Clustering Propagation for Universal Medical Image Segmentation

Yuhang Ding[1], Liulei Li[1], Wenguan Wang[2], Yi Yang[2*]

[1] ReLER, AAII, University of Technology Sydney   [2] ReLER, CCAI, Zhejiang University

https://github.com/dyh127/S2VNet

Abstract

Prominent solutions for medical image segmentation are typically tailored for automatic or interactive setups, posing challenges in facilitating progress achieved in one task to another. This also necessitates separate models for each task, duplicating both training time and parameters. To address above issues, we introduce S2VNet, a universal framework that leverages **S***lice-**to**-**V***olume propagation to unify automatic/interactive segmentation within a single model and one training session. Inspired by clustering-based segmentation techniques, S2VNet makes full use of the slice-wise structure of volumetric data by initializing cluster centers from the cluster results of previous slice. This enables knowledge acquired from prior slices to assist in the segmentation of the current slice, further efficiently bridging the communication between remote slices using mere 2D networks. Moreover, such a framework readily accommodates interactive segmentation with no architectural change, simply by initializing centroids from user inputs. S2VNet distinguishes itself by swift inference speeds and reduced memory consumption compared to prevailing 3D solutions. It can also handle multi-class interactions with each of them serving to initialize different centroids. Experiments on three benchmarks demonstrate S2VNet surpasses task-specified solutions on both automatic/interactive setups.*

1. Introduction

In the realm of medical imaging, the practice of precisely revealing anatomical or pathological structure changes in a pixel observation holds the promise to substantially advance diagnostic efficiency[1]. Depending on the presence of user interactions, it can be categorized into automatic or interactive medical image segmentation (AMIS/IMIS)[2], with the latter involving active user engagement (*e.g*., click, scribble) throughout the segmentation process[3,4].

Benefiting from the rapid development of deep learning techniques, both AMIS and IMIS have witnessed great progress in their respective field. For AMIS, the emerging
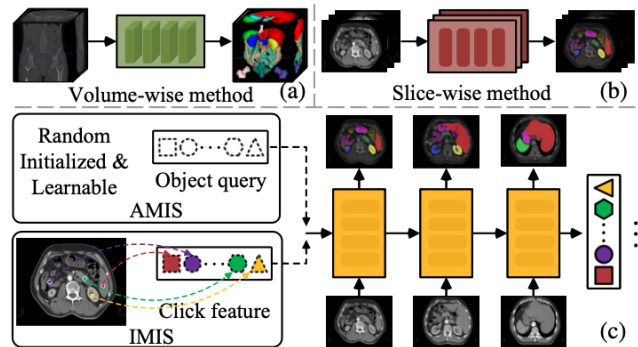
---

*Corresponding author: Yi Yang.

Figure 1. (a-b) Existing *volume-wise* and *slice-wise* solutions. (c) Our slice-to-volume solution that bridges distant slices by cluster center propagation and further unifies automatic/interactive segmentation under the same model with 2D segmentation networks.

of seminal work[5] leads the research efforts towards developing stronger backbones[6–8], harnessing multi-scale features[9–11] or incorporating attention mechanism[12–14], *etc*. Conversely, IMIS centers its primary focus on effectively integrating user inputs into segmentation models[15, 16], yielding remarkable performance. Nevertheless, such a tailored paradigm for each task greatly diffuses the research endeavors, impeding the seamless transfer of advancement made in one task to another due to the fundamental differences in model architecture and training strategy. Moreover, when working with the same dataset, current solutions necessitate the developing of two separate models for AMIS and IMIS, respectively. This results in a duplication in terms of both training time and network parameters.

In this work, we aim to formulate a ***universal*** segmentation framework capable of addressing both AMIS and IMIS within ***one unified model*** and a ***single training session***. Towards this, we first initiate a thorough exploration on the limitations commonly observed in current AMIS and IMIS solutions: **i)** the top-leading approaches for volume segmentation rely heavily on 3D networks which suffer from slow inference[9] and present significant challenges in deploying on hospital devices that usually exhibit limited parallel computation capabilities, **ii)** they prove inefficient in bridging remote slices due to the usage of sliding window inference to handle large memory consumption, which further hinders the broadcast of user inputs to entire volumes,

iii) current interactive solutions are limited to handle single foreground class, in contrast to automatic approaches which develop rapidly and excel in multi-class segmentation.

To solve the aforementioned limitations as well as reconcile AMIS and IMIS, we proposed S2VNet. It draws inspiration from the clustering-based image segmentation methods [17–20] that utilize a set of learnable queries as cluster centers to aggregate pixel feature associated with target objects, and update in an iterative manner. This insightful approach prompts us to reformulate volumetric segmentation by utilizing mere 2D segmentation models. Specifically, it is observed that objects in a volume usually manifest identical representation across different slices. This inherent consistency forms the basis for a novel slice-to-volume propagation approach that centroids after comprehensive updates in one slice can be passed forward and serve as the initial values for cluster centers in successive slices, facilitating effortless transfer of knowledge retrieved in prior segmentation process to the next round. This paradigm is simple yet powerful, harnessing both the key principle of clustering-based methodologies and the slice-wise structure of volumetric data. Moreover, this framework is readily adapted to IMIS without architectural changes by initializing centroids from backbone features at the position of user inputs, which clearly signify intended objects. Since there would be multiple clicks for a identical object, we further design an adaptive sampling strategy to reweight feature points when given new interactions. Finally, as the current pipeline may be affected by outliers and face decaying awareness to prior cues after rounds of propagation, we devise a recurrent centroid aggregation strategy to collect historic centroids and fuse them into a single vector which introduces nearly no additional cost to deliver a more robust network inference.

Taking advantage of such slice-to-volume propagation paradigm, S2VNet unveils several compelling facets: **First**, it seamlessly accommodates AMIS and IMIS into a unified model through a single training process, accomplished by initializing a subset of cluster centers from user inputs while the others are left as random, enabling both automatic and interactive segmentation learning. **Second**, in leveraging of reusing centroids, S2VNet extends user inputs or slice cues throughout the entire volume with 2D networks, contributing to a significant alleviation in computational resource (*i.e.*, 15 times faster inference speed and 48.2% memory reduction compared to 3D counterparts). **Third**, S2VNet can simultaneously accept multiple classes of user inputs, with each of them initializing one cluster center. This facilitates parallel refinement for multiple instances of ***different classes*** in a ***single network forward pass***, while prior work could only handle one foreground class [15, 21, 22]. **Fourth**, given the universal characteristic of S2VNet, we could build a diagnosis system that meets rigorous clinical requirements. Concretely, S2VNet is able to provide coarse segmentation results for multiple lesion/organ classes via AMIS. Physicians can then choose instances of interest and conduct refinement with precise feedback, saving considerable time in the initial search for lesions/organ areas.

We open a new avenue for medical image segmentation from the universal perceptive, and further provide a feasible solution via clustering-based slice-to-volume propagation. To comprehensively evaluate our method, we experiment S2VNet on three volumetric datasets, *i.e.*, WORD [23], BTCV [24], and AMOS [25]. Our empirical findings substantiate that S2VNet could consistently yield superior performance even compared to the specified solutions for each task, through the utilization of only one single model.

2. Related Work

Volumetric Medical Image Segmentation aims to segment organic or pathological structures [26–28] in 3D medical images and can be broadly grouped into two categories [29]: *slice-wise* and *volume-wise*. The *slice-wise* methods [7, 30–33] usually split 3D images into 2D slices along the z-axis, and then segment each slice separately. Since the proposal of [5], there has been a research surge based on the U-shaped architecture [13, 34–46]. Such a paradigm enjoys fast inference but makes no use of the 3D structure of images. In contrast, *volume-wise* methods [6, 8, 14, 47–49] directly process 3D images by extending 2D operations to their 3D counterparts. While capturing spatial context in three dimensions, they are inefficient in establishing meaningful connections between distant regions due to the limited receptive field of CNNs [50]. Recently, efforts have been made to leverage Transformer to capture long-range dependencies [51–58]. However, the inputs are still 3D image patches that contains only nearby slices, remaining unable to bridge remote slices.

