Few-Shot 3D Volumetric Segmentation with Multi-Surrogate Fusion

Meng Zheng[*1], Benjamin Planche[1], Zhongpai Gao[1],
Terrence Chen[1], Richard J. Radke[2], and Ziyan Wu[1]

[1] United Imaging Intelligence, Boston, MA, USA
{first.last}@uii-ai.com
[2] Rensselaer Polytechnic Institute, Troy, NY, USA
rjradke@ecse.rpi.edu

Abstract. Conventional 3D medical image segmentation methods typically require learning heavy 3D networks (*e.g.*, 3D-UNet), as well as large amounts of in-domain data with accurate pixel/voxel-level labels to avoid overfitting. These solutions are thus extremely time- and labor-expensive, but also may easily fail to generalize to unseen objects during training. To alleviate this issue, we present MSFSeg, a novel few-shot 3D segmentation framework with a lightweight multi-surrogate fusion (MSF). MSFSeg is able to automatically segment unseen 3D objects/organs (during training) provided with one or a few annotated 2D slices or 3D sequence segments, via learning dense query-support organ/lesion anatomy correlations across patient populations. Our proposed MSF module mines comprehensive and diversified morphology correlations between unlabeled and the few labeled slices/sequences through multiple designated surrogates, making it able to generate accurate cross-domain 3D segmentation masks given annotated slices or sequences. We demonstrate the effectiveness of our proposed framework by showing superior performance on conventional few-shot segmentation benchmarks compared to prior art, and remarkable cross-domain cross-volume segmentation performance on proprietary 3D segmentation datasets for challenging entities, *i.e.*, tubular structures, with only limited 2D or 3D labels.

Keywords: 3D Medical Segmentation · Few Shot Segmentation.

1 Introduction

Volumetric/3D medical image segmentation is crucial to various clinical applications, from treatment planning to disease monitoring [4, 32, 33], and has been benefiting from advances in data-centric machine-learning solutions. However, densely annotating the large-scale volumetric data required for their fully-supervised training is extremely time/labor-consuming, and the generalizability of these models to unseen object classes is uncertain.
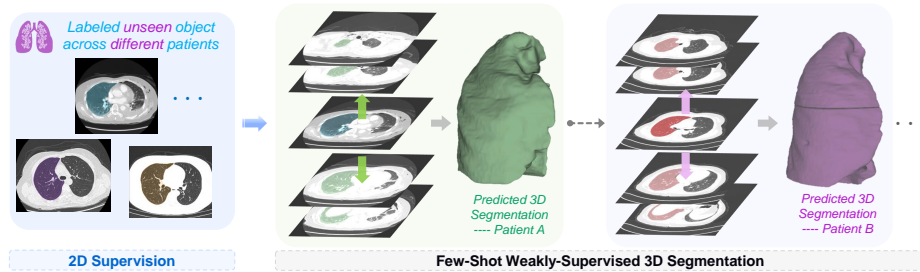
---

[*] Corresponding author.

Fig. 1: Proposed few-shot weakly-supervised 3D segmentation pipeline.

To tackle these limitations, the task of Few Shot Segmentation (FSS) has gained traction [2,3,8,14,16,17,20,21,24,25,30,40]. FSS aims to develop domain-generalizable neural networks capable of segmenting previously unseen objects from new *query* images/slices with minimal annotated examples (*i.e.*, *support* images/slices and masks), improving data annotation efficiency.

Efforts to tackle FSS tasks have focused on both class-wise [7,38,39] and pixel-wise [12,18,26,29,37] prototyping architectures that rely on computing similarity/affinity correlation between support and query data. However, few-shot 3D segmentation, which involves segmenting video or volumetric data with limited labeled 2D/3D images [10,13,24], remains underexplored as proposed networks need fully-labeled 3D supervisions, compromising generalizability and computation efficiency. Existing few-shot 3D-segmentation solutions [2,19,20,24,31] often rely on 2D FSS networks that treat each slice individually, neglecting patient-specific inter-slice morphology correlations crucial for accurate 3D mask prediction (*c.f.* useful clues and strong constraints). Moreover, they typically only consider *1-shot* cases [2,8,19,20,31] and overlook multi-support ones (*n-shot*), or use generic aggregation methods (*e.g.*, ensemble [36] or voting strategies [3,18,25]), which may compromise diverse morphology correlations across support image/mask or sequence/mask pairs. Effective aggregation of information from multiple support images or sequences is crucial for few-shot 3D segmentation, as 3D volumes contain spatially correlated 2D slices, and different patient volumes exhibit distinct morphology correlations. Utilizing both consistent and diversified inter-slice/mask morphology variations across patients, given limited labeled support, is key to solving this problem.

