machine learning & knowledge extraction

MDPI

Article

Going to Extremes: Weakly Supervised Medical Image Segmentation

Holger R. Roth *, Dong Yang, Ziyue Xu, Xiaosong Wang and Daguang Xu *

NVIDIA Corporation, Bethesda, MD 20814, USA; dongy@nvidia.com (D.Y.); ziyuex@nvidia.com (Z.X.); xiaosongw@nvidia.com (X.W.)

* Correspondence: hroth@nvidia.com (H.R.R.); daguangx@nvidia.com (D.X.)

Abstract: Medical image annotation is a major hurdle for developing precise and robust machine-learning models. Annotation is expensive, time-consuming, and often requires expert knowledge, particularly in the medical field. Here, we suggest using minimal user interaction in the form of extreme point clicks to train a segmentation model which, in effect, can be used to speed up medical image annotation. An initial segmentation is generated based on the extreme points using the random walker algorithm. This initial segmentation is then used as a noisy supervision signal to train a fully convolutional network that can segment the organ of interest, based on the provided user clicks. Through experimentation on several medical imaging datasets, we show that the predictions of the network can be refined using several rounds of training with the prediction from the same weakly annotated data. Further improvements are shown using the clicked points within a custom-designed loss and attention mechanism. Our approach has the potential to speed up the process of generating new training datasets for the development of new machine-learning and deep-learning-based models for, but not exclusively, medical image analysis.

Keywords: image segmentation; image annotation; weak supervision

Citation: Roth, H.R.; Yang, D.; Xu, Z.; Wang, X.; Xu, D. Going to Extremes: Weakly Supervised Medical Image Segmentation. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 507–524. https://doi.org/10.3390/make3020026

Academic Editors: Jaime Cardoso, Nicholas Heller, Pedro Henriques Abreu, Ivana Išgum and Diana Mateus

Received: 27 February 2021
Accepted: 18 May 2021
Published: 2 June 2021

1. Introduction

A major bottleneck for the development of novel machine-learning (ML)-based models is the annotation of datasets that are useful to train such models. This is especially true for healthcare applications, where annotation typically needs to be performed by experts with clinical domain knowledge. This bottleneck often inhibits our ability to integrate ML-based models into clinical workflows. At the same time, there is a growing demand for ML methods to improve clinical image analysis workflows, driven by the growing number of medical images taken in routine clinical practice.

In particular, volumetric analysis has shown several advantages over 2D measurements for clinical applications [1], which in turn, further increases the amount of data (a typical CT scan contains hundreds of slices) needing to be annotated to train accurate 3D models. Apart from acquiring accurate measurements, volumetric segmentation is widely desirable for visualization, 3D printing, radiomics, radiation treatment planning, image-guided surgery, and registration. Despite the increasing need for 3D volumetric training data to train accurate and efficient ML models for medical imaging, the majority of annotation tools available today are constrained to performing the annotation in multiplanar reformatted views. The annotator needs to either use a virtual paint brush or draw boundaries around organs of interest, often on a slice-by-slice basis [2]. Classical techniques such as 3D region-growing or interpolation can speed up the annotation process by starting from seed points or sparsely annotated slices, but its usability is often limited to certain types of structures. Some tools allow the user to skip certain regions of the image using interpolation between slices or cross-sectional views which can be helpful, but they often ignore the underlying image information. Hence, these approaches cannot always generalize to the varied use cases in medical imaging.

In this work, we propose to use only minimal user interaction in the form of extreme point clicks to train a deep-learning (DL)-based segmentation model with minimal annotation burden. "Extreme points" are user-provided clicks that lie on the surface of the organ of interest and define the extent of that organ along each image dimension in a three-dimensional medical image. The proposed approach integrates an iterative training and refinement scheme to gradually improve the models' performance. Starting from the user-defined extreme points along each dimension of a 3D medical image, an initial segmentation is produced based on the random walker (RW) algorithm [3]. This segmentation is then used as a noisy supervisory signal to train a fully convolutional network (FCN) that can segment the organ of interest based on the provided user clicks. Furthermore, we propose several variations on the deep-learning setup to make full use of the extreme point information provided by the user. For example, we integrate the point information into a novel point-based loss function and combine it with an attention mechanism to further guide the segmentations. Through large-scale experimentation, we show that the network's predictions can be iteratively refined using several rounds of training and prediction while only using the same weakly annotated point data as our manually provided supervision signal.

1.1. Related Work

1.1.1. Segmentation Networks

Fully convolutional networks (FCNs) [4] have established themselves as the state-of-the-art methods for medical image segmentation in recent years [5–9]. However, a major drawback is that they are very data-hungry, limiting their application in healthcare where data annotation is very expensive. To reduce the cost of labeling, semi-automated/interactive and weakly supervised methods have been proposed in the literature [10,11].

1.1.2. Interactive Segmentation

The integration of semi-automated approaches has been an active area of development [12], typically using classical methods such as graph cut [13], random walks [3], active shape models [14,15], and others [16]. Machine-learning methods have also been considered to be a viable way for interactive algorithms. In Wang et al. [17], an online random forest is used in combination with conditional random fields and 4D graph cuts to segment, in a minimally interactive framework, the human placenta in fetal MRI scans. Recently, building on advances in deep learning, several new methods have been proposed. One popular form of interaction is user-drawn scribbles. In Amrehn et al. [18], a user can iteratively add scribble hints as seed points to improve the segmentation result given by an FCN. In Wang et al. [19], the DeepIGeoS algorithm leverages geodesic distance transforms and scribbles to allow interactive segmentation. An alternative method [20] uses image-specific fine-tuning and leverages both bounding boxes and scribble-based interaction. Can et al. [21] proposes to use scribbles with random walks [3] and FCN predictions to achieve semi-automated segmentation. Scribbles are also used to generate pixel-level maximum category likelihood via propagation to their neighborhood in [22]. Instead of scribbles, point clicks is another widely practiced interaction. In Sakinis et al. [23], the authors use the clicks as Gaussian kernels and put them in a separate input channel to an FCN to model user interactions via seed-point placing. Khan et al. [24] extends the Gaussian kernel idea to a confidence map derived from extreme points that quantitatively encodes some priors. Majumder and Yao [25] transforms the positive and negative clicks into images based on superpixel and object proposals, so that image information can be used with clicks to generate a guidance map. In addition to scribbles and points, Ling et al. [26] parameterizes the segmentation boundary as polygons/splines, which are further modeled as a graph. Location shifts for each node are then predicted via Graph Convolutional Networks (GCN).

1.1.3. Weakly Supervised Segmentation

Weak supervision significantly reduced the time needed for user annotation, and therefore is an important research area for DL. One popular idea is to apply classical non-learning-based methods over a DL-generated feature map. For example, in Dias and Medeiros [27], Monte Carlo region-growing is triggered from confidence scores given by a network, and in Cerrone et al. [28], random walks is performed over learned edge weights. An "opposite" idea is to use classical unsupervised methods as initial estimate for further learning process. In Rajchl et al. [29], an initial *GrabCut* segmentation is used for this purpose, and segmentation performance is then improved with several rounds of predictions using CNN plus Dense CRF post-processing. Similarly, in Zhang et al. [30], segmentation results based on K-means are used to train a deep segmentation network on cystic lung regions. Without proper supervision, such approaches might work well if unsupervised techniques can have good enough initial performance. However, completely unsupervised techniques might fail to generalize to organs where the boundary information is not as clear. One possible way to address this issue is to add a confidence network [31] to judge the quality of additional information generated, so that unlabeled data can be included to adversarially train the segmentation network. More recently, Kervadec et al. [32] introduced inequality constraints based on target-region size and image tags in the loss function of a CNN to train the network for weakly supervised segmentation. Instead of information extracted by classical methods, weakly supervised or self-learning can also make use of measurements readily available, or use non-experts' judgements. One example is the measurements acquired during evaluation of the RECIST criteria [33] in the hospital picture archiving and communication system (PACS). However, such measurements are typically constraint to 2D and might miss adequate constraints for more complex three-dimensional shapes. Non-expert annotations can be acquired by using crowd-sourcing platform, Rajchl et al. [34] distributes superpixel weak annotation tasks and collects such annotations from a crowd of non-expert raters, and further use them as weak supervision for network training.