In this work, we adhere to the *slice-wise* pipeline, but devise a novel slice-to-volume propagation mechanism characterized by utilizing pixel clustering to facilitate the storage and reusage of knowledge acquired in prior slices. This seamlessly associates predictions across individual slices. Finally, S2VNet combines the strengths of both efficient inference offered by *slice-wise* methods and the effective segmentation achieved in *volume-wise* methods, ultimately leading to accurate and consistent 3D predictions. The propagation strategy shares a similar spirit to object association in video segmentation [59–64]. However, these work mainly targets at challenges like fast motion, occlusion, and object reappearing, while S2VNet explicitly modeling object patterns via clustering. This benefits medical segmentation that usually contains no complex contextual cues.

Interactive Medical Image Segmentation. Though achieving promising performance, the above automatic methods still face challenges [65–67] in clinical applications due to the severe biological variation present in medical im-

ages[4]. In response to this, interactive medical image segmentation (IMIS)[68–73] is emerging as a practical strategy to improve accuracy by incorporating user interactions, which includes bounding boxes[74, 75], scribbles[70, 76], clicks[15, 16, 77], and extreme points[78]. Moreover, endeavors have been striven to enhance accuracy by emphasizing the effective integration of user interactions, such as extracting informative cue maps[15, 16, 79], or adapting networks to inference images[3, 80]. Recently, alternate research[4, 22, 81–83] explores interactive segmentation in a mask propagation manner, *i.e.*, wrapping the mask of previous slice according to affinity matrix to predict the next slices. They differ from S2VNet in two aspects: **i) core idea**: they follow the mask wrapping pipeline, while S2VNet passes cluster centers to enable continuous segmentation of targets; **ii) network architecture**: they need two distinct networks (one providing 2D predictions with user inputs and the other propagating the predictions) to conduct IMIS, and cannot handle AMIS, while S2VNet can tackle both AMIS and IMIS in an unified network.

Clustering-Based Image Segmentation. Prior to the resurgence of deep learning, clustering stood out as a straightforward yet highly efficient technique for image segmentation[84–86]. However, these traditional methods rely heavily on low-level features such as texture or color, limiting their capacity to capture high-level semantics[19]. Recent studies have explored the utilization of CNNs to extract feature representations[87–97], with mask predictions are delivered by clustering pixels into semantically coherent segments in a post-processing manner[17, 98–103]. There has also been a shift towards query-based Transformer approaches[88]. For instance, [18, 19] rethinks the relation between pixel features and object queries by reformulating cross-attention as a clustering solver. On this basis, [20] introduces a recurrent cross-attention mechanism which unlocks the power of iterative clustering in pixel grouping. Inspired by these work, S2VNet further extends pixel clustering to the continuous segmentation of 3D volumetric images, achieved by propagating clustering results (*i.e.*, centroids) of previous slices to the next. This not only preserves the coherence of segmentation results over the z-axis, but also establishes a robust prior for predicting identical objects. Moreover, it seamlessly and effortlessly adapts IMIS in the same architecture, contributing to the advancement of universal segmentation for volumetric images.

3. Methodology

3.1. Preliminary: *K-Means* Cross-Attention

Inspired by DETR[104], contemporary query-based image segmentation methods[88, 105] typically introduce a set of learnable embeddings as queries to collect pixel features associated with specific objects via *cross-attention*:

$$\hat{C} = C + \text{softmax}_{HW}(Q(K)^\top)V, \tag{1}$$

where $C \in \mathbb{R}^{N \times D}$ represents N object queries with dimension size D, $\hat{C}$ denotes the updated queries, $Q \in \mathbb{R}^{N \times D}$, $K \in \mathbb{R}^{HW \times D}$, $V \in \mathbb{R}^{HW \times D}$ stand for the features for query, key, and value. Here softmax_{HW} means to conduct softmax along the spatial dimension of image features, *i.e.*, computing the probability of affiliated to a unique query across all pixels. It is crucial to note that this mechanism involves attending to a substantial number of pixels. In contrast to above, [19] devise the *k-means* cross attention:

$$\hat{C} = C + \text{argmax}_N(Q(K)^\top)V. \tag{2}$$

Here, Eq. 2 compels Q to query pixel features belonging to a specific object, and subsequently inspect which query embedding within C these features correspond to by applying argmax along the query dimension N. Such process is similar to the *k-means*[106] algorithm which proceeds by alternating between the *assignment* and *update* two steps:

$$\begin{aligned} \text{Assignment Step:} \quad &\hat{C} = AV, \\ \text{Update Step:} \quad &\hat{C} = C + \hat{C}, \end{aligned} \tag{3}$$

where $A = \text{argmax}_N(Q(K)^\top)$ is the assignment matrix (*i.e.*, attention map) where each element indicates whether a pixel feature is assigned to a particular cluster. As a results, following the execution of a succession of Transformer decoder layers composed by *k-means* cross attention, the query embeddings C can be regarded as the cluster centers, which adeptly captures the representation of target objects.

3.2. Clustering Propagation-Driven Universal Segmentation Framework

Motivation. Given a volume $V \in \mathbb{R}^{C \times H \times W}$ with a spatial size of $H \times W$ for C slices, volumetric image segmentation aims to group it into a set of segments with corresponding semantic labels. This task is distinguished by the inherent structural property of volumetric image data, *i.e.*, anatomical or pathological regions of interest often spanning across multiple consecutive slices and exhibiting consistent visual patterns. This property allows the same class of targets in distinct slices to be compressed within a shared object-centric representation. Given this context, we introduce the clustering-based methodologies into volume segmentation, as shown in Fig. 2. Specifically, our approach involves extending the dynamic evolution of cluster centers C which is originally conducted within the image-level mask decoding process (Fig. 2 (a)) to volume-level by using the same collection of C throughout the segmentation for all slices in V (Fig. 2 (b)). As such, the separate *slice-wise* segmentation for each individual slice is seamlessly integrated into a coherent segmentation process, and iteratively delivering intermediate output for each slice.
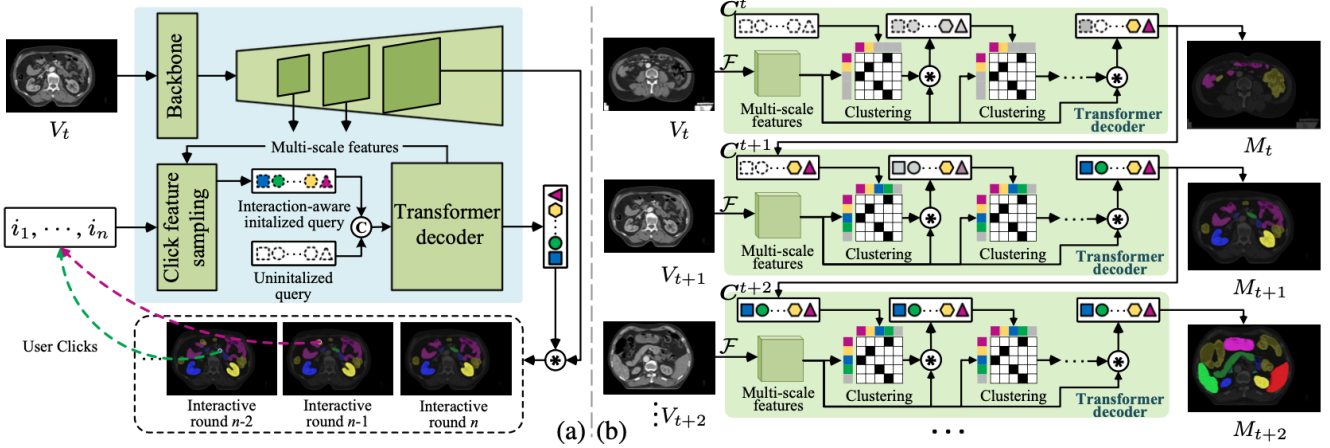
Figure 2. Our clustering propagation-driven universal segmentation framework (§3.2). (a) S2VNet adapts multi-class interactive segmentation and refinement by iteratively initializing cluster centers from user clicks. (b) Our clustering-based slice-to-volume propagation pipeline where centroids are evolved during slice-level segmentation and passed to the next slices. ©️ and ⊛ denote concatenation and dot product.