We thus introduce **MSFSeg**, a novel few-shot 3D segmentation pipeline with a Multi-Surrogate Fusion (MSF)-informed few-shot network (*c.f.* Fig. 1). MS-FSeg can segment unseen 3D objects with just a few support 2D slices or sequences. It accommodates various support formats and numbers (*n*-shot, $n \geqslant 1$), *i.e.*, taking 2D slices or sequential slice segments from different scans as support. Exploring dense foreground/background pixel relationships and mining semantic features across multiple supports from diverse patient data distributions, MS-FSeg enhances query mask prediction. We evaluate MSFSeg on conventional abdominal CT and MRI organ segmentation benchmarks, comparing it to prior FSS methods. Additionally, we demonstrate its effectiveness on few-shot weakly-

supervised 3D segmentation across 4 organs and 40 data volumes collected from a hospital. To summarize, our key contributions include:

– We introduce a new few-shot 3D segmentation pipeline for unseen 3D objects, given limited labeled 2D or 3D supports as weak supervision. It leverages intra- and inter-patient/volume morphology correlations to better predict 3D masks.

– We design a novel FSS network with multi-surrogate fusion to exhaustively explore and mine complementary/diversified semantic information across various support slices/sequences, resulting in superior accuracy in transferred domains.

– We demonstrate the effectiveness and generalizability of MSFSeg, sharing state-of-the-art FSS performance on medical benchmarks, as well as satisfying few-shot 3D segmentation performance on clinical data, particularly in segmenting challenging objects like tubular structures in cross-domain settings.

2 Methodology

2.1 MSFSeg with Multi-Surrogate Fusion

Considering the task of segmenting unseen objects in 3D medical images given a few support slices or sequences (akin to transfer learning), we base the proposed MSFSeg on existing few-shot segmentation pipelines [18,24], which takes one 2D query image $\mathbf{I}^q \in \mathbb{R}^{h \times w}$ of size $h \times w$, n support sequences $\{\mathbf{I}_i^s \in \mathbb{R}^{d_i \times h \times w}\}_{i=1}^n$ (d_i is the number of slices in each sequence, $d_i = 1$ is same as conventional 2D FSS where supports are single slices), along with their corresponding labeled segmentation masks $\{\mathbf{M}_i^s \in \mathbb{R}^{d_i \times h \times w}\}_{i=1}^n$ to infer the mask $\mathbf{M}^q$ of the query image. Unlike prior art that considers only one support (1-shot) [2,20,24,27,31,35] or applies vanilla aggregation strategies for n-shot cases at inference time only [18,25], our solution takes n support slices at training time for end-to-end optimization, leveraging the proposed multi-surrogate fusion. Inherently optimized for leveraging diverse multi-support information, our proposed MSFSeg can effectively segment unseen 3D objects across different patients (query volumes).