1.2. Contributions

This work follows our preliminary study presented in [35] which investigated a 3D extension of [36] in a weakly supervised setting and building on random walker initialization from scribbles. In this work, we extend this approach and add the following contributions:

- We use a modern network architecture shown to be very efficient for medical image segmentation tasks, namely the architecture proposed in Myronenko [9] and integrate the attention mechanism proposed by Oktay et al. [37].
- We make proper use of the point channel information not just at the input level of the network, but throughout the network, namely in the new attention gates.
- We furthermore propose a novel loss function that integrates the extreme point locations to encourage the boundary of our model's predictions to align with the clicked points.
- We extend the experimentation to a new multi-organ dataset that shows the generalizability of our approach.

2. Method

The starting point for our framework is a set of user-provided clicks on the extreme points $\{e\}$ that lie on the surface of the organ of interest. We follow the approach of Maninis et al. [36] and assume the users to provide only the extreme points along each image dimension in a three-dimensional medical image. This information is then used at several places within the network and during our iterative training scheme. The overall proposed algorithm for weakly supervised segmentation from extreme points can be divided into the steps which are detailed below:

1. Extreme point selection
2. Initial segmentation from scribbles via random walker (RW) algorithm

3. Segmentation via deep fully convolutional network (FCN), where we explore several variations on the training scheme
   (a) Without RW and Dice loss
   (b) With RW but without the extra point channel and Dice loss
   (c) With RW and Dice loss
   (d) With RW and Dice loss and point loss
   (e) With RW and Dice loss and point loss and attention
   (f) With RW and Dice loss and point loss and point attention
4. Regularization using random walker algorithm

Steps 2, 3, and 4 will be iterated until convergence. Here, convergence is defined based on a holdout validation set which is used to select the best model during training.

2.1. Step 1: Extreme Point Selection

Defining extreme points $\{e\}$ on the surface of the organ will allow the extraction of a bounding box around the organ of interest. Additional padding is typically useful to allow the network to learn some contextual information around the organ of interest.

Bounding box selection significantly reduces the image content to be analyzed and simplifies the machine-learning problem, as previous work on cascaded approaches showed [38]. The computer vision literature has extensively studied bounding boxes and extreme points on objects [36]. They give some advantages over the technical drawbacks of bounding box selection in which the user often must pick the corners of bounding boxes outside of the object of interest. This is particularly difficult to do for three-dimensional objects where users typically must traverse three multiplanar reformatted views (axial, coronal, sagittal) to accomplish the task. Recent studies also demonstrated the time savings achieved with extreme point selection instead of conventional bounding box selection [36,39]. At the same time, extreme points can provide the segmentation model with additional information which can be seen in our experimental section, Table 1 where we compare various ways of integrating the extreme point information into the model training. They lie on the surface of the object. In the basic approach, we can model them together with the image intensities as an additional input channel. This extra channel $G(\{e\})$ includes 3D Gaussians centered on each user clicked point location e. This approach is similar to Maninis et al. [36] but we have extended this approach to problems with 3D medical imaging. At the same time, we can use the point information to guide the loss function towards making predictions whose boundary aligns with the point locations (see Section 2.9) or use it as an additional signal that can be used to guide model attention mechanisms (see Section 2.7).

Figure 1 illustrates the different ways of how the extreme point information can be used by our proposed network architecture. We ask the user to click on six extreme points that describe the largest extent of the organ. Here, six click locations are shown after conversion to Gaussians in the extra input channel to the network, loss, and attention gates. These points are then used to compute a bounding box B automatically, including some padding p. In this study, we extract the extreme points automatically during training from a given ground truth mask. In order to simulate user interaction, we add some Gaussian noise σ to the x, y, z point locations at each DL training iteration as in Maninis et al. [36].

Table 1. Summary of our weakly supervised segmentation results on validation data. This table compares the random walker initialization (*rnd. walk. init.*) with the weakly supervised training approach using Dice loss (DL) (*weak. sup. dextr3D (w/o RW) DL*), without the extra point channel information as input to the network (*weak. sup. dextr3D (w RW; no extr. points. channel) DL*), when using the point channel as input (*weak. sup. dextr3D (w RW) DL*), using the proposed point loss (PL) together with DL (*weak. sup. dextr3D (w RW) Dice + PL*), integrating the attention mechanism as in Oktay et al. [37] (*weak. sup. dextr3D (w RW) Dice + PL + Attn.*), attention with point channel information at attention gates (*weak. sup. dextr3D (w RW) Dice + PL + Pt. Attn.*), and fully sup. (*dextr3D DL*) for reference on different datasets.

Dice [Mean ± std (Median)]	***MSD-Spleen***	***MO-Spleen***	***MO-Liver***	***MO-Pancreas***	***MO-L.Kidney***	***MO-Gallbladder***
rnd. walk. init.	0.922 ± 0.018 (0.922)	0.830 ± 0.144 (0.913)	0.786 ± 0.146 (0.847)	0.458 ± 0.206 (0.414)	0.741 ± 0.137 (0.815)	0.638 ± 0.195 (0.619)
weak. sup. (w/o RW) DL	0.939 ± 0.011 (0.943)	0.942 ± 0.009 (0.939)	0.924 ± 0.020 (0.924)	0.656 ± 0.089 (0.634)	0.878 ± 0.034 (0.893)	0.678 ± 0.194 (0.740)
weak. sup. (w RW; no extr. points. channel)DL	0.945 ± 0.012 (0.950)	0.942 ± 0.009 (0.937)	0.940 ± 0.011 (0.942)	0.637 ± 0.166 (0.664)	0.900 ± 0.013 (0.899)	0.677 ± 0.252 (0.787)
weak. sup. (w RW)DL	0.946 ± 0.011 (0.950)	0.944 ± 0.023 (0.945)	0.937 ± 0.013 (0.941)	0.700 ± 0.068 (0.676)	0.909 ± 0.017 (0.907)	0.701 ± 0.209 (0.795)
weak. sup. (w RW)Dice + PL	0.946 ± 0.010 (0.949)	**0.945 ± 0.019 (0.947)**	**0.939 ± 0.012 (0.940)**	**0.726 ± 0.080 (0.746)**	0.906 ± 0.024 (0.909)	**0.719 ± 0.186 (0.789)**
weak. sup. (w RW)Dice + PL + Attn.	0.945 ± 0.013 (0.948)	0.924 ± 0.053 (0.948)	0.920 ± 0.059 (0.943)	0.451 ± 0.124 (0.427)	0.905 ± 0.023 (0.907)	0.606 ± 0.256 (0.632)
weak. sup. (w RW)Dice + PL + Pt. Attn.	**0.948 ± 0.011 (0.950)**	**0.945 ± 0.021 (0.943)**	**0.939 ± 0.013 (0.939)**	0.703 ± 0.077 (0.688)	**0.913 ± 0.013 (0.916)**	0.702 ± 0.184 (0.773)
fully sup. DL	***0.958 ± 0.007 (0.959)***	***0.954 ± 0.027 (0.959)***	***0.956 ± 0.010 (0.957)***	***0.747 ± 0.082 (0.721)***	***0.942 ± 0.019 (0.946)***	***0.715 ± 0.173 (0.791)***