Slice-to-Volume Cluster Center Propagation. Denoting $\mathcal{F}$ as the feature encoder, N cluster centers $\{C_n^t\}_{n=1}^N$ are employed to extract the object-centric representation for each class within the given slice V_t by:

$$\{\hat{C}_n^t\}_{n=1}^N = \mathcal{D}(\mathcal{F}(V_t), \{C_n^t\}_{n=1}^N), \qquad (4)$$

where $\mathcal{D}$ is the Transformer decoder composed of *k-means* cross attention [19]. In the context of automatic volumetric image segmentation, the segmentation often begins from the first slice along the z-axis of the volume, which typically contains no foreground objects. It is the common case that these foreground objects usually appear in the middle part of the volume. To address the challenge that all cluster centers collect features of the background class and further impose negative impact to the segmentation of subsequent slices, only cluster centers matched with foreground classes will be propagated to the next slice. To achieve this, we perform one-to-one bipartite matching between the mask predictions $\{\hat{Y}_n^t\}_{n=1}^N$ and the ground truth $\{Y_k^t\}_{k=1}^K$ by:

$$\hat{\theta} = \arg\min_{\theta \in \Theta_N} \sum_{n=1}^N \mathcal{L}_{\text{match}}(Y_n, \hat{Y}_{\sigma(n)}). \qquad (5)$$

Here $\hat{\theta}$ represents the optimal assignment among a permutations of N elements $\theta \in \Theta_N$. Based on $\hat{\theta}$, we select cluster centers $\{\hat{C}_k^t\}_{k=1}^K$ associated with foreground classes and pass them to the next slice V_{t+1} as the initial values:

$$\{C_k^{t+1}\}_{k=1}^K = \{\hat{C}_k^t\}_{k=1}^K. \qquad (6)$$

As such, these object-centric representation could encapsulate the coherent appearances of regions across different slices, fostering a more compact and informative representation for subsequent segmentation and analysis. Note that during the inference stage, we keep elements in $\{\hat{C}_n^t\}_{n=1}^N$ only if the corresponding class $\{\hat{c}_n^t\}_{n=1}^N$ is not identified as the background class, and pass them to subsequent slices.

Interaction-Aware Cluster Center Initialization. In prior research [15, 71, 72], the user input is conventionally repre-

sented as an binary mask $M \in \{0, 1\}^{H \times W}$ where the foreground region signifies user guidance. Subsequently, M is combined with gray-scale images as inputs to segmentation networks. Though achieving promising results, such a paradigm suffers from several limitation: **i)** concatenating user inputs with images introduces architectural modifications and disrupts the integration with automatic segmentation into a unified framework, and **ii)** prior methods encounter challenges when accommodating multiple semantic classes, thereby limiting the application to more complex scenarios. To tackle above limitations, instead of explicitly incorporating user guidance as the inputs to networks, we harness the clustering-based property of S2VNet. Specifically, denoting $\{Q_k\}_{k=1}^K = \{(P_k, c_k, t_k)\}_{k=1}^K$ as a set of user inputs where each element Q_k represents a click P_k associated for one semantic class c_k annotated on the slice V_{t_k}, we initialize the cluster center C from user input by:

$$\hat{C}_k = \text{FFN}(O_k),$$
$$O_k = \text{Sample}(F_{t_k}, P_k), \qquad (7)$$

where Sample indicates retrieving the point feature O_k from backbone features F_{t_k} of slice V_{t_k} according to the click position P_k, and FFN is a simple feed forward network to project O_k to the same size as C. $\hat{C}_k$ further serves to aggregate pixel features similar to user indicated regions and will be passed to subsequent slices. This realizes user-guided segmentation across the whole volume by leveraging above automatic segmentation pipeline, while introducing no modification to the network architecture. Moreover, it can accommodate an **arbitrary number of classes** with each of them serving to initialize one cluster center, perfectly addressing all aforementioned limitations. Notably, extending beyond these benefits, such a centroid initialization-based interactive segmentation strategy offers several additional advantages: **first**, in contrast to prior work treating user interactions and images equally by concatenating them as

inputs which can not exercise the guidance ability of interactions to the fullest extent, our interaction-aware centroid initialization implicitly guarantees predictions always conforming to user highlighted regions and enhances interpretability. **Second**, our method enables unified learning for interactive/automatic segmentation, as the only difference lies in the initial states of centroids. The input data, network architecture, and training objectives remain consistent.

Adaptive Pixel Feature Sampling. Interactive segmentation commonly involves multiple rounds of refinement to improve the precision of previously segmentation results by incorporating newly provided user inputs. These iterative refinements yield multiple instances of Q_k associated with the same category label, thus calling for an adaptive strategy to initialize cluster centers for a specific semantic category from multiple user inputs. As the latest user input should play more important role in refinement compared to prior clicks, we adopt a weighted sum to combine the pixel feature O_k^r sampled from the user input at the current refinement round r with those sampled from prior rounds by:

$$
\begin{aligned}
\hat{O}_k^r &= O_k^r + \beta^1 \cdot O_k^{r-1} + \cdots + \beta^n \cdot O_k^1, \\
&= O_k^r + \beta \cdot \hat{O}_k^{r-1},
\end{aligned} \tag{8}
$$

where $\hat{O}_k^r$ is the weighted output controlled by the factor $\beta \in [0, 1]$. Then $\hat{O}_k^r$ at each round of refinement is utilized to initialize a new cluster center, delivering a pair of prediction $\{\hat{M}_k^r, \hat{c}_k^r\}$ where $\hat{M}_k^r \in \mathbb{R}^{C \times H \times W}$ is the binary mask score for all C slices in volume V and $\hat{c}_k^r \in \mathbb{R}^C$ is the score for class c_k. Consequently, multiple predictions are delivered for each semantic class. To obtain the ultimate output, we first multiply $\hat{c}_k$ with corresponding $\hat{M}_k$ and then retrieve the maximum value across all R rounds of predictions by:

$$
\hat{M}_k = \max_R (\hat{M}_k^0 \cdot \hat{c}_k^0, \cdots, \hat{M}_k^R \cdot \hat{c}_k^R). \tag{9}
$$

It is crucial to emphasize that for all refinement rounds in S2VNet, the pixel features associated with user inputs are sampled from the ***same backbone features*** which only need to be computed once. This stands in stark contrast to prior work[3, 15, 71] that repetitively combines prior results with image data and conducts a full network pass at each refinement round. This also contributes to accelerated inference and enhances the efficiency of computer-aided diagnosis.

