Modeling clinical assessor intervariability using deep hypersphere encoder–decoder networks

Joost van der Putten¹ · Fons van der Sommen¹ · Jeroen de Groof² · Maarten Struyvenberg² · Svitlana Zinger¹ · Wouter Curvers³ · Erik Schoon³ · Jacques Bergman² · Peter H. N. de With¹

Received: 12 March 2019 / Accepted: 8 November 2019 / Published online: 21 November 2019
© The Author(s) 2019

Abstract

In medical imaging, a proper gold-standard ground truth as, e.g., annotated segmentations by assessors or experts is lacking or only scarcely available and suffers from large intervariability in those segmentations. Most state-of-the-art segmentation models do not take inter-observer variability into account and are fully deterministic in nature. In this work, we propose hypersphere encoder–decoder networks in combination with dynamic leaky ReLUs, as a new method to explicitly incorporate inter-observer variability into a segmentation model. With this model, we can then generate multiple proposals based on the inter-observer agreement. As a result, the output segmentations of the proposed model can be tuned to typical margins inherent to the ambiguity in the data. For experimental validation, we provide a proof of concept on a toy data set as well as show improved segmentation results on two medical data sets. The proposed method has several advantages over current state-of-the-art segmentation models such as interpretability in the uncertainty of segmentation borders. Experiments with a medical localization problem show that it offers improved biopsy localizations, which are on average 12% closer to the optimal biopsy location.

Keywords Deep learning · Intervariability modeling · segmentation · Localization

1 Introduction

A long-standing challenge in machine learning is the usage of data sets that lack a proper gold-standard ground truth. Especially in medical imaging applications, many data set segmentations exhibit a large inter-observer variability [20, 22]. In contrast to, e.g., a traffic sign, a cancerous lesion boundary might be undefined or only clearly defined in certain regions of the lesion. Training conventional state-of-the-art algorithms for segmentation such as U-net [16] will result in segmentation proposals that do not incorporate inter-observer variability of assessors. Additionally, these networks are deterministic and therefore

incapable of producing multiple proposals with varying confidence given the same image. Therefore, it is impossible to effectively infer information from differing expert opinions from the output of the model.

In this work, we propose the hypersphere encoder–decoder network (HEDN), a novel deep learning architecture capable of modeling inherent data ambiguity, based on multi-assessor input to address the above-mentioned problem. Our framework extends an encoder–decoder network similar to U-net with an external scaling factor ρ , with which we establish a relation between the scaling factor and the size of the output segmentation. First, a feature mapping layer is added to the architecture which maps the latent variables on a hypersphere with radius ρ . Second, we propose dynamic leaky ReLUs (DLReLUs) which depend on ρ as well. Parameter ρ is used as an additional input variable to the network to control the *tightness* of the segmentation. A small ρ will result in the most probable segmentation location, while a large ρ results in a segmentation prediction that encompasses a larger relevant area based on the intervariability of

✉ Joost van der Putten
j.a.v.d.putten@tue.nl

¹ VCA Group, Eindhoven University of Technology, Eindhoven, The Netherlands

² Amsterdam University Medical Center, Amsterdam, The Netherlands

³ Catharina Hospital, Eindhoven, The Netherlands

assessors. A key feature of this algorithm is the capability to generate multiple segmentation proposals of a given image, using a single network by changing ρ . This feature provides a measure of interpretability, since a confidence estimation based on the modeled inter-observer variability can be generated by integrating over ρ . Summarizing, broadly accepted models for ground-truth annotations are rather limited in the current state of the art. Therefore, we propose a tunable model that intuitively incorporates the uncertainty of multiple assessors, by employing a specific variable ρ which controls the output size of the prediction.

In work by Kohl et al. [7], the authors present a multi-proposal framework by combining U-net with a conditional variational auto-encoder [6, 19]. Although the presented results look promising, the proposed architecture requires two auxiliary networks in order to generate new samples. Additionally, the authors do not present a real-world example in which their model improves a metric on a data set with ambiguous ground truth compared to standard methods. A simpler way to produce multiple hypotheses is by creating an ensemble of models [8]. However, the outputs of ensemble models are not necessarily diverse and require a large amount of computational resources. In order to overcome the diversity problem, several approaches train models jointly using the oracle set loss [4], i.e., a loss that only takes the closest prediction to the ground truth into account. In the same category, Lee et al. [9] explore ensembles of deep networks, while other work [17] investigates a single model with multiple heads. Two disadvantages of these approaches are the requirement of fixing the number of allowed hypotheses at training time and the inability of these models to scale to a large number of hypotheses.