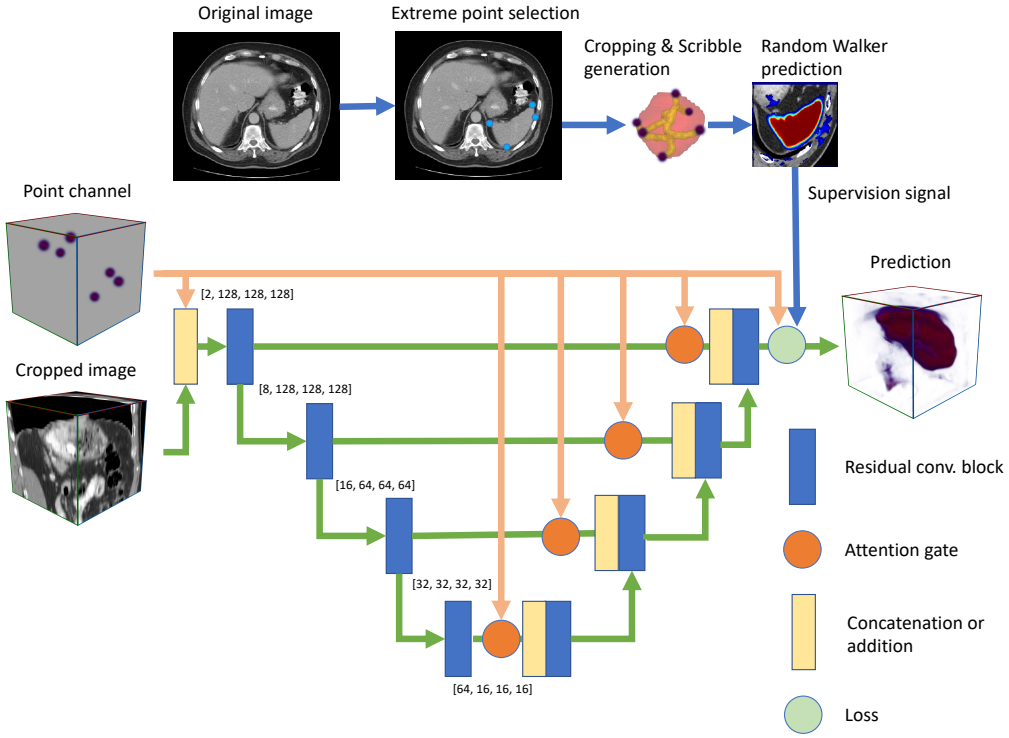


Figure 1. A high-level overview of our proposed network architecture and weakly supervised learning approach. First, the user selects six extreme points that define the 3D bounding box around the organ of interest and are used to automatically generate the scribbles that are the input (together with the image) to the random walker algorithm, whose output prediction serves as the supervision signal for training the segmentation network. The network receives both a *image channel* input and a *point channel* input that represents the user-provided extreme points. The point channel is then used throughout the network to further guide the segmentation training, i.e., as an additional input to attention gates and in the loss function. The feature shapes of each layer of the encoder are shown for the setting used in this study.

After cropping the image based on B, we resize each bounding box region to a constant size $S = s_x \times s_y \times s_z$. In all our experiments we set $s_x = s_y = s_z = 128$ and choose p = 20 mm which can include enough contextual information for typical applications of clinical CT scanning (see Section 3).

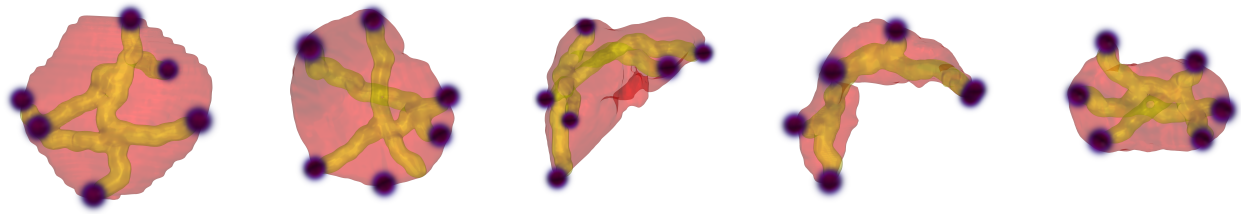
2.2. Step 2: Initial Segmentation from Scribbles via Random Walker Algorithm

In this step, we generate a set of scribbles or seed points $\mathcal{S} = \{s_0, s_1\}$ as input for the random walker algorithm [3] whose output is a probability map $\hat{Y}$. This map $\hat{Y}$ then acts as a pseudo-dense or "noisy" label map used to supervise a 3D deep network learning the segmentation task. The foreground scribbles s_1 and background scribbles s_0 are automatically generated from the initial set of extreme points $\{e\}$ by computing the shortest path between each extreme point pair along each image axis via the Dijkstra algorithm [40]. Here, we model the distance between neighboring voxels by their gradient magnitude

$$D = \sqrt{\left(\frac{\partial I}{\partial x}\right)^2 + \left(\frac{\partial I}{\partial y}\right)^2 + \left(\frac{\partial I}{\partial z}\right)^2}, \quad (1)$$

where I denotes the image intensity. The resulting path can be seen as an approximation of the geodesic distance between the two extreme points in each dimension [19] with respect to the content of the image. Figure 2 displays the foreground scribbles to be used for the random walker algorithm used as input seeds and shows the ground truth surface information for reference. Please note that this ground truth is not used to compute the

scribbles (apart from simulating the extreme points). To increase the number of foreground seeds s_1 for the random walker, each path will also be dilated with a 3D ball structure element of radius $r_1 = 2$. The background seeds s_0 are estimated as the dilated and inverted version of the input scribbles. The amount of dilation needed for successful initialization depends on the size of the organ of interest. We typically dilate the scribbles with a ball structure element of radius $r_0 = 30$, which achieves good initial seeds for organs such as the spleen and liver (see Figure 3).



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 2. Examples of automatically created foreground "scribbles" (yellow) from extreme point clicks, modeled as 3D Gaussians in our network learning. We use a geodesic shortest path algorithm to compute a scribble based on the image information alone that connects two opposing extreme points across one of the three image dimensions. (**a**–**e**) are showing examples from *MSD-Spleen*, *MO-Spleen*, *MO-Liver*, *MO-Pancreas*, *MO-L.Kidney*, and *MO-Gallbladder*, respectively. The surface rendering show the ground truth segmentations for reference in red. Best viewed in color.

2.3. Random Walker

Next, the random walker algorithm [3] is used to produce an initial prediction map $\hat{Y}$ based on the background s_0 and foreground s_1 scribbles mentioned above. The random walker basically solves the diffusion equation between voxels by turning the scribbles $\mathcal{S} = \{s_0, s_1\}$ into a source and sink. The 3D volume here is defined as a $G(E, V)$ graph with $e \in E$ edges and $v \in V$ vertices. Each edge between two vertices of v_i and v_j is referred to as e_{ij} and a weight of w_{ij} can be assigned based on gradients of the image intensities. In addition, $d_i = \sum w_{ij}$ defines the degree of a given vertex. To get a probability $p(\omega|x)$ of whether each vertex v_i belongs to the foreground ω_1, we solve the diffusion equation. L is the Laplacian matrix of the weighted image graph G with each element of the matrix defined as:

$$L_{ij} = \begin{cases} d_i, & \text{if } i = j, \\ -w_{ij}, & \text{if } i \text{ and } j \text{ are adjacent voxels,} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