Recurrent Centroid Aggregation. Though the cluster centers undergo continuous evolution during the mask decoding so as to effectively associate successive slices, they tend to be drifted by outliers such as foreign objects and artifacts commonly encountered in clinical practice[107], and lose track of distant structural cues with the slice-wise segmentation process iterates. To deliver a robust inference and retain enduring cues of remote slices, we propose to accumulate historic centroids of each slice and fuse them into a consolidated entity in a recurrent manner. As shown in Fig. 3,
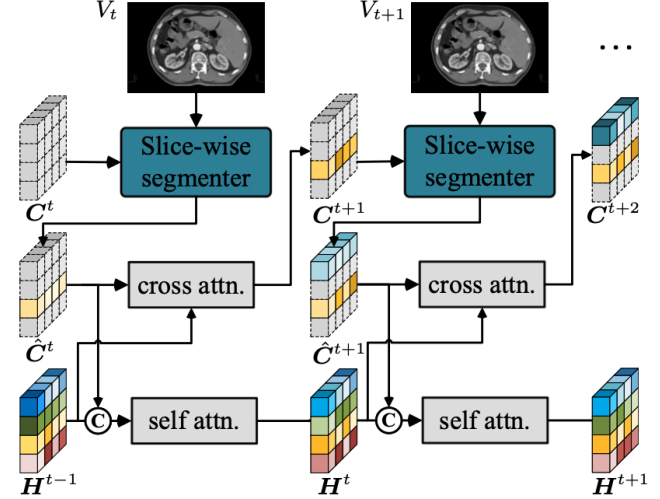

Figure 3. Illustration of **recurrent centroid aggregation** (§3.2). After clustering within the slice-wise segmentation for each slice, the centroids are recurrently merged with the historic ones to assist in the initialization of centroids belonging to the subsequent slice.

H_k^{t-1} is denoted as the fused vector, aggregating information from slice V_0 to V_{t-1}. When given new centroid $\hat{C}_k^t$ after mask decoding for slice V_t, we fuse it with H_k^{t-1}:

$$
H_k^t = \text{FFN}(\text{SelfAttn}([H_k^{t-1}; \hat{C}_k^t])), \tag{10}
$$

where $[;]$ means concatenation. Here the self-attention (*i.e.*, SelfAttn) is employed to identify the most relevant regions within the concatenated vector $[H_k^{t-1}; \hat{C}_k^t]$, and FFN is subsequently used to project it into the same dimension as $\hat{C}_k^t$. In this way, rather than introducing a memory bank which would impose additional GPU memory and computational time overhead, we efficiently store historic structural cues by recurrently merging new centroids into the existing one. Then, when initializing the centroid C_k^{t+2} for slice V_{t+2}, we incorporate not only the cluster center obtained after mask decoding at slice V_{t+1} (*i.e.*, $\hat{C}_k^{t+1}$), but also query the centroids from the previous t slices stored in H_k^t by:

$$
C_k^{t+2} = \hat{C}_k^{t+1} + \text{CrossAttn}(\hat{C}_k^{t+1}, H_k^t). \tag{11}
$$

Here CrossAttn refers to the standard cross-attention.

3.3. Implementation details

Network Configuration. S2VNet is constructed upon the clustering-based image segmenter. Specifically, for the slice-wise segmentation, we adopt Mask2Former[88] and integrate *k-means* cross attention[19] to replace the standard ones in the Transformer decoder. Other setups remain consistent to the default configuration. The positional encoding in Transformer is reserved to help capture the location of diseases. In order to align S2VNet with the most recent top-leading solutions[53, 58] for medical image segmentation that favor Transformer-based backbones, we employ Swin-B[108] for feature extraction. We empirically set the weighted factor β as 0.8 in adaptive pixel feature sampling.

Interaction Simulation. To evaluate S2VNet under the interactive setup, we opt for click as the primary mode of user interaction which is generally more accessible and can accommodate various input devices like mice, touchscreens, and styluses. Following conventions [15, 22, 109], we adhere to the automatic evaluation pipeline wherein the clicks are simulated based on ground truth and current segmentation results. Specifically, the initial click is sampled near the center of the target object, while subsequent clicks aimed at refinement are generated iteratively from the the most significant error regions by comparing the current prediction mask with the ground truth. The user clicks comprises both positive and negative ones with the former targeting foreground objects and the latter being applied to background.

Unified Segmentation Learning. To facilitate the slice-to-volume propagation learning, we randomly sample three slices from each volume and use clustering results obtained in previous slice to initialize centroids for the next slice. We designate 20 cluster centers for each semantic class, with each click serving as the trigger to initialize one of them, *i.e.*, allowing up to 20 clicks. Notably, for classes presenting in the inputs, there exists a 50% probability that the cluster centers are initialized from simulated user clicks, while the left are randomized initialized from empty, so as to enable both automatic and interactive segmentation learning. Following prior work [15, 49, 53, 58], the final learning target is the combination of the Cross Entropy loss and Dice loss.

4. Experiment

4.1. Experimental Setup

Datasets. Our experiments are conducted on three datasets:
- **WORD** [23] is a large-scale real clinical abdomen benchmark, providing high-quality annotations for up to 16 organs in the abdominal region. It contains 100/20/30 CT images for `train`/`val`/`test`, respectively.
- **BTCV** [24] consists of 30 CT volumes which is divided into 24 and 6 volumes for `train` and `val`. This dataset provides careful annotation for 13 organs, including 8 of them from Synapse. Following existing work [53, 58], We report the DSC score on all 13 abdominal organs.
- **AMOS** [25] is a large-scale diverse dataset collected from multiple centers and provides voxel-level annotations for 15 abdominal organs. It covers CT and MRI two modalities with each of them containing 200/100/200 and 40/20/40 scans for `train`/`val`/`test`.

Training. We train our network for 20k iterations with a batch size of 8. The AdamW [110] optimizer with a initial learning rate 0.0002 and weight decay 0.02 is adopted. The learning rate is scheduled following the step policy, *i.e.*, decaying by a factor of 10 at 14K and 18K steps, respectively. A learning rate multiplier of 0.1 is applied to the backbone which is initialized with ImageNet [111] pre-trained weights. After adapting the volumetric data in to 2D slices,

we employ z-score normalization to rescale image intensities within the range of 0 to 255. The remaining setups are determined following [15, 22, 49, 50, 53] for fair comparison. Specifically, for data augmentation, we use standard large-scale jittering (LSJ) augmentation with a random scaling sampled from range 0.5 to 1.75, followed by a fixed-size crop of 512×512 for all datasets [23–25]. The random horizontal flipping are also applied to enhance diversity.

Testing. The inference steps are tailored to optimize the usage of user inputs. Please note that we adopt the identical network architecture and model weight for both two tasks.
- *Automatic.* Inference starts from the first slice along the z-axis, proceeding sequentially till the final slice.
- *Interactive.* Inference is initiated from the slice with user inputs and broadcast bidirectionally throughout the entire volume, emphasizing the significance of user interactions.

For fair comparison, we follow prior work [56, 58] to use the input resolution of 512×512 for all datasets [23–25].

Evaluation Metric. Following the standard evaluation protocol [23, 58, 112], We employ Dice Similarity Coefficient (DSC) [113], Hausdorff Distance (HD) [114] and normalized surface dice (NSD) [25] to assess the performance under both automatic and interactive setups. DSC quantifies the overlap between predictions and ground-truths, whereas HD functions for measuring the 3D surface distance between them. To eliminate the impact of outliers, we employ HD95, which captures the 95% distance of all points in one surface to the other. For NSD, it scores the category-wise segmentation quality for evaluating precision of boundaries.

Reproducibility. S2VNet is implemented in PyTorch and trained on four NVIDIA Tesla A100 GPU. Evaluation for all methods is conducted on the same machine.

IMIS Comparison. As existing interactive approaches [4, 76, 81, 115] are limited to binary segmentation with single foreground class, we train an independent model for each target class while considering remaining classes as background. To render a more comprehensive comparison, we adapt the top-leading automatic work into the interactive setup by concatenating user clicks and prior round predictions with image data. Given this substantial workload, we only report performance for several representative classes with relatively lower performance across each dataset.

