Few-shot Medical Image Segmentation via Supervoxel Transformer

Anonymous submission

Abstract

Addressing the challenge of segmenting volumetric medical data with limited annotations and patient variability, few-shot learning emerges as a pivotal approach. Prototype-based methods, which leverage support images to diminish intra-class variability, typically process volumetric data as sequential 2D slices, thereby ignoring their inherent 3D structure. This common oversight stems from the challenges associated with managing the high dimensionality of 3D data, particularly when implementing transformer architectures that are characterized by their quadratic computational complexity with respect to input size. In this work, we introduce the first 3D Transformer-based few-shot framework that utilizes supervoxel representations instead of traditional voxel cubes. We propose a novel clustering method, Supervoxel Cross Attention (SCA), to extract flexible supervoxel representations which effectively reduce feature redundancy while preserving rich 3D semantic details. Building upon the structural priors established by SCA, we develop a supervoxel-based prototypical segmentation technique that generates interpretable 3D prototypes by aligning supervoxels with target organs. The effectiveness of SVFormer has been validated across three public datasets—Abdominal-CT, Abdominal-MRI, and Cardiac-MRI—where it consistently outperformed state-of-the-art methods, demonstrating clear superiority and potential in real-world applications.

Introduction

The application of fully supervised methods in medical image segmentation is often limited by the substantial demand for expert annotations and the variability in imaging techniques, which hamper scalability and generalization. Inspired by the human ability to learn from few examples (Shaban et al. 2017), few-shot segmentation (FSS) is a promising solution for training models with limited data.

Recent advancements in FSS have notably explored prototype-based metric learning. In this approach, features corresponding to the target class from the support set are aggregated into a compact prototype. Subsequently, each pixel in the query image is classified based on its proximity to these prototypes (Liu et al. 2020; Wang et al. 2019; Snell, Swersky, and Zemel 2017). A critical aspect of this methodology is the generation of an accurate and representative prototype—a challenge that remains at the forefront of ongoing research. Along this research direction, various strategies have been proposed. For instance, Ouyang et al. (2020) employed non-overlapping pooling windows to generate multiple local prototypes from support features to capture more diverse features. Similarly, Yu et al. (2021) introduced prototype arrays with grid constraints for location-guided comparisons for more representative representations. While these studies generally show promise in few-shot segmentation, they largely rely on the premise that 2D prototypes are sufficient to encapsulate the complex 3D structure of organs. Such oversimplification can lead to significant loss of semantic information, thereby compromising the quality of the segmentation results.

However, transitioning few-shot segmentation from 2D to 3D presents considerable challenges, and only a few studies have delved into 3D few-shot medical segmentation. These explorations are primarily centered around CNN architectures. Notable examples include ADNet, which utilizes supervoxels in preprocessing stages and a 3D-CNN for volumetric data processing (Hansen et al. 2022). Additionally, Q-Net builds upon ADNet's framework by incorporating a dual-path feature extraction module to refine feature capture (Shen et al. 2023). However, these CNN-based approaches typically view volumes as grids of densely packed voxels, limiting contextual integration due to the constrained scope of their receptive fields. Moreover, directly processing dense volumetric data is challenging, as the presence of adjacent organs from other classes can interfere with target features, resulting in suboptimal performance (Tian et al. 2020).

In light of these limitations, the attention mechanism (Vaswani et al. 2017), recognized for improving medical vision tasks such as segmentation (Xie et al. 2023), offers a potential solution. This mechanism enables the network to focus selectively on pertinent areas of the input, thereby enriching the class-related prototype with more semantic accuracy. Powered by this, transformers have been applied to few-shot tasks such as CAT-NET (Lin et al. 2023) and RPT (Zhu et al. 2023b). In spite of these advancements, the inherent quadratic complexity of the attention mechanism often necessitates the division of extensive 3D volumes into disjointed 2D slices, which are then subdivided further into smaller patches for processing. This segmentation approach can severely fragment continuous organ structures, thereby impeding the comprehensive semantic understanding of class prototypes. Furthermore, this patch-based pro-

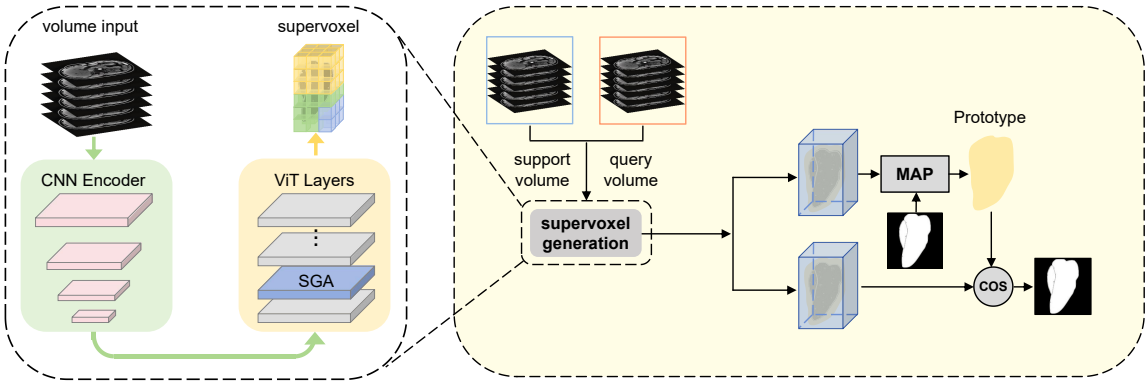


Figure 1: **Overview of SVFormer.** The framework consists of two components: Supervoxel Extraction and Supervoxel-based Prototypical Segmentation. Initially, both query and support volumes are processed through a shared feature encoder, transforming the data into supervoxel representations. These supervoxels from the support volumes are converted into a foreground prototype. For each supervoxel in the query feature, an anomaly score is calculated using cosine similarity with the prototype. Segmentation is then achieved by thresholding these scores with a learned anomaly threshold to delineate the query volume.

cessing can inadvertently mix features from different organs, leading to reduced specificity and potential inaccuracies in the prototype formation.