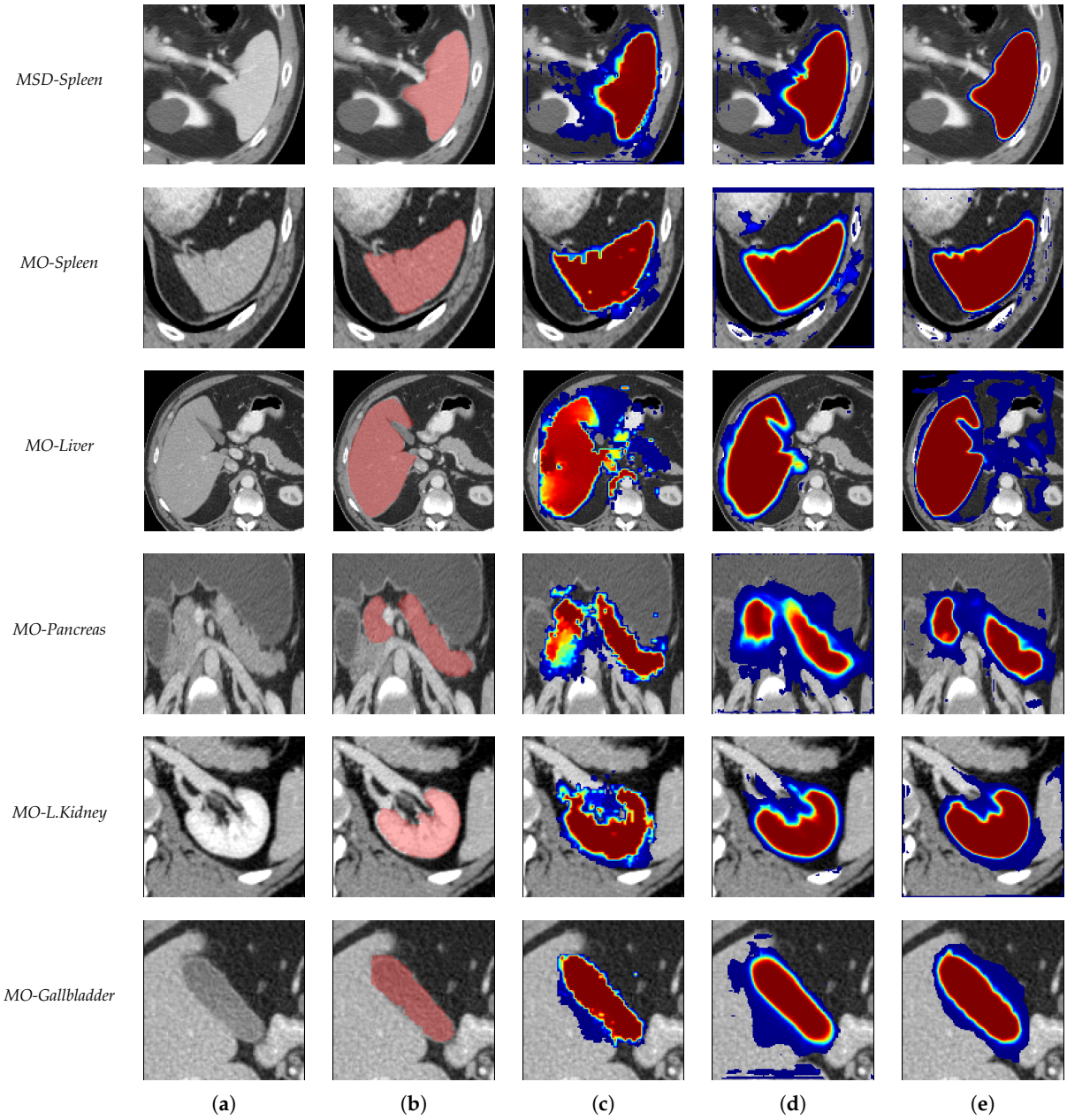


Figure 3. Our results on six different segmentation tasks on example cases from the validation set. We show (**a**) the image after cropping based on extreme points, (**b**) overlaid (full) ground truth (used for evaluation only), (**c**) initial random walker prediction, (**d**) our final segmentation result produced by the weakly supervised segmentation scheme, (**e**) the fully supervised result for reference. Specifically, we compare example cases for *weak. sup. dextr3D (w RW) Dice + Point loss + Point Attn* and *fully sup. dextr3D Dice loss* for (**d**,**e**), respectively. The probability maps are scaled between 0 and 1 and we show all non-zero probabilities.

The weights between adjacent voxels can be defined as $w_{ij} = e^{-\beta|z_j - z_i|^2}$. This will make the diffusion between similar voxel intensities z_i and z_j easier and hence allow them to be assigned to the same class. Here, β is a tunable hyperparameter that controls the amount of diffusion. We keep $\beta = 130$ in all our experiments. By separating the voxels marked by scribbles S and unmarked voxels, the Laplacian matrix L can be decomposed into blocks.

$$L = \begin{bmatrix} L_M & B \\ B^T & L_U \end{bmatrix} \tag{3}$$

Here, M corresponds to voxels marked by scribbles S and U to unmarked voxels. This can be formulated as a system of equations which can be analytically solved as:

$$L_U X = -B^T M, \tag{4}$$

where M is made of elements m_j^ω which are 1 for marked voxels of s_ω for the given class ω, and 0 otherwise. Solving Equation (4), results in a probability for each voxel $p(\omega|x) = x_i^\omega$, resulting in our pseudo-label $\hat{Y}$.

2.4. Step 3: Segmentation via Deep Fully Convolutional Network

Next, we can train a fully convolutional neural network to segment the given foreground class with $P(X) = f(X)$ with pairs of X and pseudo-labels $\hat{Y}$. Our preferred network architecture follows the encoder-decoder network proposed in Myronenko [9] (without the VAE part), using 3D convolutions throughout the network.

2.5. Encoder

The encoder uses residual blocks [41], where each block consists of two convolutions with normalization and ReLU, followed by additive skip connection. Here, we use *group normalization* (GN) [42], which typically shows better performance than batch normalization [43] when batch size is small (in our case batch size 4). We adopt a standard FCN approach to slowly decrease the number of image dimensions by 2 and simultaneously increase the number of features by 2 as in Ronneberger et al. [5]. For downsizing, we use strided convolutions with a stride of 2. All conversions are $3 \times 3 \times 3$ with an initial filter number equal to 8 in the input layer of the network.

2.6. Decoder

The design of the decoder is identical to the one of the encoders, but with a single residual block per each spatial level of the network. Each level of decoders starts with upsampling that involves reducing the number of features by a factor of 2 (using $1 \times 1 \times 1$ convolutions) and doubling the spatial dimension using trilinear upsampling. This is followed by adding or concatenating the features from the equivalent spatial level encoder. In this study we use addition due to the lower memory consumption of the resulting network. At the end of the decoder, the features have the same spatial size as the original image and the number of features equal to the size of the initial input function. This is followed by $1 \times 1 \times 1$ conversion into one output channel followed by a final *sigmoid* activation as we are assuming the binary segmentation case in this work. As shown in Figure 1, we use four different resolution levels inside the segmentation network. Therefore, we require three downsizing and upsampling operations each in the encoder/decoder branch, respectively. We use trilinear upsampling here as it has been shown to be efficient for medical imaging tasks [9] and requires no trainable parameters compared to alternative operations such as transposed convolution.

2.7. Attention

We follow the approach of Oktay et al. [37] to implement attention gates in the decoder part of our segmentation network. The attention gates help the model to focus on the structures of interest. Attention gates can encourage the model to suppress regions that are irrelevant to the segmentation task and highlight the regions of interest most relevant to the segmentation task (see Figure 1).

The attention gate can be further augmented by the point channel information available from extreme point selection. We propose to add the extreme point channels $G(\{e\})$ at each level of the decoder to further guide the network to learn the relevant information. In practice, we downsample the initial input point channel to match the resolution of each decoder level and concatenate it with the gating features from the encoder path of the network in each attention gate.

2.8. Dice Loss

The Dice loss [6] is a popular objective function for segmentation tasks in medical imaging. Its properties allow it to automatically scale to unbalanced labeling problems. At the same time, it also naturally adapts without any changes to the original formulation to learn from probability maps:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} \hat{y}_i^2} \tag{5}$$

Here, y_i is the predicted probability from our network $f(X)$ and $\hat{y}_i$ is the weak label probability from our pseudo-label map $\hat{Y}$ at voxel i.

