Medical Image Analysis 89 (2023) 102870

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

ADNet++: A few-shot learning framework for multi-class medical image volume segmentation with uncertainty-guided feature refinement

Stine Hansen *, Srishti Gautam, Suaiba Amina Salahuddin, Michael Kampffmeyer, Robert Jenssen

Department of Physics and Technology, UiT The Arctic University of Norway, NO-9037 Tromsø, Norway

ARTICLE INFO

Keywords:
Few-shot segmentation
Medical image segmentation
Uncertainty estimation

ABSTRACT

A major barrier to applying deep segmentation models in the medical domain is their typical data-hungry nature, requiring experts to collect and label large amounts of data for training. As a reaction, prototypical few-shot segmentation (FSS) models have recently gained traction as data-efficient alternatives. Nevertheless, despite the recent progress of these models, they still have some essential shortcomings that must be addressed. In this work, we focus on three of these shortcomings: (i) the lack of uncertainty estimation, (ii) the lack of a guiding mechanism to help locate edges and encourage spatial consistency in the segmentation maps, and (iii) the models' inability to do one-step multi-class segmentation. Without modifying or requiring a specific backbone architecture, we propose a modified prototype extraction module that facilitates the computation of uncertainty maps in prototypical FSS models, and show that the resulting maps are useful indicators of the model uncertainty. To improve the segmentation around boundaries and to encourage spatial consistency, we propose a novel feature refinement module that leverages structural information in the input space to help guide the segmentation in the feature space. Furthermore, we demonstrate how uncertainty maps can be used to automatically guide this feature refinement. Finally, to avoid ambiguous voxel predictions that occur when images are segmented class-by-class, we propose a procedure to perform one-step multi-class FSS. The efficiency of our proposed methodology is evaluated on two representative datasets for abdominal organ segmentation (CHAOS dataset and BTCV dataset) and one dataset for cardiac segmentation (MS-CMRSeg dataset). The results show that our proposed methodology significantly (one-sided Wilcoxon signed rank test, $p < 0.05$) improves the baseline, increasing the overall dice score with +5.2, +5.1, and +2.8 percentage points for the CHAOS dataset, the BTCV dataset, and the MS-CMRSeg dataset, respectively.

1. Introduction

Accurate image segmentation is an essential prerequisite for various clinical applications, such as radiotherapy treatment planning (Gonzalez et al., 2021), tissue quantification (Militello et al., 2019), and diagnostics (Tsochatzidis et al., 2021). Prototypical few-shot segmentation (FSS) models have recently shown promise as data efficient alternatives to solving this task by using a small set of labeled examples to extract class-wise prototypes that can be leveraged to segment objects in new images (Tang et al., 2021; Yu et al., 2021; Ouyang et al., 2022; Hansen et al., 2022). These models thus eliminate the need to collect and annotate large amounts of images, which is a key challenge for the application of deep learning models in the medical domain (Shen et al., 2020). In particular, Hansen et al. (2022) propose ADNet, an anomaly detection-inspired approach to FSS that simplifies the problem by refraining from explicitly modeling the difficult background class. This results in a model that is robust to the large and inhomogeneous background class, thus for the first time enabling one-step volume-wise prototypical FSS, yielding state-of-the-art performance.

When trained, the FSS models mentioned above can generalize from a *few* labeled samples to solve new segmentation tasks during inference. Specifically, a few labeled examples are exploited to extract class-wise prototypes that are used to make predictions on the unlabeled test data. However, despite their recent advances, current FSS models have some fundamental shortcomings that need to be addressed to approach clinical application.

Firstly, existing medical FSS models do *not* provide any measure of uncertainty for their predictions, which limits their trustworthiness. Knowing when the model is uncertain and therefore more likely to make mistakes is important information that should accompany the prediction in a safety-critical application such as medical image segmentation (Kompa et al., 2021).

Secondly, in current methods, the segmentation is performed directly on the spatially compressed feature representation, without any

* Corresponding author.
E-mail address: s.hansen@uit.no (S. Hansen).

https://doi.org/10.1016/j.media.2023.102870
Received 30 October 2022; Received in revised form 23 April 2023; Accepted 12 June 2023
Available online 26 June 2023

mechanism to guide the precise location of edges and structures in the image. The final segmentation map is simply obtained by re-sampling the output via bi-/tri-linear up-sampling, resulting in segmentation masks that typically struggle to accurately locate edges.

Finally, in medical image segmentation, there are often *multiple* foreground classes of relevance, e.g. a number of different organs. However, current medical FSS methods only focus on *binary* foreground/background segmentation and are forced to segment the images class-by-class. In addition to unnecessary forward passes, this can lead to regions with ambiguous predictions as voxels might get classified as "foreground" for multiple classes.

In this work, we focus on the inference phase to address the above-mentioned shortcomings. Without requiring modification or re-training of the network parameters, we develop methods to better exploit the available information in order to provide more *trustworthy* and more *accurate* predictions. Specifically, to facilitate the computation of uncertainty maps in prototypical FSS models we propose a modified prototype extraction module that introduces a Bernoulli distributed variable for each voxel location in the feature representation. Uncertainty maps are then based on the predictive distribution estimated from a set of prototypes extracted by this proposed module. Further, to alleviate the loss of spatial details and encourage spatial consistency in the predictions, we propose a novel feature refinement module that leverages supervoxels in the inference phase. Supervoxels are collections of voxels that represent compact regions of coherent voxel intensities and/or textures in the image volume. By utilizing supervoxels, we are able to encourage spatial consistency in the prediction, and help locate edges accurately in the segmentation map. Additionally, we show how uncertainty maps can be used to automatically guide this feature refinement. Finally, to avoid the problem of ambiguous voxel predictions, we propose a procedure to perform one-step multi-class FSS.

Exploiting its ability to perform volume-wise one-step FSS, we illustrate the benefit of the proposed methodology in the context of the current state-of-the-art 3D medical FSS model, ADNet (Hansen et al., 2022), and refer to the modified model as ADNet++.

To summarize, our contributions are as follows:

1. We propose a novel prototype extraction module that, with negligible computational overhead, can produce uncertainty maps for prototypical FSS models.
2. We propose a novel feature refinement module that leverages supervoxels to encourage spatial consistency and to locate edges in the segmentation masks. We also show how uncertainty maps can be used to guide the feature refinement.
3. We propose a one-step multi-class segmentation procedure to avoid ambiguous voxel predictions.

2. Related work

2.1. Medical few-shot segmentation

Lately, few-shot learning models have demonstrated promising segmentation performance on medical images (Roy et al., 2020; Tang et al., 2021; Yu et al., 2021; Ouyang et al., 2022; Hansen et al., 2022). Previous works can be categorized into methods that require labeled data during the training phase (Roy et al., 2020; Tang et al., 2021; Yu et al., 2021) and methods that are trained in a self-supervised fashion on unlabeled data (Ouyang et al., 2022; Hansen et al., 2022). In the former category, as the first medical FSS model, Roy et al. (2020) proposed a two-branched architecture, where the support features are used to implicitly guide the query segmentation through multiple interaction blocks. The succeeding works build on prototypical ideas (Snell et al., 2017), with a direct comparison between the query features and computed support prototypes. In Yu et al. (2021), the authors proposed a prototype network that leverages strong spatial priors by dividing the input images into grids and solving the segmentation problem for each grid-element separately via multiple local prototypes. Tang et al. (2021) proposed a prototype network with a recurrent mask refinement, where the previous query prediction is used to refine the query features in an iterative manner.

The few-shot learning models discussed above are only few-shot in the sense that a *trained* few-shot model only needs a few labeled instances to segment a new class. During the training phase, the models still require abundant labeled data in order to avoid over-fitting. However, the availability of labeled data is often limited in the medical setting, and to overcome this challenge, Ouyang et al. (2022) proposed a self-supervised few-shot segmentation model. The network itself, ALPNet, is a prototype based network that introduce adaptive local prototype pooling where local prototypes are computed on a regular grid to preserve local information. As opposed to Yu et al. (2021), Ouyang et al. (2022) do not divide the input images into grids, but segment the images as one segmentation problem. To train the network, Ouyang et al. (2022) proposed a new self-supervision task for segmentation by utilizing superpixels. The authors construct a pseudo-labeled support/query pair based on *one* unlabeled image slice and its unsupervised superpixel segmentation. The support label is then generated by randomly selecting a superpixel from the support image's superpixel segmentation and binarizing it to obtain a binary mask. Then the query image and label are created by applying random spatial and intensity transformations to the support image-label pair. Hansen et al. (2022) built further on this work and extend the self-supervision task to supervoxels, utilizing the 3D information in the image volumes. Further, they proposed an anomaly detection-inspired prototypical segmentation network, ADNet, where they avoid modeling the large and inhomogeneous background class with prototypes. While previous methods are limited to slice-by-slice segmentation of the image volumes, Hansen et al. (2022) were the first to extend prototypical FSS to one-step volume-wise 3D segmentation.