4.2. Comparison to State-of-the-Arts

WORD [23]. As shown in Table 1, S2VNet yields remarkable performance on the automatic setup, *i.e.*, surpassing SwinMM [56] by **1.18%** in terms of DSC and outperforming all 3D solutions in terms of HD95 which emphasizes on the coherence of predictions across slices. This demonstrates the effectiveness of our 2D slice-to-volume propagation strategy in bridging distance cues. Under the interactive setup, S2VNet achieve a **4.05%** average improvement in DSC compared to the automatic setup, verifying the supe-

| Methods | Average | | Liv | Spl | Kid L | Kid R | Sto | Gal | Eso | Pan | Duo | Col | Int | Adr | Rec | Bla | Fem L | Fem R |
	HD95↓	DSC↑																
Automatic Setup																		
UNETR [53]	17.34	79.77	94.67	92.85	91.49	91.72	85.56	65.08	67.71	74.79	57.56	74.62	80.40	60.76	74.06	85.42	89.47	90.17
CoTr [54]	12.83	84.66	95.58	94.90	93.26	93.63	89.99	76.40	74.37	81.02	63.58	84.14	86.39	69.06	80.00	89.27	91.03	91.87
Swin UNETR [55]	14.24	84.34	96.08	95.32	94.20	94.00	90.32	74.86	76.57	82.60	65.37	84.56	87.37	66.84	79.66	92.05	86.40	83.31
ESPNet [116]	15.02	79.92	95.64	93.90	92.24	94.39	87.37	67.19	67.91	75.78	62.03	78.77	72.80	60.55	74.32	78.58	88.24	89.04
DMFNet [117]	7.52	85.10	95.96	94.64	94.70	94.96	89.88	79.84	74.10	81.66	66.66	83.51	86.95	66.73	79.62	88.18	91.99	92.55
LCOVNet [118]	9.11	85.82	95.89	95.40	95.17	95.78	90.86	78.87	74.55	82.59	**68.23**	84.22	87.19	69.82	79.99	88.18	92.48	93.23
SwinMM [56]	9.35	86.18	96.30	95.46	93.83	94.47	91.43	80.08	76.59	83.60	67.38	**86.42**	**88.58**	69.12	80.48	90.56	92.16	92.40
S2VNet (Ours)	**4.64**	**87.36**	**96.72**	**96.01**	**95.84**	**95.93**	**91.80**	**82.96**	**77.28**	**85.10**	67.07	86.19	88.46	**72.40**	**83.27**	**91.73**	**93.30**	**93.75**
Interactive Setup																		
iSegFormer† [81]	-	-	-	92.14†	91.07†	93.86†	-	72.01†	73.37†	-	69.52†	-	-	69.91†	48.13†	-	-	-
Mem3D† [4]	-	-	-	94.88†	93.55†	93.96†	-	77.38†	80.61†	-	76.29†	-	-	74.57†	73.37†	-	-	-
SwinMM† [56]	-	-	-	95.78†	94.27†	95.11†	-	82.26†	80.33†	-	78.54†	-	-	72.96†	85.12†	-	-	-
S2VNet (Ours)	**3.28**	**91.41**	**96.91**	**96.37**	**96.15**	**96.22**	**94.79**	**87.23**	**86.32**	**88.51**	**83.91**	**90.50**	**91.17**	**77.73**	**90.73**	**94.35**	**95.85**	**95.82**

†: An independent model is trained for each class as prior work can only handle binary segmentation. Given substantial workload, we evaluate 8 classes.

Table 1. **Quantitative segmentation results** with comprehensive scoring for each organ on WORD[23] `test`.

Method	Avg DSC	Gal	Eso	IVC	PSV	RAG	LAG
Automatic Setup							
TransUNet [50]	76.72	59.84	70.96	77.23	71.47	65.24	64.06
TransBTS [57]	81.31	68.38	75.61	82.48	74.21	67.23	67.03
UNETR [53]	76.00	58.23	71.21	76.51	70.37	66.25	63.04
Swin-UNETR [55]	80.44	65.37	75.43	81.61	76.30	68.23	66.02
nnFormer [58]	81.62	65.29	76.22	80.80	75.97	**70.20**	66.05
3D-UX-Net [119]	80.76	64.32	75.17	80.42	75.39	69.52	65.77
S2VNet (Ours)	**83.81**	**65.63**	**78.29**	**84.41**	**79.77**	68.38	**72.28**
Interactive Setup							
iSegFormer† [81]	-	-	69.37†	72.78†	-	64.40†	66.89†
Mem3D† [4]	-	-	74.84†	79.52†	-	68.45†	67.88†
nnFormer† [58]	-	-	82.47†	83.65†	-	70.41†	67.34†
S2VNet (Ours)	**86.11**	**69.94**	**87.92**	**89.96**	**81.64**	**72.23**	**73.22**

†: An independent model is trained for each target class.

Table 2. **Quantitative segmentation results** on BTCV[24] `val`.

| Method | Average | | CT | | MRI | |
	DSC↑	NSD↑	DSC↑	NSD↑	DSC↑	NSD↑
Automatic Setup						
CoTr [54]	77.31	67.12	77.13	64.15	77.50	70.10
UNETR [53]	76.81	63.40	78.33	61.49	75.30	65.3
TransUNet [50]	-	-	85.05	73.86	-	-
TransBTS [57]	-	-	86.52	75.49	-	-
nnFormer [58]	83.12	74.07	85.63	74.15	80.60	**74.00**
Swin UNETR [55]	81.04	70.60	86.37	75.32	75.70	65.80
3D-UX-Net [119]	-	-	87.28	76.48	-	-
S2VNet (Ours)	**86.22**	**77.23**	**87.80**	**82.89**	**84.64**	71.57
Interactive Setup						
S2VNet (Ours)	**88.75**	**80.94**	**89.65**	**85.27**	**87.84**	**76.61**

Table 3. **Quantitative segmentation results** on AMOS[25] `val`.

#	S2V Propagation	Interaction Initialization	Adaptive Sampling	Recurrent Aggregation	HD95↓	DSC↑
1					16.63	78.67
2	✓				5.03	86.19
3	✓			✓	4.64	87.36
4	✓	✓			4.30	89.70
5	✓	✓	✓		3.79	90.64
6	✓	✓	✓	✓	**3.28**	**91.41**

Table 4. **Analysis of essential component** on WORD[23] `test`.

riority of our interaction-aware centroid initialization strategy. Especially, our approach boosts the performance up to **83.91%** for the class 'Duo.', surpassing both existing interactive and adapted automatic approaches by a large margin.

BTCV [24]. Table 2 compares our method against several top-leading approaches on BTCV[24] `val`. As seen, S2VNet achieves the best performance on both automatic and interactive setups. In particularly, compared with nn-Former [58] which is the previous SOTA, our approach earns **2.19%** improvement in terms of averaged DSC score for the automatic setup. This indicates that S2VNet can generalize well to different datasets with various challenging scenarios. We also provide detailed scores for six representative organs with poor performance, where S2VNet gives **2%~6%** performance gain compared to prior work.

AMOS[25]. Table 3 confirms again the exceptional performance of S2VNet in the segmentation of both CT and MRI images. Specifically, our algorithm achieves an improvement of **0.52%/6.41%** over 3D-UX-Net [119] in terms of DSC/NSD. Moreover, with the incorporation of interaction-aware query initialization, S2VNet consistently surpasses existing methods across all modalities and metrics.

4.3. Qualitative Comparison Result

Fig. 4 depicts visual comparison on WORD[23] `test`. As seen, S2VNet yields more accurate results compared to SwinMM[56], and the interactive mode can handle various challenging cases with small objects or distortions.