2.9. Point Loss

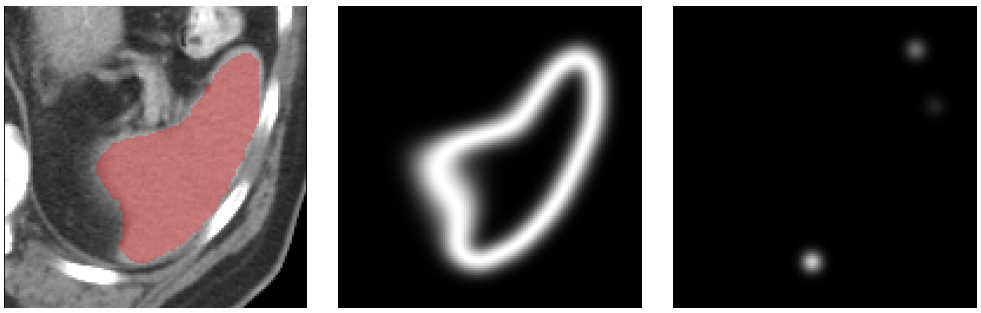
The extreme points selected by the user for weak annotation cannot only be used for generating initial scribbles but also in an additional loss function during the training of our deep neural network. We add an additional constraint to the deep-learning training making use of the extreme points the user has already selected. This new loss L_{points} penalizes the distance between the boundary of our model's predicted segmentation mask $P = f(X)$ and the location of the extreme points. To compute it, we apply a Gaussian filter $\mathcal{G}(\cdot)$ to our models prediction $P(X)$ which can be easily implemented using standard 3D convolutional operations with a constant $n \times n \times n$ kernel with each element being $1/n^3$. The resulting point distance loss between the filtered prediction $\mathcal{G}(P(X))$ and the extreme points channel $G(\{e\})$ (which includes a Gaussian kernel placed over each extreme point) is therefore

$$L_{\text{points}} = -\frac{1}{N}\sum_{i=1}^{N} g_i \mathrm{g}_i, \tag{6}$$

where N are the number of voxels i in the image and $g_i \in \mathcal{G}(P(X))$ and $\mathrm{g}_i \in G(\{e\})$, respectively. The point loss computation is illustrated in Figure 4. This results in a new total loss used for training:

$$L = L_{\text{Dice}} + \alpha L_{\text{points}}. \tag{7}$$

Here, α is hyperparameter weight that controls the influence of the point distance loss.



(**a**) (**b**) (**c**)

Figure 4. Visualization of the boundary enhancement map computed in Equations (6)–(8) in the paper. In this example, we show (**a**) the ground truth overlaid on the image, (**b**) the boundary enhancement map $b(P)$, and (**c**) the point channel $G(\{e\})$ on the corresponding axial slice of the 3D volume used in the computation of the point loss L_{points} (see Equation (6)). The loss is minimized if the prediction's boundary $b(P)$ aligns with the center of each clicked extreme point e in $G(\{e\})$. Note that during training, we compute the boundary on the model's prediction P but here we show it computed on the ground truth for illustration purpose.

2.10. Point Loss Implementation

We implement the Gaussian filter $\mathcal{G}(\cdot)$ using a set of standard 3D convolutions. First, we use B convolution operations to enhance the boundary of the prediction $P(X)$:

$$\begin{aligned} \mathcal{G}_0(P(X)) &= conv_B(...(conv_3(conv_2(conv_1(P(X)))))...) \\ \mathcal{G}_1(P(X)) &= (\mathcal{G}_0(P(X)) - 0.5)^2 \\ \mathcal{G}(P(X)) &= e^{-\mathcal{G}_1(P(X))} \end{aligned} \tag{8}$$

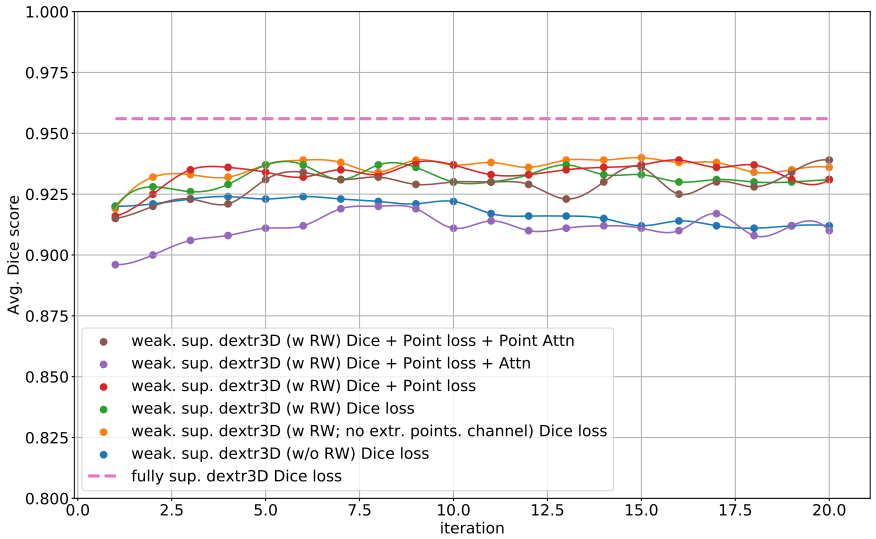
Here, the convolutional kernel in each *conv* operation is set to be constant $n \times n \times n$ kernel with each element being $1/n^3$. B should be adjusted depending on the size of input image and the extent of organ of interest. In our setting, we use $B = 25$ to achieve a good boundary enhancement at the scale of the images and targeted organs.

The resulting point distance loss between the filtered prediction $\mathcal{G}(P(X))$ and the extreme points channel $G(\{e\})$ (which includes a Gaussian kernel placed over each extreme point) is therefore as in Equation (6).

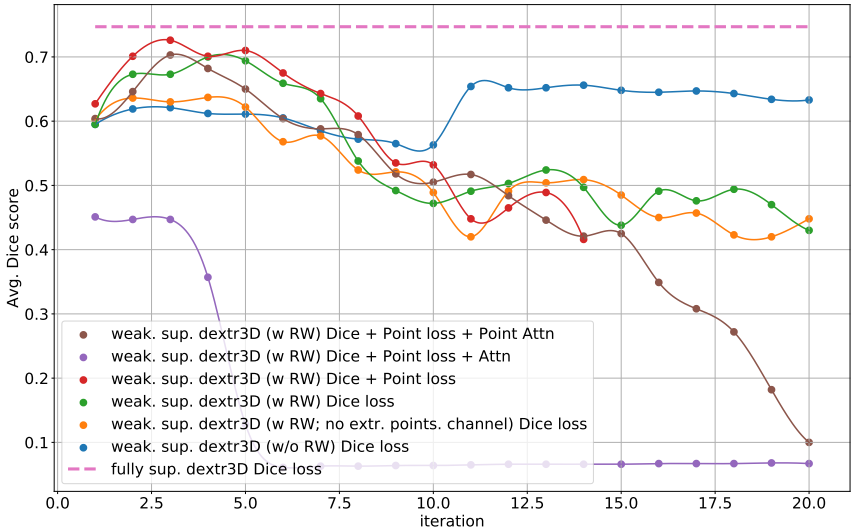
2.11. Step 4: Regularization Using Random Walker Algorithm

We could stop learning after the above segmentation network $f(X)$ is trained on the $\hat{Y}$ pseudo-labels for the first time. Nevertheless, we note that an additional regularization step by an additional random walker segmentation as mentioned above may be of great benefit to the convergence of our weakly supervised segmentation approach. This approach is close in spirit to Rajchl et al. [29], where a DenseCRF is used as the post-processing step during iterative refinement.

To increase the amount of regularization that the random walker can give to the predictions $P(X)$ of the network, we define an area of uncertainty $\mathcal{U}(P(X))$. The foreground and background in the prediction map can be defined as $P(X) >= 0.5$ and $P(X) < 0.5$, respectively. Here, we chose a ball structure element of radius $r_{\text{randomwalker}} = 4$ to erode both the foreground and background regions in all our segmentation tasks to compute $\mathcal{U}$ which in turn is acting as the unmarked voxels in the random walker algorithm. This allows the random walker to generate new predictions around the foreground object's boundary that differ from previous 3D network's predictions and, in turn, help the next deep-learning training iteration to learn new features from the same set of training images and to not become stuck in a poor local minimum. Moreover, we find that our weakly supervised segmentation framework becomes unstable without this step and does not converge as easily to a satisfactory result (see Figure 5).



