Uncertainty Quantification in Medical Image Segmentation with Normalizing Flows

Raghavendra Selvan[1(✉)], Frederik Faye[2], Jon Middleton[2], and Akshay Pai[1,2]

[1] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
raghav@di.ku.dk

[2] Cerebriu A/S, Copenhagen, Denmark

Abstract. Medical image segmentation is inherently an ambiguous task due to factors such as partial volumes and variations in anatomical definitions. While in most cases the segmentation uncertainty is around the border of structures of interest, there can also be considerable inter-rater differences. The class of conditional variational autoencoders (cVAE) offers a principled approach to inferring distributions over plausible segmentations that are conditioned on input images. Segmentation uncertainty estimated from samples of such distributions can be more informative than using pixel level probability scores. In this work, we propose a novel conditional generative model that is based on conditional Normalizing Flow (cFlow). The basic idea is to increase the expressivity of the cVAE by introducing a cFlow transformation step after the encoder. This yields improved approximations of the latent posterior distribution, allowing the model to capture richer segmentation variations. With this we show that the quality and diversity of samples obtained from our conditional generative model is enhanced. Performance of our model, which we call *cFlow Net*, is evaluated on two medical imaging datasets demonstrating substantial improvements in both qualitative and quantitative measures when compared to a recent cVAE based model.

Keywords: Segmentation · Uncertainty · Normalizing flow · cVAE · chest CT · Vessels

1 Introduction

Medical image segmentation is inherently an ambiguous task and segmentation methods capable of quantifying uncertainty by inferring distributions over segmentations are therefore of substantial interest to the medical imaging community [7,8,25]. Estimating uncertainty from distributions over segmentations is closer to the clinical settings, than obtaining pixel-wise uncertainty estimates, where *whenever feasible* multiple expert opinions are used to ascertain downstream clinical decisions. Such consensus based decisions not only account

M. Liu et al. (Eds.): MLMI 2020, LNCS 12436, pp. 80–90, 2020.
https://doi.org/10.1007/978-3-030-59861-7_9

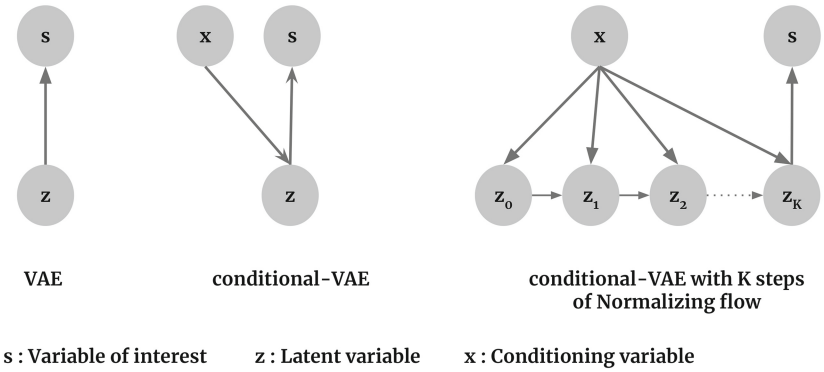

Fig. 1. Graphical model view of VAE and variations to it including the proposed cFlow Net (right)

for the aleatoric (inherent) and epistemic (modeling) uncertainties but also explain the inter-rater variability that is largely inevitable in medical image segmentation.

Remarkable strides in supervised medical image segmentation have been made with deep learning methods [4,20,26]. These methods, however, provide point estimates of segmentations – meaning a single segmentation mask per image – which limits our ability to quantify the uncertainty of said segmentations.

Bayesian deep learning methods offer a natural setting to infer distributions over segmentations. This has been explored to some extent for medical image segmentation in the spirit of Monte Carlo estimation where multiple hypotheses are explored by predicting segmentation masks with different dropout rates [5] or with an ensemble of models [21]. These methods can output a fixed number of samples with pixel level probability scores which can be a limitation.