A drawback of all the methods discussed above is that they only perform binary image segmentation and are forced to segment multi-class segmentation problems in a class-by-class manner. Further, due to the loss of spatial detail during the encoding of the images, the models have difficulty with accurately locating edges. Finally, these models do not provide any measure of uncertainty of their predictions, which is important to build trustworthy models. In this work, we build further on the branch of self-supervised models and propose a framework for one-step multi-class medical image segmentation that provides uncertainty maps to accompany the model predictions and that involves a feature refinement that addresses the loss of spatial detail during encoding.

2.2. Uncertainty estimation

In critical decision-making processes, such as medical image segmentation, there is a need to quantify model uncertainty. That is, in addition to the model prediction, a measure of model uncertainty should be conveyed to the user to improve both safety and the reliability of the model (Kompa et al., 2021).

In medical image segmentation, Bayesian approximation (Gal and Ghahramani, 2016) and ensemble learning techniques (Lakshminarayanan et al., 2017) are often used for uncertainty quantification. While ensemble approaches (Karimi et al., 2019; Mehrtash et al., 2020) are conceptually simpler than Bayesian methods (Wickstrøm et al., 2020; Harper and Southern, 2020; van Hespen et al., 2021), they typically require training of multiple models, making them computationally expensive.

In few-shot segmentation outside the medical domain, Johnander et al. (2021) proposed a few-shot learner formulated as a deep Gaussian process. The Gaussian process works as a layer in the network that predicts the mean and covariance of the conditional probability distribution of the query mask given the query image and support set. This

information is then fed to a decoder that produces the final output. The model is thus able to model the uncertainty and uses the information to improve the segmentation performance. Concurrently, Kim et al. (2021) proposed another Gaussian process inspired technique to few-shot segmentation by using a network to estimate the uncertainty. They then use the uncertainty maps to exclude samples with high prediction uncertainty for pseudo label construction in a semi-supervised setting. While these approaches provide uncertainty maps in the FSS setting, they are model-specific and thus not directly applicable to the current state-of-the-art medical FSS models, raising the need for architecture-agnostic approaches.

3. Methods

We start by briefly describing the FSS problem setting in Section 3.1 before introducing the details of the proposed ADNet++ in Section 3.2. Specifically, we present the multi-class extension in Section 3.2.1, the proposed uncertainty estimation in Section 3.2.2, and the feature refinement in Section 3.2.3. Finally, in Section 3.3 we describe the supervoxel generation.

3.1. Problem definition

The goal of FSS is to obtain a model that, based on only a few labeled samples can generalize to new object classes. More specifically, given a training dataset with base classes $\mathcal{C}_{train}$, we learn a model that can segment novel target classes $\mathcal{C}_{test}$ from few annotated examples. The model is trained and tested in episodes, where a support set consisting of k labeled support images is used to predict the segmentation of N classes in the unlabeled query image. The support set is defined as $\mathcal{S} = \{(\mathbf{X}_1^s, \mathbf{Y}_1^s), \dots, (\mathbf{X}_k^s, \mathbf{Y}_k^s)\}$ and the query set as $\mathcal{Q} = \{\mathbf{X}^q\}$, where $\mathbf{X}^* \in \mathbb{R}^{C\times H\times W}$ represents an image volume and $\mathbf{Y}^* \in \mathbb{R}^{C\times H\times W}$ the corresponding voxel-wise annotation.[1]

3.2. ADNet++

3.2.1. Multi-class anomaly detection-inspired segmentation

As demonstrated in Hansen et al. (2022), an anomaly detection-inspired approach to few-shot medical image segmentation results in a model that is less sensitive to variations in the background class, thus enabling one-step volume-wise 3D segmentation (as opposed to slice-by-slice 2D segmentation). As a consequence, this framework facilitates the extraction of all class-prototypes simultaneously, thereby making it suitable for multi-class segmentation.

Similar to the original ADNet, ADNet++ uses a backbone network $f_\theta : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^{C\times H\times W\times d}$ to encode the support images $\{\mathbf{X}_i^s\}_{i=1}^k$ and query images $\{\mathbf{X}_i^q\}_{i=1}^k$ into deep feature maps, $\mathbf{F}_i^s = f_\theta(\mathbf{X}_i^s)$ and $\mathbf{F}_i^q = f_\theta(\mathbf{X}_i^q)$, respectively. Note that due to max-pooling operations and strided convolutions in the backbone network, the spatial resolution of these feature maps is compressed, compared to the input, and the feature maps are therefore up-sampled to original size (C, H, W). Let $\Omega = \{\mathbf{r}_j\}_{j=1}^{C\cdot H\cdot W}$ denote the set of all voxel positions $\mathbf{r} = (x, y, z)$ in the image. Prototype $\mathbf{p}_c \in \mathbb{R}^d$, representing class c, is defined as:

$$\mathbf{p}_c = \frac{\sum_{i=1}^k \sum_{\mathbf{r}\in\Omega} \mathbf{F}_i^s(\mathbf{r}) \cdot \mathbf{Y}_c^s(\mathbf{r})}{\sum_{i=1}^k \sum_{\mathbf{r}\in\Omega} \mathbf{Y}_c^s(\mathbf{r})}, \tag{1}$$

where $\mathbf{Y}_c^s = \mathbb{1}(\mathbf{Y}_i^s = c)$ is the ground-truth mask of class c. Unlike ADNet, which only performs binary segmentation and thus only extracts *one* class-prototype at a time, we propose a procedure to perform one-step multi-class segmentation. In a N-class segmentation problem, this results in a set of N prototypes $\mathcal{P} = \{\mathbf{p}_c\}_{c=1}^N$, for which we compute a set of N anomaly score maps $\mathcal{S} = \{\mathbf{S}_c\}_{c=1}^N$, computed as:

$$\mathbf{S}_c(\mathbf{r}) = -\alpha \cos\left(\mathbf{F}^q(\mathbf{r}), \mathbf{p}_c\right), \tag{2}$$

where $\alpha = 20$ is a commonly used scaling factor (Wang et al., 2019; Ouyang et al., 2022; Hansen et al., 2022). The resulting anomaly score maps represent the dissimilarity between each voxel feature vector $\mathbf{F}^q(\mathbf{r})$ and each of the class-prototypes in $\mathcal{P}$. The soft foreground predictions for each foreground class $c = 1, \dots, N$ are then found by thresholding the anomaly score maps with a learned threshold T:

$$\hat{\mathbf{Y}}_c^q(\mathbf{r})' = 1 - \sigma(\mathbf{S}_c(\mathbf{r}) - T), \tag{3}$$

where σ is the Sigmoid function. For a general number of N foreground classes, the soft background mask is then computed as:

$$\hat{\mathbf{Y}}_{c=0}^q(\mathbf{r})' = 1 - \max\left\{\hat{\mathbf{Y}}_c^q(\mathbf{r})' : c = 1, \dots, N\right\}. \tag{4}$$

Finally, the class probabilities are obtained by scaling the scores with a softmax function:

$$\hat{\mathbf{Y}}_i^q(\mathbf{r}) = \frac{\exp\left(\hat{\mathbf{Y}}_i^q(\mathbf{r})'\right)}{\sum_{j=0}^N \exp\left(\hat{\mathbf{Y}}_j^q(\mathbf{r})'\right)}. \tag{5}$$

This assures that no voxel can be assigned to more than one class, thereby preventing the ambiguous voxel predictions in binary class-by-class segmentation, occurring when a voxel lies within the threshold of multiple class-prototypes.

The network is then trained as in Hansen et al. (2022), in an end-to-end manner to optimize a loss function consisting of three terms:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_T + \mathcal{L}_{PAR}, \tag{6}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss between the query prediction and the query label:

$$\mathcal{L}_{CE} = -\frac{1}{|\Omega|} \sum_{\mathbf{r}\in\Omega} \sum_{c=0}^N \hat{\mathbf{Y}}_c^q(\mathbf{r}) \log \mathbf{Y}_c^q(\mathbf{r}), \tag{7}$$

where $|\cdot|$ indicates the cardinality of the set, $\mathcal{L}_T = T/\alpha$ is a loss on the threshold to encourage a compact embedding of the foreground classes via a smaller learned threshold, and $\mathcal{L}_{PAR}$ is the prototype alignment regularization loss from Wang et al. (2019), obtained by reversing the roles of the support and query.[2] The predicted query mask is used to segment the support image, and the loss is computed as the cross-entropy loss between the predicted support mask and the support ground-truth mask:

$$\mathcal{L}_{PAR} = -\frac{1}{|\Omega_s|} \sum_{\mathbf{r}\in\Omega_s} \sum_{c=0}^N \hat{\mathbf{Y}}_c^s(\mathbf{r}) \log \mathbf{Y}_c^s(\mathbf{r}), \tag{8}$$

where Ω_s is the set of voxel positions in the support image.

After the model is trained, the weights (θ, T) are frozen and the inference episodes are sampled from $\mathcal{C}_{test}$.