MO-Liver



MO-Pancreas

Figure 5. Weakly supervised training from random walker initialization. For illustration, we only show the *MO-Liver* and *MO-Pancreas* segmentation tasks with the varying training settings as shown in our ablation study of Table 1 at each round of deep network training. Although the performance of the *MO-Liver* models generally improves with the number of iterations, it can also be noticed that for *MO-Pancreas*, a poor initialization by the random walker can cause the models to degrade quickly. Notice that adding the point channel information results in a more stable training behavior.

3. Experiments and Results

3.1. Datasets

We use the training datasets (as they include ground truth annotations) from public resources, specifically, from the multi-organ (MO) segmentation study in Gibson et al. [44] (https://zenodo.org/record/1169361#.XcsiOHFKi90, accessed on 28 May 2021) which provided annotation for abdominal CT data from previously published datasets: Roth et al. [45] (https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT, accessed on 28 May 2021) and BTCV [46] (https://www.synapse.org/#!Synapse:syn3193805/wiki/217752, accessed on 28 May 2021). Furthermore, we use data from the *Medical Segmentation Decathlon* (MSD) challenge [47] (http://medicaldecathlon.com, accessed on 28 May 2021). From MO, we use the spleen, liver, pancreas, left kidney, and gallbladder segmentation masks, denoted as *MO-Spleen*, *MO-Liver*, *MO-Pancreas*, *MO-L.Kidney*, and *MO-Gallbladder*, respectively. From MSD, we include the spleen mask, denotes as *MSD-Spleen*. Qualitative results are shown in Figure 3 for each segmentation task on example cases from the validation set. For MO, we use a constant data split of 81 training and 9 validation cases, respectively. For MSD, there are 32 training and 9 validation cases, respectively, available. In Table 1, we only evaluate the different variations of our proposed approach on the validation set as our method is aimed at building models to help the creation of annotated data.

For testing, we evaluate three of the trained annotation models using the proposed approach (*weak. sup. (w RW) Dice + PL + Pt. Attn.*) on completely unseen datasets. Specifically, we use all cases available in CT-ORG [48] (https://wiki.cancerimagingarchive.net/display/Public/CT-ORG%3A+CT+volumes+with+multiple+organ+segmentations, accessed on 28 May 2021) to evaluate the *MO-Liver*and *MO-L.Kidney*on 139 and 137 test cases, respectively. We excluded cases that did not have liver annotations or surgically removed left kidneys. The *MO-Pancreas*model is evaluated on all the training cases from MSD-Pancreas set, including 281 cases with public annotations. These CT images were acquired from patients undergoing resection of pancreatic masses [47]. Results on unseen testing are reported in Table 2.

3.2. Experiments

In all cases, we iterate our algorithm for a maximum of 20 iterations as shown in Figure 5, with 300 epochs per iteration. In Table 1, we compare training with and without using random walker (RW) regularization after each round of 3D FCN learning. The Gaussian noise added to the extreme points during training was set to be $\sigma = 3$ voxels. During standard inference, it was set as $\sigma = 0$ voxels. An analysis of the effect of this noise parameter during inference is given in Section 3.5. In addition, by running the framework with RW regularization but without the extreme points channel, we quantify the benefit of modeling the extreme points as an extra input channel to the network versus only using the bounding box as in Rajchl et al. [29]. It can be further observed that the greatest changes occur after initial random walker segmentation in the first round of FCN training. Although the average dice score is not always enhanced by random walker regularization alone, it helps to incorporate enough "novelty" into our learning system to boost the overall Dice score in later iterations as shown in Figure 5. We furthermore show the average Dice scores on the validation set after convergence when using the proposed point loss, point loss plus attention gates as in Oktay et al. [37], and a setting when using the point information as an additional guiding feature to the attention gates. The weight in Equation (7) is set to be $\alpha = 0.2$ in all experiments. The fully supervised case using Dice loss with the strong label ground truth masks are shown for reference. It can be observed that using the point channel information in the point loss function and the attention gates generally improves the performance of the model. The addition of point loss and point attention works best in four out of six weakly supervised cases, while the addition of point loss alone showed an advantage in two out of the six tasks. Notice that the average Dice score in the *MO-Gallbladder* task even outperforms the fully supervised setting.

3.3. Implementation

The training and evaluation of the deep neural networks used in the proposed framework were implemented based on the *NVIDIA Clara Train SDK* (https://developer.nvidia.com/clara, accessed on 28 May 2021) using 4 NVIDIA Tesla V100 GPUs with 16 GB memory for each round of training. All models were trained using the deterministic training setup in *Tensorflow* (https://github.com/NVIDIA/tensorflow-determinism, accessed on 28 May 2021) with the same random seed initialization to guarantee comparable results between the different variations of training. For the random walker algorithm, we use the default parameters (https://scikit-image.org/docs/dev/auto_examples/segmentation/plot_random_walker_segmentation.html, accessed on 28 May 2021).

3.4. Analysis of Point Loss

An analysis of the impact of the point loss on our weakly supervised models' predictions is shown in Figure 6.

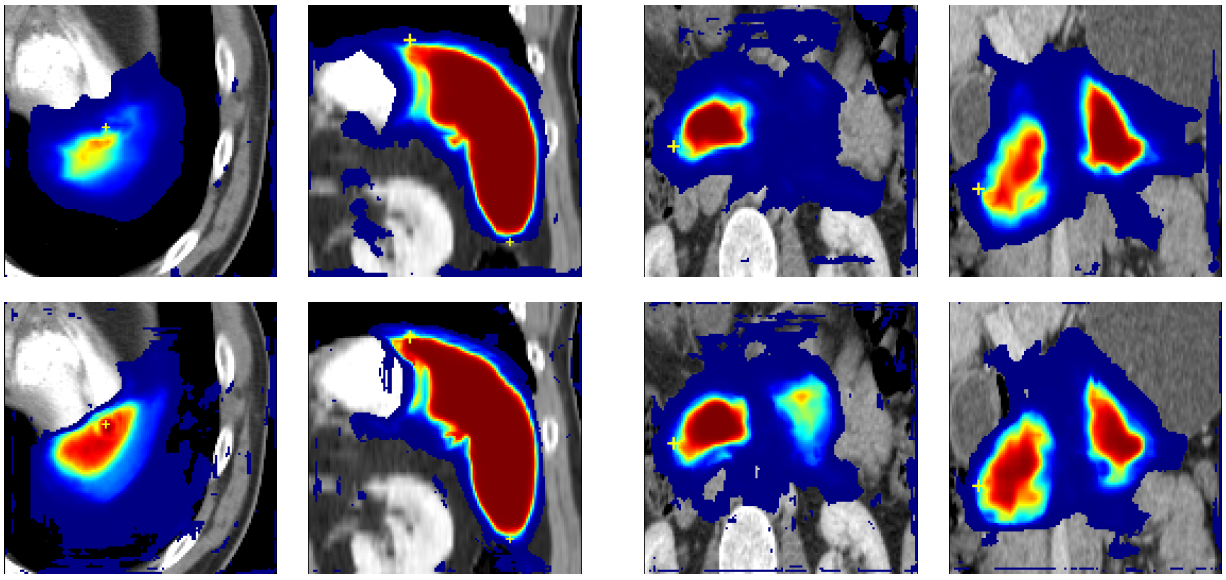


Figure 6. The impact of adding the point loss and point attention to our weakly supervised models. We show the results of the (**top**): *weak. sup. dextr3D (w RW) Dice*; and (**bottom**): *weak. sup. dextr3D (w RW) Dice + Point loss + Point Attn* settings. Examples from the *MSD-Spleen* (**left**) and *MO-Pancreas* (**right**) datasets are shown, respectively. The clicked extreme points are shown by a yellow cross. Best viewed in color. The predictions learned together with the point loss do lie markedly closer to the clicked point locations.

3.5. Effect of Extreme Point Noise during Inference

Figure 7 shows the effect of the noise level σ added to the extreme points to simulate user interaction during inference. Here, we show the average Dice score as a value of σ when evaluating the *MO-Liver*model on the CT-ORG data (N = 139) unseen during training and validation. One can observe that the model's performance is relatively robust up to a noise level of $\sigma = 10$.