Conditional variational autoencoders (cVAE) [23] belong to the class of conditional generative models. cVAEs can be used to obtain an unlimited number of predictions by sampling from a latent space conditioned on the input images. This model was adapted for medical image segmentation as the probabilistic U-Net (Prob. U-Net) [13] demonstrating the possibility of generating large number of plausible segmentations. The Prob. U-Net model fuses an additional channel obtained from the latent space to the final layer (at the highest resolution) of U-Net to obtain a variety of albeit less diverse and blurry segmentations when compared to the raters [3]. Quite recently, two models have sought to improve upon the Prob. U-Net [3,14]. Both these methods hypothesize that the blurriness and lack of diversity observed in samples obtained from Prob. U-Net is caused by the use of a single latent variable at the highest resolution. They propose using latent variables in a hierarchical fashion operating at different resolutions to make the model more expressive and demonstrate this to be helpful.

In this work, we focus on obtaining expressive latent representations that can yield diverse segmentations within the cVAE setting. While we agree with [3,14]

that Prob. U-Net suffers from using the latent representation at a single resolution, we argue that it can be alleviated by using a more expressive latent posterior distribution instead of using multiple latent variables in a hierarchical setting. This arises from the fact that all cVAE type models, including Prob. U-Net, use an axis aligned Gaussian as the latent distribution which can be limiting when approximating a complex latent posterior distribution [11,19]. We propose to improve the approximation of the latent posterior distribution with conditional normalizing flows (cFlow) which can yield arbitrarily complex distributions starting from simple ones. We demonstrate that these complex distributions operating at a single resolution are able to capture richer diversity of realistic segmentations. We propose a novel conditional normalizing flow model – cFlow Net – and demonstrate the use of two types of normalizing flow transformations: Planar flows [19] and Generative Flows [10]. We evaluate the method on two medical imaging datasets: LIDC-IDRI [2] for detecting lesions in lungs from chest CT and for detecting retina blood vessels from on a new Retinal Vessel dataset created from three older datasets [6,16,24]. We compare the performance of our model with deterministic U-Net [20] and Prob. U-Net demonstrating significant improvements on both quantitative (generalized energy distance and dice) and qualitative measures.

2 Background and Problem Formulation

Image segmentation tasks can be formulated in a conditional generative model setting with the objective of estimating the conditional distribution $p(\mathbf{s}|\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{s} \in \{0,1\}^{H\times W}$ are the input images and corresponding binary segmentations, respectively, of dimensions H, W with C channels. This has been approached using the conditional VAE formulation where the conditional distribution $p(\mathbf{s}|\mathbf{x})$ is approximated by introducing dependency on a d-dimensional latent variable $\mathbf{z} \in \mathbb{R}^d$ [13,23], as shown in Fig. 1 (center).

The cVAE objective minimizes the KL divergence between the true latent posterior distribution $p(\mathbf{z}|\mathbf{s},\mathbf{x})$ and its variational approximation $q(\mathbf{z}|\mathbf{s},\mathbf{x})$ resulting in an objective of the form [23]:

$$\mathcal{L}_{\text{cVAE}} = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{s},\mathbf{x})}\big[\log p_\theta(\mathbf{s}|\mathbf{z},\mathbf{x})\big] + \text{KL}\big[q_\phi(\mathbf{z}|\mathbf{s},\mathbf{x})||p_\psi(\mathbf{z}|\mathbf{x})\big] \tag{1}$$

The first term is the expected conditional log-likelihood (CLL) under the variational distribution $q_\phi(\mathbf{z}|\mathbf{s},\mathbf{x})$ and the second term can be seen as the regularization forcing the posterior distribution to match the conditional prior distribution $p_\psi(\mathbf{z}|\mathbf{x})$. In cVAE, the posterior density is modeled as a diagonal Gaussian density for tractability reasons: $q_\phi(\mathbf{z}|\mathbf{s},\mathbf{x}) = N(\mathbf{z};\boldsymbol{\mu}_\phi,\boldsymbol{\sigma}^2_\phi)$. The mean $\boldsymbol{\mu}_\phi$ and variance $\boldsymbol{\sigma}^2_\phi$ are predicted using an encoder network parameterized by ϕ. The decoder and prior networks are parameterized by θ and ψ, respectively.

Normalizing flows can be used to transform simple base distributions into complex ones using a sequence of bijective transformations (the flow chain) with easy to compute Jacobians [17,19]. They basically extend the change of variable rule to transform a base distribution into a target distribution in K successive

steps. Normalizing flows can transform a simple base distribution $p(\mathbf{z}_0)$ into an arbitrarily complex target distribution, $p(\mathbf{z}_K)$, by composing complex flow transformations with simpler flow steps [17].