In this paper, we introduce a novel framework that harnesses the power of 3D Transformer technology to facilitate few-shot learning, a domain where leveraging 3D prior structures is challenging. To overcome these obstacles, this study pioneers the first 3D Transformer-based few-shot framework, incorporating supervoxel representations as a substitute for traditional voxel cubes and a brand new clustering method Supervoxel Cross Attention (SCA) to enhance few-shot segmentation. The framework is featured by:

Flexible Representation Supervoxels offer inherent flexibility, allowing a bidirectional process in which voxels are compressed into supervoxels and, conversely, supervoxels can be expanded back to their original voxel form. During feature extraction, similar voxels are grouped into supervoxels using SCA, forming semantically enriched regions along organ boundaries and reducing feature redundancy. Subsequently, a structural fidelity upsampling technique restores supervoxel data to voxel space. This flexibility allows us to leverage the compactness and efficiency of supervoxels while preserving the detailed information within the voxel space.

Interpretable Prototype Building on supervoxels, we develop the supervoxel-based prototypical segmentation, a novel approach that compresses supervoxels aligning target organs into a comprehensive, interpretable 3D prototype. The adaptiveness of supervoxels allows them to distinguish between different organs and group them accordingly, unlike traditional voxels, which inherently mix different organs. This ability to group similar regions leads to prototypes that align more accurately with anatomical structures, making them more interpretable. As a result, supervoxel-based prototypes are semantically richer than those derived from discrete 2D slices or fixed-size 3D cubes.

Additionally, we propose centroidal sampling for selecting a compact and representative volume as the support set in few-shot learning, thereby improving the quality and representativeness of the 3D representation.

In summary, our contributions in this work are threefold:

1. We introduce a brand new clustering method Supervoxel Cross Attention (SCA) to extract supervoxel representations to enhance few-shot 3D medical segmentation. By clustering neighboring voxels into supervoxels during the processing phase and subsequently restoring them to voxel format prior to prediction, our approach effectively reduces feature redundancy while meticulously preserving rich semantic details inherent in 3D medical images.
2. Building upon supervoxel representation, we have developed the supervoxel-based prototypical segmentation, a novel framework that uses supervoxel-based prototypes as medical priors. This technique significantly enhances the efficacy of 3D few-shot learning, providing a robust alternative to traditional methods that rely on generating prototypes from discrete 2D slices or fixed-size 3D cubes.
3. We present the first 3D Transformer-based few-shot framework, named Supervoxel Transformer (SVFormer), by incorporating supervoxel representations as a substitute for traditional voxel cubes. SVFormer outperforms the previous state-of-the-art on benchmark datasets, achieving Dice scores of 84.8%, 79.1%, and 79.7% for Abd-MRI, Abd-CT, and Card-MRI, respectively.

Related Works

Prototypical Few-shot Medical Image Segmentation Few-Shot Segmentation (FSS) in medical imaging often relies on 2D frameworks using prototypical networks (Snell, Swersky, and Zemel 2017). These methods extract prototypes from the support set to predict query set segmentation via similarity. However, a common issue with prototypical FSS lies in the reliance on accurately extracting and compressing target class representations during prototype ex-

traction. To address this, numerous strategies are designed to enhance the representation by computing diverse additional prototypes. For instance, Ouyang et al. (2020) innovated the use of non-overlapping pooling windows to create a set of local prototypes from support features, while the AAAS-DCL approach (Ding et al. 2023) integrates contrastive learning to improve prototype diversity. Additionally, PMMs (Yang et al. 2020a) leverages an Expectation-Maximization (EM) algorithm for generating multiple prototypes, and PPNet (Liu et al. 2020) adopts a clustering method for the development of part-aware prototypes. Despite these advancements, a common limitation among these methods is their dependence on intricate prototype-learning algorithms that are anchored in 2D slice-based representations, potentially overlooking the comprehensive 3D context of medical data.

Supervoxel Segmentations Supervoxels extend superpixels (Ren and Malik 2003) to 3D, grouping voxels into meaningful regions, which efficiently represent local volume features. In medical imaging, non-deep learning superpixel or supervoxel methods have been applied to few-shot tasks. For instance, SSL-ALPNet (Ouyang et al. 2020) uses superpixels in a self-supervised learning framework for stable pseudo-labels, while ADNet (Hansen et al. 2022) employs supervoxels to enhance segmentation accuracy. However, these methods don't utilize clustering to reduce spatial redundancy and are offline and non-differentiable, limiting their use in end-to-end neural network training. Recently, in natural imaging, superpixels have been integrated with deep learning frameworks, such as CNNs (Jampani et al. 2018; Zhu et al. 2023a) and vision transformers (ViTs) (Huang et al. 2023; Mei et al. 2024), but the integration of differentiable 3D supervoxels remains unexplored. No studies have yet applied differentiable superpixel or supervoxel techniques in the medical domain. Inspired by the efficiency of superpixels, we extend this approach to 3D supervoxels, offering a compact solution for 3D few-shot learning.