3.2.2. Uncertainty estimation

To obtain a measure of uncertainty for the model's predictions, we take inspiration from Gal and Ghahramani (2016), who exploit dropout layers in the network architecture to be able to represent the model uncertainty. As illustrated in Fig. 1, we suggest an architecture-agnostic approach to generate uncertainty maps by randomizing the masked average pooling during prototype generation in Eq. (1). Instead of applying deterministic masked average pooling to obtain one prototype per class, we propose to perform masked randomized average pooling

[1] Superscript * denotes support (s) or query (q).

[2] The influence of the sub-losses was evaluated in Hansen et al. (2022), and the results were reported in the ablation study.

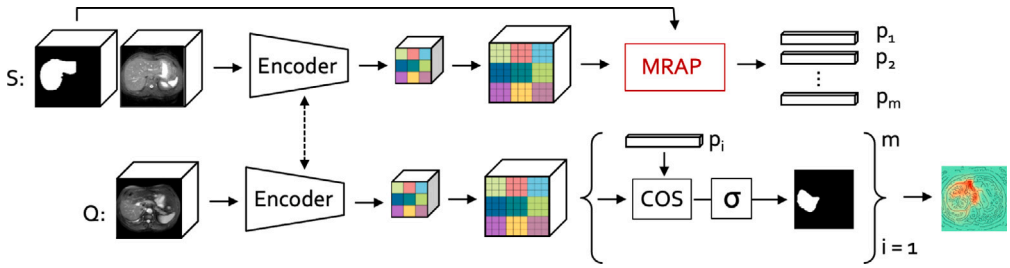

Fig. 1. We facilitate the estimation of uncertainty maps in prototypical FSS models by replacing the deterministic masked average pooling module with a randomized alternative, the masked randomized average pooling (MRAP), denoted in red. This allows us to generate a set of prototypes, and thereby a set of query predictions that can be used to estimate the model uncertainty.

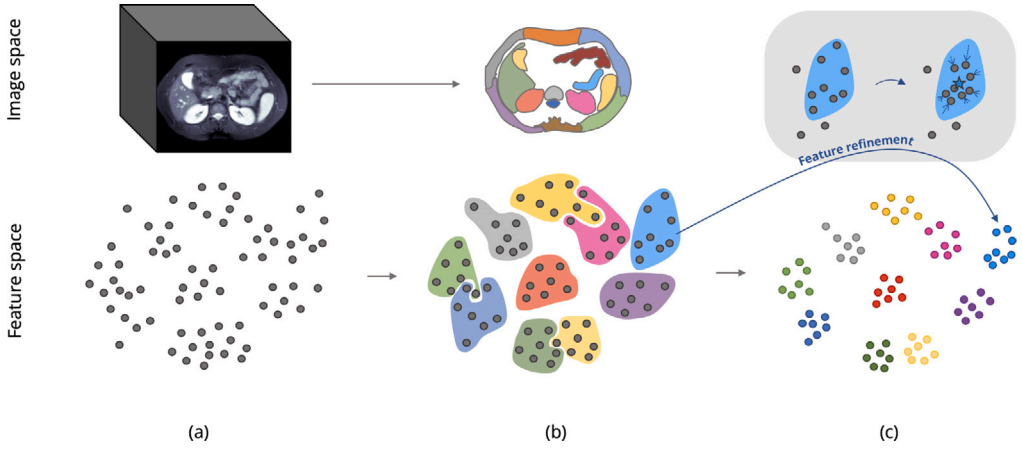

Fig. 2. Conceptual illustration of the feature refinement process. (a) In the encoding process, the input image is transformed into a set of feature vectors (gray dots). (b) Supervoxels are generated in the input space and can thus be used to locate feature vectors that "belong" together in the input space. (c) The refinement process consists in moving the feature vectors within the supervoxel towards its center (indicated with blue star), leading to a more compact embedding where the edges defined in the input space are respected.

(MRAP) to obtain a set of P prototypes per class c $\mathcal{P}_c = \{\boldsymbol{p}_j\}_{j=1}^{P}$ from the support set as:

$$\mathbf{p}_j = \frac{\sum_{i=1}^{k}\sum_{\mathbf{r}\in\Omega}\mathbf{F}_i^s(\mathbf{r})\cdot\mathbf{Y}_c^s(\mathbf{r})\cdot\mathbf{M}_i(\mathbf{r})}{\sum_{i=1}^{k}\sum_{\mathbf{r}\in\Omega}\mathbf{Y}_c^s(\mathbf{r})\cdot\mathbf{M}_i(\mathbf{r})}, \tag{9}$$

where $\mathbf{M}_i(\mathbf{r})$ is sampled from a Bernoulli(ρ) distribution. ρ is the probability of $\mathbf{M}_i(\mathbf{r})$ taking the value one and is set to 0.5. From this set of prototypes, we can obtain a set of anomaly scores $\{\mathbf{S}_j\}_{j=1}^{P}$, and thereby predictions $\left\{\hat{\mathbf{Y}}_j^q\right\}_{j=1}^{P}$ for the query image. These predictions can be considered samples from an approximate predictive distribution, and the model uncertainty map can be estimated as the predictive entropy (Gal, 2016). Therefore, by computing the voxel-wise predictive entropy of the P predictions, we obtain the uncertainty map as:

$$\mathbf{U}(\mathbf{r}) = -\sum_c \bar{\mathbf{Y}}_c(\mathbf{r})\log\bar{\mathbf{Y}}_c(\mathbf{r}), \tag{10}$$

where $\bar{\mathbf{Y}}_c = \frac{1}{P}\sum_{j=1}^{P}\hat{\mathbf{Y}}_j^q$ is the average (soft) prediction map of class c. To ensure an accurate uncertainty assessment, the number of sampled prototypes P (and thereby predictions) must be efficiently large. In our experiments we set $P = 10$ as a trade-off between uncertainty quality and computational complexity. Overall, the computational overhead due to this uncertainty estimation can be considered negligible as the costly feature extraction only needs to be performed once per volume.

The obtained uncertainty maps can be used to visualize and assess the voxel-wise uncertainty of the model's predictions. Further, in the next section, we show how these uncertainty maps can be leveraged to guide the proposed feature refinement.

3.2.3. Supervoxel-informed feature refinement module

Assuming that supervoxels capture voxels that belong together in the input space, it follows that a segmentation model should assign consistent class labels for all voxels within the same supervoxel. To encourage this spatial consistency, we propose a supervoxel-informed feature refinement module that refines the embedded image representations to respect edges as defined by the supervoxels. If a supervoxel defines a set of voxels that belong together in the input space, it consequently also defines a set of feature vectors that should belong together in the feature space, and as the encoding of images involves a spatial compression with loss of spatial details, the supervoxel-informed refinement can thus act as a mechanism to guide the precise location of edges and structures in the output. The concept of the proposed supervoxel-informed feature refinement (SFR) module is illustrated in Fig. 2.

To refine the query features during inference, the up-sampled feature maps are refined as follows. Each query image $\mathbf{x}^q$ is clustered into a set of M non-overlapping supervoxels $\pi = \{\pi_1, \ldots, \pi_M\}$, representing homogeneous regions in the input image. Overlaying this supervoxel segmentation on top of the up-sampled query feature map, each supervoxel π_i defines a set of voxel feature vectors, corresponding to a homogeneous region in the input image. For a feature vector $\mathbf{F}^q(\mathbf{r}) \in \pi_i$, we propose a refined voxel feature vector $\mathbf{F}^q(\mathbf{r})'$, computed as:

$$\mathbf{F}^q(\mathbf{r})' = \beta\mathbf{F}^q(\mathbf{r}) + (1-\beta)\boldsymbol{\mu}_i, \tag{11}$$

where $\boldsymbol{\mu}_i$ is the center of π_i, given by:

$$\boldsymbol{\mu}_i = \frac{1}{|\pi_i|}\sum_{\mathbf{F}^q(\mathbf{r})\in\pi_i}\mathbf{F}^q(\mathbf{r}), \tag{12}$$

and β is a refinement parameter controlling the size of the feature vectors' movement, ranging from $\beta = 1$ with no movement to $\beta = 0$ where the feature vector moves all the way to its supervoxel center. However, choosing β in this way, as a fixed constant for all voxels, is quite restrictive. A dynamic $\beta(\mathbf{r})$, on the other hand, would increase the module's flexibility by allowing different regions in the feature map to

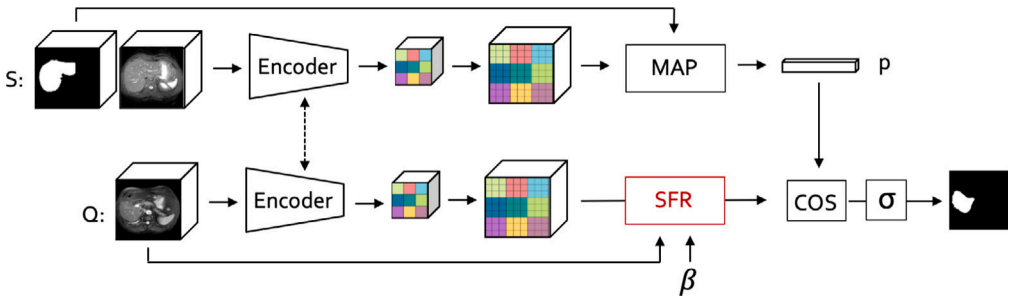

Fig. 3. Workflow of the proposed feature refinement module. The module acts to refine the features before entering the classifier. The original features, the input image and a choice of β is input to the module. The refined features then follow the ordinary pipeline to produce the end segmentation result.

experience different degrees of refinement. One possible approach to obtain a dynamic refinement is to utilize the uncertainty map[3]:

$$\beta(\mathbf{r}) = 1 - \mathbf{U}(\mathbf{r}). \tag{13}$$

In this way, uncertain voxels, typically on the boundaries between classes, get a lower β and rely more on the sharp edge information in the supervoxels, and vice versa.[4] Exploiting the uncertainty map has the additional advantage that no labeled data is required to determine β.[5] Fig. 3 illustrates the module in the FSS framework.