Consider one such bijective transformation T composed of K steps:

$$T = T_K \circ T_{K-1} \circ \cdots T_1. \tag{2}$$

Forward evaluation of this flow chain, transforming $\mathbf{z}_0 \rightarrow \mathbf{z}_K$, can be written as:

$$\mathbf{z}_k = T_k(\mathbf{z}_{k-1}) \quad \text{for } k = 1 \ldots K \tag{3}$$

where $\mathbf{z}_0$ is distributed according to the base distribution $p(\mathbf{z}_0)$.
Reverse evaluation of the flow chain, transforming $\mathbf{z}_K \rightarrow \mathbf{z}_0$, can be written as:

$$\mathbf{z}_{k-1} = T_k^{-1}(\mathbf{z}_k) \quad \text{for } k = K \ldots 1. \tag{4}$$

The transformed distribution, $p(\mathbf{z}_K)$, is obtained from the base distribution, $p(\mathbf{z}_0)$, adjusted by the inverse absolute Jacobian determinant of the flow transformation. For a single flow step k:

$$p(\mathbf{z}_k) = p(\mathbf{z}_{k-1}) \left| \frac{\partial T_k(\mathbf{z}_{k-1})}{\partial \mathbf{z}_{k-1}} \right|^{-1} = p(\mathbf{z}_{k-1}) \left| J_{T_k}(\mathbf{z}_{k-1}) \right|^{-1} \tag{5}$$

where $J_{T_k}(\mathbf{z}_{k-1})$ denotes the Jacobian determinant. The complete transformation using the full flow chain in log domain is given by

$$\log p(\mathbf{z}_K) = \log p(\mathbf{z}_0) + \log \left| \prod_{k=1}^{K} J_{T_k}^{-1}(\mathbf{z}_{k-1}) \right| = \log p(\mathbf{z}_0) - \sum_{k=1}^{K} \log \left| J_{T_k}(\mathbf{z}_{k-1}) \right|, \tag{6}$$

where the last equality follows from $\log \left| J_{T_k}^{-1} \right| = \log |J_{T_k}|^{-1} = -\log |J_{T_k}|$.

3 Methods

When using cVAE-like models for medical image segmentation tasks, it is assumed that the diversity of segmentations is captured with the latent posterior distribution. However, using a simple distribution such as an axis-aligned Gaussian to approximate the latent posterior distribution can be too restrictive and might not be sufficiently expressive to capture richer variations. This is noticeable in the Prob. U-Net model [13] where the segmentations are blurry and lack diversity [3,14]. It is in this context that normalizing flows can be used to improve the flexibility of the approximate posterior density to capture a richer diversity of high quality segmentations.

If we denote the approximate posterior density output by the encoder network as the base distribution, $q(\mathbf{z}_0|\mathbf{s}, \mathbf{x})$, using the latent variable $\mathbf{z}_0$, then using the idea of normalizing flows in Sect. 2 can yield more expressive posterior densities. If the base distribution is transformed using a flow chain of K steps according