Table 2. Results of our weakly supervised segmentation models on unseen testing data. N indicates the number of test cases.

Dice	CT-ORG-Liver	CT-ORG-l.kidney	MSD-Pancreas
Mean	0.932	0.897	0.678
Std	0.075	0.126	0.111
Median	0.947	0.930	0.706
N	139	137	281

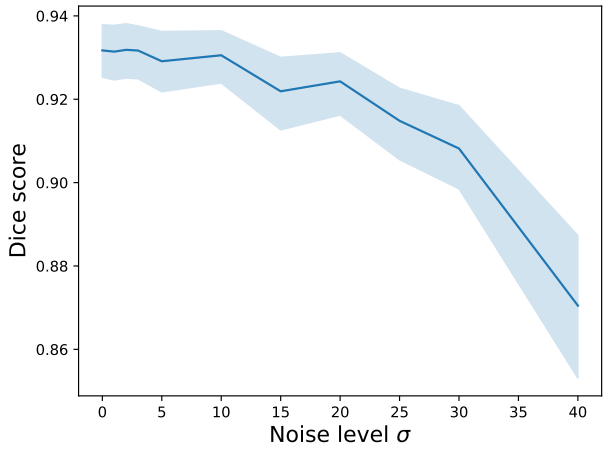


Figure 7. Dice score (mean and 95% confidence interval) as a value of the Gaussian noise level σ added to extreme points when simulating user interaction. Here, we evaluate the *MO-Liver*model on the CT-ORG data (N = 139) unseen during training and validation.

4. Discussion

We provided a method for weakly supervised 3D segmentation from extreme points. Asking the user to simply click on the surface of the organ in each spatial dimension can drastically reduce the cost of labeling. The point clicks can simultaneously identify the region of interest and simplify the 3D machine-learning task. The extreme points can also be used to create an initial noisy pseudo-label based on the extreme points using the random walker algorithm. From our experiments, it can be observed that this initialization is relatively robust for six different tasks from medical image segmentation.

Previous work primarily used boundary box annotations for weakly supervised learning in 2D/3D medical imaging, such as Rajchl et al. [29]. We consider, however, that selecting extreme points on the surface of the organ is more natural than selecting corners of a bounding box outside the organ of interest and more efficient than adding scribbles within and around the organ [19,21]. This is consistent with recent findings in the computer vision literature [39]. Extreme points have also shown to be useful for further down-stream tasks such as domain adaptation [49]. An application of our proposed approach to the 2D tasks would be straightforward. Therefore, it could likely be used for weakly supervised annotation and segmentation of other biomedical imaging data, such as X-rays, microscopy, digital pathology, or even natural images.

We conducted a comprehensive ablation study of our proposed method in Table 1. Some of these settings are similar to previous work. For example, performing the network training without the extra point channel is equivalent to studies using bounding boxes alone such as in Rajchl et al. [29]. From Table 1, we can see that adding the additional point-click information in the loss and as attention mechanism is however beneficial while not increasing the labeling cost.

The test results (see Table 2) on completely unseen datasets from different data sources show that our trained weakly supervised annotation models can generalize to unseen data. The performance on unseen liver and kidney cases is very comparable to the validation performance of these models (see Table 1). Given that unseen test cases for the pancreas model only include patients with malignant pancreatic masses, the drop in performance compared to the validation score is understandable. The training set did not contain such cases, but patients were typically healthy or had diseases outside the pancreas [44–46].

Limitations

Occasionally, the random walker may lack robustness for organs with very diverse interior textures or highly concave curved shapes, for example, the pancreas (see *MO-Pancreas* task in Table 1). In this situation, the shortest path result might sometimes lie

outside the organ. A boundary search algorithm might provide a better initial segmentation here. Still, the initial segmentation can be significantly enhanced by the first round of FCN training. In this study, we used one dataset (*MSD-Spleen*) as our development set and kept the hyperparameters of the full approach constant across different segmentation tasks and datasets. One might achieve better performance when optimizing the hyperparameters, especially for the initial random walker, based on the task at hand. In practice, we performed model selection for each round of training in our approach based on the pseudo-labels $\hat{Y}$ alone. However, we do need a fully annotated validation set to practically evaluate the overall convergence of our iterative approach for it to be clinically useful. One could use the predictions of the first round of FCN training to build an ML-based annotation tool that could speed up the creation of such a holdout "gold standard" validation dataset and reduce the amount of manual labeling and editing needed in total.

5. Summary

In summary, we proposed a weakly supervised 3D segmentation framework based on extreme point clicks. Experimentation on six datasets showed that the approach can achieve performance close to the fully supervised setting in four tasks and even outperforms the fully supervised training in one of them (*MO-Gallbladder*). In the future, an automatic proposal network could assist the user with the region of interest and extreme point selection to further reduce the manual burden of medical image annotation.

Author Contributions: Conceptualization, H.R.R., D.Y., Z.X., X.W. and D.X.; Formal analysis, H.R.R.; Investigation, D.X.; Methodology, H.R.R.; Software, H.R.R.; Supervision, D.X.; Writing—original draft, H.R.R.; Writing—review & editing, D.Y., Z.X., X.W. and D.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: A pre-print is available on arxiv https://arxiv.org/abs/2009.11988, accessed on 28 May 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Devaraj, A.; van Ginneken, B.; Nair, A.; Baldwin, D. Use of Volumetry for Lung Nodule Management: Theory and Practice. *Radiology* **2017**, *284*, 630–644. [CrossRef]
2. Yushkevich, P.A.; Piven, J.; Cody Hazlett, H.; Gimpel Smith, R.; Ho, S.; Gee, J.C.; Gerig, G. User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability. *Neuroimage* **2006**, *31*, 1116–1128. [CrossRef] [PubMed]
3. Grady, L. Random walks for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1768–1783. [CrossRef]
4. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
5. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
6. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
7. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Proceedings of the 19th International Conference on Medical Image Computing and Computer Assisted Intervention, Athens, Greece, 17–21 October 2016; Springer: Cham, Switzerland, 2016; pp. 424–432
8. Liu, S.; Xu, D.; Zhou, S.K.; Pauly, O.; Grbic, S.; Mertelmeier, T.; Wicklein, J.; Jerebko, A.; Cai, W.; Comaniciu, D. 3D Anisotropic Hybrid Network: Transferring Convolutional Features from 2d Images to 3d Anisotropic Volumes. In Proceedings of the International Conference on Medical Image Computing & Computer Assisted Intervention, Granada, Spain, 16–20 September 2018; Springer: Cham, Switzerland, 2018; pp. 851–858.
9. Myronenko, A. 3D MRI brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 311–320.