The contribution of this work is threefold. (1) We propose a novel deep learning architecture, hypersphere encoder–decoder networks, capable of providing the user with multiple region proposals based on the intervariability of assessors. (2) We show a proof of concept of the model by evaluating it on three distinct data sets. (3) After evaluating a segmentation and localization case study which has inherent ambiguity, we exhibit that the proposed architecture improves results in both cases.

The remainder of this paper is structured as follows. First, Sect. 2.1 discusses the data sets used for this study. Second, an in-depth explanation of the proposed deep learning architecture is discussed in Sect. 2.2. Third, the performed experiments are presented in Sect. 2.3. Fourth, in Sect. 3 results of the experiment are discussed. Additionally, visual examples of the output proposals are presented. Fifth, Sect. 4 concludes the paper with discussion and future work.

2 Methods

2.1 Data sets

The proposed methodology is evaluated on three distinct data sets with increasing complexity that illustrate different use cases for our approach. First, a toy data set shows a proof of concept of the model. Second, in a segmentation data set (VLE), regions with both clear and unclear borders are segmented and visualizations by our model provide better insights into the data. Lastly, on a localization data set (ARGOS), we address the biopsy localization problem when assessor annotations vary greatly from image to image. In all three data sets, multiple annotations are available.

2.1.1 Vanishing shapes data set

The vanishing shapes data set was specifically created for this study. In this data set, multiple rectangles, circles and triangles with an outward vanishing gradient are generated on a cluttered background of 128×128 pixels. The pixels of the vanishing shapes are made more transparent further from the centroid by lowering the alpha channel, thereby creating an ambiguous problem as the edges are not clearly defined. However, since the images with vanishing shapes are computer generated, the location of the intersection and the union are a priori known and defined as the same shape on the same location with 30% and 80% of the original vanishing shape size, respectively. Since infinite examples can be generated, there is no distinction between training and test sets. Instead, evaluation is performed on newly generated images. Several examples from the vanishing shapes data set including union and intersection are shown in Fig. 1.

2.1.2 VLE data set

Volumetric laser endomicroscopy (VLE) is a second-generation optical coherence tomography (OCT) imaging modality, capable of making a full circumferential scan of the esophageal wall up to a tissue depth of 3 mm [2]. This novel modality allows physicians to analyze the deeper underlying tissue layers. However, due to the nature of the imaging procedure, the signal-to-noise ratio decreases for deeper layers of the tissue. As a result, the lower boundary of relevant tissue is ambiguous. Previous work has shown that tissue segmentation is a crucial preprocessing step for these types of images [14], but since the lower boundaries are not clearly defined, multiple proposals can greatly benefit the user.

Fig. 1 Top: examples of vanishing shapes images. Middle: corresponding union. Bottom: corresponding intersection

The VLE data set was acquired in a prospective single-center study in which 23 Barrett's patients with and without early neoplasia were included. In total, 131 regions of interest (ROIs) were extracted from the patients, where an ROI is defined as a selected section of the entire scan where histology was proven. The original ROIs have a resolution of 1342×1024 pixels, which were resized to 256×256 pixels for computational efficiency. Three VLE experts with more than three years of experience with this modality were asked to delineate the tissue of interest (TOI). The TOI is defined as the region in the image which potentially contains relevant information for further dysplasia detection. Examples of a VLE image with corresponding union and intersection are shown in Fig. 2. This figure clearly shows that all assessors include the top layers, while lower layers are ambiguous due to the decreasing signal-to-noise ratio. Even though attempts were made to account for the decreasing signal-to-noise ratio, the resulting prediction masks from other related work [14] show no nuance in their predictions.

2.1.3 ARGOS data set

Patients suffering from Barrett's esophagus (BE) have a greatly increased risk of developing dysplasia and eventually esophageal adenocarcinoma. Therefore, it is of paramount importance that dysplasia associated with BE is detected at an early stage. However, early neoplasia is very difficult to detect during surveillance endoscopies due to the subtle appearance and rarity of the disease. Leading experts only attain a sensitivity and specificity of merely 80% and 89% in diagnosis of dysplasia, respectively [18]. Furthermore, the current biopsy protocol is prone to sampling error and requires an abundance of biopsies in order to avoid missing the dysplastic region [5]. Due to the subtle nature of the dysplastic region, experts do not always agree on the location of the lesion, thereby making it even more difficult to properly train a computer-aided detection algorithm. Especially when the union in the image is relatively large, choosing an appropriate biopsy location can be challenging.