4.4. Diagnostic Experiments

To evaluate the core designs and gain further insights, we conduct a series of ablative studies on WORD[23] `test`.
Key Component Analysis. We first examine the efficacy of each component in Table 4, where the *row* #1 indicates directly segmenting each slice using 2D networks without any form of association. Upon the integration of clustering-based slice-to-volume propagation (*i.e.*, *row* #2), both DSC
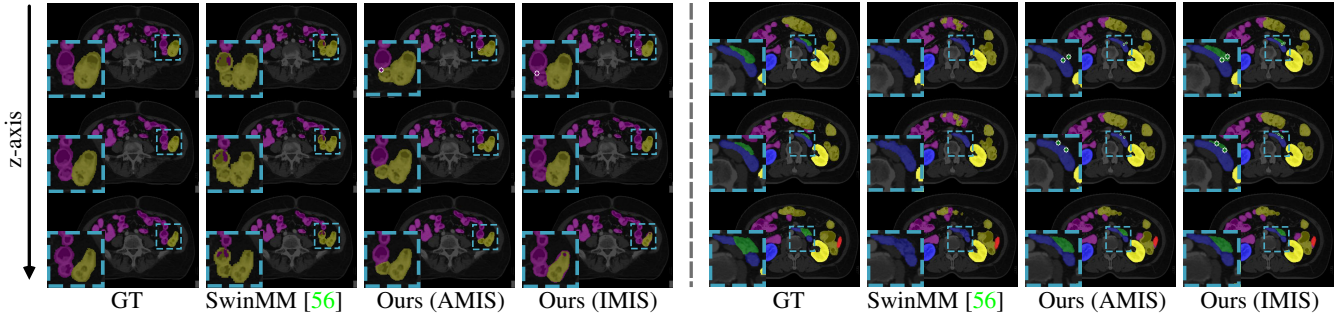
| | GT | SwinMM [56] | Ours (AMIS) | Ours (IMIS) | | GT | SwinMM [56] | Ours (AMIS) | Ours (IMIS) |

Figure 4. **Visual comparison results** on WORD[23] `test`. See §4.3 for detailed analysis.

Method	Memory (G) ↓	Volume Per Minute ↑	HD95 ↓	DSC ↑
CoTr [54]	26	0.18	12.83	84.66
Swin UNTER [55]	23	0.21	14.24	84.34
SwinMM [56]	27	0.15	9.35	86.18
Baseline	11	2.69	16.63	78.67
S2VNet	**14**	**2.33**	**4.64**	**87.36**

Table 5. **Comparison of running efficiency** on WORD[23] `test`.
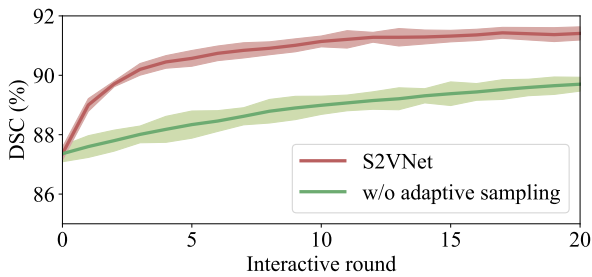
Figure 5. **Convergence analysis** on WORD[23] `test`. We report the DSC score with different round of user interactions.

Figure 6. **Analysis of unified training** on WORD[23] `test`.

and HD95 exhibit noteworthy improvement which demonstrate the effectiveness of our design. For interactive segmentation, as seen in *row* #4, our interaction-aware centroid initialization strategy can bring up to **3.51%** performance gains in DSC. With adaptive pixel-feature sampling (*i.e.*, *row* #5) to fuse different rounds of user interactions, the performance further boosts to **90.64%**. Finally, after incorporating recurrent centroid aggregation, S2VNet obtains the best performance on both setups (*i.e.*, *row* #3 and #6), underscoring the general compatibility of this module.

Run-Time Analysis. Next we probe the running efficiency of S2VNet during inference. Here 'Baseline' represents 2D segmentation network without association. As evidenced in Table 5, S2VNet achieves nearly **15** times faster inference speed in terms of FPS and saves **48.2%** memory usage compared to the previous state-of-the-art (*i.e.*, SwinMM [56]). Moreover, our association strategy incurs minor additional cost compared to the baseline method while elevating the performance by an impressive **8.69%** in DSC scores. All of the above confirms the urgency of shifting the traditional 3D segmentation paradigm to a more efficient one, with S2VNet providing a pragmatic and effective answer.

Convergence Analysis. We study the correlation between the number of refinement rounds and resulting DSC scores
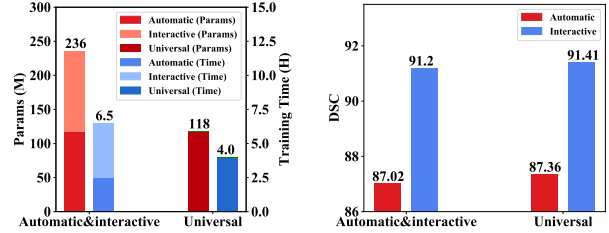
on WORD[23] `val`. As seen in Fig. 5, the performance of S2VNet exhibits a stable improvement as the rounds of refinement increase, and consistently outperforms the variant that without adaptive feature sampling to consider interactions in prior rounds. To strike a balance between accuracy and efficiency, we constrain the average refinement rounds to 15 from which there is no significant gain in performance.

Unified Training. We provide the network parameters and training time comparison between task-specific models for automatic/interactive segmentation and the universal model in Fig. 6. As seen, our universal model requires only half of parameters and training times. Furthermore, the performance under such a unified training paradigm even enjoys improvement compared to task-specific training strategies.

5. Conclusion

We present S2VNet, a unified framework to tackle automatic/interaction medical image segmentation in a slice-to-volume propagation manner. It makes use of the clustering-based methods, wherein the knowledge pertaining to targets is compressed within centroids and passed to next slices to produce coherent and robust predictions with merely 2D segmentation networks. On this basis, we realize interactive segmentation by initialize the cluster centers with respect to user guidance. This also facilitates concurrent interaction across multiple classes which overcomes the limitation of prior work confined to binary setups. Finally, to eliminate the impact of outliers and enhance the awareness to preceding slice cues, a recurrent aggregation approach is proposed to collect historic centroids. All of the above contributes to a flexible solution for volumetric image segmentation characterized by remarkable speed and state-of-the-art accuracy.

References

[1] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. Medical image segmentation using deep learning: A survey. *IET Image Processing*, 16(5):1243–1267, 2022. 1

[2] Dzung L Pham, Chenyang Xu, and Jerry L Prince. Current methods in medical image segmentation. *Annual Review of Biomedical Engineering*, 2(1):315–337, 2000. 1

[3] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE TMI*, 37(7):1562–1573, 2018. 1, 3, 5

[4] Tianfei Zhou, Liulei Li, Gustav Bredell, Jianwu Li, Jan Unkelbach, and Ender Konukoglu. Volumetric memory network for interactive medical image segmentation. *Medical Image Analysis*, 83:102599, 2023. 1, 3, 6, 7

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1, 2

[6] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *IEEE 3DV*, 2016. 1, 2

[7] JMJ Valanarasu and VM Patel. Unext: Mlp-based rapid medical image segmentation network. In *MICCAI*, 2022. 1, 2

[8] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *MICCAI*, 2016. 1, 2

[9] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multiscale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017. 1

[10] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE TMI*, 39(6):1856–1867, 2019. 1

[11] Abhishek Srivastava, Debesh Jha, Sukalpa Chanda, Umapada Pal, Håvard D Johansen, Dag Johansen, Michael A Riegler, Sharib Ali, and Pål Halvorsen. Msrf-net: a multiscale residual fusion network for biomedical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(5):2252–2263, 2021. 1

[12] Yan Wang, Yuyin Zhou, Wei Shen, Seyoun Park, Elliot K Fishman, and Alan L Yuille. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *Medical Image Analysis*, 55:88–102, 2019. 1

[13] Fei Ding, Gang Yang, Jinlu Liu, Jun Wu, Dayong Ding, Jie Xv, Gangwei Cheng, and Xirong Li. Hierarchical attention networks for medical image segmentation. *arXiv preprint arXiv:1911.08777*, 2019. 1, 2

[14] Xudong Wang, Shizhong Han, Yunqiang Chen, Dashan Gao, and Nuno Vasconcelos. Volumetric attention for 3d medical image segmentation and detection. In *MICCAI*, 2019. 1, 2

[15] Guotai Wang, Maria A Zuluaga, Wenqi Li, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE TPAMI*, 41(7):1559–1572, 2018. 1, 2, 3, 4, 5, 6