Ultimately, the feature refinement module is determined by two parameters: (i) the number of supervoxels M (or effectively the supervoxel *size*) and (ii) the feature refinement parameter, β. The choice of these parameters is explored in Section 4.4.2.

3.3. Supervoxel generation

Supervoxels are computed offline for all the query images using a 3D extension[6] of the Felzenszwalb's efficient graph-based segmentation algorithm (Felzenszwalb and Huttenlocher, 2004). This is the same algorithm that is used to generate pseudo-labels for the self-supervised training in Ouyang et al. (2022) and Hansen et al. (2022), and is known to produce superpixels with irregular shapes and sizes that adhere well to image boundaries (Achanta et al., 2012).

Similarly to prior supervoxel- and superpixel-based approaches (Ouyang et al., 2022; Hansen et al., 2022), the final segmentation results depend on the size of the generated supervoxels and their adherence to class boundaries in the image: Choosing supervoxels that are too small can lead to loss of representativeness, and thereby noisy results, while choosing supervoxels that are too large can result in overlapping anatomical areas and inaccurate segmentation. Felzenszwalb's algorithm has a parameter controlling the minimum supervoxel size, and effect of this parameter on the final segmentation result is explored in Section 4.4.2.

4. Experiments

In this Section we evaluate the proposed methodology in the context of the state-of-the-art self-supervised 3D FSS model, ADNet (Hansen et al., 2022),[7] and illustrate the benefit of the proposed multi-class procedure, the uncertainty maps, and the uncertainty guided feature refinement. The modified ADNet is referred to as ADNet++.

4.1. Experiment setup

Datasets. We demonstrate the properties and performance of the proposed ADNet++ by conducting experiments on three publicly available benchmark datasets[8] in medical image segmentation: (i) the bSSFP fold from the Multi-sequence Cardiac MRI Segmentation (**MS-CMRSeg**) challenge from MICCAI 2019 (Zhuang, 2016, 2018), (ii) task 5 from the Combined Healthy Abdominal Organ Segmentation (**CHAOS**) Challenge from ISBI 2019 (Kavur et al., 2019, 2020, 2021), and (iii) the abdomen dataset from the Beyond the Cranial Vault (**BTCV**) Challenge from MICCAI 2015 (Landman et al., 2015). Both the MS-CMRSeg and the CHAOS dataset consist of volumetric magnetic resonance imaging (MRI) scans, whereas the BTCV dataset consists of volumetric computed tomography (CT) scans. The MS-CMRSeg dataset contains 20 cardiac MRIs with ground-truth segmentations for left-ventricle blood pool (LV-BP), left-ventricle myocardium (LV-MYO), and right ventricle (RV). The CHAOS dataset and the BTCV dataset contain 20 abdominal T2-SPIR MRIs and 30 abdominal CT scans, respectively, with ground-truth segmentations for left kidney (L. kid.), right kidney (R. kid.), spleen, and liver.

Prior to training, the data is pre-processed following common practice (Ouyang et al., 2022; Hansen et al., 2022): First, we cut the top 0.5% intensities. Then, we re-sample and crop the image volumes such that the short-axis slices in the MS-CMRSeg dataset and the axial slices in the CHAOS dataset and the BTCV dataset have the same size (256×256).

Evaluation metric. To compare model predictions to ground-truth segmentation masks, we employ two widely used evaluation metrics in medical image segmentation: The dice similarity coefficient (DSC), which is a overlap measure, and the Hausdorff distance (HD), which is a surface distance measure.

The DSC between a model prediction $\hat{Y}$ and the ground-truth Y is computed as:

$$\mathrm{DSC}(Y, \hat{Y}) = 2\frac{|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \cdot 100\%. \tag{14}$$

Thus, the DSC varies from 100%, indicating perfect overlap between the segmentations, to 0%, when the segmentations have no overlap.

[3] In this case, the prediction step is done twice: The uncertainty map is generated using the output of the segmentation network before the feature refinement module is applied. The beta values are then derived from these uncertainty maps and used to guide the feature refinement, leading to the final prediction. Note, only one feature extraction step needs to be performed as the randomization lies in the prototype generation.

[4] The use of more complex functions, such as the Sigmoid function, for this scaling was considered. However, preliminary results did not show significant improvements over linear scaling. Nonetheless, this possibility remains an interesting avenue for future exploration.

[5] Determination of the "ideal" fixed β requires a line-search on an annotated validation set.

[6] https://github.com/sha168/Felzenszwalb-supervoxel-segmentation.

[7] Note, the benefit of one-step 3D segmentation, compared to the 2D slice-by-slice approach commonly found in previous work, was demonstrated in Hansen et al. (2022). However, to illustrate the general applicability of the proposed modules, experiments in the context of a 2D state-of-the-art approach are provided in Section 4.5.

[8] Links to the MS-CMRSeg dataset https://zmiclab.github.io/zxh/0/mscmrseg19/, the CHAOS dataset https://chaos.grand-challenge.org, and the BTCV dataset https://doi.org/10.7303/syn3193805.

Table 1
Quantitative evaluation of the proposed method in the context of ADNet. Mean 95 HD and mean DSC with standard deviations are reported for three runs per fold. * indicates that the increase in mean DSC, compared to the ADNet baseline, is statistically significant ($p < 0.05$).

Abdominal MRI										
Method	95 HD (↓)					DSC (↑)				
	L. kid.	R. kid.	Spleen	Liver	Mean	L. kid.	R. kid.	Spleen	Liver	Mean
ADNet	10.56 (±4.22)	6.63 (±3.34)	27.08 (±12.14)	16.60 (±7.38)	15.22 (±10.83)	79.57 (±7.55)	81.41 (±10.17)	68.03 (±24.05)	74.29 (±23.39)	75.82 (±5.20)
ADNet++	**6.85 (±3.35)**	**5.14 (±3.03)**	**22.08 (±13.94)**	**15.93 (±5.42)**	**12.50* (±10.41)**	**86.80 (±6.01)**	**86.62 (±10.37)**	**75.69 (±26.21)**	**74.85 (±23.82)**	**80.99* (±5.73)**
Abdominal CT										
Method	95 HD (↓)					DSC (↑)				
	L. kid.	R. kid.	Spleen	Liver	Mean	L. kid.	R. kid.	Spleen	Liver	Mean
ADNet	20.96 (±5.82)	30.65 (±10.16)	**22.63 (±4.26)**	23.22 (±7.09)	24.36 (±8.08)	47.89 (±11.94)	40.30 (±14.71)	59.25 (±10.11)	**75.88 (±10.0)**	55.83 (±11.69)
ADNet++	**20.40 (±8.02)**	**26.66 (±9.35)**	24.15 (±6.94)	**23.21 (±4.67)**	**23.60* (±7.78)**	**53.47 (±14.03)**	**50.29 (±15.68)**	**65.76 (±12.26)**	74.24 (±4.67)	**60.94* (±16.16)**
Cardiac MRI										
Method	95 HD (↓)					DSC (↑)				
	LV-BP	LV-MYO	RV	Mean		LV-BP	LV-MYO	RV	Mean	
ADNet	3.96 (±0.53)	6.57 (±0.41)	8.46 (±1.58)	6.33 (±2.09)		80.95 (±1.42)	53.68 (±2.13)	66.12 (±1.34)	66.92 (±11.27)	
ADNet++	**3.69 (±0.72)**	**6.24 (±0.55)**	**8.31 (±1.32)**	**6.08* (±2.11)**		**82.79 (±0.89)**	**58.67 (±1.86)**	**67.57 (±1.82)**	**69.68* (±10.08)**	

The maximum HD defines the maximum distance of a set to the nearest point in the other set, and is computed as:

$$\mathrm{HD}(Y, \hat{Y}) = \max \left\{ \sup_{a \in Y} \inf_{b \in \hat{Y}} d(a,b), \sup_{b \in \hat{Y}} \inf_{a \in Y} d(a,b) \right\}, \tag{15}$$

where sup denotes the supremum, inf denotes the infimum, and $d(\cdot,\cdot)$ is the Euclidean distance between two points. The maximum HD thus quantifies the largest segmentation error between the model prediction and the ground-truth. To reduce the effect of outliers, the 95th percentile HD (95 HD) is employed, where the measure is based on the 95th percentile of the distances between points in $\hat{Y}$ and Y.