Fig. 2 From left to right: VLE image, corresponding union, and corresponding intersection

Fig. 3 From left to right: ARGOS image, corresponding union, and corresponding intersection

The ARGOS¹ [3] data set consists of 630 endoscopic BE images from 198 patients with histologically proven dysplasia with corresponding ground-truth annotations from two BE experts. The original images had varying resolutions of between 680×556 pixels and $1,400 \times 1,072$ pixels, but all images were normalized to 256×256 pixels to reduce the computing time. An example of an ARGOS image with corresponding intersection and union is shown in Fig. 3. The large difference between union and intersection clearly visualizes a large assessor intervariability in this image. Images were excluded in the cases where the assessors provided disjoint annotations. This was done for the following three reasons. (1) A corroborated ground truth is important in this problem since the intervariability is already very large. Without consensus, the affected samples cannot be used as absolute truth. (2) During training, the algorithm relies on an interpolation between the intersection and the union (Sect. 2.2.2) to calculate the loss; however, since the interpolation is not defined in these cases, a proper calculation is impossible. (3) Two out of three evaluation metrics described in Sect. 2.3.3 cannot be calculated without a nonzero intersection of the expert annotations. Including these samples would make evaluation of the models more difficult. Note that inference is still possible in these cases. This limitation is further discussed in Sect. 4.

2.2 Hypersphere encoder–decoder network

2.2.1 Model definition

The proposed HEDN architecture uses U-net [16] as the base model, but is extended with an additional feature mapping layer (purple block in Fig. 4), as well as an activation function which can dynamically be changed for each input. An overview of the proposed architecture is

shown in Fig. 4. In this section, model $M_\theta(Y|X, \rho)$ is discussed, which produces segmentation mask Y conditioned on input image X and parameter ρ . Model $M_\theta(Y|X, \rho)$ is comprised of encoder network $q_\theta(\mathbf{z}|X, \rho)$, which maps the input to latent variable vector $\mathbf{z} \in \mathbb{R}^N$, and decoder network $p_\theta(Y|\mathbf{z}, \rho)$, which produces a segmentation mask depending on latent variable $\mathbf{z}$ and size factor ρ . Encoder q_θ and decoder p_θ are symmetrical, meaning that they both have the same number of convolutions and resolutions. Max-pooling is used to downsample in the encoder, while transposed convolutions are used to upsample feature maps in the decoder. Standard ReLUs are used for all activation function except for the activation functions after the final convolutional layer in each encoder level. Furthermore, batch normalization is used after each convolutional layer. The output of the final layer is fed through a sigmoid function to generate output values between 0 and 1. Finally, the original skip connections from standard U-net are replaced with fully residual connections to reduce the amount of learnable parameters in the decoder.

2.2.2 Mask interpolation

Since the goal of the proposed algorithm is to make the output prediction size dependent on ρ , one of the most important learning aspects is determining a ground truth that reflects this property. In an ideal case with an abundance of assessors, one could simply use a level map obtained from all assessors and take the area that is most agreed upon when ρ is small and the area corresponding with minimal agreement when ρ is large. However, as expert labeling is expensive, the amount of expert annotators is usually limited. In our case, the VLE data set is annotated by three experts and the ARGOS data set only by two. In order to obtain intermittent ground-truth masks, a linear interpolation is generated between the intersection and the union in the following manner:

¹ Automatic Recognition of irreguLarities in the OesophaguS.

Fig. 4 Overview of the proposed architecture

1. Obtain intersection and union perimeter $\text{inter}_{\text{perim}}$, and $\text{union}_{\text{perim}}$,
2. Compute distance of each pixel to $\text{inter}_{\text{perim}}$, and $\text{union}_{\text{perim}}$, using the distance transform [13],
3. Invert values of distance maps inside the perimeters,
4. Interpolate between the distance maps at ‘distance’ $0 \leq \rho \leq 1$,
5. Threshold interpolation at zero crossing.

An example of an image and its intersection and union with three interpolations ($\rho = \{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\}$) is shown in Fig. 5.

After an interpolated mask is generated based on ρ , the loss is calculated using a soft dice score:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_i \hat{y}_i y_i + s}{\sum_i \hat{y}_i + \sum_i y_i + s}, \quad (1)$$

where real value $\hat{y}_i \in [0, 1]$ is the i th output of the last network layer passed through a sigmoid nonlinearity and

$y_i \in \{0, 1\}$ is the corresponding binary label. A smoothing parameter $s = 1$ [12] is added to the loss function for regularization purposes.

