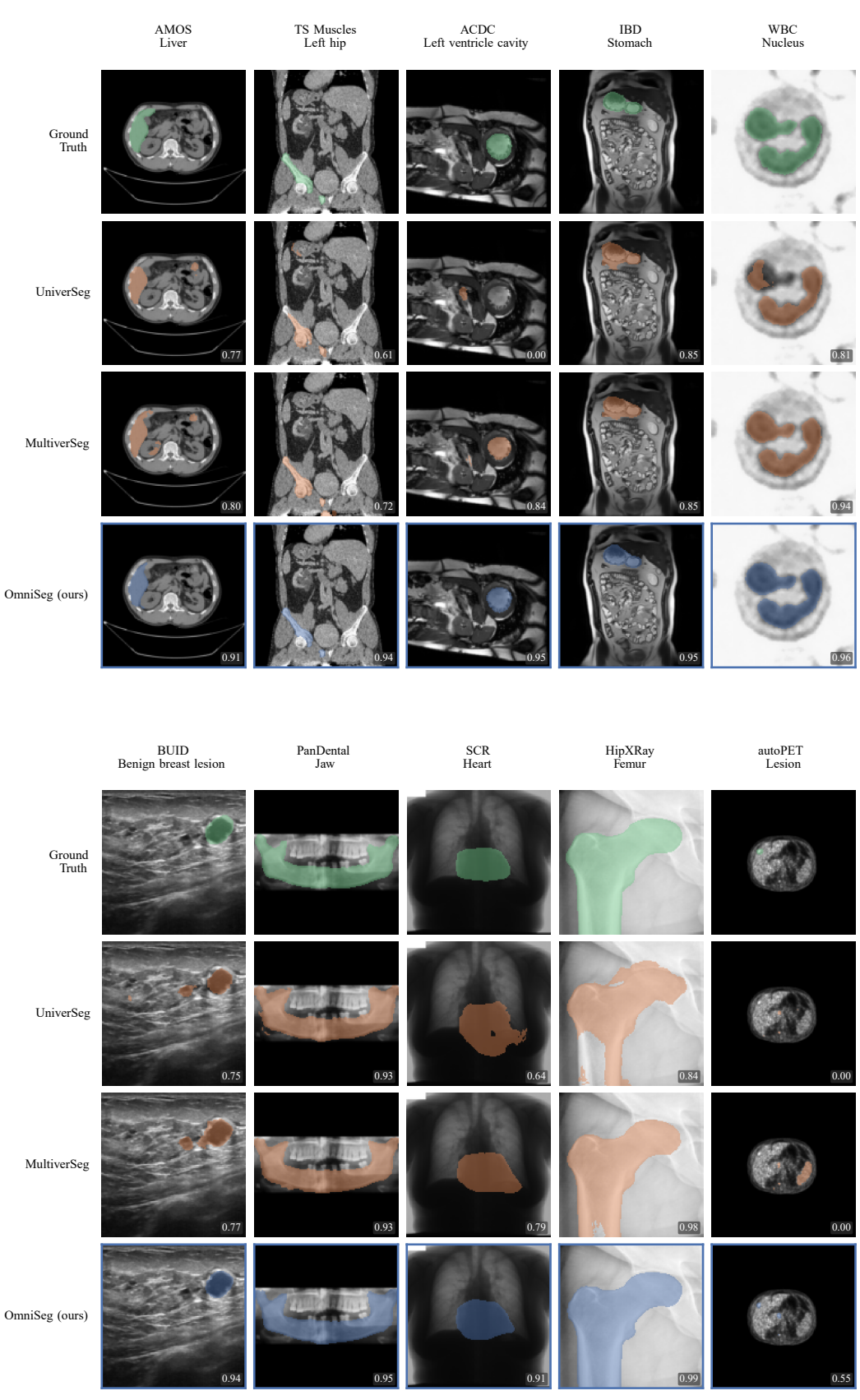

Figure 21: **Visual comparison of OmniSeg vs. baselines.**