Transformer in Few-shot Medical Image Segmentation ViTs (Dosovitskiy et al. 2020) are increasingly popular in natural imaging but less explored in few-shot medical segmentation. CAT-NET (Lin et al. 2023) uses a Cross Attention Transformer to mine correlations between support and query images, while RPT (Zhu et al. 2023b) employs a Region-enhanced Prototypical Transformer to address intra-class diversity. However, ViTs face challenges in medical imaging: their computational complexity ($O(n^3)$ for 3D data) limits them to processing 2D slices, missing cross-slice attention. Additionally, fixed-size cube embeddings from ViTs can merge multiple organs into a single token, reducing semantic and anatomical clarity, especially for small targets. To address these challenges, we propose SVFormer, which leverages flexible supervoxel representations to reduce complexity and utilizes 3D supervoxel-based prototypes for more accurate organ representation.

Method

In this section, we introduce our supervoxel representation and implement it as Supervoxel Cross Attention. Building upon it, we then present the first 3D Transformer-based few-shot framework, SVFormer.

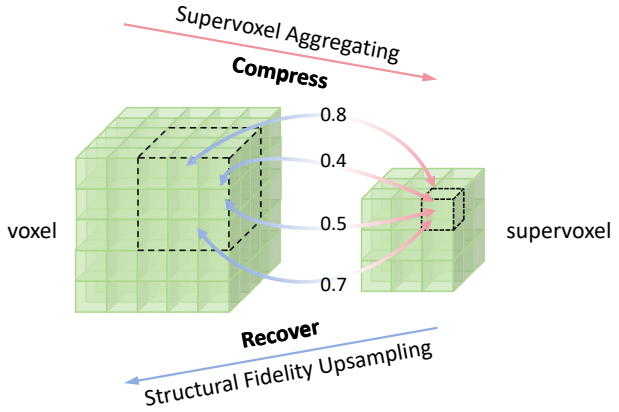


Figure 2: **Bidirectional data flow between voxels and supervoxels.** Voxels are aggregated into supervoxels via cross-attention, while supervoxels can be structurally upsampled back into voxels.

Traditional approaches in few-shot learning typically rely on 2D prototypes derived from individual slices within a volume, which often neglects the essential 3D context of the data crucial for capturing the complete structural integrity of organs. Moreover, the processing of raw 3D data can be challenging due to the voluminous nature of voxel data, where each voxel represents a value on a consistent grid in three-dimensional space. Conventional methods attempt to simplify this complexity by crudely dividing the voxels into cubes, a process that can inadvertently mix elements from different organs and obscure finer details.

Supervoxel Representation To effectively manage the complexity of 3D data while preserving detailed information, our approach is inspired by established techniques that incorporate superpixels into neural networks (Jampani et al. 2018; Yang et al. 2020b; Huang et al. 2023; Zhu et al. 2023a; Mei et al. 2024). We transform voxels into a more manageable unit called supervoxels—a supervoxel aggregates closely situated voxels that share similar characteristics. Supervoxels aptly conform to the contours of organs or structures within the volume, thereby facilitating the clustering of organ volumes into meaningful segments that closely mirror actual 3D organ structures, as shown in Fig. 7. This strategy not only enhances the interpretability of our model by aligning it more closely with the 3D manifestations of organs but also boosts processing efficiency by simplifying the redundancy inherent in the data.

Supervoxel Cross Attention Our method incrementally refines the supervoxel representation to enhance its semantic richness by iteratively adjusting voxel-to-supervoxel assignments and their features. We represent supervoxels by $\mathbf{S} \in \mathbb{R}^{s_h \times s_w \times s_d \times s_c}$ and voxels by $\mathbf{V} \in \mathbb{R}^{v_h \times v_w \times v_d \times v_c}$, where $\mathbf{S}$ and $\mathbf{V}$ denote the dimensions of supervoxels and voxels in terms of height (s_h and v_h), width (s_w and v_w), depth (s_d and v_d), and channels (s_c and v_c), respectively.

To dynamically refine the relationship between voxel i and its supervoxel neighbors, the potential voxel-to-supervoxel assignments are evaluated within a $3 \times 3 \times 3$ vicinity, denoted as $\mathcal{N}_i$. Similarly, each supervoxel v encompasses a $3\Delta \times 3\Delta \times 3\Delta$ neighborhood of voxels, represented by $\mathcal{W}_v$. We define the stride Δ as the ratio of voxel dimensions to supervoxel dimensions, explicitly $\Delta = \frac{v_h}{s_h} = \frac{v_w}{s_w} = \frac{v_d}{s_d}$.

The supervoxel feature $\mathbf{S}_v$ is updated by aggregating information from the voxels assigned within its vicinity:

$$\mathbf{S}_v^t = \mathbf{S}_v^{t-1} + \sum_{i \in \mathcal{W}_v} \text{softmax}\left(\mathbf{q}_{\mathbf{S}_v^{t-1}} \cdot \mathbf{k}_{\mathbf{V}_i^{t-1}}\right) \mathbf{v}_{\mathbf{V}_i^{t-1}}, \quad (1)$$

where $\mathbf{q}, \mathbf{k}$ and $\mathbf{v}$ represent the query, key, and value tensors, respectively, generated by linear transformations applied to the respective features of supervoxel ($\mathbf{S}_v^{t-1}$) and voxel ($\mathbf{V}_i^{t-1}$) from the prior iteration $t-1$.