2.2.3 Hypersphere loss

Other work by Kingma et al. [6] or Cheng et al. [1] aims to shape the feature space into a structure that reflects certain properties of the input data. In contrast, this work aims to shape the feature space such that it reflects the variability of the *output* segmentation, so that this structure can be exploited to learn the variability in the network predictions. In order to achieve this, the output prediction is made dependent on ρ with a proposed loss termed the hypersphere loss (HSL), a special case of the least-squares error. The HSL is defined by:

Fig. 5 An ARGOS image with five ground-truth masks corresponding to $\rho = \{0, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1\}$

$$\text{HSL} = \frac{1}{n} \sum_{i=1}^n (\|\mathbf{z}_i\|_2 - 1)^2, \quad (2)$$

where n is the batch size, i is the batch index and $\|\mathbf{z}_i\|_2$ is the dot product of $\mathbf{z}_i$ with itself. The resulting HSL is minimal when the latent vector $\mathbf{z} \in \mathbb{R}^N$ is a point on an N -dimensional hypersphere $\mathcal{S}$ with unity radius. The quadratic form of HSL is used as it penalizes outliers more compared to, e.g., a sum of absolute distances error. Additionally, the squared error is less computationally expensive to calculate compared to an absolute error. Subsequently, $\mathbf{z}$ is scaled such that $\mathbf{z}_{\text{scaled}} = \mathbf{z}(\chi + \rho)$ with $0 \leq \rho \leq 1$ prior to being used as input to the decoder. The offset $\chi = 0.5$ is included to prevent multiplication by 0, which otherwise would produce a null feature vector while the output is a nonzero mask. The normalized vector $\|\mathbf{z}_{\text{scaled}}\|$ is now directly dependent on the desired output annotation size, where we define that $\rho = 0$ corresponds to the intersection and $\rho = 1$ corresponds to the union. As a secondary benefit, the hypersphere loss essentially acts as a regularizer for the model as it penalizes low-dimensional representations of an input image ($\mathbf{z}$) to lie on $\mathcal{S}$ instead of anywhere in the feature space. The total loss of the network is defined by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Dice}} + 0.5 \cdot \text{HSL}, \quad (3)$$

where $\mathcal{L}_{\text{Dice}}$ enforces spatial overlap of the ground-truth and the predicted segmentations. HSL enforces that latent vector $\mathbf{z}$ is a point on hypersphere $\mathcal{S}$ with unity radius, so it can subsequently be scaled with ρ . The HSL is scaled by a factor of 0.5 as a means to balance the two separate loss functions. This value was determined empirically in the validation phase.

2.2.4 Dynamic leaky ReLU

A second way to make the output dependent on ρ is by introducing the dynamic leaky ReLU (DLReLU). The DLReLU is very similar to the original leaky rectified linear unit (LReLU) first introduced in [11]. Mathematically, the LReLU is defined as

$$y = \begin{cases} x, & \text{if } x \geq 0; \\ \frac{x}{a}, & \text{if } x < 0, \end{cases} \quad (4)$$

where a is a constant within the range $(1, +\infty)$. For a DLReLU, instead of fixed constant a , the negative slope is dependent on size factor ρ , such that

$$y = \begin{cases} x, & \text{if } x \geq 0; \\ 0.5\rho \cdot x, & \text{if } x < 0. \end{cases} \quad (5)$$

It should be noted that the value of ρ can be different per pass through the network. Hence, this dynamically changes

the network based on ρ . Since $0 \leq \rho \leq 1$, the slope of the activation function for negative inputs will be 0 when the output annotation should correspond to the intersection and 0.5 when it corresponds to the union. This means that for large values of ρ , the negative activations are “squashed” less compared to small values of ρ . In this manner, the network can learn that larger negative activations correspond to larger segmentations. Similarly, when ρ is small, the negative activations are also small. The decoder can then learn to use the magnitude of the negative activations to encode the general size of the segmentation.

2.3 Experimental setup

2.3.1 Vanishing shapes: proof of concept

The vanishing shapes data set is introduced for validating a proof of concept. Since an infinite number of unique vanishing shapes images can be generated, we show that the algorithm converges to a point where the output segmentation prediction is dependent on ρ without the possibility of overfitting to the data. Additionally, the general performance of the model is evaluated 1000 new vanishing shapes images with the dice score.

2.3.2 VLE: explainable AI and modeled border ambiguity

In previous work [14], several fully residual U-net models were trained with varying ground truths (different assessors, weighted ground-truth masks). Although the resulting segmentation masks were very similar to their respective ground truths, the prediction masks gave no additional information about ambiguity in the images. In the final segmentation prediction results, the upper boundary was predicted with the same confidence as the lower boundary, although the signal-to-noise ratio is much lower and less well defined [Fig. 2 (left)]. We show that the proposed HEDN not only classifies the more important strict upper boundary of the VLE tissue with a high confidence estimate, it also simultaneously decreases the prediction score in ambiguous regions. This property makes the algorithm more interpretable and explainable compared to, e.g., U-net which, in this example, does not give lower confidence in areas where experts disagree. In conclusion, the HEDN approach learns the location of specific observer variability.