A) Feature Extraction and Self-Attention. We follow existing 1-shot FSS pipelines [2,6,20,24,27,31,35] and apply a pretrained ResNet-101 [5] to extract query and support image features. We collect the sets of multi-scale query and support feature maps, $\{\mathbf{f}_j^q\}_{j=1}^b$ and $\{\mathbf{f}_{1..N,j}^s\}_{j=1}^b$ ($N = \sum_{i=1}^n d_i$ is the total number of slices from n supports), out of the b ResNet blocks, from query image $\mathbf{I}^q$ and support sequences $\mathbf{I}_{1..n}^s$ respectively. At each feature scale level, we then compute the multi-head attention (c.f. Transformer strategy [28]) to correlate query and support pixel-wise morphology and predict n coarse query mask features $\{\hat{\mathbf{M}}_{1..n,j}^q\}_{j=1}^b$ (at b different scales) given query image and support image/mask pairs. Specifically, each support mask $\mathbf{M}_i^s$, used as attention value vector, is first downsampled to the same size $h_j \times w_j$ as the considered feature map (c.f. scaling by j^{th} ResNet block resulting in $h_j \times w_j \times c_j$ features) and then flattened into a 1-dimensional vector to form $V_{i,j}$. We similarly resample the feature maps $\mathbf{f}_j^q$ and $\mathbf{f}_{1..n,j}^s$ into $(h_j * w_j) \times c_j$ matrices, with each pixel treated as a token. Adding positional encoding and linear projections, we respectively get the attention query Q_j and key $K_{i,j}$ matrices. Note that d_i support feature maps and masks from the same support sequence $\mathbf{I}_i^s$ are concatenated and treated as different tokens
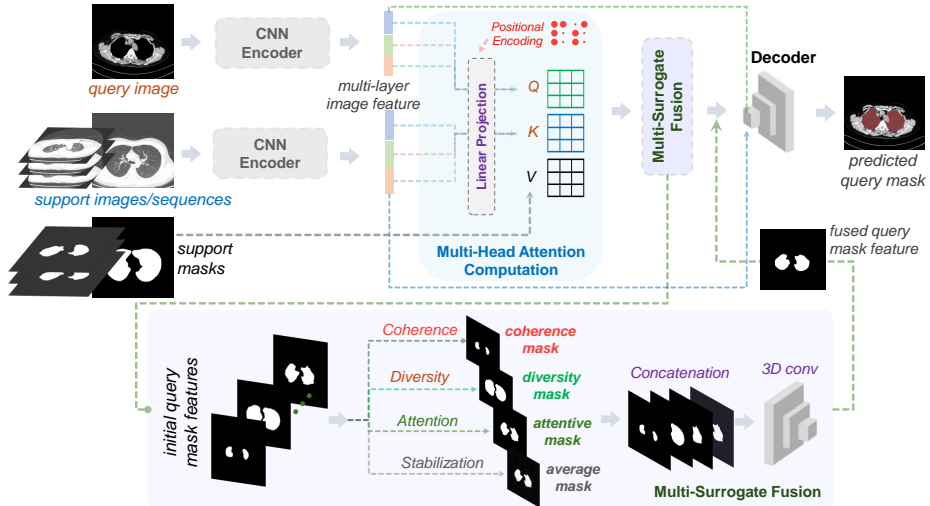
Fig. 2: Proposed MSF solution for few-shot 3D segmentation.

to form the key and value matrices. This way, the following multi-head attention mechanism can internally aggregate token-wise information across d_i slices via scaled dot-product attention and enable flexible support sequence length.

The multi-head attention is finally computed for each support pair i and scale j to generate $n \times b$ initial multi-scale query masks $\{\hat{\mathbf{M}}^{\mathrm{q}}_{1..n,j}\}_{j=1}^{b}$: Attention$(Q_j,$ $K_{i,j}, V_{i,j}) = $ Softmax$\left(Q_j K_{i,j}^{\intercal}/\sqrt{d}\right)V_{i,j}$, with d dimension of Q_j. The intuition behind is that by multiplying Q and $K^{\intercal}$, pixel-wise correlations between query and support images are densely captured, to enhance and weigh the support mask towards predicting an accurate query mask. The inner product of Q and $K^{\intercal}$ acts like a projection to bridge the pixel-wise image correlations in support-query data with the mask correlations. We finally aggregate the b multi-scale results via upsampling and summing, followed by cascaded 2D convolutions (details in appendix). This produces n multi-scale-informed query mask features $\hat{\mathbf{M}}^{\mathrm{q}}_{1..n}$.

B) Fusion via Mixture of Surrogates. Efficiently and effectively utilizing multi-support information has been a challenging, long-standing problem, especially for cross-domain cross-patient/volume segmentation in few-shot setting, where learning consistent and patient-specific morphology correlations across different patient data could greatly benefit FSS pipelines in generalizing better to unseen patient data and objects. We thus propose a Multi-Surrogate Fusion (MSF) module (Fig. 2, bottom) based on a ResNet-101 backbone, to effectively distill complementary morphology information across multiple supports.

Specifically, our MSF takes the n multi-scale query mask features $\hat{\mathbf{M}}^{\mathrm{q}}_{1..n}$, and assign 4 specialized surrogates to explore and undermine local and global morphology information that is (1) coherent, (2) diverse, (3) cross-attentive, and (4) stabilized across n query mask features predicted from n different support images/masks. Specifically, we deploy a coherence surrogate to distill consistent information by performing soft intersection operation (element-wise product of

soft-maxed masks) to generate coherence masks $\hat{\mathbf{M}}^{q}_{coh} = \bigcap_{i=1}^{n} \hat{\mathbf{M}}^{q}_{i}$, and extract diversified information by performing soft union operation over n query mask features (sum of soft-maxed masks) to generate diversity masks, *i.e.* $\hat{\mathbf{M}}^{q}_{div} = \bigcup_{i=1}^{n} \hat{\mathbf{M}}^{q}_{i}$. We then perform channel-wise attention across all query mask features, each query treated as one channel, aiming to find attentive masks:

