Few-shot Medical Image Segmentation via Differentiable Supervoxel Transformer

Anonymized Authors

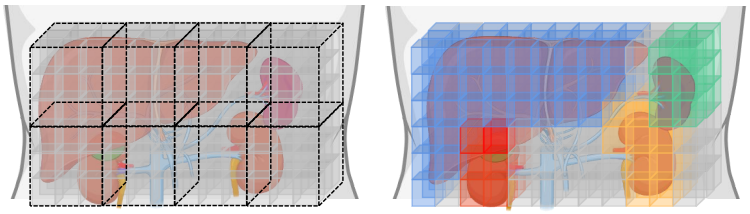
Anonymized Affiliations
email@anonymized.com

Abstract. Few-shot learning emerges as a pivotal approach to address the challenge of segmenting volumetric medical data with limited annotations. Prototype-based networks have become a popular strategy in this field. However, existing methods typically separate volumetric data into 2D slices individually, thereby ignoring the inherent 3D anatomical structure. And existing fixed-size tokenization method may divide continuous structures into different tokens, leading to disrupted semantic continuity. In this work, we propose a new representation termed the differentiable supervoxel, which naturally follows anatomical structures and provides more accurate semantics than voxels in CNNs and patches in ViT. Based on it, we introduces Similarity-Guided Upsampling, recovering supervoxels to voxel scale for better prototype computation. Building upon these, we introduce a novel 3D Transformer-based few-shot framework called Differentiable Supervoxel Transformer (DSVFormer) that utilizes high-fidelity supervoxel representations to generate discriminative prototypes. The effectiveness of DSVFormer has been validated across two public datasets—Abdominal-MRI and Cardiac-MRI—where it consistently outperformed state-of-the-art methods, demonstrating clear superiority and potential in real-world applications.

Keywords: Few-Shot Learning · Medical Image Segmentation · Transformer.

1 Introduction

Fully supervised medical image segmentation methods often require extensive expert annotations and face challenges from limited scalability and generalization. Few-Shot Segmentation (FSS) offers a promising solution. Recent advances in FSS have notably utilized prototype-based learning [19,27,25,7,11,2,23,9,18]. Most of these methods follow two stages: (1) representation learning and (2) prototype generation and segmentation. Existing approaches for representation learning in FSS mainly develop different networks to extract features from medical volumes. However, they primarily rely on 2D networks [25,2,23,12,15,17,21] that process each slice individually, overlooking the 3D spatial structure of organs. While some 3D FSS methods [4,26] exist, they either employ 3D CNNs [4] with limited long-range context modeling or use 2D ViTs [26] that divide volumes



(a) Fixed-size cubes (b) Learnable Supervoxels

Fig. 1: **Conceptual illustration of supervoxels and cubes.** Human anatomical structures are inherently 3D with irregular shape. In Fig. 1a, the ViT tokenizer divides voxels into fixed-size patches (cubes), causing semantic confusion. In contrast, in Fig. 1b, our proposed supervoxels aligns with anatomical boundaries by dynamically clustering natural sub-regions that represent different organs.

into fixed-size patches. This fixed tokenization breaks anatomical continuity by separating adjacent voxels into different tokens, limiting the model's capacity to dynamically attend to relevant organ regions within 3D structures. Moreover, other ViT variants with dynamic tokenization [28,14] are not directly applicable to prototype-based FSS frameworks and thus cannot be easily adopted for few-shot segmentation.

To address these representation learning limitations, we propose a novel differentiable supervoxels representation. As illustrated in Fig. 1, supervoxels tend to follow the boundaries of anatomical structures and dynamically cluster natural sub-regions representing different organ semantic regions. Thus, supervoxels are more semantically meaningful than fixed-size cubes. For prototype generation, conventional prototype-based FSS methods typically apply mask average pooling (MAP) to downsampled feature maps [17,23] from backbones. Because this feature map is downsampled, the resulting prototype loses information. Our method introduces Similarity-Guided Upsampling to preserve spatial information during representation learning. The differentiable nature of supervoxels enables recovery from downsampled supervoxels features to the original voxel scale through supervoxel-to-voxel similarity mapping. This bidirectional data flow between supervoxels and voxels maintains high-fidelity supervoxel features for improved prototype generation.