2.3.3 ARGOS: biopsy localization

In the final experiment, biopsy localization performance is evaluated by using a multi-proposal approach facilitated by the HEDN model. In clinical practice, computer-aided biopsy localization has the most patient benefit because

many lesions are missed due to the current sampling protocol. Since the current sampling protocol takes multiple biopsies every few cm as well, the models are evaluated for $i \in \{1, 2, 3\}$ biopsy proposals per image. The centroids of i connected regions with the largest confidence estimates are chosen as biopsy locations. The proposed architecture is evaluated against other standard segmentation methods and evaluated with the following metrics:

$$d_{\text{biopsy}} = \min_{i=1 \dots n} \|p - b_i\|_2, \quad (6)$$

$$\text{inter-flag} = (n_{\text{inter}} * 100) / n_{\text{total}}, \quad (7)$$

$$\text{union-flag} = (n_{\text{union}} * 100) / n_{\text{total}}, \quad (8)$$

where n_{total} is the total number of images. Parameters n_{inter} and n_{union} are the amount of images where the indicated biopsy location is inside the intersection and union, respectively. Finally, d_{biopsy} is defined as the minimum of the distances between the intersection center p and biopsy location proposals b_i . Since p and b_i are coordinates, distance is calculated using the Euclidean norm: $\|p - b_i\|_2$.

For comparison, we train five models in addition to the proposed HEDN architecture until convergence is reached. Four fully residual U-net models (state of the art in many medical image segmentation problems) are trained with the segmentation masks of Assessor 1, Assessor 2, the intersection and the union as ground truths and DICE as a loss function. A naive way to model uncertainty in the network is to generate a *gradient image* from the intersection and the union, where the pixel labels gradually decrease from 1 to 0 in the area between the intersection and the union, thereby attempting to solve the problem with regression. The gradient image model is trained with a mean-squared error loss instead of the dice loss and a gradient image generated from the intersection and union as ground truths. Lastly, in clinical practice, a pseudo-random biopsy protocol (Seattle protocol [15]) is the standard for Barrett's esophagus surveillance endoscopies. In an attempt to simulate this protocol, $i \in \{1, 2, 3\}$ random biopsy locations are proposed per image. This experiment attempts to show that there is a clear benefit compared to the random biopsy approach used in clinical practice (emulating the Seattle protocol).

2.3.4 Implementation details

The algorithm was trained using Adam and AMS-grad with a weight decay of 10^{-5} . A cyclic cosine learning rate scheduler was used to control the learning rate [10]. Additionally, batch normalization was used to regularize the network. For the vanishing shapes and VLE experiments, no data augmentation was used. For the ARGOS experiments, the model was further regularized with data augmentation. Images were randomly rotated with $\theta \in \{0, 90, 180, 270\}$ degrees and random flipped along the x - and y -axis with probability $p = 0.5$. Additionally, random permutations were made to the color, contrast and brightness of the images. Furthermore, images were randomly translated by up to 10% of image width. Both the VLE and the ARGOS experiments were performed with a 60%–20%–20% train–validation–test split. All images from a given patient were always contained in the same split.

3 Results

3.1 Vanishing shapes results

Figure 6 shows inference results for increasing values of ρ for a newly generated vanishing shapes image [Fig. 7 (left)]. These results show the proof of concept of the HEDN architecture as the size of the inference results clearly depends on ρ . The two vanishing shapes at the top right of Fig. 7 (left) are partly overlapping, which makes this particular image more complex. As a result, localization of the two separate shapes is more difficult and a normal U-net model would have trouble with distinguishing the two shapes. With the HEDN approach, the centroids of the two separate shapes are still found, due to ability of the model to generate multiple differently sized segmentation proposals as can be observed from the heatmap in Fig. 7. Several additional examples of segmentation results of the vanishing shapes experiment are shown in Fig. 10 in the “Appendix.”

3.2 VLE results

Figure 8 shows visual results of the hypersphere encoder–decoder network architecture on VLE data. As shown in

Fig. 6 Inference results of the HEDN model with $\rho = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$

Fig. 7 From left to right: vanishing shapes image, corresponding union, corresponding intersection and heatmap generated from inference results from Fig. 6

Fig. 8 From left to right: VLE inference image, corresponding union, corresponding intersection, heatmap generated from inference results with six levels and standard U-net prediction

Fig. 8 (right), a standard U-net prediction contains no information about the assessor uncertainty, whereas the prediction heatmap in Fig. 8 clearly shows a decrease in prediction certainty for deeper layers in the tissue where the signal-to-noise ratio degrades. Additionally, similar to the union and the intersection, the predicted heatmap shows no ambiguity near the top layer of the tissue where the boundary is clearly defined and there is hardly any inter-observer variability. These results highlight that the network does not simply increase the general size of the prediction depending on the value of ρ , similar to what a binary dilation operator normally would perform. Instead, the network learns to increase the size of the prediction based on the modeled uncertainty. Figure 8 clearly shows the decreasing confidence in the heatmap at lower tissue layers. In clinical practice, this model can assist the physician with informed decision making based on the modeled multiple expert opinions. In this way, the

methodology facilitates explainable AI to the clinical users. Several additional examples of segmentation results of the VLE experiment are shown in Fig. 11 in the “Appendix.”