$$\mathbf{M}^{q}_{ca} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{M}}^{q}_{i} \otimes \mathrm{Softmax}\left(\mathrm{AvgPool3d}(\hat{\mathbf{M}}^{q}_{i})\right), \tag{1}$$

where "$\otimes$" is the element-wise product. $\mathbf{M}^{q}_{ca}$ extracts and fuses complementary foreground information while suppressing distracting background information from multiple experts. We also average over n query mask features, *i.e.* $\mathbf{M}^{q}_{avg} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{M}}^{q}_{i}$, to obtain stabilized global information across $\hat{\mathbf{M}}^{q}_{1..n}$. Finally, the coherence, diversity, attentive, and average masks—resp. $\hat{\mathbf{M}}^{q}_{coh}$, $\hat{\mathbf{M}}^{q}_{div}$, $\hat{\mathbf{M}}^{q}_{ca}$, $\hat{\mathbf{M}}^{q}_{avg}$—are concatenated together channel-wise and input into a 3D conv. layer to fuse them into query mask features. While our coherence and diversity surrogates aim to explore complementary local, pixel-wise morphology query-support and support-support correlations, our attention and stability surrogates distill global information across supports for improved query mask prediction.

C) Mask Upsampling and Refinement. We finally concatenate the fused query mask to the image features $\mathbf{f}^{q}_{1}$ and $\mathbf{f}^{s}_{1..n,1}$ (skip-connection operations similar to U-Net [23]) and send them through two deconvolutional blocks (decoder) to refine and upsample into the final $h \times w$ query mask prediction.

2.2   Few-Shot 3D Segmentation Workflow

Leveraging MSFSeg (Sec. 2.1), we define an automated few-shot 3D segmentation workflow, explained in Alg. 1. Specifically, given B 3D volumes with an unseen object unlabeled, we first annotate one or a few slices/sequences (from one or different data volumes) as supports, then segment each of the B data volumes by sending the query slice and labeled support slices/sequences to our MSFSeg for few-shot 3D segmentation. Note that we are able to constantly add segmented slices to the support pool with quality control filtering, to better segment the unlabeled slices/volumes by selecting reliable n support slices (*i.e.*, support slices with high-level features extracted from ResNet101 backbone which have highest cosine similarity to the query feature) out of the support pool.

3   Experiments

We evaluate MSFSeg through various experiments. First, we apply MSFSeg to conventional few-shot segmentation (FSS) with 1-shot setting, in order to compare to prior art, as methods inherently tackling 3D few-shot segmentation are lacking. For 1-shot case, we randomly apply data augmentation techniques, e.g. horizontal flipping, affine transformations or color jittering and input "n" number of augmented support images (as pseudo n-shot case) to explore comprehensive morphology information from the single support. For 3D FSS, we highlight the

---

Algorithm 1 MSFSeg-based few-shot 3D segmentation workflow.

Input Unseen object, unlabeled B 3D volumes $\{V_k\}_{k=1}^B$, empty support pool $\mathcal{P}$.
Output B labeled 3D volumes.
 1: Label n sequences $\mathbf{I}_{1..n}^s$ from $\{V_k\}_{k=1}^B$ as exemplar/support, and generate label $\mathbf{M}_{1..n}^s$.
 2: Add labeled slices to support pool, $\mathcal{P} \leftarrow (\mathbf{I}_{1..n}^s, \mathbf{M}_{1..n}^s)$.
 3: **for** volume $V_k \in V_{1..B}$ **do**
 4:      **for** 2D slice $\mathbf{I}_{k,i}^q \in$ volume V_k **do**
 5:          Identify n most similar support slices from pool: $(\mathbf{I}_{1..n}^s, \mathbf{M}_{1..n}^s) \in \mathcal{P}$.
 6:          Generate label for $\mathbf{I}_{k,i}^q$: $\mathbf{M}_{k,i}^q = \text{MSFSeg}(\mathbf{I}_{k,i}^q, \mathbf{I}_{1..n}^s, \mathbf{M}_{1..n}^s)$.
 7:      **end for**
 8:      Add labeled volume $\hat{V}_k$ to support pool: $\mathcal{P} \leftarrow \hat{V}_k$.
 9: **end for**

---

satisfying performance of MSFSeg on our proprietary and challenging dataset of anatomical tubular structures.