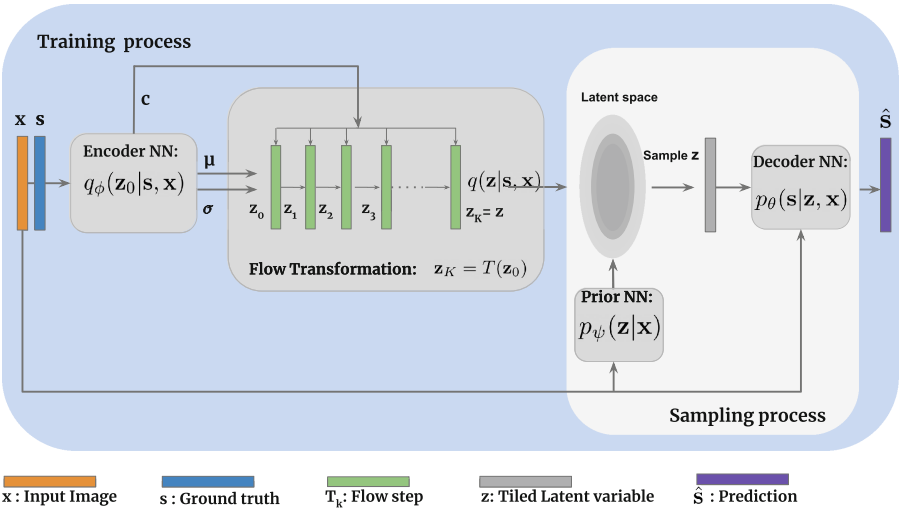

Fig. 2. Proposed cFlow Net model. (Training) The training process takes the reference segmentations $\mathbf{s}$ and the image data $\mathbf{x}$ as input to the encoder, which predicts the mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ of the base distribution along with the context vector $\mathbf{c}$ for the flow transformation. The flow transformation block transforms the base distribution, $q_\phi(\mathbf{z}_0|\mathbf{s},\mathbf{x})$ to an approximation of the target posterior distribution $q(\mathbf{z}|\mathbf{s},\mathbf{x})$ in K steps. The latent space is jointly learned by minimizing the KL divergence between the transformed posterior distribution $q(\mathbf{z}|\mathbf{s},\mathbf{x})$ and the conditional prior $p_\psi(\mathbf{z}|\mathbf{x})$. (Sampling) The sampling process involves obtaining samples from the conditional prior which is used with the input image together to be decoded in the decoder $p_\theta(\mathbf{s}|\mathbf{z},\mathbf{x})$ to obtain the segmentation $\hat{\mathbf{s}}$. After training the model, only the sampling part of the network is used for inference.

to Eq. (2), then the transformed distribution after K steps with $\mathbf{z} = \mathbf{z}_K$ can be written using Eq. (6) as:

$$\log q(\mathbf{z}|\mathbf{s},\mathbf{x}) = \log q(\mathbf{z}_K|\mathbf{s},\mathbf{x}) = \log q(\mathbf{z}_0|\mathbf{s},\mathbf{x}) - \sum_{k=1}^{K} \log \left| J_{T_k}(\mathbf{z}_{k-1}|\mathbf{x}) \right|. \quad (7)$$

It can be shown that the modified objective for the conditional flow-based model becomes (see Sect. 6.1 in the online supplementary material [22]):

$$\mathcal{L}_{\text{cFlow}} = -\mathbb{E}_{q_\phi(\mathbf{z}_0|\mathbf{s},\mathbf{x})}\Big[\log p_\theta(\mathbf{s}|\mathbf{z}_K,\mathbf{x})\Big] + \mathrm{KL}\Big[q_\phi(\mathbf{z}_0|\mathbf{s},\mathbf{x})||p_\psi(\mathbf{z}_K|\mathbf{x})\Big] - \mathbb{E}_{q_\phi(\mathbf{z}_0|\mathbf{s},\mathbf{x})}\Big[\sum_{k=1}^{K} \log \left| J_{T_k}(\mathbf{z}_{k-1}|\mathbf{x}) \right|\Big]. \quad (8)$$

Note that the expectation is with respect to the *base* distribution of the normalizing flow $q_\phi(\mathbf{z}_0|\mathbf{s},\mathbf{x})$. The KL divergence is similar to the term for cVAE in Eq. 1 except for an additional term due to the log determinant of the Jacobian terms in Eq. (7).

Planar Flows: In this work we use planar flows introduced in [19] modified to be conditioned on the input image $\mathbf{x}$ with each step of the flow:

$$\mathbf{u}_k, \mathbf{w}_k, b_k = f_k(\mathbf{x}) \tag{9}$$

$$T(\mathbf{z}_k|\mathbf{x}) = \mathbf{z}_{k-1} + \mathbf{u}_k h(\mathbf{w}_k^T \mathbf{z}_{k-1} + b_k) \tag{10}$$

where $\{\mathbf{u}_k, \mathbf{w}_k \in \mathbb{R}^L, b_k \in \mathbb{R}\}$ are learnable parameters predicted by a conditioning neural network $f_k(\cdot)$ similar to the conditioning network used in [15], L is the dimensionality of the latent space and $h(\cdot)$ is an element-wise non-linearity such as *tanh* with derivative $h'(\cdot)$. The Jacobian determinant for the planar flow step T_k is given by

$$\left|J_{T_k}(\mathbf{z}_{k-1}|\mathbf{x})\right| = \left|1 + \mathbf{u}_k^T \psi_k(\mathbf{z}_{k-1})\right| \quad \text{where } \psi_k(\mathbf{z}_{k-1}) = h'(\mathbf{w}_k^T \mathbf{z}_{k-1} + b_k)\mathbf{w}_k. \tag{11}$$