Similarly, voxel features $\mathbf{V}_i$ are updated by gathering information from its corresponding supervoxels, weighted by the attention score, which corresponds to the voxel-to-supervoxel assignments:

$$\mathbf{V}_i^t = \mathbf{V}_i^{t-1} + \sum_{v \in \mathcal{N}_i} \text{softmax}\left(\mathbf{q}_{\mathbf{V}_i^{t-1}} \cdot \mathbf{k}_{\mathbf{S}_v^{t-1}}\right) \mathbf{v}_{\mathbf{S}_v^{t-1}}, \quad (2)$$

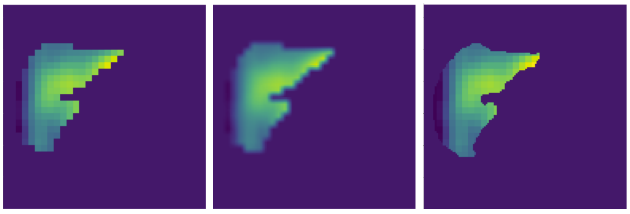
Our approach redefines supervoxel generation using sliding window cross-attention, enabling seamless integration of supervoxel representation into the network for end-to-end training. SCA effectively consolidates nearby and similar voxels into cohesive supervoxels, precisely delineating organ boundaries, as shown in Fig. 7.

Structural Fidelity Upsampling Traditional few-shot segmentation models often downsample features to a lower resolution, as illustrated in Fig. 3(a), leading to the generation of coarse prototypes and significant information loss. In contrast, our method employs a bidirectional data flow, which aggregates data from voxels to supervoxels and subsequently reconstructs it from supervoxels back to a finer voxel scale. This bilateral hierarchy is depicted in Fig. 2 and effectively restores features to their original resolution, producing more accurate prototypes than those based on low-resolution features, as demonstrated in Fig. 3(c).

The essence of this method is anchored in the attention scores defined in Eq. (1). These scores generate association matrices that elucidate the intricate relationships between supervoxels and neighboring voxels. Leveraging the final iteration attention score $\mathbf{A}_{s_v \to v}$ as a measure of similarity between voxels and supervoxels, features are systematically upscaled from the supervoxel level ($\mathbf{S}_v \in \mathbb{R}^{s_h \times s_w \times s_d}$) back to their original voxel scale ($\mathbf{V} \in \mathbb{R}^{v_h \times v_w \times v_d}$):

$$\begin{aligned} \mathbf{A}_{s_v \to v} &= \sum_{v \in \mathcal{N}_i} \text{softmax}\left(\mathbf{q}_{\mathbf{V}_i} \cdot \mathbf{k}_{\mathbf{S}_v}\right), \\ \mathbf{V} &= \mathbf{A}_{s_v \to v} \cdot \mathbf{S}_v \end{aligned} \quad (3)$$

This procedural flow intricately enhances prototype generation within a more detailed support feature space, mitigating the spatial information loss typically incurred by resizing the support mask. It ensures a higher fidelity to the original data when creating prototypes, enhancing the quality and accuracy as shown in Fig. 3.



(a) Traditional Low-Resolution Prototype (b) Bilinear Upsampled Prototype (c) Our Prototype

Figure 3: **Comparison of prototype visualizations.** Traditional low-resolution methods suffer from significant semantic loss, even when bilinear upsampling is applied, often resulting in suboptimal and overly smooth outcomes. In contrast, structural fidelity upsampling consistently produces high-resolution prototypes with sharper boundaries and greatly improved preservation of semantic details.

Supervoxel Transformer Architecture

Building on the supervoxel representation, we introduce the Supervoxel Transformer (SVFormer) for few-shot medical semantic segmentation, as shown in Fig. 1. The model begins with a supervoxel extractor network that generates supervoxel representations for 3D organ structures from both support and query volumes. This is followed by a 3D similarity-based prototypical segmentation strategy inspired by the methods used in ADNet (Hansen et al. 2022). Finally, we present a centroid sampling method for selecting representative support slices.

Supervoxel Extraction Both support and query volumes are initially processed by a 3D CNN to extract voxel-wise features, which are then averaged to initialize supervoxel-wise features. These initial features are input into the SCA module, which aggregates proximate and similar voxels into supervoxels. In our method, SCA blocks are strategically placed before the first and third self-attention layers of the ViT, clustering features to delineate regions within each organ class.

The resulting compact supervoxel representation facilitates efficient global interactions within the ViT layers, enabling deeper semantic interactions and more comprehensive volumetric data analysis.

Additionally, a lightweight masked cross attention module updates the supervoxels using organ-specific masks to ensure only features from corresponding organ classes are enhanced, effectively excluding dissimilar features. Further details on the masked cross attention module are provided in the supplementary material.

Supervoxel-Based Prototypical Segmentation After obtaining the refined support and query supervoxels, we implement the supervoxel-based prototypical segmentation. Unlike the low-resolution prototypes used in ADNet, our prototypes are derived from detailed support supervoxels by applying structural fidelity upsampling, followed by masked

average pooling. This process, visualized in Fig. 1, shows a significant enhancement in prototype detail through structural fidelity upsampling, as illustrated in Fig. 3.