Building upon these innovations, we present Differentiable Supervoxel Transformer (DSVFormer), a supervoxel-based Transformer framework for few-shot segmentation. In summary, our contributions are as follows:

- We propose differentiable supervoxels that aligns more effectively with 3D anatomical structures in few-shot segmentation, offering better semantics than standard voxels or fixed-size cubes.
- We introduce Similarity-Guided Upsampling, which recovers supervoxels to the original voxel scale with high fidelity, thereby enhancing supervoxel-based prototype computation.

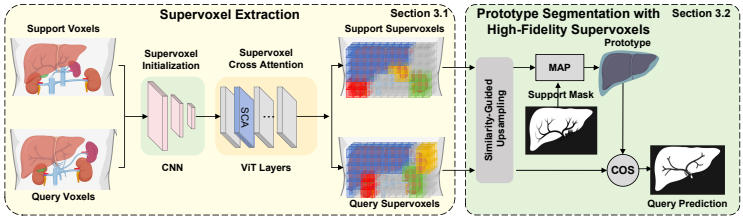


Fig. 2: **Overview of DSVFormer.** The framework comprises two stages: (1) Supervoxel Extraction and (2) Prototypical Segmentation with High-Fidelity Supervoxels.

– We present a supervoxel-based 3D Transformer few-shot framework, DSVFormer, that surpasses previous state-of-the-art methods, achieving Dice scores of 84.8% on Abdominal-MRI and 79.7% on Cardiac-MRI.

2 Method

In this section, we introduce our DSVFormer, which consists of two stages: Supervoxel Extraction (see Sect. 2.1) and Prototypical Segmentation with High-Fidelity Supervoxels (see Sect. 2.2). The pipeline is shown in Fig. 2.

2.1 Supervoxel Extraction

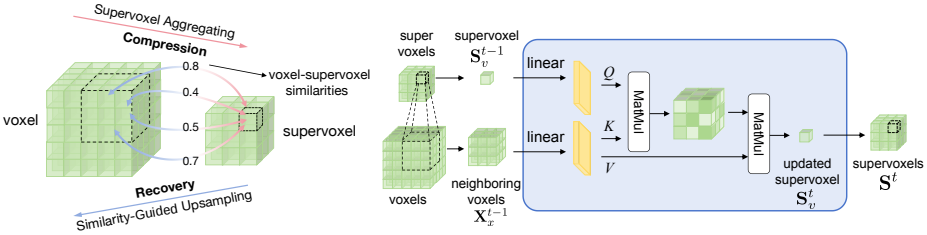
Supervoxel Initialization. To initialize supervoxels as clustering centers, we first use a 3D ResNeXt-101 model [5] as an encoder backbone to downsample the original 3D input **Volume** to $\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}$ of the input size and extract hypercolumn features [6], resulting in voxels $\mathbf{X} \in \mathbb{R}^{X_h \times X_w \times X_d \times X_c}$. Then, we apply simple average pooling to voxels, obtaining the supervoxel representation $\mathbf{S} \in \mathbb{R}^{S_h \times S_w \times S_d \times S_c}$, which is $\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}$ of the voxel size:

$$\mathbf{X}^0 = \text{Encoder}(\textbf{Volume}), \quad \mathbf{S}^0 = \text{AvgPool}(\mathbf{X}^0)$$

Supervoxel Cross Attention. After initializing supervoxels, we iteratively refine their representations to enhance semantic richness by utilizing Supervoxel Cross Attention (SCA) to compute voxel-to-supervoxel similarities and update features accordingly, as illustrated in Fig. 3b.