3.1   Evaluation Setup and Protocols

Implementation Details. We apply a ResNet-101 [5] pretrained on ImageNet-1K [1] as feature extractor for fair comparison with prior art, with input image size set to 384×384. Please see supplementary material for more architecture and implementation details, experimental results and model parameter analysis.
Datasets. To evaluate conventional FSS on medical scans, we follow the same train/test split and evaluation protocol as [2, 20, 24], and train our network on Abdomen-CT [11] and CHAOS-MRI [9] datasets. Accordingly, we test our proposed MSFSeg in 2 different settings, with *setting 1* using all training/testing data, and *setting 2* enforcing testing classes to be completely unseen by removing any training image that contains a testing class object. While state-of-the-art FSS methods [2,20,31] typically only consider 1-shot setting, we further perform 5-shot evaluations on Abdomen-CT and CHAOS-MRI, randomly selecting 5 support slices across the dataset. For few-shot 3D segmentation on our proprietary dataset, we train MSFSeg on domain-irrelevant natural [15] and medical [9,11] datasets. Our collected test set contains 40 CT data volumes collected from hospitals, with 4 objects of interest, *i.e.*, *hepatic artery*, *renal artery*, *renal vein*, and *carotid artery*, manually annotated for ground-truth. Note that these objects of interest are all challengingly small tubular structures, unseen during training.
Evaluation Metrics. We report Dice scores [2,19,20,31] on Abdomen-CT/CHAOS-MRI for comparison to prior FSS methods on medical images. We also report Region Jaccard ($\mathcal{J}$), Boundary F measure ($\mathcal{F}$), and their average ($\mathcal{J}\&\mathcal{F}$) additional to Dice scores, commonly used as video object segmentation metrics [22,34] and are thus suitable for few-shot 3D segmentation evaluation. We report the average over 5 runs for all our experimental results.

3.2   Results and Discussion

Conventional 1-/5-shot Segmentation. Tab. 1 highlights the superiority of our solution compared to the state-of-the-art for 1-shot FSS tasks, in terms of

Table 1: Comparison to prior art (Dice score) on various datasets and settings. * indicates method using 3D scribble inputs and full 3D labels for network training. Evaluations are performed on left kidney(LK), right kidney(RK), spleen, liver.

	Methods	Super-pixel or -voxel	Abdomen-CT [11]					CHAOS-MRI [9]				
			LK	RK	Spleen	Liver	avg	LK	RK	Spleen	Liver	avg
Setting 1	SE-Net [24]	✗	24.42	12.51	43.66	35.42	29.00	45.78	47.96	47.30	29.02	42.51
	Vanilla PANet [30]	✗	20.67	21.19	36.04	49.55	31.86	30.99	32.19	40.58	50.40	38.53
	ALPNet [20]	✗	29.12	31.32	41.00	65.07	41.63	44.73	48.42	49.61	62.35	51.28
	VQV [8]	✗	-	-	-	-	-	60.03	68.94	79.08	81.72	72.44
	SSL-PANet [20]	✓	56.52	50.42	55.72	60.86	57.88	58.83	60.81	61.32	71.73	63.17
	SSL-ALPNet [20]	✓	72.36	71.81	70.96	78.29	73.35	81.92	85.18	72.18	76.10	78.84
	CRAPNet [2]	✓	74.69	74.18	70.37	75.41	73.66	81.95	86.42	74.32	76.46	79.79
	RPT [40]	✓	77.05	79.13	72.58	82.57	77.83	80.72	89.82	76.37	**82.86**	82.44
	SSL-VQV [8]	✓	-	-	-	-	-	**89.54**	**91.56**	77.21	79.92	**84.56**
	PRNet* [13] (1 scribble)	✗	-	-	-	-	-	85.10	82.30	83.50	78.10	82.30
	Ours − 1-shot (1 slice)	✗	**81.11**	**78.41**	**73.64**	**78.91**	**78.02**	84.18	88.10	77.12	76.11	81.38
	Ours − 5-shot (5 slices)	✗	**87.22**	**85.62**	**82.71**	82.57	**84.53**	88.63	90.94	**82.73**	82.10	**86.10**
Setting 2	ALPNet-init [20]	✗	13.90	11.61	16.39	41.71	20.90	19.28	14.93	23.76	37.73	23.93
	SE-Net [24]	✗	32.83	14.34	0.23	0.27	11.91	62.11	61.32	51.80	27.43	50.66
	Vanilla PANet [30]	✗	32.34	17.37	29.59	38.42	29.43	53.45	38.64	50.90	42.26	46.33
	ALPNet [20]	✗	34.96	30.40	27.73	43.37	35.11	53.21	58.99	52.18	37.32	50.43
	CAT [19]	✗	-	-	-	-	-	74.92	71.89	75.43	64.57	71.70
	CAT-Net [16]	✗	63.36	60.05	67.65	75.31	66.59	-	-	-	-	-
	SSL-PANet [20]	✓	37.58	34.69	43.73	61.71	44.42	47.71	47.95	58.73	64.99	54.85
	SSL-ALPNet [20]	✓	63.34	54.82	60.25	73.65	63.02	73.63	78.39	67.02	73.05	73.02
	SSL-RPNet [27]	✓	65.14	66.73	64.01	72.99	67.22	71.46	81.96	73.55	75.99	75.74
	CRAPNet [2]	✓	70.91	67.33	70.17	70.45	69.72	74.66	82.77	70.82	73.82	75.52
	RPT [40]	✓	72.99	67.73	70.80	75.24	71.69	78.33	86.01	75.46	**76.37**	79.04
	SRCL [31]	✓	67.39	63.37	67.36	73.63	67.94	77.07	84.24	73.73	75.55	77.65
	Ours − 1-shot (1 slice)	✗	**79.24**	**77.36**	**75.21**	**76.73**	**77.14**	**82.83**	**86.98**	**78.07**	76.14	**81.01**
	Ours − 5-shot (5 slices)	✗	**85.73**	**84.51**	**81.60**	**81.22**	**83.27**	**87.70**	**90.62**	**81.97**	**82.52**	**85.70**