[16] Xiangde Luo, Guotai Wang, Tao Song, Jingyang Zhang, Michael Aertsen, Jan Deprest, Sebastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning. *Medical Image Analysis*, 72:102102, 2021. 1, 3

[17] Davy Neven, Bert De Brabandere, Marc Proesmans, and Luc Van Gool. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In *CVPR*, 2019. 2, 3

[18] Qihang Yu, Huiyu Wang, Dahun Kim, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In *CVPR*, 2022. 2, 3

[19] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *ECCV*, 2022. 2, 3, 4, 5

[20] James Liang, Tianfei Zhou, Dongfang Liu, and Wenguan Wang. Clustseg: Clustering for universal segmentation. In *ICML*, 2023. 2, 3

[21] Wentao Liu, Chaofan Ma, Yuhuan Yang, Weidi Xie, and Ya Zhang. Transforming the interactive segmentation for medical imaging. In *MICCAI*, 2022. 2

[22] Tianfei Zhou, Liulei Li, Gustav Bredell, Jianwu Li, and Ender Konukoglu. Quality-aware memory network for interactive volumetric image segmentation. In *MICCAI*, 2021. 2, 3, 6

[23] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N Metaxas, Guotai Wang, and Shaoting Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642, 2022. 2, 6, 7, 8

[24] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015. 2, 6, 7

[25] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *NeurIPS*, 2022. 2, 6, 7

[26] Yuhang Ding, Xin Yu, and Yi Yang. Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *CVPR*, 2021. 2

[27] Yuhang Ding, Xin Yu, and Yi Yang. Modeling the probabilistic distribution of unlabeled data for one-shot medical image segmentation. In *AAAI*, 2021. 2

[28] Yuxing Wang, Wenguan Wang, Dongfang Liu, Wenpin Hou, Tianfei Zhou, and Zhicheng Ji. Genesegnet: a deep learning framework for cell segmentation by integrating gene expression and imaging. *Genome Biology*, 24(1):235, 2023. 2

[29] Zhuotun Zhu, Yingda Xia, Wei Shen, Elliot Fishman, and Alan Yuille. A 3d coarse-to-fine framework for volumetric medical image segmentation. In *IEEE 3DV*, 2018. 2

[30] Rodney LaLonde, Ziyue Xu, Ismail Irmakci, Sanjay Jain, and Ulas Bagci. Capsules for biomedical image segmentation. *Medical Image Analysis*, 68:101889, 2021. 2

[31] Duo Wang, Ming Li, Nir Ben-Shlomo, C Eduardo Corrales, Yu Cheng, Tao Zhang, and Jagadeesan Jayender. Mixed-supervised dual-network for medical image segmentation. In *MICCAI*, 2019. 2

[32] Xiaowei Xu, Qing Lu, Lin Yang, Sharon Hu, Danny Chen, Yu Hu, and Yiyu Shi. Quantization of fully convolutional networks for accurate biomedical image segmentation. In *CVPR*, 2018. 2

[33] Wenxuan Zou, Xingqun Qi, Wanting Zhou, Muyi Sun, Zhenan Sun, and Caifeng Shan. Graph flow: Cross-layer graph flow distillation for dual efficient medical image segmentation. *IEEE TMI*, 42(4):1159–1171, 2022. 2

[34] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *MICCAI Workshop*, 2018. 2

[35] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *IEEE ICASSP*, 2020. 2

[36] Sihang Zhou, Dong Nie, Ehsan Adeli, Jianping Yin, Jun Lian, and Dinggang Shen. High-resolution encoder–decoder networks for low-contrast medical image segmentation. *IEEE TIP*, 29:461–475, 2019. 2

[37] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE TMI*, 38(10):2281–2292, 2019. 2

[38] Jeya Maria Jose Valanarasu, Vishwanath A Sindagi, Ilker Hacihaliloglu, and Vishal M Patel. Kiu-net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation. *IEEE TMI*, 41(4):965–976, 2021. 2

[39] Sheng He, Yanfang Feng, P Ellen Grant, and Yangming Ou. Segmentation ability map: Interpret deep features for medical image segmentation. *Medical Image Analysis*, 84:102726, 2023. 2

[40] Yanhao Zhu, Zhineng Chen, Shuai Zhao, Hongtao Xie, Wenming Guo, and Yongdong Zhang. Ace-net: biomedical image segmentation with augmented contracting and expansive paths. In *MICCAI*, 2019. 2

[41] Hong Joo Lee, Jung Uk Kim, Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Structure boundary preserving segmentation for medical image with ambiguous boundary. In *CVPR*, 2020. 2

[42] Zhijie Zhang, Huazhu Fu, Hang Dai, Jianbing Shen, Yanwei Pang, and Ling Shao. Et-net: A generic edge-attention guidance network for medical image segmentation. In *MICCAI*, 2019. 2

[43] Ran Gu, Guotai Wang, Tao Song, Rui Huang, Michael Aertsen, Jan Deprest, Sébastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE TMI*, 40(2):699–711, 2020. 2

[44] Yuyu Guo, Lei Bi, Euijoon Ahn, Dagan Feng, Qian Wang, and Jinman Kim. A spatiotemporal volumetric interpolation network for 4d dynamic medical image. In *CVPR*, 2020. 2

[45] Xinyi Wang, Tiange Xiang, Chaoyi Zhang, Yang Song, Dongnan Liu, Heng Huang, and Weidong Cai. Bix-nas: Searching efficient bi-directional architecture for medical image segmentation. In *MICCAI*, 2021. 2

[46] Xingang Yan, Weiwen Jiang, Yiyu Shi, and Cheng Zhuo. Ms-nas: Multi-scale neural architecture search for medical image segmentation. In *MICCAI*, 2020. 2

[47] Yuanfeng Ji, Ruimao Zhang, Zhen Li, Jiamin Ren, Shaoting Zhang, and Ping Luo. Uxnet: Searching multi-level feature aggregation for 3d medical image segmentation. In *MICCAI*, 2020. 2

[48] Qihang Yu, Dong Yang, Holger Roth, Yutong Bai, Yixiao Zhang, Alan L Yuille, and Daguang Xu. C2fnas: Coarse-to-fine neural architecture search for 3d medical image segmentation. In *CVPR*, 2020. 2

[49] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 2, 6

[50] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2, 6, 7

[51] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *MICCAI*, 2021. 2

[52] Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *MICCAI*, 2021. 2

[53] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *WACV*, 2022. 2, 5, 6, 7

[54] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *MICCAI*, 2021. 2, 7, 8

[55] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, 2021. 2, 7, 8

[56] Yiqing Wang, Zihan Li, Jieru Mei, Zihao Wei, Li Liu, Chen Wang, Shengtian Sang, Alan L Yuille, Cihang Xie, and Yuyin Zhou. Swinmm: masked multi-view with swin transformers for 3d medical image segmentation. In *MICCAI*, 2023. 2, 6, 7, 8

[57] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *MICCAI*, 2021. 2, 7

[58] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021. 2, 5, 6, 7

[59] Rajat Koner, Tanveer Hannan, Suprosanna Shit, Sahand Sharifzadeh, Matthias Schubert, Thomas Seidl, and Volker Tresp. Instanceformer: An online video instance segmentation framework. In *AAAI*, volume 37, pages 1188–1195, 2023. 2

[60] Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. A generalized framework for video instance segmentation. In *CVPR*, pages 14623–14632, 2023. 2

[61] Wenguan Wang, Jianbing Shen, Fatih Porikli, and Ruigang Yang. Semi-supervised video object segmentation with super-trajectories. *IEEE TPAMI*, 41(4):985–998, 2018. 2

[62] Liulei Li, Tianfei Zhou, Wenguan Wang, Lu Yang, Jianwu Li, and Yi Yang. Locality-aware inter-and intra-video reconstruction for self-supervised correspondence learning. In *CVPR*, 2022. 2

[63] Liulei Li, Wenguan Wang, Tianfei Zhou, Jianwu Li, and Yi Yang. Unified mask embedding and correspondence learning for self-supervised video segmentation. In *CVPR*, 2023. 2