Segmentation is performed by comparing the prototype from the support volume ($\mathbf{p}$) with the supervoxels from the query volume. The similarity score $\mathcal{S}$ for a supervoxel at coordinates (x, y, z) in the query volume $\mathbf{S}_v^q$ is calculated using the negative cosine distance:

$$\mathcal{S}(x,y,z) = -\alpha \frac{\mathbf{S}_v^q(x,y,z) \cdot \mathbf{p}}{\|\mathbf{S}_v^q(x,y,z)\| \|\mathbf{p}\|}, \tag{4}$$

where α is a predefined scaling factor. The segmentation process identifies the final foreground mask by thresholding these similarity scores with a learnable parameter T. Supervoxels with scores below T are classified as foreground, and those above it as background. The predicted foreground mask is obtained through soft thresholding:

$$\hat{y}_{fg}^q(x,y,z) = 1 - \sigma\left(\mathcal{S}(x,y,z) - T\right), \tag{5}$$

where $\sigma(\cdot)$ denotes the Sigmoid function, modulating the steepness of the threshold response.

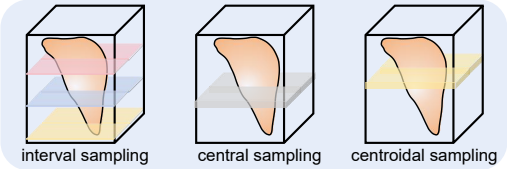


Figure 4: **Illustration of different sampling methods.** Interval sampling divides the volume into chunks and selects the middle slices; central sampling chooses the exact middle slice of the entire volume; centroidal sampling computes the most representative slice based on the volume-weighted centroid of the organ, accommodating its irregular shape.

Centroidal Sampling for Enhanced 3D Prototype In few-shot tasks, only a limited number of annotated slices from the 3D support volume are available. Prior methods, such as interval sampling depicted in Fig. 4, tackled this by dividing the volume into chunks and annotating only the middle slices. However, the discontinuity disrupts continuous organ structures and compromises the semantic integrity of class prototypes.

To address these challenges, we introduce centroidal sampling, as shown in Fig. 4. This method calculates the organ's centroid from pseudo labels to guide the selection of support slices, ensuring a more accurate representation without additional annotations.

Pseudo labels are generated following the method described by Ouyang et al. (2020), which segments the entire volume into sub-regions that represent semantic organ regions and their volume distributions. We then compute a weighted average of these volumes to determine the centroid position and its corresponding slice. This selected slice, being at the "center" of the organ, better reflects the overall structure and is more representative for analysis.

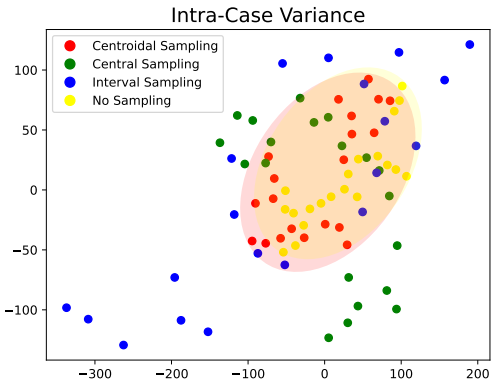


Figure 5: **t-SNE visualization of support sets derived from different sampling methods.** Our centroidal sampling aligns more closely with features from unsampled data compared to either interval sampling or central sampling.

Specifically, the centroidal slice C_d along the depth dimension is calculated by the formula:

$$C_d = \frac{\sum_{i=1}^{D} i \cdot A_i}{\sum_{i=1}^{D} A_i} \tag{6}$$

where A_i is the area of the organ in the i-th slice, determined by counting pixels labeled as organ presence:

$$A_i = \sum_{h,w} (label_{ihw} \neq 0) \tag{7}$$

Here, D represents the total number of slices.

Unlike central sampling, which selects based solely on slice position and does not consider variance in organ area across slices, centroidal sampling prioritizes slices with larger organ areas. This method ensures that more representative slices influence centroid determination, leading to a selection that is both more representative and anatomically coherent.

To validate that centroidal sampling provides a more comprehensive selection, we present both qualitative and quantitative analyses. Quantitatively, centroidal sampling outperforms interval sampling by 11.3% and central sampling by 2.6%. An ablation study comparing these methods is detailed in the supplementary material. Qualitatively, t-SNE visualization in Fig. 5 shows that centroidal sampling closely aligns with features from the entire unsampled volume, indicating its superiority in preserving information and providing representative prototypes for segmentation tasks.

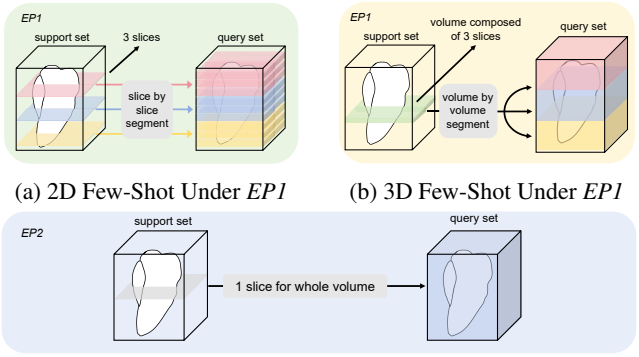
Experiments

Datasets

The proposed method is comprehensively evaluated on three public datasets, including Abdominal-MRI, Abdominal-CT and Cardiac-MRI. Concretely, Abd-MRI (Kavur et al. 2021) is an abdominal MRI dataset used in the

ISBI 2019 Combined Healthy Abdominal Organ Segmentation Challenge. Abdominal-CT (Landman et al. 2015) is an abdominal CT dataset from MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge. Cardiac-MRI (Zhuang 2018) is a cardiac MRI dataset from MICCAI 2019 Multi-Sequence Cardiac MRI Segmentation Challenge. We follow all preprocess scheme in ADNet (Hansen et al. 2022).