Dice scores, for nearly all 4 organs over both benchmarks and settings. Specifically, our proposed 1-shot solution outperforms existing solutions with a reasonable margin under setting 2, *e.g.*, +5.45% and +1.97% on Abdomen-CT and CHAOS-MRI dataset respectively over [40] (2nd-best) on average Dice score over 4 organs. Note that numerous existing methods (with ✓) in Tab.1 relies on an off-the-shelf superpixel or supervoxel model as auxiliary network to pre-group image pixels. Our proposed MSFSeg, *without using any prior knowledge*, still outperforms them on Abdomen-CT (setting 1 and 2) and CHAOS-MRI (setting 2) datasets. Additionally, Tab. 1 and Fig. 3 presents 5-shot FSS results (5 support slices for each query) on Abdomen-CT and CHAOS-MRI, highlighting the ability of our proposed MSF to learn comprehensive query-support morphology correlations across multiple supports (*e.g.*, +6.51/+6.13% average Dice score improvement for 5-shot over 1-shot on Abdomen-CT, for setting 1/2 respectively). Specifically, compared to the few-shot 3D segmentation method PRNet [13], which requires 3D scribbles on data volumes, full 3D labels as data pairs for training, and a heavier architecture; MSFSeg performs significantly better with 5 support 2D slices and comparably with 1 support slice on CHAOS-MRI.

Few-Shot 3D Segmentation. We evaluate our few-shot 3D segmentation workflow on both intra-volume and inter-volume cases and report the scores over all data volumes for each object of interest in Tab. 2. To contrast MSFSeg's superiority, we apply a state-of-the-art FSS method, VAT [6] (retrained/evaluated

Table 2: Evaluation of intra-/inter-volume 3D segmentation on challenging tubular targets in proprietary CT data, compared to baseline [6].

| Target↓ \ n-shot propag.	Metrics→	$\mathcal{J}$				$\mathcal{F}$				$\mathcal{J}\&\mathcal{F}$				Dice			
		baseline	$n{=}1$	$n{=}5$	$n{=}5^2$	base.	$n{=}1$	$n{=}5$	$n{=}5^2$	base.	$n{=}1$	$n{=}5$	$n{=}5^2$	base.	$n{=}1$	$n{=}5$	$n{=}5^2$
Hep. A.	intra-vol.	.293	.559	.629	.640	.394	.656	.672	.678	.344	.607	.651	.659	.393	.709	.757	.767
	inter-vol.	.283	.580	.625	.635	.387	.634	.662	.667	.335	.606	.643	.651	.383	.688	.753	.760
Ren. A.	intra-vol.	.155	.610	.608	.647	.266	.745	.814	.830	.210	.677	.711	.738	.212	.678	.726	.763
	inter-vol.	.137	.510	.580	.638	.248	.780	.798	.826	.192	.645	.689	.732	.193	.635	.696	.753
Ren. V.	intra-vol.	.137	.395	.487	.519	.287	.569	.700	.706	.212	.482	.593	.612	.213	.508	.624	.655
	inter-vol.	.128	.348	.502	.491	.277	.595	.686	.697	.203	.471	.594	.594	.203	.474	.608	.627
Car. A.	intra-vol.	.233	.393	.430	.479	.313	.554	.570	.609	.273	.473	.500	.544	.295	.491	.536	.583
	inter-vol.	.232	.389	.393	.459	.312	.516	.552	.599	.272	.452	.472	.529	.294	.447	.501	.569
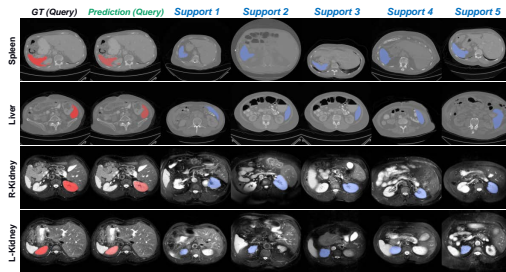
Fig. 3: Qualitative results of 5-shot FSS.