[64] Yurong Zhang, Liulei Li, Wenguan Wang, Rong Xie, Li Song, and Wenjun Zhang. Boosting video object segmentation via space-time correspondence learning. In *CVPR*, 2023. 2

[65] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Adversarial style mining for one-shot unsupervised domain adaptation. *NeurIPS*, 33:20612–20623, 2020. 2

[66] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, pages 2507–2516, 2019. 2

[67] Yawei Luo, Ping Liu, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Category-level adversarial adaptation for semantic segmentation using purified features. *IEEE TPAMI*, 44(8):3940–3956, 2021. 2

[68] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062, 2021. 3

[69] Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *ICCV*, 2001. 3

[70] Guotai Wang, Maria A Zuluaga, Rosalind Pratt, Michael Aertsen, Tom Doel, Maria Klusmann, Anna L David, Jan Deprest, Tom Vercauteren, and Sébastien Ourselin. Slicseg: A minimally interactive segmentation of the placenta from sparse and motion-corrupted fetal mri in multiple views. *Medical Image Analysis*, 34:137–147, 2016. 3

[71] Ruiwei Feng, Xiangshang Zheng, Tianxiang Gao, Jintai Chen, Wenzhe Wang, Danny Z Chen, and Jian Wu. Interactive few-shot learning: Limited supervision, better medical image segmentation. *IEEE TMI*, 40(10):2575–2588, 2021. 3, 4, 5

[72] Guotai Wang, Michael Aertsen, Jan Deprest, Sébastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Uncertainty-guided efficient interactive refinement of fetal brain segmentation from stacks of mri slices. In *MICCAI*, 2020. 3, 4

[73] Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N Straehle, Bernhard X Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier, Markus Rudy, et al. Ilastik: interactive machine learning for (bio) image analysis. *Nature Methods*, 16(12):1226–1232, 2019. 3

[74] Martin Rajchl, Matthew CH Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Mellisa Damodaram, Mary A Rutherford, Joseph V Hajnal, Bernhard Kainz, et al. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE TMI*, 36(2):674–683, 2016. 3

[75] Yuyin Zhou, Lingxi Xie, Wei Shen, Yan Wang, Elliot K Fishman, and Alan L Yuille. A fixed-point model for pancreas segmentation in abdominal ct scans. In *MICCAI*, 2017. 3

[76] Hyeonsoo Lee and Won-Ki Jeong. Scribble2label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency. In *MICCAI*, 2020. 3, 6

[77] Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajadin, and Nasir Rajpoot. Nuclick: a deep learning framework for interactive segmentation of microscopic images. *Medical Image Analysis*, 65:101771, 2020. 3

[78] Shadab Khan, Ahmed H Shahin, Javier Villafruela, Jianbing Shen, and Ling Shao. Extreme points derived confidence map as a cue for class-agnostic interactive segmentation using deep neural network. In *MICCAI*, 2019. 3

[79] Zdravko Marinov, Rainer Stiefelhagen, and Jens Kleesiek. Guiding the guidance: A comparative analysis of user guidance signals for interactive segmentation of volumetric images. *arXiv preprint arXiv:2303.06942*, 2023. 3

[80] Bhavani Sambaturu, Ashutosh Gupta, CV Jawahar, and Chetan Arora. Efficient and generic interactive segmentation framework to correct mispredictions during clinical evaluation of medical images. In *MICCAI*, 2021. 3

[81] Qin Liu, Zhenlin Xu, Yining Jiao, and Marc Niethammer. isegformer: Interactive segmentation via transformers with application to 3d knee mr images. In *MICCAI*, 2022. 3, 6, 7

[82] Luyue Shi, Xuanye Zhang, Yunbi Liu, and Xiaoguang Han. A hybrid propagation network for interactive volumetric image segmentation. In *MICCAI*, 2022. 3

[83] Qin Liu, Meng Zheng, Benjamin Planche, Zhongpai Gao, Terrence Chen, Marc Niethammer, and Ziyan Wu. Exploring cycle consistency learning in interactive volume segmentation. *arXiv preprint arXiv:2303.06493*, 2023. 3

[84] Nameirakpam Dhanachandra, Khumanthem Manglem, and Yambem Jina Chanu. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764–771, 2015. 3

[85] SM Aqil Burney and Humera Tariq. K-means cluster analysis for image segmentation. *International Journal of Computer Applications*, 96(4), 2014. 3

[86] Siddheswar Ray and Rose H Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *International Conference on Advances in Pattern Recognition and Digital Techniques*, 1999. 3

[87] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 3

[88] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 3, 5

[89] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. Gmmseg: Gaussian mixture based generative semantic segmentation models. *NeurIPS*, 35:31360–31375, 2022. 3

[90] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *CVPR*, pages 2582–2593, 2022. 3

[91] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, pages 7303–7313, 2021. 3

[92] Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. In *CVPR*, pages 1246–1257, 2022. 3

[93] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. Logic-induced diagnostic reasoning for semi-supervised semantic segmentation. In *ICCV*, 2023. 3

[94] Liulei Li, Wenguan Wang, and Yi Yang. Logicseg: Parsing visual semantics with neural logic learning and reasoning. In *ICCV*, 2023. 3

[95] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023. 3

[96] Liulei Li, Wenguan Wang, Tianfei Zhou, Ruijie Quan, and Yi Yang. Semantic hierarchy-aware segmentation. *IEEE TPAMI*, 2023. 3

[97] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *CVPR*, 2024. 3

[98] Shu Kong and Charless C Fowlkes. Recurrent pixel embedding for instance grouping. In *CVPR*, 2018. 3

[99] Tuo Feng, Wenguan Wang, Xiaohan Wang, Yi Yang, and Qinghua Zheng. Clustering based point cloud representation learning for 3d analysis. In *ICCV*, 2023. 3

[100] Junbo Yin, Dingfu Zhou, Liangjun Zhang, Jin Fang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Proposal-contrast: Unsupervised pre-training for lidar-based 3d object detection. In *ECCV*, 2022. 3

[101] Guikun Chen, Xia Li, Yi Yang, and Wenguan Wang. Rethinking visual feature extraction: Modeling representatives from a neural clustering view. In *CVPR*, 2024. 3

[102] Ruijie Quan, Wenguan Wang, Fan Ma, Hehe Fan, and Yi Yang. Clustering for protein representation learning. In *CVPR*, 2024. 3

[103] Tuo Feng, Ruijie Quan, Xiaohan Wang, Wenguan Wang, and Yi Yang. Interpretable3d: An ad-hoc interpretable classifier for 3d point clouds. In *AAAI*, 2024. 3

[104] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[105] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 3

[106] Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. 3

[107] Marcel Prastawa, Elizabeth Bullitt, Sean Ho, and Guido Gerig. A brain tumor segmentation framework based on outlier detection. *Medical Image Analysis*, 8(3):275–283, 2004. 5

[108] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5

[109] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *CVPR*, 2016. 6

[110] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[111] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[112] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 6

[113] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 6

[114] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE TPAMI*, 15(9):850–863, 1993. 6

[115] Xuan Liao, Wenhao Li, Qisen Xu, Xiangfeng Wang, Bo Jin, Xiaoyun Zhang, Yanfeng Wang, and Ya Zhang. Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning. In *CVPR*, 2020. 6

[116] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *ECCV*, 2018. 7

[117] Chen Chen, Xiaopeng Liu, Meng Ding, Junfeng Zheng, and Jiangyun Li. 3d dilated multi-fiber network for real-time brain tumor segmentation in mri. In *MICCAI*, 2019. 7

[118] Qianfei Zhao, Huan Wang, and Guotai Wang. Lcov-net: A lightweight neural network for covid-19 pneumonia lesion segmentation from 3d ct images. In *International Symposium on Biomedical Imaging*, 2021. 7

[119] Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A Landman. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. In *ICLR*, 2023. 7