Evaluation protocol. The trained models are evaluated in a five-fold cross-validation scheme where the test fold is held out during training. During inference, we sample all possible support/query combinations for the volumes in the fold to make the evaluation unbiased towards specific choices of support and query. In the tables, we report mean dice (with standard deviations) over all folds, where each fold is repeated thrice to account for the stochasticity in the optimization. To indicate statistically significant improvements, one-sided Wilcoxon signed rank tests (Wilcoxon, 1992) are performed to compare the mean scores across all runs.

Implementation details. The implementation of ADNet++ is based on the PyTorch (v1.7.1) implementation of 3D ADNet,[9] and the training phase is identical to Hansen et al. (2022): We optimize the weights using stochastic gradient descent over 25k iterations with momentum 0.9, learning rate 1e-3, decay rate 0.98 per 1k iterations, and a weight decay of 5e-4. To account for the class imbalance, a weighted cross-entropy loss is employed, where the foreground and background weights are set to 1.0 and 0.1, respectively. We adopt the data augmentation scheme in Ouyang et al. (2022) and Hansen et al. (2022), which applies the following random transformations to the support and query images during training: image rotation, translation, shearing, scaling, and gamma transformations.

[9] https://github.com/sha168/ADNet.

4.2. Quantitative results in the context of state-of-the-art model ADNet

Table 1 presents the quantitative results of the proposed method in the context of ADNet, which is the state-of-the-art model for one-step volume-wise medical FSS.

ADNet++ significantly ($p < 0.05$) improves the overall mean DSC of ADNet by +5.2, +5.1, and +2.8 percentage points for the CHAOS, BTCV and MS-CMRSeg datasets, respectively. Similarly, ADNet++ significantly ($p < 0.05$) reduces the overall mean 95 HD across all datasets. These results thus indicate that the proposed method improves both the overall overlap between the predicted masks and the ground truth masks, as well as reduces the largest segmentation mistakes. In Section 4.4 we further demonstrate the contribution of the individual proposed modules of our method.

4.3. Uncertainty maps

Fig. 4 visualizes three example slices from the CHAOS dataset with corresponding ground-truths, predictions, and uncertainty maps obtained by sampling $P = 10$ prototypes per class in the proposed prototype extraction module. From these examples, we can see that the model uncertainty typically is higher for voxels close to and on the boundaries between classes. Furthermore, when the model makes mistakes (e.g. where it over-segments the liver), we can see how the uncertainty map highlights these areas as uncertain.

Following Kampffmeyer et al. (2016), to quantify the fidelity of the estimated uncertainty maps, we start by removing all voxels in the predictions and successively add voxels according to their estimated uncertainty, starting with the least uncertain voxels. Fig. 5 shows how the segmentation performance decreases for all classes as more uncertain voxels are included.[10] This illustrates that voxels that are indicated by the uncertainty maps to be *certain* in fact are more probable of being correctly classified, whereas *uncertain* voxels have a higher probability of being falsely segmented. This means that the uncertainty maps can be used to quantify how much a prediction can be trusted.

[10] Note that the measure of the segmentation performance is accuracy (and not DSC) in this experiment. This because the denominator in Eq. (14) varies as we include more and more voxels, making comparisons difficult.

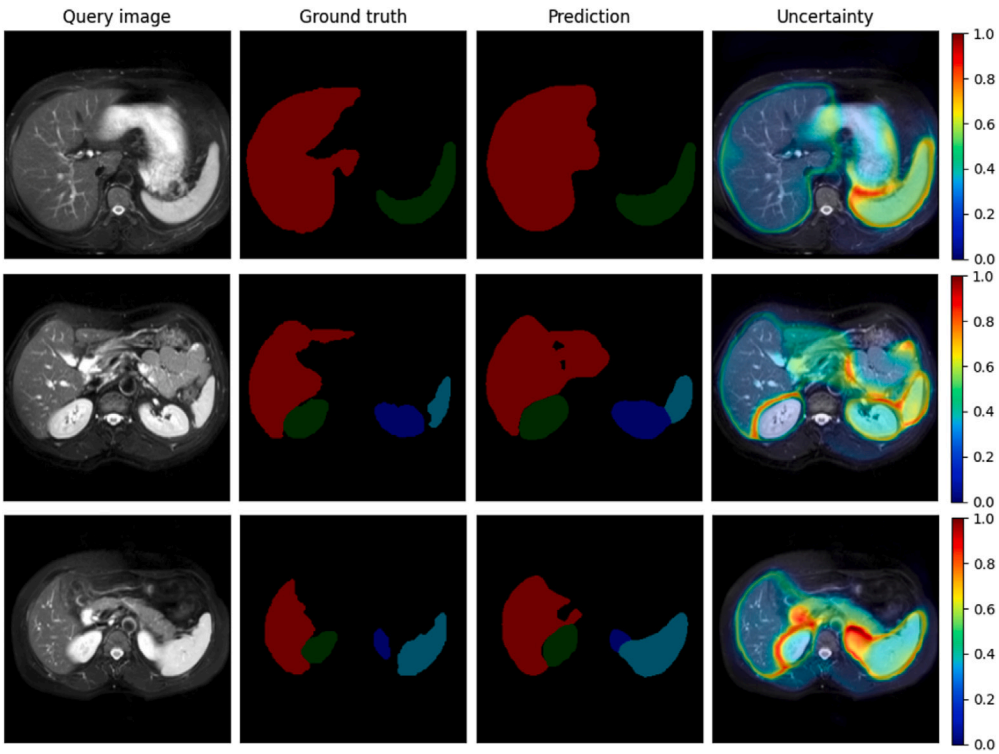

Fig. 4. Illustration of example slices from the CHAOS dataset with corresponding ground-truths, predictions and uncertainty maps. The uncertainty maps typically highlight the boundary regions between classes. (red = liver, green = right kidney, dark blue = left kidney, and light blue = spleen).

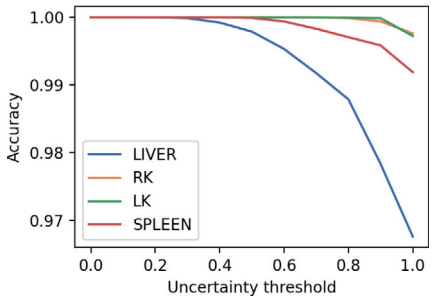

Fig. 5. Relationship between accuracy and estimated uncertainty. By successively including more uncertain voxels, the segmentation accuracy decreases.

Interestingly, the accuracy of the liver class decreases faster than for the other classes. This is related to the tendency of the model to over-segment this class. As seen in Fig. 4, the relatively large over-segmented regions are uncertain, and when they are included in the computation of the accuracy (from a uncertainty threshold around 0.4), the accuracy drops.

4.4. Ablation study

In the following, we analyze the contribution of the different proposed components to the improved DSC and 95 HD, compared to the ADNet baseline. Table 2 summarizes the quantitative ablation results for the three datasets.

4.4.1. Binary vs. multi-class segmentation

In the first two rows for each dataset in of Table 2, we analyze the effect of moving from binary to multi-class segmentation. Any differences in segmentation results here are caused by the resolving of ambiguous voxel predictions, i.e. voxels previously assigned to multiple classes are now forced to choose one. In the CHAOS dataset, the issue of ambiguous voxels are most prominent between left kidney and spleen. This is because these organs share a boundary that often appears weak in the MRI scans. While the overall performance is only slightly improved when moving to the multi-class segmentation setting, the performance gains for left kidney and spleen are more visible (on average +0.99 and +0.82 percentage points, respectively). In the BTCV dataset, the performance on the individual classes changes considerably when introducing the multi-class segmentation. While large improvements are obtained for the left and right kidney classes, the performance for the liver class decreases. This can be explained by the interaction between the liver and right kidney prototypes. While the model generally over-segments the right kidney in the binary setting, this over-segmentation is reduced when introducing the liver-prototype in the multi-class setting. On the other hand, the liver segmentation is impacted negatively as some of the previously correctly segmented liver voxels now are incorrectly being assigned to the right kidney class. In the MS-CMRSeg dataset, all three classes share boundaries with one or both other organs. In the binary setting, the model typically over-segment all three classes, particularly hurting the performance of the LV-MYO because of its high surface-to-volume ratio. When the model is forced to choose, the performance increases for all classes, especially for LV-MYO with +3.8 percentage points, yielding an average overall improvement of +1.7 percentage points. Fig. 6 illustrates how the over-segmentation in the binary setting is improved in the multi-class setting for one example slice in the MS-CMRSeg dataset.