3.3 ARGOS results

Segmentation results of a particularly difficult sample from the ARGOS data set are shown in Fig. 9. Figure 9 (left) shows a dysplastic lesion in an area of the image where lighting cannot sufficiently reach the lesion. In most cases, photographs of lesions are taken with optimal lighting, which introduces an inherent bias in most image-based endoscopy data sets. However, in practice, most lesions are not clearly illuminated at all times during an endoscopic video surveillance. Despite this lighting problem, Fig. 9 shows that the HEDN algorithm detects a lesion at the proper location, while all other models mentioned in Table 1 missed the lesion in this particular case. An

Fig. 9 From left to right: ARGOS inference image, corresponding union, corresponding intersection, heatmap generated from inference results with six levels and standard U-net prediction

Table 1 Results of the ARGOS experiment: for the inter-flag, and union-flag metrics, higher is better

Model	d_{biopsy} (pixels)			Inter-flag (%)			Union-flag (%)		
	1	2	3	1	2	3	1	2	3
N biopsies									
Assessor 1	53.0 (4.6)	36.0 (2.7)	32.0 (2.1)	49.0 (6.2)	60.0 (5.3)	64.0 (4.9)	70.0 (5.5)	83.0 (3.6)	88.0 (2.3)
Assessor 2	51.0 (4.8)	35.0 (3.2)	31.0 (3.1)	49.0 (5.8)	60.0 (6.7)	63.0 (7.4)	70.0 (5.1)	83.0 (3.6)	86.0 (4.1)
Union	50.0 (2.4)	36.0 (2.5)	33.0 (2.1)	49.0 (5.1)	58.0 (5.9)	61.0 (4.7)	72.0 (4.0)	83.0 (2.7)	85.0 (2.9)
Intersection	50.0 (5.0)	34.0 (3.9)	29.0 (2.9)	54.0 (4.5)	67.0 (5.3)	71.0 (4.3)	74.0 (4.3)	86.0 (3.7)	90.0 (3.1)
Gradient mask	48.0 (3.3)	33.0 (2.5)	30.0 (2.2)	57.0 (4.4)	68.0 (4.2)	71.0 (3.8)	76.0 (3.5)	87.0 (2.8)	90.0 (3.1)
Random	95.0 (4.6)	71.0 (2.1)	57.0 (2.4)	13.0 (1.7)	25.0 (4.5)	31.0 (5.8)	25.0 (2.3)	45.0 (3.7)	55.0 (4.1)
HEDN (Ours)	43.0 (4.9)	31.0 (2.8)	28.0 (2.6)	60.0 (4.3)	69.0 (4.0)	71.0 (4.8)	79.0 (4.9)	88.0 (3.3)	89.0 (3.0)

For the d_{biopsy} metric, lower is better. Reported values are the averages over 10 repeats of the same experiment. The standard deviation is reported between brackets. The bold font indicates the best scores per metric

example of this is the prediction obtained by the model trained with Assessor 1, as shown in Fig. 9 (right). The proposed algorithm delineation also detects dysplasia in a region not indicated by experts; however, dysplasia cannot be ruled out in that region from this image alone due to a lack of proven histology. In contrast, this assertion can confidently be made in the section annotated by experts, and it is therefore paramount that the region indicated by experts is not missed by a computer-aided detection algorithm. Several additional examples of segmentation results of the ARGOS experiment are shown in Fig. 12 in the “Appendix.”

Table 1 shows the results of the experiment described in Sect. 2.3.3. Each model was retrained ten times with different random data splits, the mean results of the ten experiments are reported, and the standard deviation of the ten trials is reported between brackets. The proposed HEDN model scores higher or on par on all metrics save for a marginally lower score on the union-flag metric with three biopsy location proposals. Most importantly, the d_{biopsy} metric is consistently approximately four pixels closer to the ideal biopsy location on average compared to other models, which corresponds to 12%. This clearly increases the chance of a biopsy containing cancerous cells. The standard deviation of the reported metrics of our model is lower in most cases, especially compared against the assessor models and the *intersection* model. Only the *gradient mask* model consistently scores better with lower standard deviations compared to the proposed approach. However, the difference in standard deviation between the two approaches is usually smaller or slightly larger compared to the difference in mean. This indicates that the increase in average performance is greater than the increase in variability. Noteworthy is the sizeable increase in performance for all models when more than one biopsy is taken. The performance increase in the inter-flag and union-flag metric increases by at least 9% for all models when two biopsy locations are proposed instead of one.