Let $\mathbf{S}$ represent supervoxel features and $\mathbf{X}$ represent voxel features. $\mathbf{S}_v^{t-1}$ denote the feature of a supervoxel v from iteration $t-1$. At iteration t, SCA aggregates information from neighboring voxels within a local $3 \times 3 \times 3$ window of the supervoxel, updating its representation as:

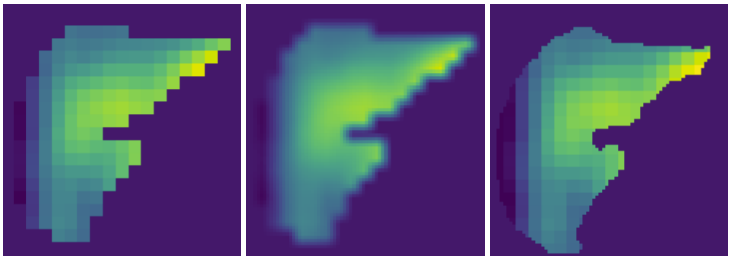
$$\mathbf{S}_v^t = \mathbf{S}_v^{t-1} + \sum_{x \in \mathcal{W}_v} \text{softmax}_{x \in \mathcal{W}_v} \left(\mathbf{q}_{\mathbf{S}_v^{t-1}} \cdot \mathbf{k}_{\mathbf{X}_x^{t-1}} \right) \mathbf{v}_{\mathbf{X}_x^{t-1}} \tag{1}$$



(a) A bidirectional data flow between voxels and supervoxels.

(b) Illustration of aggregating from voxels to supervoxels using Supervoxel Cross Attention (SCA).

Fig. 3: **Bidirectional aggregation (Fig. 3b) and recovery between voxels and supervoxels (Fig. 3a).**



(a) Low-Resolution Prototype

(b) Bilinear Upsampled Prototype

(c) Our Prototype

Fig. 4: **Comparison of prototype visualizations.** Similarity-Guided Upsampling (Fig. 4c) generates high-resolution prototypes with sharper boundaries and preserved details better than prior methods (Fig. 4a and 4b).

where $\mathcal{W}_v$ denotes the set of voxels within the local window of supervoxel v. The query $\mathbf{q}$, key $\mathbf{k}$, and value $\mathbf{v}$ are linear projections of the features from the previous iteration $t-1$.

Similarly, SCA updates the representation of each voxel X_x^{t-1} by aggregating information from neighboring supervoxels within a local $3 \times 3 \times 3$ window of the voxel:

$$\mathbf{X}_x^t = \mathbf{X}_x^{t-1} + \sum_{v \in \mathcal{W}_x} \text{softmax}_{v \in \mathcal{W}_x} \left(\mathbf{q}_{\mathbf{X}_x^{t-1}} \cdot \mathbf{k}_{\mathbf{S}_v^{t-1}} \right) \mathbf{v}_{\mathbf{S}_v^{t-1}} \tag{2}$$

where $\mathcal{W}_x$ is the set of supervoxels within the local window of voxel x.

In our approach, we strategically insert SCA layers into the vanilla ViT architecture before its first and third self-attention layers. This enables the model to first locally aggregate voxels into supervoxels, and then use the efficient supervoxels to further capture global information within the self-attention layers.

2.2 Prototype Segmentation with High-Fidelity Supervoxels

Similarity-Guided Upsampling. Previous prototype-based FSS models often compute prototypes based on downsampled features from the backbone. When segmenting a target organ, they often generate the target prototype by applying mask average pooling (MAP) to the feature map, which requires resizing to align the mask and feature map. Prior methods [17,23] either resize the mask to low resolution, producing low-resolution prototypes (Fig. 4a(a)), or use bilinear upsampling [4,15], resulting in over-smoothed prototypes (Fig. 4b). Both approaches yield coarse prototypes and information loss. Thanks to our differentiable supervoxels, which allows a bidirectional data flow that aggregates voxels into supervoxels and then recovers them back to the finer voxel scale before computing prototypes, as depicted in Fig. 3a.

The core of our method is to preserve information during downsampling to facilitate information recovery during upsampling. We utilize supervoxel-to-voxel similarities to project supervoxels back onto the voxel spatial space. Specifically, considering a supervoxel v, SCA in Eq. 1 computes the attention scores between supervoxel v and its neighboring $3 \times 3 \times 3$ voxels $\mathcal{W}_v$, denoted as $\mathbf{A}_{v \to x}$. By leveraging the attention scores from the final iteration $\mathbf{A}^t_{v \to x}$ as similarity, the supervoxel $\mathbf{S}^t_v$ from the last iteration can be upsampled in a similarity-guided manner back to its original $3 \times 3 \times 3$ neighboring voxel space $\mathbf{X}^t_x$:

$$\mathbf{X}^t_x = \sum_{v \in \mathcal{W}_x} \mathbf{A}^t_{v \to x} \, \mathbf{S}^t_v, \quad \text{where} \quad \mathbf{A}^t_{v \to x} = \text{softmax}_{v \in \mathcal{W}_x}\big(\mathbf{q}_{\mathbf{X}^t_x} \cdot \mathbf{k}_{\mathbf{S}^t_v}\big). \tag{3}$$