Implementation Details



(a) 2D Few-Shot Under *EP1* (b) 3D Few-Shot Under *EP1*

(c) 2D and 3D Few-Shot Under *EP2*

Figure 6: **Illustration of Different Evaluation Protocols (EP)**. In *EP1*: (a) For 2D few-shot methods, each support and query volume is divided into three sub-chunks; the middle slice of each support sub-chunk is labeled, which then guides the segmentation of corresponding query sub-chunks. (b) For our 3D few-shot method, three centroidal slices are used uniformly across all sub-chunks. (c) In *EP2*, only the middle slice of the support volume is labeled, which is then used to segment the entire query volume.

Training Our model undergoes training for $30,000$ iterations with a batch size of 1, starting with a learning rate of 1×10^{-3}, which decreases by a factor of 0.8 every 1,000 iterations, aligning with standard few-shot medical segmentation practices (Ouyang et al. 2020; Hansen et al. 2022; Lin et al. 2023; Zhu et al. 2023b). During training, since we utilize pseudo labels, all slices within the volume are available, eliminating the need for a sampling method. Due to memory constraints, we select only m slices from the total n slices of the volume as the support and query volumes for network prediction each time. For Abdominal-MRI, Abdominal-CT, and Cardiac-MRI, m is set to 8, 10, and 10, respectively, according to the total number of slices in each dataset's volume.

Evaluation During evaluation, given the limited availability of human-annotated labels in a few-shot setting, we use only a few labeled slices as the support volume while predicting the entire query volume. The generalization of our method is evaluated using two protocols. Evaluation Protocol 1 (EP1) employs centroidal sampling to select three slices as the support set for predicting the entire volume. This approach ensures a fair comparison with 2D methods by using the same number of labels without requiring additional annotations, despite our method being 3D. Evaluation Protocol 2 (EP2), inspired by ADNet (Hansen et al. 2022), forgoes centroidal sampling and uses a single central slice from the support set to segment the query volume, posing a more challenging test for model performance and better reflecting real-world scenarios. All methods undergo 5-fold cross-validation to ensure reliability, with results reported as mean values to emphasize the effectiveness and generalization capabilities of the methodology across different scenarios.

Comparisons with State-of-the-art Methods

The efficacy of our method is demonstrated through comparisons with classical and state-of-the-art FSS methods under two evaluation protocols, EP1 and EP2 (Table 1, Table 2). Under EP1, as shown in Table 1 and Table 2, our model surpasses the previous best, RPT (Zhu et al. 2023b), on Abdominal-MRI, Abdominal-CT, and Cardiac-MRI by 2.3%, 1.3%, and 0.5%, respectively. This improvement, especially in segmenting smaller organs like the spleen, is attributable to the utilization of more precise and flexible supervoxel representation. In pursuit of validating our model under a more stringent few-shot scenario—EP2, where only a single slice is available in the support set for reference—we found that our approach surpasses the latest state-of-the-art by substantial margins of 5.1% and 1.2% on the Abdominal-MRI and Cardiac-MRI datasets, respectively, as shown in Table 1. This underscores the practical viability of our model in clinical settings, where extensive manual annotation is often unfeasible. This finding highlights the importance of leveraging the intrinsic 3D structure of medical images in few-shot learning, an aspect that has been largely neglected.

Beyond surpassing previous state-of-the-art methods, our model significantly outperforms ADNet, one of the few existing 3D few-shot models. Unlike ADNet, which relies solely on CNN methods with limited contextual integration due to constrained receptive fields, our transformer network facilitates global interaction with supervoxels, thereby enhancing both representation accuracy and efficiency.

Table 2: Comparison of dice scores (%) for abdominal CT datasets under evaluation protocol *EP1*.

Method	Abdominal CT (EP1)				
	L kid.	R kid.	Spleen	Liver	Mean
AAS-DCL(Wu, Xiao, and Liang 2022)	74.6	73.2	72.3	78.0	74.5
SR&CL(Wang, Zhou, and Zheng 2022)	73.5	71.2	73.4	76.1	73.5
CRAPNet(Ding et al. 2023)	74.7	74.2	70.4	75.4	73.7
RPT (Zhu et al. 2023b)	**77.1**	**79.1**	72.6	**82.6**	77.8
ADNet(Hansen et al. 2022)	72.1	79.1	63.5	77.2	73.0
Ours	75.5	77.2	**82.4**	81.5	**79.1**

Qualitative Analysis

Analysis of the visualizations reveals that the learned supervoxels, as shown in Fig. 7b, generally align with the boundaries of multiple organs in the ground truth Fig. 7a. This indicates that supervoxels can effectively segment images into irregular 3D regions that are aware of organ seman-

Table 1: Comparison of dice scores (%) for abdominal MRI and cardiac datasets under evaluation protocols *EP1* and *EP2*, categorized by 2D and 3D methodologies.