However, increasing the amount of biopsy locations to three increases the performance only slightly. It should be noted that only dysplastic images have been used for this experiment, which limits the extrapolation of these results to a classification problem that also includes non-dysplastic images. However, a separate classification algorithm can be used prior to localization with HEDN, or the proposed algorithm can be retrained with additional non-dysplastic data (outside of the scope for this pilot experiment). Furthermore, it is noteworthy that the *gradient mask* model outperforms the single ground-truth models, showing that incorporating ground-truth uncertainty during training is beneficial with multiple approaches. Lastly, the random biopsy approach consistently performs worse on all metrics compared to all other models, especially when only a single biopsy is taken. In conclusion, the experiments show that localization with the HEDN model clearly improves compared to other state-of-the-art models.

4 Discussion and conclusion

Many medical imaging data sets contain ground-truth segmentations with a large inter-observer variability. Most state-of-the-art segmentation models do not take this drawback into account and are fully deterministic in nature. To this end, we have proposed the hypersphere encoder-decoder network (HEDN), which represents an architecture for explicitly incorporating multi-assessor intervariability for multiple segmentation proposals into the segmentation model. More specifically, the proposed model $M_{\theta}(Y|X, \rho)$ maps a latent vector to a hypersphere to generate segmentation proposals of varying sizes. We have shown that making the model dependent on parameter ρ effectively facilitates the generation of confidence estimate heatmaps that visualize the underlying clinical model uncertainty along segmentation boundaries. This heatmap with visualized confidence margins is a valuable assisting tool to

physicians for inferring information trained with multiple expert opinions. Additionally, we provide a proof of concept on randomly generated images as well as improved interpretability of results on two real-life data sets. Lastly, we show improved ideal biopsy location performance in patients with dysplastic Barrett's esophagus.

The output of $M_\theta(Y|X, \rho)$ was made dependent on ρ in two ways. First, we proposed dynamic leaky ReLUs as a way of propagating size information through the negative activations of the network. Second, a feature mapping layer was added which scales the length of the low-dimensional feature vector proportional to the value of ρ . This feature mapping is inspired by variational encoder–decoder networks which perform a similar operation to generate new samples representative of the input data instead of segmentation proposals. In future work, variational methods such as in [6] can be explored to sample, e.g., coarseness in addition to size. This facilitates the generation of infinite unique region proposals for even better heatmap generation. Additionally, this method only models the intervariability between assessors. Another interesting research direction would be the intravariability of a single assessor with respect to the entire data set. This would allow for an aleatoric uncertainty estimate without the need of multiple assessors.

As mentioned in Sect. 2.1.3, images were omitted when the expert annotations did not overlap at all. One of the reasons why images without consensus were removed is attributed to the current implementation of the algorithm. Without a nonzero intersection, the interpolation cannot be calculated which is required to compute the loss. However, there is useful information in these samples that is currently not employed. In future work, this limitation can be addressed to allow for these cases as well by, e.g., creating a synthetic intersection or by using a different loss function for these images.

In other medical fields, inter-observer variability is also a relevant problem. For example, Response Evaluation Criteria In Solid Tumors (RECIST) is the standard measurement for tumor extent to evaluate treatment responses in cancer patients. For this reason, RECIST annotations must be accurate to avoid wrong conclusions which could impact the patient severely. In related work by Tang et al.

[21], an algorithm was developed that automatically generates RECIST annotations from image patches containing a tumor. The proposed model has less intervariability compared to the other assessors which shows model stability. However, the intervariability between assessors might contain useful information, which is currently not used. The model proposed in this work could easily be adapted to work with RECIST annotations as well. By incorporating the intervariability during training, the model proposed by Tang et al. could be improved further.

Upon acceptance, the code for hypersphere encoder–decoder networks will be made publicly available.

Acknowledgements We thank Nvidia for supplying a Titan X for this research through their GPU grant program. Additionally, we thank KWF for funding this research.

Funding This work was funded by a joint KWF (Dutch Cancer Society) (Grant No. 15191) STW Grant for the development of technology in oncology.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval For the VLE data set, written informed consent was obtained from all patients and they were included under protocol NTR 6728, registered at <http://www.trialregister.nl>. Official approval was obtained by the Institutional Review Board of the Amsterdam University Medical Centers. In the case of the ARGOS data set, only anonymized imagery was used. The Medical Research Involving Human Subjects Act did not apply to this study. Official approval of this study was therefore waived by the Medical Ethics Review Committees of the participating centers.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix

See Figs. 10, 11 and 12.