Thus, by utilizing all similarities between supervoxels and their corresponding voxels, we upsample the supervoxel representation $\mathbf{S} \in \mathbb{R}^{S_h \times S_w \times S_d \times S_c}$ back to the voxel resolution $\mathbf{X} \in \mathbb{R}^{X_h \times X_w \times X_d \times X_c}$, recovering information with high fidelity. As shown in Fig. 4c, this upsampling step recovers the information lost during downsampling and yields a more fine-grained feature map, enhancing prototype generation within a more detailed support feature space and mitigating spatial information loss.

Supervoxel-based Prototype Segmentation. Having acquired high-fidelity support and query supervoxels based on similarity-guided upsampling, we compute a 3D prototype p by masked average pooling (MAP) over support supervoxels ($S^s(x,y,z)$) weighted by the foreground mask $\mathbf{y}^{fg}(x,y,z)$ (Eq.4). This accumulation is normalized by the mask sum to create a class-specific embedding:

$$p = \frac{\sum_{x,y,z} S^s(x,y,z) \odot \mathbf{y}^{fg}(x,y,z)}{\sum_{x,y,z} \mathbf{y}^{fg}(x,y,z)}, \tag{4}$$

Here, $\odot$ denotes element-wise multiplication, and $\mathbf{y}^{fg} = \mathbb{1}(\mathbf{y} = c)$ is the binary mask for class c. Next, we compare p with query supervoxels, computing a similarity score $\mathcal{S}(x,y,z)$ using negative cosine distance (Eq. 5):

$$\mathcal{S}(x,y,z) = -\alpha \, \frac{\mathbf{S}^q(x,y,z) \cdot p}{\|\mathbf{S}^q(x,y,z)\| \|p\|}, \tag{5}$$

where α is a predefined scaling factor [16] that regulates gradient updates and improves few-shot adaptability. We threshold $\mathcal{S}$ with a learnable parameter T to separate foreground from background. The predicted mask $\hat{y}fg^q(x, y, z)$ is then computed by soft thresholding (Eq. 6):

$$\hat{y}^q_{fg}(x, y, z) = 1 - \sigma\big(\mathcal{S}(x, y, z) - T\big), \tag{6}$$

where $\sigma(\cdot)$ is the Sigmoid function. This approach isolates relevant structures and provides robust segmentation in few-shot settings, retaining essential class details under limited training data.

3 Experiments

3.1 Datasets and Implementation Details

The proposed method is comprehensively evaluated on two public datasets, including Abdominal-MRI [8] and Cardiac-MRI [30]. Abdominal-MRI is an abdominal MRI dataset used in the ISBI 2019 Combined Healthy Abdominal Organ Segmentation Challenge. Cardiac-MRI is a cardiac MRI dataset from MICCAI 2019 Multi-Sequence Cardiac MRI Segmentation Challenge. We follow the same pre-process steps like resampling and cropping, in previous methods [17,4].

We evaluate with a limited set of labeled slices from support volume to predict the entire query volume, following [4] for two protocols. *Evaluation Protocol 1 (EP1)* uses three support slices to segment the entire query volume. *Evaluation Protocol 2 (EP2)*, inspired by ADNet [4], uses one central slice as support, presenting a more challenging scenario with fewer annotations. Because our approach is 3D and requires continuous volumetric input, we adopt centroidal sampling in *EP1* to select the three support slices. Specifically, for a given organ to be segmented in the query volume, we first compute its centroid in the support volume using pseudo-labels. The centroid is defined as the average slice index (along the depth dimension) of the organ's volume. We then take three consecutive slices centered at this computed centroid as the support. This strategy ensures a fair comparison with 2D methods, which also rely on three labeled slices for support. We use 5-fold cross-validation and report mean performance across folds. We trained our model for 30,000 iterations with a batch size of 1. We started with a learning rate of 1×10^{-3} and decreased it by a factor of 0.8 every 1,000 iterations, following standard practices in few-shot medical segmentation [17,4,12,29].