The conditioning on the flow chain is introduced through the context vector $\mathbf{c}$ which is dependent on $\mathbf{x}$. The context vector $\mathbf{c} \in \mathbb{R}^H$ of dimension H is also predicted by the encoder network. The proposed cFlow Net model is visualized in Fig. 2.

Note that at inference, to sample multiple segmentations only the *Sampling Process* part of the model is used. Given an image $\mathbf{x}$, the prior network can be used to obtain multiple latent variable samples $\mathbf{z}$ which are then decoded by the decoder network to output multiple segmentations for the input image.

4 Experiments and Results

4.1 Data

All experiments are performed on two publicly available datasets. Both datasets comprise labels from at least two raters used to quantify the performance of all models. We use a training-validation-test split of 60:20:20 for both datasets.

LIDC-IDRI Dataset: The LIDC-IDRI dataset consists of 1018 thoracic CT scans with four raters annotating the lesions in them [2]. We use patches of size 128×128 centered on lesions similar to the procedures followed in [3,13] to obtain $15,096$ patches in total. The preprocessed data is obtained from [12].

Retinal Vessel Dataset: As a secondary dataset we create a new dataset derived from three older retinal vessel segmentation datasets: DRIVE [24], STARE [6] and CHASE [16]. Each of these datasets has a subset of images with labels from two raters. We collected images with two raters from these three datasets, extracted retinal masks when there were none and resized them such that all images are of height 512 px. This yields 68 images of which 20 are of size 620×512 px and the remaining 48 are 512×512 px. All images have vessel annotations from two raters. (Figure 4 in online supplementary material [22]).

Table 1. Performance comparison of all models. Higher is better for Dice and lower is better for -CLL and d^2_{GED}. Significant differences are shown in bold.

Models	LIDC dataset					Retina dataset				
	All raters		Single rater			All raters		Single rater		
	-CLL	d^2_{GED}	-CLL	d^2_{GED}	Dice	-CLL ($\times 10^3$)	d^2_{GED}	-CLL ($\times 10^3$)	d^2_{GED}	Dice
Det.U-Net [20]	–	–	–	–	**0.727**	–	–	–	–	0.624
Prob.U-Net [13]	52.1	0.279	238.9	0.579	0.698	4.738	0.905	4.495	0.946	0.616
cFlow Net (Planar)	**47.3**	**0.204**	**89.0**	**0.288**	0.713	**4.436**	**0.884**	**4.482**	0.877	0.632
cFlow Net (Glow)	49.2	0.302	217.0	0.547	0.704	4.482	0.901	4.488	0.878	0.620

4.2 Experiments and Results

The proposed cFlow Net model is compared with the probabilistic U-Net [13], and additionally with the deterministic U-Net [20] for the single rater setting. Other than the cFlow Net model described in Sect. 3 with planar flows [19], we additionally report the cFlow model with conditional generative flow model which uses the Glow transformation steps [10,15] (Sect. 6.2 in the online supplementary material [22]).

Performance of the models in the multiple annotator setting is evaluated based on the generalized energy distance (d^2_{GED}) which captures the diversity of samples obtained from the generative models when compared to the annotators. It is given by

$$d^2_{\text{GED}}(P_R, P_M) = 2\mathbb{E}\Big[d(\mathbf{s}, \hat{\mathbf{s}})\Big] - \mathbb{E}\Big[d(\mathbf{s}, \mathbf{s}')\Big] - \mathbb{E}\Big[d(\hat{\mathbf{s}}, \hat{\mathbf{s}}')\Big], \tag{12}$$

where $\mathbf{s}, \mathbf{s}'$ are samples from the ground truth distribution, P_R, comprising different raters, $\hat{\mathbf{s}}, \hat{\mathbf{s}}'$ are samples from the generative distribution, P_M, learned by the model and $d(\cdot)$ is 1-IoU (intersection-over-union) measure. Additionally, we report the negative conditional log likelihood (-CLL $= -\log p(\mathbf{s}|\mathbf{x})$) approximated with 128 samples (Sect. 6.3 in the online supplementary material [22]) and the dice accuracy for the single rater settings.