10. Guo, Y.; Liu, Y.; Georgiou, T.; Lew, M.S. A review of semantic segmentation using deep neural networks. *Int. J. Multimed. Inf. Retr.* **2018**, *7*, 87–93. [CrossRef]
11. Tajbakhsh, N.; Jeyaseelan, L.; Li, Q.; Chiang, J.N.; Wu, Z.; Ding, X. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Med. Image Anal.* **2020**, *63*, 101693. [CrossRef] [PubMed]
12. An, G.; Hong, L.; Zhou, X.B.; Yang, Q.; Li, M.Q.; Tang, X.Y. Accuracy and efficiency of computer-aided anatomical analysis using 3D visualization software based on semi-automated and automated segmentations. *Ann. Anat. Anat. Anz.* **2017**, *210*, 76–83. [CrossRef] [PubMed]
13. Boykov, Y.; Funka-Lea, G. Graph cuts and efficient ND image segmentation. *IJCV* **2006**, *70*, 109–131. [CrossRef]
14. van Ginneken, B.; de Bruijne, M.; Loog, M.; Viergever, M.A. Interactive shape models. *Med. Imaging 2003 Image Process. Int. Soc. Opt. Photonics* **2003**, *5032*, 1206–1216.
15. Schwarz, T.; Heimann, T.; Wolf, I.; Meinzer, H.P. 3D heart segmentation and volumetry using deformable shape models. In Proceedings of the 2007 Computers in Cardiology, Durham, NC, USA, 30 September–3 October 2007; pp. 741–744.
16. Dougherty, G. *Medical Image Processing: Techniques and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2011.
17. Wang, G.; Zuluaga, M.A.; Pratt, R.; Aertsen, M.; Doel, T.; Klusmann, M.; David, A.L.; Deprest, J.; Vercauteren, T.; Ourselin, S. Slic-Seg: A minimally interactive segmentation of the placenta from sparse and motion-corrupted fetal MRI in multiple views. *Med. Image Anal.* **2016**, *34*, 137–147. [CrossRef] [PubMed]
18. Amrehn, M.; Gaube, S.; Unberath, M.; Schebesch, F.; Horz, T.; Strumia, M.; Steidl, S.; Kowarschik, M.; Maier, A. UI-Net: Interactive artificial neural networks for iterative image segmentation based on a user model. *Eurographics Workshop Vis. Comput. Biol. Med.* **2017**, arXiv:1709.03450.
19. Wang, G.; Zuluaga, M.A.; Li, W.; Pratt, R.; Patel, P.A.; Aertsen, M.; Doel, T.; Divid, A.L.; Deprest, J.; Ourselin, S.; et al. DeepIGeoS: A deep interactive geodesic framework for medical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1559–1572. [CrossRef] [PubMed]
20. Wang, G.; Li, W.; Zuluaga, M.A.; Pratt, R.; Patel, P.A.; Aertsen, M.; Doel, T.; David, A.L.; Deprest, J.; Ourselin, S.; et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Trans. Med. Imaging* **2018**, *37*, 1562–1573. [CrossRef]
21. Can, Y.B.; Chaitanya, K.; Mustafa, B.; Koch, L.M.; Konukoglu, E.; Baumgartner, C.F. Learning to Segment Medical Images with Scribble-Supervision Alone. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 236–244.
22. Dias, P.A.; Shen, Z.; Tabb, A.; Medeiros, H. FreeLabel: A Publicly Available Annotation Tool Based on Freehand Traces. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019; pp. 21–30.
23. Sakinis, T.; Milletari, F.; Roth, H.; Korfiatis, P.; Kostandy, P.; Philbrick, K.; Akkus, Z.; Xu, Z.; Xu, D.; Erickson, B.J. Interactive segmentation of medical images through fully convolutional neural networks. *arXiv* **2019**, arXiv:1903.08205.
24. Khan, S.; Shahin, A.H.; Villafruela, J.; Shen, J.; Shao, L. Extreme Points Derived Confidence Map as a Cue for Class-Agnostic Interactive Segmentation Using Deep Neural Network. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*; Springer International Publishing: Cham, Switzerlans, 2019; pp. 66–73.
25. Majumder, S.; Yao, A. Content-Aware Multi-Level Guidance for Interactive Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
26. Ling, H.; Gao, J.; Kar, A.; Chen, W.; Fidler, S. Fast Interactive Object Annotation With Curve-GCN. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
27. Dias, P.A.; Medeiros, H. *Semantic Segmentation Refinement by Monte Carlo Region Growing of High Confidence Detections*; Computer Vision—ACCV 2018; Jawahar, C.V., Li, H., Mori, G., Schindler, K., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 131–146.
28. Cerrone, L.; Zeilmann, A.; Hamprecht, F.A. End-To-End Learned Random Walker for Seeded Image Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
29. Rajchl, M.; Lee, M.C.; Oktay, O.; Kamnitsas, K.; Passerat-Palmbach, J.; Bai, W.; Damodaram, M.; Rutherford, M.A.; Hajnal, J.V.; Kainz, B.; et al. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE Trans. Med. Imaging* **2017**, *36*, 674–683. [CrossRef]
30. Zhang, L.; Gopalakrishnan, V.; Lu, L.; Summers, R.M.; Moss, J.; Yao, J. Self-learning to detect and segment cysts in lung CT images without manual annotation. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 1100–1103.
31. Nie, D.; Gao, Y.; Wang, L.; Shen, D. ASDNet: Attention Based Semi-supervised Deep Networks for Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*; Springer International Publishing: Cham, Switzerland, 2018; pp. 370–378.
32. Kervadec, H.; Dolz, J.; Tang, M.; Granger, E.; Boykov, Y.; Ayed, I.B. Constrained-CNN losses for weakly supervised segmentation. *Med Image Anal.* **2019**, *54*, 88–99. [CrossRef] [PubMed]

33. Cai, J.; Tang, Y.; Lu, L.; Harrison, A.P.; Yan, K.; Xiao, J.; Yang, L.; Summers, R.M. Accurate weakly-supervised deep lesion segmentation using large-scale clinical annotations: Slice-propagated 3D mask generation from 2D RECIST. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 396–404.
34. Rajchl, M.; Lee, M.C.; Schrans, F.; Davidson, A.; Passerat-Palmbach, J.; Tarroni, G.; Alansary, A.; Oktay, O.; Kainz, B.; Rueckert, D. Learning under distributed weak supervision. *arXiv* **2016**, arXiv:1606.01100.
35. Roth, H.; Zhang, L.; Yang, D.; Milletari, F.; Xu, Z.; Wang, X.; Xu, D. Weakly supervised segmentation from extreme points. In *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS) and Hardware Aware Learning (HAL) for Medical Imaging and Computer Assisted Intervention (MICCAI)*; Springer: Cham, Swizerland, 2019.
36. Maninis, K.K.; Caelles, S.; Pont-Tuset, J.; Van Gool, L. Deep Extreme Cut: From Extreme Points to Object Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 616–625.
37. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. In Proceedings of the 1st Conference on Medical Imaging with Deep Learning (MIDL), Amsterdam, The Netherlands, 4–6 July 2018.
38. Roth, H.R.; Lu, L.; Lay, N.; Harrison, A.P.; Farag, A.; Sohn, A.; Summers, R.M. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. *Med Image Anal.* **2018**, *45*, 94–107. [CrossRef] [PubMed]
39. Papadopoulos, D.P.; Uijlings, J.R.; Keller, F.; Ferrari, V. Extreme clicking for efficient object annotation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4930–4939.
40. Dijkstra, E.W. A note on two problems in connexion with graphs. *Numer. Math.* **1959**, *1*, 269–271. [CrossRef]
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
42. Wu, Y.; He, K. Group normalization. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
43. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
44. Gibson, E.; Giganti, F.; Hu, Y.; Bonmati, E.; Bandula, S.; Gurusamy, K.; Davidson, B.; Pereira, S.P.; Clarkson, M.J.; Barratt, D.C. Automatic multi-organ segmentation on abdominal CT with dense v-networks. *IEEE Trans. Med. Imaging* **2018**, *37*, 1822–1834. [CrossRef]
45. Roth, H.R.; Lu, L.; Farag, A.; Shin, H.C.; Liu, J.; Turkbey, E.B.; Summers, R.M. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 556–564.
46. BTCV. Multi-Atlas Labeling Beyond the Cranial Vault—MICCAI Workshop and Challenge. 2015. Available online: https://www.synapse.org/#!Synapse:syn3193805 (accessed on 28 May 2021).
47. Simpson, A.L.; Antonelli, M.; Bakas, S.; Bilello, M.; Farahani, K.; van Ginneken, B.; Kopp-Schneider, A.; Landman, B.A.; Litjens, G.; Menze, B.; et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv* **2019**, arXiv:1902.09063.
48. Rister, B.; Yi, D.; Shivakumar, K.; Nobashi, T.; Rubin, D.L. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. *Sci. Data* **2020**, *7*, 1–9. [CrossRef] [PubMed]
49. Raju, A.; Ji, Z.; Cheng, C.T.; Cai, J.; Huang, J.; Xiao, J.; Lu, L.; Liao, C.; Harrison, A.P. User-Guided Domain Adaptation for Rapid Annotation from User Interactions: A Study on Pathological Liver Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 457–467.