3.2 Comparisons with State-of-the-art Methods

The effectiveness of our method is validated by comparing it with classical and recent FSS approaches under two evaluation protocols, EP1 and EP2. In EP1, our model outperforms the previous state-of-the-art method, RPT [29], by 2.3% and 0.5% on the Abdominal-MRI and Cardiac-MRI datasets, respectively. This

Table 1: **Quantitative comparison.** Our method achieves leading result on both abdominal MRI and Cardiac MRI datasets under evaluation protocols *EP1* and *EP2*. The results are extracted directly from the reported value in each method's original paper. "-" indicates that no results were provided.

Method	Abdominal MRI (EP1)					Abdominal MRI (EP2)					Cardiac MRI (EP1)				Cardiac MRI (EP2)			
	L kid	R kid	Spleen	Liver	Mean	L kid	R kid	Spleen	Liver	Mean	LV-BP	LV-MYO	RV	Mean	LV-BP	LV-MYO	RV	Mean
2D methods																		
PANet [20]	63.1	66.1	63.9	72.1	66.3	32.8	30.2	34.8	53.9	37.9	80.2	45.7	66.9	64.3	68.3	38.6	55.2	54.0
ALPNet [17]	62.1	71.8	66.6	73.1	68.4	56.4	50.4	44.7	56.7	52.0	87.5	60.2	76.1	74.6	80.6	53.3	69.2	67.7
PPNet [13]	62.1	71.8	66.6	73.1	68.4	43.4	56.9	43.1	56.3	49.9	67.8	42.6	60.8	57.0	56.7	34.8	47.6	46.4
CANet [24]	69.5	77.2	67.1	72.9	71.7	50.2	69.9	48.8	64.0	58.2	79.0	43.6	61.1	61.1	74.5	35.1	47.6	52.4
AAS-DCL [22]	80.4	86.1	76.2	72.3	78.8	-	-	-	-	-	85.2	64.0	79.1	76.1	-	-	-	-
SR&CL [21]	79.3	87.4	76.0	80.2	80.8	-	-	-	-	-	84.7	65.8	78.4	76.3	-	-	-	-
CRAPNet [3]	82.0	86.4	74.3	76.5	79.8	-	-	-	-	-	83.0	65.5	78.3	75.6	-	-	-	-
GMRD [2]	73.6	**90.0**	76.0	65.4	76.3	-	-	-	-	-	60.8	73.7	80.8	71.8	-	-	-	-
SSM-SAM [10]	81.7	80.3	78.8	77.5	79.6	-	-	-	-	-	-	-	-	-	-	-	-	-
DSPNet [18]	75.1	85.4	81.9	70.9	78.3	-	-	-	-	-	87.8	64.9	79.7	77.5	-	-	-	-
ADINet [1]	73.8	78.9	68.1	74.0	73.7	-	-	-	-	-	85.2	67.0	81.1	77.8	-	-	-	-
RPT [29]	80.7	89.8	76.4	82.9	82.4	76.4	89.1	70.5	76.5	78.1	89.9	66.9	80.8	79.2	**86.7**	58.0	**73.1**	72.6
3D methods																		
ADNet [4]	73.9	85.8	72.3	82.1	78.5	77.9	73.5	75.0	75.5	75.5	87.5	62.4	77.3	75.8	81.3	56.5	66.2	68.0
MSFSeg [26]	-	-	-	-	-	84.2	88.1	77.1	76.1	81.4	-	-	-	-	-	-	-	-
Ours	**85.5**	89.9	**81.2**	81.7	**84.8**	**86.5**	**89.4**	**78.2**	**80.9**	**83.8**	**90.4**	**67.3**	**81.5**	**79.7**	84.1	**64.6**	72.7	**73.8**

(a) Ground Truth (b) Learned Supervoxels (c) Fixed-size cube embeddings

Fig. 5: **Visualizations of the ground truth, cubes, and supervoxels.** The learned supervoxels align more closely with the boundaries of various organs in the ground truth (Fig. 5a) than cubes.

improvement, especially for smaller organs like spleen, arises from the adaptability and flexibility of differentiable supervoxels. In the more challenging EP2 scenario, where only a single support slice is available, our approach surpasses the state-of-the-art by 2.4% and 1.2% on the same datasets, highlighting potential clinical applicability when manual labeling is limited. These results underscore the advantage of leveraging the intrinsic 3D structure of medical images for few-shot segmentation.