Method	Abdominal MRI (EP1/EP2)					Cardiac (EP1/EP2)			
	L kid.	R kid.	Spleen	Liver	Mean	LV-BP	LV-MYO	RV	Mean
PANet(Wang et al. 2019)	- / 32.8	- / 30.2	- / 34.8	- / 53.9	- / 37.9	- / 68.3	- / 38.6	- / 55.2	- / 54.0
ALPNet(Ouyang et al. 2020)	- / 56.4	- / 50.4	- / 44.7	- / 56.7	- / 52.0	- / 80.6	- / 53.3	- / 69.2	- / 67.7
PPNet(Liu et al. 2020)	- / 43.4	- / 56.9	- / 43.1	- / 56.3	- / 49.9	- / 56.7	- / 34.8	- / 47.6	- / 46.4
CANet(Zhang et al. 2019)	- / 50.2	- / 69.9	- / 48.8	- / 64.0	- / 58.2	- / 74.5	- / 35.1	- / 47.6	- / 52.4
AAS-DCL(Wu, Xiao, and Liang 2022)	80.4 / -	86.1 / -	76.2 / -	72.3 / -	78.8 / -	85.2 / -	64.0 / -	79.1 / -	76.1 / -
SR&CL(Wang, Zhou, and Zheng 2022)	79.3 / -	87.4 / -	76.0 / -	80.2 / -	80.8 / -	84.7 / -	65.8 / -	78.4 / -	76.3 / -
CRAPNet(Ding et al. 2023)	82.0 / -	86.4 / -	74.3 / -	76.5 / -	79.8 / -	83.0 / -	65.5 / -	78.3 / -	75.6 / -
RPT (Zhu et al. 2023b)	80.7 / 76.4	89.8 / 89.1	76.4 / 70.5	82.9 / 76.5	82.4 / 78.1	89.9 / 86.7	66.9 / 58.0	80.8 / 73.1	79.2 / 72.6
ADNet(Hansen et al. 2022)	73.9 / 77.9	85.8 / 73.5	72.3 / 75.0	82.1 / 75.5	78.5 / 75.5	87.5 / 81.3	62.4 / 56.5	77.3 / 66.2	75.8 / 68.0
Ours	**85.5** / **86.5**	**89.9** / **89.4**	**81.2** / **78.2**	81.7 / **80.9**	**84.8** / **83.8**	**90.4** / 84.1	**67.3** / **64.6**	**81.5** / 72.7	**79.7** / **73.8**

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

Figure 7: Visualizations of supervoxels and cubes.

tics. In contrast, fixed-size cube embeddings Fig. 7c may merge multiple organs into a single token, thereby reducing both semantic and anatomical clarity.

We have also provided qualitative segmentation results, comparing our model with others, in the supplementary material.

Table 3: Ablation study on the 3D representation for the abdominal MRI dataset under *EP1*.

Method	L kid.	R kid.	Spleen	Liver	Mean	ΔMean
Cube	79.5	85.9	69.8	79.5	77.9	-
Superpixel	82.9	89.3	77.3	79.7	82.2	+4.3
Supervoxel	85.5	89.9	81.2	81.7	**84.8**	+6.9

Ablation Studies

Supervoxel Effectiveness In the initial phase, we evaluate the effectiveness of supervoxel (supervoxels) representation (Table 3). Switching from supervoxels to cube representation, as in the vanilla 3D ViT model (Chen et al. 2023), leads to a 6.9% performance drop, particularly a 11.4% decline in spleen segmentation. This underperformance stems from 3D ViT's use of fixed-size cubic embeddings to create 3D visual tokens, causing semantic ambiguities when different organs are contained within the same cube. This issue is more pronounced with smaller organs. Conversely, substituting cubes with superpixels, which cluster similar pixels within 2D slices, yields a 4.3% improvement. We further enhance this by adopting supervoxels representation, enabling voxel-level grouping rather than slice-level aggregation, and leveraging volumetric information to achieve an additional 2.6% improvement. These findings highlight the critical role of supervoxels representations in more accurately capturing 3D anatomical details.

Table 4: Ablation study of component effectiveness on the abdominal MRI dataset under *EP1*.

3D-Resnet	ViT	Supervoxel Generation	Masked Cross Attention	Mean	ΔMean
✓				79.8	-
✓		✓		82.9	+3.1
✓	✓	✓		84.2	+4.4
✓	✓	✓	✓	**84.8**	+5.0

Component Effectiveness In our evaluation, Table 4 shows the impact of each component. Shifting from voxel-based predictions to supervoxels representations, without ViT integration, boosts the 3D ResNet model's performance by 3.1%, demonstrating the value of coherent voxel grouping. Adding ViT further enhances performance by 1.3%, enabling global interactions across the volume. Our approach also introduces masked cross-attention, enriching query features using support slices, improving accuracy beyond traditional distance-based classification methods.

Conclusion

In this study, we introduced SVFormer, an innovative 3D transformer model that leverages supervoxel representation for the efficient processing of volumetric medical data. Moving beyond conventional slice-based segmentation approaches, SVFormer harnesses the full spatial context of 3D imagery, significantly enhancing prototype sampling and generation. This shift to a comprehensive 3D analysis framework has enabled our model to set new benchmarks across three public datasets, demonstrating the potential of our method for advanced 3D medical image processing and analysis.

References

Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y.; et al. 2023. 3d transunet: Advancing medical image segmentation through vision transformers. *arXiv preprint arXiv:2310.07781*.

Ding, H.; Sun, C.; Tang, H.; Cai, D.; and Yan, Y. 2023. Few-shot medical image segmentation with cycle-resemblance attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2488–2497.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Hansen, S.; Gautam, S.; Jenssen, R.; and Kampffmeyer, M. 2022. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. *Medical Image Analysis*, 78: 102385.

Huang, H.; Zhou, X.; Cao, J.; He, R.; and Tan, T. 2023. Vision Transformer with Super Token Sampling. In *CVPR*.