Table 3: MSF ablation study on CHAOS-MRI (5-shot, setting 2).

$\hat{\mathbf{M}}_{coh}^q$	$\hat{\mathbf{M}}_{div}^q$	$\hat{\mathbf{M}}_{ca}^q$	$\hat{\mathbf{M}}_{avg}^q$	LK	RK	Spleen	Liver
✓	✗	✗	✗	84.81	88.64	73.94	81.47
✗	✓	✗	✗	81.11	87.16	71.58	82.14
✓	✓	✗	✗	85.74	90.20	78.75	82.33
✗	✗	✓	✗	86.28	89.46	81.14	70.66
✗	✗	✗	✓	87.12	89.78	79.85	79.33
✗	✗	✓	✓	87.17	90.11	81.30	81.90
✓	✓	✓	✓	**87.70**	**90.62**	**81.97**	**82.52**

with same training/testing splits as our protocol), as baseline. For 1-shot intra-volume inference, we consider the central slice and corresponding ground-truth label as support and sequentially infer the other masks in the volume, whereas for 5-shot segmentation, we randomly select 5 support slices for each query slice. We further demonstrate MSFSeg ability to process 3D/sequential supports by randomly selecting 5 sequences as support, each sequence containing 5 consecutive slices ("$n{=}5^2$" in Tab. 2). The inter-volume segmentation is performed as described in Alg. 1, using 1 or 5 most similar sequences in the labeled pool. Despite its similar backbone, VAT cannot properly propagate the fine labels of new tubular objects, whereas our method yields satisfying results w.r.t. both intra- and inter-volume 3D segmentation thanks to its fusion module and cross-volume workflow. Tab. 2 also shows how MSF leverages additional support data for improved accuracy ($\geq 6\%$ mean Dice score increase from $n = 1$ to 5). The small margin performance gap between intra- and inter-volume segmentation also proves the generalizability of MSFSeg; *i.e.*, support across sequences from different patient scans does not compromise segmentation results compared to label support from ground truth (user).

Ablation Study of Multi-Surrogate Fusion. We demonstrate the role of each MSF surrogate (coherence, diverse, attention and stabilization operations) through an ablation study on CHAOS-MRI in Tab. 3 (5-shot, setting 2). The study confirms the positive impact of each surrogate function and the combination of local (coherence/diverse) and global (attention/stabilization) expertise.

4   Conclusion

We proposed MSFSeg, a novel solution for segmenting unseen (cross-domain) 3D objects given only a few labeled slices. Our solution rely on a lightweight 2D low-shot segmentation network with multi-surrogate fusion to aggregate information from multiple supports for 3D segmentation in a transferred domain, and on a pool-based workflow to select the most relevant 2D/3D candidates for cross-volume generalization. MSFSeg can be integrated into actual 3D medical annotation tools to greatly reduce the burden of medical data annotators by proposing quick and valid initial segmentation masks across whole patient populations, as demonstrated on unseen datasets with challenging anatomical targets.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database
2. Ding, H., Sun, C., Tang, H., Cai, D., Yan, Y.: Few-shot medical image segmentation with cycle-resemblance attention. In: WACV. pp. 2487–2496 (2023)
3. Feng, Y., Wang, Y., Li, H., Qu, M., Yang, J.: Learning what and where to segment: A new perspective on medical image few-shot segmentation. Medical Image Analysis **87**, 102834 (2023)
4. Hatamizadeh, A., et al.: Unetr: Transformers for 3d medical image segmentation. In: WACV (2022)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
6. Hong, S., Cho, S., Nam, J., Lin, S., Kim, S.: Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In: ECCV (2022)
7. Hu, T., Yang, P., Zhang, C., Yu, G., Mu, Y., Snoek, C.G.M.: Attention-based multi-context guiding for few-shot semantic segmentation. In: AAAI (2019)
8. Huang, S., Xu, T., Shen, N., Mu, F., Li, J.: Rethinking few-shot medical segmentation: A vector quantization view. In: CVPR (2023)
9. Kavur, A.E., Gezer, N.S., Barış, M., et al.: CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. Medical Image Analysis (2021), `http://www.sciencedirect.com/science/article/pii/S1361841520303145`
10. Kim, S., An, S., Chikontwe, P., Park, S.H.: Bidirectional rnn-based few shot learning for 3d medical image segmentation. In: AAAI (2021)
11. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault-workshop and challenge (2015). DOI: https://doi. org/10.7303/syn3193805 (2015)
12. Lang, C., Cheng, G., Tu, B., Han, J.: Learning what not to segment: A new perspective on few-shot segmentation. In: CVPR (2022)
13. Lei, W., et al.: One-shot weakly-supervised segmentation in 3d medical images. IEEE Transactions on Medical Imaging (2023)
14. Li, G., Jampani, V., Sevilla-Lara, L., Sun, D., Kim, J., Kim, J.: Adaptive prototype learning and allocation for few-shot segmentation. In: CVPR (2021)