Our model also surpasses existing 3D prototype approaches such as ADNet [4] and MSFSeg [26]. While ADNet relies on CNNs alone, lacking global attention, MSFSeg applies attention only to 2D slices, missing 3D volumetric context. By integrating both local and global attention volumetrically, our DSVFormer enables comprehensive information interaction among supervoxels and improves representation accuracy and efficiency.

Table 2: **Supervoxel ablation experiments.** The experiment is conducted on abdominal MRI dataset under *EP1*.

Method	Mean	ΔMean
w/o differentiable	77.9	-
w differentiable	**84.8**	+6.9

Table 3: **Component ablation experiments.** The experiment is conducted on abdominal MRI dataset under *EP1*.

3D-Resnet	ViT	Supervoxel Extraction	Similarity-Guided Upsampling	Mean	ΔMean
✓				79.8	-
✓		✓		82.9	+3.1
✓	✓	✓		83.8	+4.0
✓	✓	✓	✓	**84.8**	+5.0

3.3 Qualitative Analysis

Visualizations indicate that learned supervoxels (Fig. 5b) align with organ boundaries (Fig. 5a), effectively segmenting volumes into 3D regions that match organ structure. In contrast, fixed-size cubes (Fig. 5c) often merge multiple organs into one token or split a single organ across tokens, limiting organ-level representation and reducing both semantic and anatomical clarity, thereby losing region-level detail.

3.4 Ablation Studies

To evaluate the effectiveness of differentiable supervoxels compared with non-differentiable methods, we conducted an ablation study (see Table 2). Using the same 3D ViT architecture and implementation details described in Section 3.1, we first employed the non-differentiable algorithm from [4] to establish a baseline. Replacing it with our proposed differentiable supervoxels resulted in a notable 6.9% performance improvement, highlighting the advantages of a differentiable approach over conventional, hand-crafted non-differentiable methods.

To evaluate the impact of each component, we performed an ablation study in Table 3. Shifting from voxel representations to supervoxels representations, without ViT integration, boosts the 3D ResNet model's performance by 3.1%. Incorporating the ViT to model global interactions among supervoxels added a further 0.9% improvement. Applying similarity-guided upsampling contributed an additional 1.0% gain. These results confirm that each component systematically improves model capability.

4 Conclusion

In this study, we introduced DSVFormer, a 3D transformer model replacing cubes with differentiable supervoxels, which aligns with 3D volumetric medical data for better flexibility and fidelity. We also propose a supervoxel-based prototypical segmentation approach that uses high-fidelity supervoxels to improve the discriminative prototype. Our experiments on challenging medical FSS datasets show promising results. We hope this inspires research on 3D feature representation.