Fig. 10 From left to right: additional vanishing shapes images, corresponding unions, corresponding intersections and heatmaps generated from inference results

Fig. 11 From left to right: VLE inference images, corresponding unions, corresponding intersections and heatmaps generated from inference results with six levels

Fig. 12 From left to right ARGOS inference images, corresponding unions, corresponding intersections and heatmap generated from inference results with six levels

References

- Cheng D, Gong Y, Zhou S, Wang J, Zheng N (2016) Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1335–1344
- Gonzalo N, Tearney GJ, Serruys PW, van Soest G, Okamura T, García-García HM, van Geuns RJ, van der Ent M, Ligthart J, Boum BE et al (2010) Second-generation optical coherence tomography in clinical practice. High-speed data acquisition is highly reproducible in patients undergoing percutaneous coronary intervention. *Revista Española de Cardiología (English Edition)* 63(8):893–903
- de Groof J, van der Putten J, van der Sommen F, Zinger S, Curvers W, Bisschops R, Pech O, Meining A, Neuhaus H, Schoon EJ et al (2018) The argos project: first results of the development of a computer aided detection system for Barrett's neoplasia. *Gastroint Endosc* 87(6):AB270
- Guzman-Rivera A, Batra D, Kohli P (2012) Multiple choice learning: learning to produce multiple structured outputs. In: Advances in neural information processing systems, pp 1799–1807
- Kariv R, Plesec TP, Goldblum JR, Bronner M, Oldenburgh M, Rice TW, Falk GW (2009) The Seattle protocol does not more reliably predict the detection of cancer at the time of esophagectomy than a less intensive surveillance protocol. *Clin Gastroenterol Hepatol* 7(6):653–658
- Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint [arXiv:1312.6114](https://arxiv.org/abs/1312.6114)
- Kohl SA, Romera-Paredes B, Meyer C, De Fauw J, Ledsam JR, Maier-Hein KH, Eslami S, Rezende DJ, Ronneberger O (2018) A probabilistic u-net for segmentation of ambiguous images. arXiv preprint [arXiv:1806.05034](https://arxiv.org/abs/1806.05034)
- Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in neural information processing systems, pp 6402–6413
- Lee S, Purushwalkam S, Cogswell M, Crandall D, Batra D (2015) Why m heads are better than one: training a diverse ensemble of deep networks. arXiv preprint [arXiv:1511.06314](https://arxiv.org/abs/1511.06314)
- Loshchilov I, Hutter F (2016) SGDR: Stochastic gradient descent with warm restarts. arXiv preprint [arXiv:1608.03983](https://arxiv.org/abs/1608.03983)
- Maas AL, Hannun AY, Ng AY (2013) Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of ICML, vol 30
- Manning C, Raghavan P, Schütze H (2010) Introduction to information retrieval. *Nat Lang Eng* 16(1):260
- Maurer CR, Qi R, Raghavan V (2003) A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Trans Pattern Anal Mach Intell* 25(2):265–270
- van der Putten J, van der Sommen F, Struyvenberg M, de Groof J, Curvers W, SchoonSchoon E, Bergman J, de With P (2019) Tissue segmentation in volumetric laser endomicroscopy data using FusionNet and a domain-specific loss function. *SPIE Medical Imaging*

15. Reid BJ, Blount PL, Feng Z, Levine DS (2000) Optimizing endoscopic biopsy detection of early cancers in Barrett's high-grade dysplasia. *Am J Gastroenterol* 95(11):3089
16. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, Springer, New York, pp 234–241
17. Rupprecht C, Laina I, DiPietro R, Baust M, Tombari F, Navab N, Hager GD (2017) Learning in an uncertain world: representing ambiguity through multiple hypotheses. In: International conference on computer vision (ICCV)
18. Sharma P, Bergman JJ, Goda K, Kato M, Messmann H, Alsop BR, Gupta N, Vennalaganti P, Hall M, Konda V et al (2016) Development and validation of a classification system to identify high-grade dysplasia and esophageal adenocarcinoma in Barrett's esophagus using narrow-band imaging. *Gastroenterology* 150(3):591–598
19. Sohn K, Lee H, Yan X (2015) Learning structured output representation using deep conditional generative models. In: Advances in neural information processing systems, pp 3483–3491
20. Sørensen J, Klee M, Palshof T, Hansen H (1993) Performance status assessment in cancer patients. an inter-observer variability study. *Br J Cancer* 67(4):773
21. Tang Y, Harrison AP, Bagheri M, Xiao J, Summers RM (2018) Semi-automatic RECIST labeling on CT scans with cascaded convolutional neural networks. In: International conference on medical image computing and computer-assisted intervention. Springer, Berlin, pp 405–413
22. Williamson SR, Rao P, Hes O, Epstein JI, Smith SC, Picken MM, Zhou M, Tretiakova MS, Tickoo SK, Chen YB et al (2018) Challenges in pathologic staging of renal cell carcinoma: a study of interobserver variability among urologic pathologists. *Am J Surg Pathol* 39(9):1242–1249

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.