4.4.2. Feature refinement vs. no feature refinement

In rows three and four for each dataset in Table 2, we investigate the effectiveness of the proposed feature refinement module. With a fixed β (dynamic $\beta(\boldsymbol{r})$), the module is able to improve the overall performance for all three datasets, with +4.7 (+4.7), +3.7 (+3.8), and +1.3 (+1.4) percentage points for the CHAOS dataset, BTCV dataset, and the MS-CMRSeg dataset, respectively. Note that the fixed β is set to the optimal choice for the respective datasets, requiring a grid-search for parameter-tuning. Thus, while the improvement in overall

Table 2
Quantitative evaluation of the proposed components' contribution on the three datasets. Mean DSC with standard deviations are reported for three runs per fold. * indicates that the increase in mean DSC, compared to the ADNet baseline, is statistically significant ($p < 0.05$)..

Method	Multi-class	β	Abdominal MRI L. kid.	R. kid.	Spleen	Liver	Mean
ADNet	✗	-	79.57 ± 7.55	81.41 ± 10.17	68.03 ± 24.05	74.29 ± 23.39	75.82 ± 5.20
ADNet++	✓	1.0	80.56 ± 5.89	81.42 ± 10.03	68.85 ± 24.24	74.49 ± 23.35	76.33 ± 5.08*
	✓	0.3	**86.94 ± 6.19**	**86.95 ± 10.60**	75.69 ± 26.35	74.40 ± 24.08	80.99 ± 5.97*
	✓	$1 - \mathbf{U}(\mathbf{r})$	86.80 ± 6.01	86.62 ± 10.37	**75.69 ± 26.21**	**74.85 ± 23.82**	**80.99 ± 5.73***
Method	Multi-class	β	Abdominal CT L. kid.	R. kid.	Spleen	Liver	Mean
ADNet	✗	-	47.89 ± 11.94	40.30 ± 14.71	59.25 ± 10.11	**75.88 ± 10.0**	55.83 ± 11.69
ADNet++	✓	1.0	60.09 ± 13.47	48.46 ± 21.10	59.63 ± 21.69	58.31 ± 31.71	57.18 ± 24.28*
	✓	0.0	**65.15 ± 21.99**	**58.96 ± 27.37**	63.09 ± 28.64	60.79 ± 32.84	60.84 ± 28.41*
	✓	$1 - \mathbf{U}(\mathbf{r})$	53.47 ± 14.03	50.29 ± 15.68	**65.76 ± 12.26**	74.24 ± 9.03	**60.94 ± 16.16***
Method	Multi-class	β	Cardiac MRI LV-BP	LV-MYO	RV	Mean	
ADNet	✗	-	80.95 ± 5.50	53.68 ± 5.52	66.12 ± 10.14	66.92 ± 11.15	
ADNet++	✓	1.0	81.29 ± 6.43	57.44 ± 6.47	67.18 ± 10.71	68.64 ± 9.79*	
	✓	0.7	**82.61 ± 6.55**	59.66 ± 5.64	**67.46 ± 11.17**	69.91 ± 9.53*	
	✓	$1 - \mathbf{U}(\mathbf{r})$	82.57 ± 6.55	**60.02 ± 5.66**	67.44 ± 11.37	**70.01 ± 9.38***	

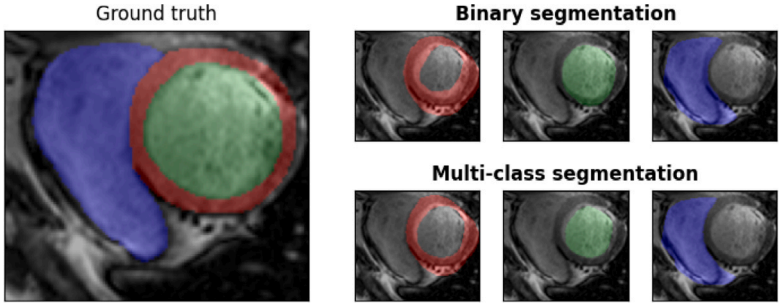

Fig. 6. Qualitative evaluation of resolving ambiguous voxel predictions in a cropped example slice from the MS-CMR dataset. Where the model in the binary segmentation setting over-segments all three classes (red = LV-MYO, green = LV-BP, and blue = RV), it is in the multi-class setting forced to choose one class per voxel, resulting in less over-segmentation and higher DSC.

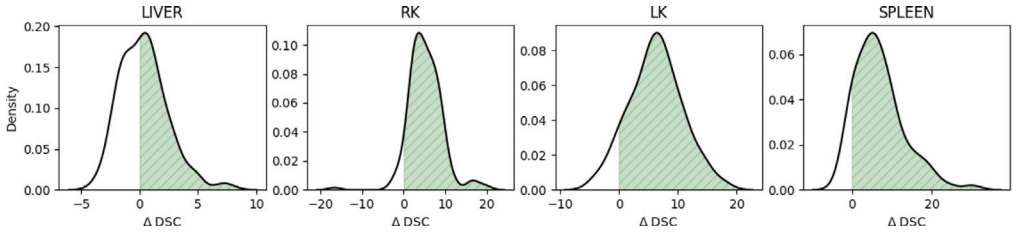

Fig. 7. Distribution of Δ DSC for the segmentation results on the CHAOS dataset with and without feature refinement.

segmentation performance is similar, the dynamic $\beta(\boldsymbol{r}) = 1 - \mathbf{U}(\boldsymbol{r})$ has the important advantage that it is computed automatically and does not need further fine-tuning.

Fig. 7 shows the distribution of the difference in DSC (Δ DSC) for predictions *with* and *without* feature refinement (with a dynamic $\beta(\boldsymbol{r})$), for each class in the CHAOS dataset. For most cases, the feature refinement improves the DSC (green regions). However, the effect is split for the liver class, resulting in no overall improvement. This is related to the difficulty in capturing the liver (especially its left lobe) with supervoxels.

In the following section, we investigate the choice of β and how it effects the final segmentation results for different supervoxel settings.

Choice of β. The choice of β controls the extent of the feature refinement, from no refinement at $\beta = 1.0$ to moving the features all the way to their corresponding supervoxel center at $\beta = 0.0$. Fig. 8 shows the prediction results for one example slice in the CHAOS dataset as we adjust the value of a fixed β from 0.0 to 1.0. For $\beta = 1.0$, we see that the model has difficulty with locating the exact class boundaries, even when the edges in the input image are strong (e.g. the boundaries between right kidney and the background). As we reduce β, we see that the segmentation boundaries become gradually sharper. However, as β approaches 0.0, the prediction relies completely on the supervoxel segmentation, which might be faulty, especially in regions where boundaries in the input image are weak.

As discussed in Section 3.2.3, a dynamic $\beta(\mathbf{r}) = 1 - \mathbf{U}(\mathbf{r})$ has the potential to increase the flexibility of the feature refinement by allowing different voxels to move with different step lengths, depending on the model's uncertainty: For voxels in regions where the model is unsure about its initial prediction, we will pay more attention to the region information in the input space.

To systematically examine the effect of the choice β (fixed and dynamic) for different supervoxel sizes, we perform a grid search. Figs. 9, 10, and 11 show the results for the CHAOS dataset, BTCV dataset and MS-CMRSeg dataset, respectively. The top row in both

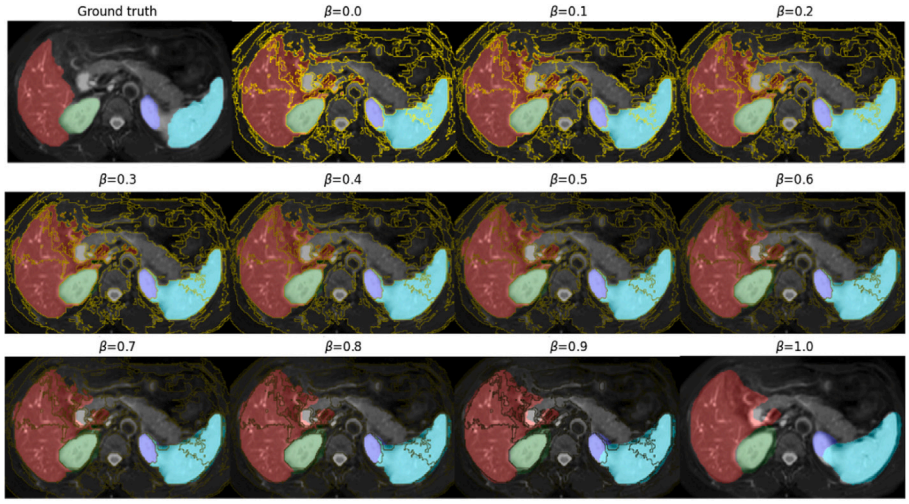

Fig. 8. Qualitative evaluation of the feature refinement with β as a fixed constant for all voxels. The example slice is taken from the CHAOS dataset and is overlaid by the corresponding supervoxel boundaries (yellow) and the resulting segmentation masks (red = liver, green = right kidney, purple = left kidney, and blue = spleen).