References

1. Chen, J., Li, X., Zhang, H., Cho, Y., Hwang, S.H., Gao, Z., Yang, G.: Adaptive dynamic inference for few-shot left atrium segmentation. Medical Image Analysis **98**, 103321 (2024) 7
2. Cheng, Z., Wang, S., Xin, T., Zhou, T., Zhang, H., Shao, L.: Few-shot medical image segmentation via generating multiple representative descriptors. IEEE Transactions on Medical Imaging (2024) 1, 7
3. Ding, H., Sun, C., Tang, H., Cai, D., Yan, Y.: Few-shot medical image segmentation with cycle-resemblance attention. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2488–2497 (2023) 7
4. Hansen, S., Gautam, S., Jenssen, R., Kampffmeyer, M.: Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. Medical Image Analysis **78**, 102385 (2022) 1, 5, 6, 7, 8
5. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6546–6555 (2018) 3
6. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 447–456 (2015) 3
7. He, W., Zhang, Y., Zhuo, W., Shen, L., Yang, J., Deng, S., Sun, L.: Apseg: Auto-prompt network for cross-domain few-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23762–23772 (2024) 1
8. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al.: Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (2021) 6
9. Lei, W., Su, Q., Jiang, T., Gu, R., Wang, N., Liu, X., Wang, G., Zhang, X., Zhang, S.: One-shot weakly-supervised segmentation in 3d medical images. IEEE Transactions on Medical Imaging (2023) 1
10. Leng, T., Zhang, Y., Han, K., Xie, X.: Self-sampling meta sam: enhancing few-shot medical image segmentation with meta-learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7925–7935 (2024) 7
11. Li, J., Shi, K., Xie, G.S., Liu, X., Zhang, J., Zhou, T.: Label-efficient few-shot semantic segmentation with unsupervised meta-training. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3109–3117 (2024) 1
12. Lin, Y., Chen, Y., Cheng, K.T., Chen, H.: Few shot medical image segmentation with cross attention transformer. arXiv preprint arXiv:2303.13867 (2023) 1, 6
13. Liu, Y., Zhang, X., Zhang, S., He, X.: Part-aware prototype network for few-shot semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 142–158. Springer (2020) 7
14. Mei, J., Chen, L.C., Yuille, A., Xie, C.: Spformer: Enhancing vision transformer with superpixel representation (2024) 2
15. Niu, Y., Li, Z., Li, S.: Cross attention with transformer for few-shot medical image segmentation. In: 2022 12th International Conference on Information Technology in Medicine and Education (ITME). pp. 644–648. IEEE (2022) 1, 5
16. Oreshkin, B., Rodríguez López, P., Lacoste, A.: Tadam: Task dependent adaptive metric for improved few-shot learning. Advances in neural information processing systems **31** (2018) 6

17. 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: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16. pp. 762–780. Springer (2020) 1, 2, 5, 6, 7
18. Tang, S., Yan, S., Qi, X., Gao, J., Ye, M., Zhang, J., Zhu, X.: Few-shot medical image segmentation with high-fidelity prototypes. Medical Image Analysis **100**, 103412 (2025) 1, 7
19. Wang, J., Li, J., Chen, C., Zhang, Y., Shen, H., Zhang, T.: Adaptive fss: a novel few-shot segmentation framework via prototype enhancement. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5463–5471 (2024) 1
20. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: proceedings of the IEEE/CVF international conference on computer vision. pp. 9197–9206 (2019) 7
21. Wang, R., Zhou, Q., Zheng, G.: Few-shot medical image segmentation regularized with self-reference and contrastive learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 514–523. Springer (2022) 1, 7
22. Wu, H., Xiao, F., Liang, C.: Dual contrastive learning with anatomical auxiliary supervision for few-shot medical image segmentation. In: European Conference on Computer Vision. pp. 417–434. Springer (2022) 7
23. Xie, W., Willems, N., Patil, S., Li, Y., Kumar, M.: Sam fewshot finetuning for anatomical segmentation in medical images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3253–3261 (2024) 1, 2, 5
24. Zhang, C., Lin, G., Liu, F., Yao, R., Shen, C.: Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5217–5226 (2019) 7
25. Zhang, Y., Li, H., Gao, Y., Duan, H., Huang, Y., Zheng, Y.: Prototype correlation matching and class-relation reasoning for few-shot medical image segmentation. IEEE Transactions on Medical Imaging (2024) 1
26. Zheng, M., Planche, B., Gao, Z., Chen, T., Radke, R.J., Wu, Z.: Few-shot 3d volumetric segmentation with multi-surrogate fusion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 286–296. Springer (2024) 1, 7
27. Zhou, T., Wang, W.: Prototype-based semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024) 1
28. Zhu, A.Z., Mei, J., Qiao, S., Yan, H., Zhu, Y., Chen, L.C., Kretzschmar, H.: Superpixel transformers for efficient semantic segmentation. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7651–7658. IEEE (2023) 2
29. Zhu, Y., Wang, S., Xin, T., Zhang, H.: Few-shot medical image segmentation via a region-enhanced prototypical transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 271–280. Springer (2023) 6, 7
30. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE transactions on pattern analysis and machine intelligence **41**(12), 2933–2946 (2018) 6