Jampani, V.; Sun, D.; Liu, M.; Yang, M.; and Kautz, J. 2018. Superpixel Sampling Networks. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *ECCV*.

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. 2021. CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69: 101950.

Landman, B.; Xu, Z.; Igelsias, J.; Styner, M.; Langerak, T.; and Klein, A. 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, 12.

Lin, Y.; Chen, Y.; Cheng, K.-T.; and Chen, H. 2023. Few Shot Medical Image Segmentation with Cross Attention Transformer. *arXiv preprint arXiv:2303.13867*.

Liu, Y.; Zhang, X.; Zhang, S.; and He, X. 2020. 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*, 142–158. Springer.

Mei, J.; Chen, L.-C.; Yuille, A.; and Xie, C. 2024. SP-Former: Enhancing Vision Transformer with Superpixel Representation. arXiv:2401.02931.

Ouyang, C.; Biffi, C.; Chen, C.; Kart, T.; Qiu, H.; and Rueckert, D. 2020. 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*, 762–780. Springer.

Ren; and Malik. 2003. Learning a classification model for segmentation. In *Proceedings ninth IEEE international conference on computer vision*, 10–17. IEEE.

Shaban, A.; Bansal, S.; Liu, Z.; Essa, I.; and Boots, B. 2017. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*.

Shen, Q.; Li, Y.; Jin, J.; and Liu, B. 2023. Q-net: Query-informed few-shot medical image segmentation. In *Proceedings of SAI Intelligent Systems Conference*, 610–628. Springer.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Tian, Z.; Zhao, H.; Shu, M.; Yang, Z.; Li, R.; and Jia, J. 2020. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 44(2): 1050–1065.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, K.; Liew, J. H.; Zou, Y.; Zhou, D.; and Feng, J. 2019. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, 9197–9206.

Wang, R.; Zhou, Q.; and Zheng, G. 2022. Few-shot Medical Image Segmentation Regularized with Self-reference and Contrastive Learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 514–523. Springer.

Wu, H.; Xiao, F.; and Liang, C. 2022. Dual Contrastive Learning with Anatomical Auxiliary Supervision for Few-Shot Medical Image Segmentation. In *European Conference on Computer Vision*, 417–434. Springer.

Xie, Y.; Yang, B.; Guan, Q.; Zhang, J.; Wu, Q.; and Xia, Y. 2023. Attention mechanisms in medical image segmentation: A survey. *arXiv preprint arXiv:2305.17937*.

Yang, B.; Liu, C.; Li, B.; Jiao, J.; and Ye, Q. 2020a. Prototype mixture models for few-shot semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, 763–778. Springer.

Yang, F.; Sun, Q.; Jin, H.; and Zhou, Z. 2020b. Superpixel segmentation with fully convolutional networks. In *CVPR*.

Yu, Q.; Dang, K.; Tajbakhsh, N.; Terzopoulos, D.; and Ding, X. 2021. A location-sensitive local prototype network for few-shot medical image segmentation. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, 262–266. IEEE.

Zhang, C.; Lin, G.; Liu, F.; Yao, R.; and Shen, C. 2019. 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*, 5217–5226.

Zhu, A. Z.; Mei, J.; Qiao, S.; Yan, H.; Zhu, Y.; Chen, L.-C.; and Kretzschmar, H. 2023a. Superpixel transformers for efficient semantic segmentation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7651–7658. IEEE.

Zhu, Y.; Wang, S.; Xin, T.; and Zhang, H. 2023b. Few-Shot Medical Image Segmentation via a Region-Enhanced Prototypical Transformer. In *International Conference on Medi-*

cal Image Computing and Computer-Assisted Intervention, 271–280. Springer.

Zhuang, X. 2018. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE transactions on pattern analysis and machine intelligence*, 41(12): 2933–2946.

Appendix

Reproducibility Checklist

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA): yes
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no): yes
- Provides well marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no): yes

Does this paper make theoretical contributions? (yes/no): yes

If yes, please complete the list below.

- All assumptions and restrictions are stated clearly and formally. (yes/partial/no): yes
- All novel claims are stated formally (e.g., in theorem statements). (yes/partial/no): yes
- Proofs of all novel claims are included. (yes/partial/no): yes
- Proof sketches or intuitions are given for complex and/or novel results. (yes/partial/no): yes
- Appropriate citations to theoretical tools used are given. (yes/partial/no): yes
- All theoretical claims are demonstrated empirically to hold. (yes/partial/no/NA): yes
- All experimental code used to eliminate or disprove claims is included. (yes/no/NA): yes

Does this paper rely on one or more datasets? (yes/no): yes

If yes, please complete the list below.

- A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA): yes
- All novel datasets introduced in this paper are included in a data appendix. (yes/partial/no/NA): NA
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes/partial/no/NA): NA
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes/no/NA): yes
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (yes/partial/no/NA): yes
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (yes/partial/no/NA): NA

Does this paper include computational experiments? (yes/no): yes

If yes, please complete the list below.

- Any code required for pre-processing data is included in the appendix. (yes/partial/no): yes
- All source code required for conducting and analyzing the experiments is included in a code appendix. (yes/partial/no): yes
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes/partial/no): yes
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no): yes
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes/partial/no/NA): yes
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes/partial/no): partial
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes/partial/no): yes
- This paper states the number of algorithm runs used to compute each reported result. (yes/no): yes
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (yes/no): no
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (yes/partial/no)
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (yes/partial/no/NA)
- This paper states the number and range of values tried per (hyper-)parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (yes/partial/no/NA): no