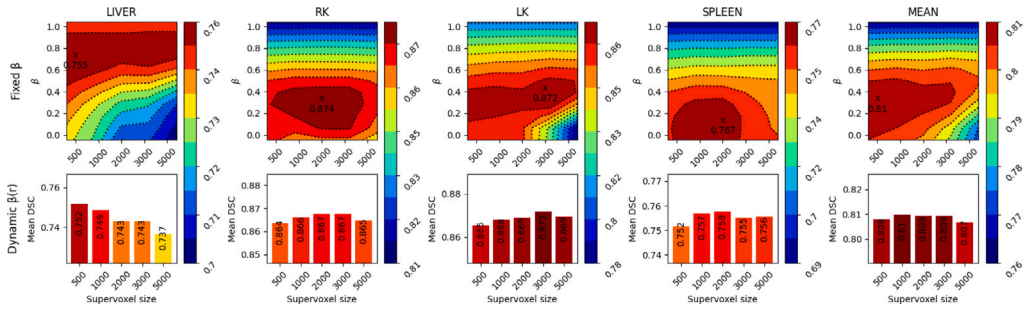

Fig. 9. Parameter sensitivity of feature-refinement module on the *CHAOS dataset*. Top: Grid-search over supervoxel sizes and a range of *fixed* betas. Bottom: Line-search over supervoxel sizes with a *dynamic* beta automatically computed from uncertainty maps.

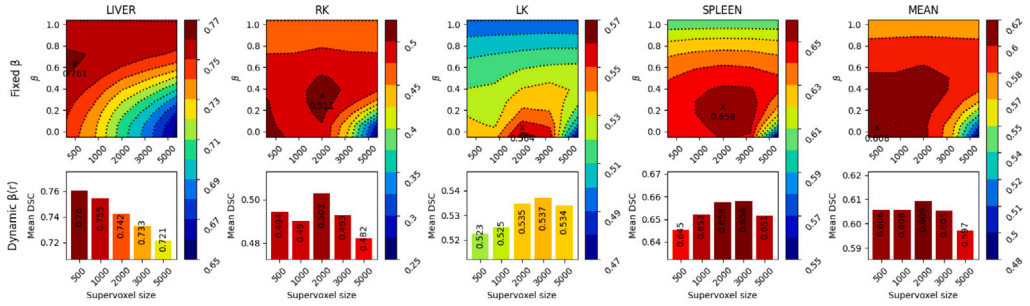

Fig. 10. Parameter sensitivity of feature-refinement module on the *BTCV dataset*. Top: Grid-search over supervoxel sizes and a range of *fixed* betas. Bottom: Line-search over supervoxel sizes with a *dynamic* beta automatically computed from uncertainty maps.

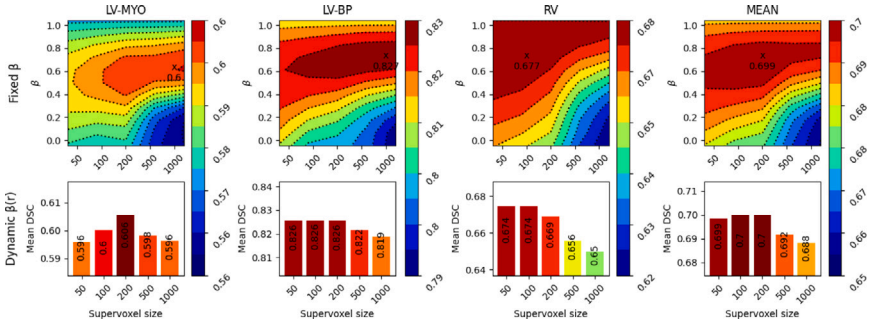

Fig. 11. Parameter sensitivity of feature-refinement module on the *MS-CMRSeg dataset*. Top: Grid-search over supervoxel sizes and a range of *fixed* betas. Bottom: Line-search over supervoxel sizes with a *dynamic* beta automatically computed from uncertainty maps.

Table 3
Quantitative evaluation of the proposed method in the 2D setting in the context of ALPNet. Mean 95 HD and mean DSC with standard deviations are reported for three runs per fold. * indicates that the increase in mean DSC (decrease in 95 HD), compared to the ALPNet baseline, is statistically significant ($p < 0.05$)..

Method	95 HD (↓)					DSC (↑)				
Abdominal MRI										
	L. kid.	R. kid.	Spleen	Liver	Mean	L. kid.	R. kid.	Spleen	Liver	Mean
ALPNet	42.59 (±13.77)	50.21 (±14.96)	40.67 (±10.43)	**25.47 (±5.71)**	39.74 (±14.81)	51.30 (±11.61)	47.66 (±10.31)	42.02 (±16.71)	56.12 (±7.00)	49.29 (±5.17)
ALPNet++	**33.46 (±9.41)**	**44.28 (±14.62)**	**36.63 (±8.47)**	26.08 (±7.08)	**35.11* (±12.20)**	**53.04 (±6.72)**	**50.09 (±6.33)**	**43.99 (±10.61)**	**57.20 (±2.30)**	**51.08* (±8.60)**
Abdominal CT										
Method	95 HD (↓)					DSC (↑)				
	L. kid.	R. kid.	Spleen	Liver	Mean	L. kid.	R. kid.	Spleen	Liver	Mean
ALPNet	58.24 (±14.10)	53.54 (±7.83)	46.98 (±4.70)	**40.49 (±2.80)**	49.81 (±10.84)	24.93 (±9.16)	24.20 (±10.04)	25.52 (±13.20)	47.63 (±8.74)	30.57 (±9.86)
ALPNet++	**51.03 (±6.64)**	**50.78 (±8.42)**	**45.51 (±3.95)**	40.54 (±2.58)	**46.97* (±7.27)**	**28.99 (±4.96)**	**29.73 (±8.59)**	**29.59 (±3.88)**	**50.85 (±4.25)**	**34.79* (±10.91)**
Cardiac MRI										
Method	95 HD (↓)					DSC (↑)				
	LV-BP	LV-MYO	RV	Mean		LV-BP	LV-MYO	RV	Mean	
ALPNet	29.74 (±12.01)	16.64 (±4.26)	13.55 (±3.84)	19.98 (±10.41)		81.30 (±1.42)	54.87 (±1.90)	68.38 (±2.47)	68.18 (±10.97)	
ALPNet++	**26.72 (±11.44)**	**15.69 (±4.12)**	**12.94 (±3.85)**	**18.45 (±9.47)**		**81.42 (±1.36)**	**55.44 (±1.83)**	**68.62 (±2.53)**	**68.49* (±10.79)**	

figures display the grid-search over supervoxel size and a range of *fixed* betas, $\beta \in [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$, while the bottom rows display a line-search over supervoxel size with a *dynamic* beta computed via Eq. (13).

For the CHAOS dataset, in particular, we see that the optimal combination of supervoxel size and fixed beta varies a lot between the classes (top row, Fig. 9). For instance, the liver class prefers smaller supervoxels and a high β, whereas the spleen class prefers somewhat larger supervoxels and a lower β. This can be connected to the typical supervoxel quality for these organs: Weak edges in the liver result in unreliable supervoxels, making it "safer" to go with small supervoxels and rely more on the original representation. The spleen, on the other hand, is easier captured by the supervoxels and the confusion between left kidney and spleen in the feature space can be resolved by relying more on the supervoxels.

The line-searches over supervoxel size with a dynamic β (bottom rows in Figs. 9, 10, and 11) show that with a dynamic β, the results across different supervoxel sizes are more stable for both datasets. They further illustrate that exploiting the uncertainty map is an efficient approach to automatically decide $\beta(\boldsymbol{r})$.

Fig. 12 shows the distribution of $\beta(\mathbf{r})$ for each class c in the CHAOS dataset, illustrating how the features of the different organs are refined with a greater or lesser influence of the supervoxel information. For instance, most of the voxels belonging to right kidney get a high value of beta, meaning that they are experiencing a lower degree of feature refinement. This is because the prediction of the right kidney class typically is quite certain, with the exception of the edge voxels, which contribute to the long tail of the distribution in Fig. 12.

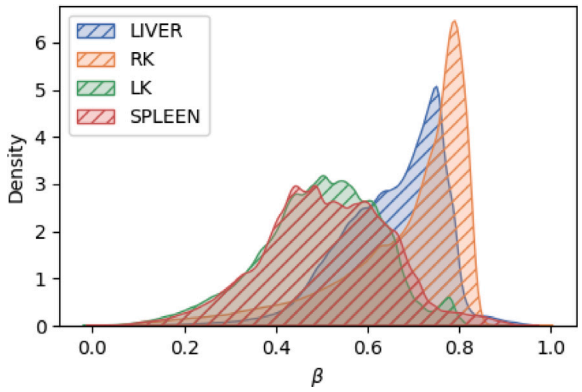

Fig. 12. Distribution of β for the different classes in the CHAOS dataset, when decided automatically from the uncertainty maps: $\beta(\mathbf{r}) = 1 - \mathbf{U}(\mathbf{r})$.