15. Lin, T., Maire, M., Belongie, S.J., et al.: Microsoft COCO: common objects in context. CoRR (2014)
16. Lin, Y., Chen, Y., Cheng, K.T., Chen, H.: Few shot medical image segmentation with cross attention transformer (2023)
17. Liu, W., Zhang, C., Lin, G., Liu, F.: Crnet: Cross-reference networks for few-shot segmentation. In: CVPR (2020)
18. Min, J., Kang, D., Cho, M.: Hypercorrelation squeeze for few-shot segmentation. In: ICCV (2021)
19. Niu, Y., Li, Z., Li, S.: Cross attention with transformer for few-shot medical image segmentation. In: International Conference on Information Technology in Medicine and Education (2022)
20. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In: ECCV (2020)
21. Pandey, P., Chasmai, M., Sur, T., Lall, B.: Robust prototypical few-shot organ segmentation with regularized neural-odes. IEEE Transactions on Medical Imaging **42**(9), 2490–2501 (2023)
22. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv:1704.00675 (2017)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
24. Roy, A.G., Siddiqui, S., Pölsterl, S., Navab, N., Wachinger, C.: 'squeeze & excite' guided few-shot segmentation of volumetric images. Medical image analysis **59**, 101587 (2020)
25. Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B.: One-shot learning for semantic segmentation. In: BMVC (2017)
26. Shi, X., Wei, D., Zhang, Y., Lu, D., Ning, M., Chen, J., Ma, K., Zheng, Y.: Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In: ECCV (2022)
27. Tang, H., Liu, X., Sun, S., Yan, X., Xie, X.: Recurrent mask refinement for few-shot medical image segmentation. In: ICCV (2021)
28. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: NeurIPS (2017)
29. Wang, H., Zhang, X., Hu, Y., Yang, Y., Cao, X., Zhen, X.: Few-shot semantic segmentation with democratic attention networks. In: ECCV (2020)
30. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: ICCV (2019)
31. Wang, R., Zhou, Q., Zheng, G.: Few-shot medical image segmentation regularized with self-reference and contrastive learning. In: MICCAI (2022)
32. Wang, X., Han, S., Chen, Y., Gao, D., Vasconcelos, N.: Volumetric attention for 3d medical image segmentation and detection. In: MICCAI (2019)
33. Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In: MICCAI (2021)
34. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.S.: Youtube-vos: A large-scale video object segmentation benchmark. ArXiv **abs/1809.03327** (2018)
35. Yu, Q., Dang, K., Tajbakhsh, N., Terzopoulos, D., Ding, X.: A location-sensitive local prototype network for few-shot medical image segmentation. In: IEEE 18th International Symposium on Biomedical Imaging (ISBI) (2021)

36. Zhang, B., Xiao, J., Qin, T.: Self-guided and cross-guided learning for few-shot segmentation. In: CVPR (2021)
37. Zhang, C., Lin, G., Liu, F., Guo, J., Wu, Q., Yao, R.: Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In: ICCV (2019)
38. Zhang, C., Lin, G., Liu, F., Yao, R., Shen, C.: Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In: CVPR (2019)
39. Zhang, X., Wei, Y., Yang, Y., Huang, T.S.: Sg-one: Similarity guidance network for one-shot semantic segmentation. IEEE Transactions on Cybernetics (2020)
40. Zhu, Y., Wang, S., Xin, T., Zhang, H.: Few-shot medical image segmentation via a region-enhanced prototypical transformer. In: MICCAI 2023 (2023)