4.5. Quantitative results in the context of ALPNet

To verify the architecture-agnostic nature of the proposed method, we extend ALPNet (Ouyang et al., 2022) (which is the closest competing model to ADNet in Hansen et al. (2022)) with our proposed uncertainty-guided feature refinement. The extended ALPNet is referred to as ALPNet++. Note that due to the 2D slice-wise segmentation approach employed by ALPNet (as opposed to one-step 3D segmentation in 3D ADNet), the presence of all classes in the support slice is not guaranteed, making multi-class segmentation infeasible. Therefore, for the experiments in this section, our focus is on the effect of the uncertainty-guided feature refinement in the binary setting only.

The 2D slice-wise segmentation approach requires a scheme for support-query matching during the inference episodes. In the original paper, this was solved by assuming availability of weak label information on the query volumes during inference. In this work we do *not* assume the availability of such weak labels and follow the more realistic evaluation protocol 2 in Hansen et al. (2022) where the middle slice in the support target volume is used to segment the entire query volume slice-by-slice. Table 3 presents the quantitative results for ALPNet and ALPNet++.[11] Similarly to the ADNet++ results in Section 4.2, we observe that ALPNet++ consistently outperforms the ALPNet baseline, obtaining significant ($p < 0.05$) performance improvements. This further illustrates that the proposed modules in this work can be leveraged in an architecture-agnostic setting to improve medical few-shot segmentation performance.

[11] Note: ALPNet is designed for 2D slice-wise segmentation, and relies on weak label information to locate the query target volumes in order to achieve state-of-the-art performance (Hansen et al., 2022). Without the weak labels, this model performs poorly on the CHAOS dataset and the BTCV dataset, compared to ADNet (and ADNet++), which are designed to handle a large, inhomogeneous background class and performs well in this setting.

5. Conclusion and outlook

Prototypical few-shot learning is an emerging research direction within medical image segmentation that offers promising results without requiring large labeled datasets. In this work, we identify three weaknesses of current prototypical FSS models for medical image segmentation and propose new methodology to overcome these. Specifically, we propose the ADNet++, the first model that performs one-step multi-class segmentation and that provides uncertainty maps to accompany its predictions. In addition to indicate the model's confidence in the predictions, thereby increasing the models trustworthiness, the uncertainty maps are further exploited to guide the proposed feature refinement that leverages structural information in the input space to provide more accurate segmentation results. The proposed model significantly improves the current state-of-the-art 3D FSS model for the tasks of MRI-based abdominal organ segmentation and cardiac segmentation, as well as CT-based abdominal organ segmentation.

In future work, it would be interesting to explore methods that can make the feature-refinement module more robust to supervoxel quality, as its success largely depends on it. Instead of relying on *one* set of supervoxels, a potential approach could be to explore *multi-scale* supervoxels, e.g. supervoxels of different sizes. Furthermore, while we do demonstrate the benefit of our proposed modules also for the ALP-Net framework, given their model-agnostic nature, future work should implement and evaluate their fidelity in even more FSS frameworks.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The experiments have been conduced on publicly available datasets. The links to these datasets are provided in the paper.

Acknowledgments

This work was supported by The Research Council of Norway (RCN), through its Centre for Research-based Innovation funding scheme [grant number 309439] and Consortium Partners; RCN FRIPRO [grant number 315029]; RCN IKTPLUSS [grant number 303514]; and the UiT Thematic Initiative.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34 (11), 2274–2282.

Felzenszwalb, P.F., Huttenlocher, D.P., 2004. Efficient graph-based image segmentation. Int. J. Comput. Vis. 59 (2), 167–181.

Gal, Y., 2016. Uncertainty in deep learning (Ph.D. thesis). University of Cambridge.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. PMLR, pp. 1050–1059.

Gonzalez, Y., Shen, C., Jung, H., Nguyen, D., Jiang, S.B., Albuquerque, K., Jia, X., 2021. Semi-automatic sigmoid colon segmentation in CT for radiation therapy treatment planning via an iterative 2.5-D deep learning approach. Med. Image Anal. 68, 101896.

Hansen, S., Gautam, S., Jenssen, R., Kampffmeyer, M., 2022. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. Med. Image Anal. 78, 102385.

Harper, R., Southern, J., 2020. A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat. IEEE Trans. Affect. Comput.

Johnander, J., Edstedt, J., Danelljan, M., Felsberg, M., Khan, F.S., 2021. Deep Gaussian processes for few-shot segmentation. arXiv preprint arXiv:2103.16549.

Kampffmeyer, M., Salberg, A.B., Jenssen, R., 2016. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–9.

Karimi, D., Zeng, Q., Mathur, P., Avinash, A., Mahdavi, S., Spadinger, I., Abolmaesumi, P., Salcudean, S.E., 2019. Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. Med. Image Anal. 57, 186–196.

Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.-H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., Baydar, B., Lachinov, D., Han, S., Pauli, J., Isensee, F., Perkonigg, M., Sathish, R., Rajan, R., Sheet, D., Dovletov, G., Speck, O., Nürnberger, A., Maier-Hein, K.H., Bozdağı Akar, G., Ünal, G., Dicle, O., Selver, M.A., 2021. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. Med. Image Anal. 69, 101950. http://dx.doi.org/10.1016/j.media.2020.101950, URL http://www.sciencedirect.com/science/article/pii/S1361841520303145.

Kavur, A.E., Gezer, N.S., Barış, M., Şahin, Y., Özkan, S., Baydar, B., Yüksel, U., Kılıkçıer, Ç., Olut, Ş., Bozdağı Akar, G., Ünal, G., Dicle, O., Selver, M.A., 2020. Comparison of semi-automatic and deep learning based automatic methods for liver segmentation in living liver transplant donors. Diagn. Int. Radiol. 26, 11–21. http://dx.doi.org/10.5152/dir.2019.19.

Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S., 2019. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data. Zenodo, http://dx.doi.org/10.5281/zenodo.3362844.

Kim, S., Chikontwe, P., Park, S.H., 2021. Uncertainty-aware semi-supervised few shot segmentation. arXiv preprint arXiv:2110.08954.

Kompa, B., Snoek, J., Beam, A.L., 2021. Second opinion needed: communicating uncertainty in medical machine learning. npj Digit. Med. 4 (1), 1–6.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv. Neural Inf. Process. Syst. 30.

Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., 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, Vol. 5. p. 12.

Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T., 2020. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans. Med. Imaging 39 (12), 3868–3878.

Militello, C., Rundo, L., Toia, P., Conti, V., Russo, G., Filorizzo, C., Maffei, E., Cademartiri, F., La Grutta, L., Midiri, M., et al., 2019. A semi-automatic approach for epicardial adipose tissue segmentation and quantification on cardiac CT scans. Comput. Biol. Med. 114, 103424.

Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D., 2022. Self-supervised learning for few-shot medical image segmentation. IEEE Trans. Med. Imaging.

Roy, A.G., Siddiqui, S., Pölsterl, S., Navab, N., Wachinger, C., 2020. "Squeeze & excite" guided few-shot segmentation of volumetric images. Med. Image Anal. 59, 101587.

Shen, C., Nguyen, D., Zhou, Z., Jiang, S.B., Dong, B., Jia, X., 2020. An introduction to deep learning in medical physics: advantages, potential, and challenges. Phys. Med. Biol. 65 (5), 05TR01.

Snell, J., Swersky, K., Zemel, R., 2017. Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. pp. 4077–4087.

Tang, H., Liu, X., Sun, S., Yan, X., Xie, X., 2021. Recurrent mask refinement for few-shot medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3918–3928.

Tsochatzidis, L., Koutla, P., Costaridou, L., Pratikakis, I., 2021. Integrating segmentation information into CNN for breast cancer diagnosis of mammographic masses. Comput. Methods Programs Biomed. 200, 105913.

van Hespen, K.M., Zwanenburg, J.J., Hendrikse, J., Kuijf, H.J., 2021. Subvoxel vessel wall thickness measurements of the intracranial arteries using a convolutional neural network. Med. Image Anal. 67, 101818.

Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J., 2019. Panet: Few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9197–9206.

Wickstrøm, K., Kampffmeyer, M., Jenssen, R., 2020. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Med. Image Anal. 60, 101619.

Wilcoxon, F., 1992. Individual comparisons by ranking methods. In: Breakthroughs in Statistics. Springer, pp. 196–202.

Yu, Q., Dang, K., Tajbakhsh, N., Terzopoulos, D., 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, IEEE, pp. 262–266.

Zhuang, X., 2016. Multivariate mixture model for cardiac segmentation from multi-sequence MRI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 581–588.

Zhuang, X., 2018. Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Trans. Pattern Anal. Mach. Intell. 41 (12), 2933–2946.