Both variants of the cFlow Net models use $K = 4$ flow steps. The *decoder* network in the cFlow Net and Prob. U-Net was a deterministic U-Net with 4 resolutions identical to the ones used in [13]. Architectures of both *encoder* and *prior* networks were similar to the encoding path of the decoder network. In addition to predicting the mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}^2$, the encoder network in the cFlow Net model outputs a context vector $\mathbf{c}$ of dimension $H = 128$ which is input to the *flow transformation* block as illustrated in Fig. 2. The conditioning network $f_k(\cdot)$ is a three layered multi-layer perceptron (MLP) with 8 hidden units. Latent space dimension of $L = 6$ was used for the Prob. U-Net and the cFlow Net models. All the models were trained using a batch size of 96 and a learning rate of 10^{-4} with the Adam optimizer [9]. The models were trained for a maximum of 300 epochs and training convergence was assumed when there was no improvement in validation loss for 20 epochs. Models with the best validation

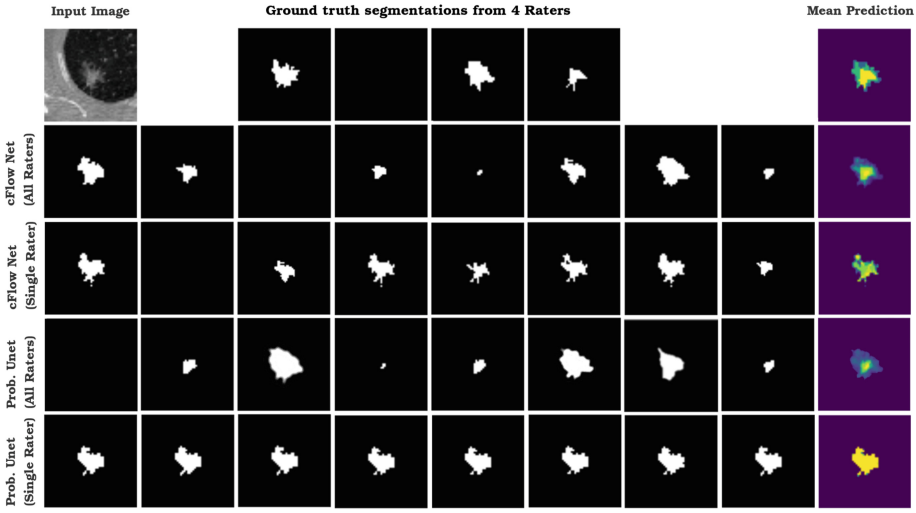

Fig. 3. Qualitative results showing the segmentation diversity of the cFlow Net model and Prob. U-Net for one scan from LIDC-IDRI test set. First row shows the input image, segmentation masks from the four raters; Rows 2 and 3 are samples from cFlow Net model when trained with all and a single (first) rater; Rows 4 and 5 show samples from the Prob. U-Net model for all and single rater setting. Mean prediction over all samples are shown in the last column (brigher regions correspond to higher probability).

loss was used to evaluate the performance on test set reported in Table 1. The experiments were run using PyTorch [18] on a single Tesla K80 GPU with 12GB memory. The computation time for both variants of the cFlow Net models on LIDC dataset was 250s, and about 30s on the Retinal Vessel dataset per training epoch. The average CO_2 footprint of *d*eveloping and training the baseline and proposed models is estimated to be 22.3 kg or equivalently about 180 km traveled by a car, measured using Carbontracker [1].

4.3 Results and Discussion

Performance of all the models on test set of both the datasets are reported in Table 1. Within each dataset we report the performance when compared to *All Raters* and a *Single Rater*. Statistically significant improvement in performance (based on paired sample t-tests with $p < 0.05$) when compared to other models are highlighted in bold.

The proposed cFlow Net (Planar) model is consistently better than the baseline Prob. U-Net model on the LIDC dataset in d^2_{GED} and -CLL measures. The performance of the cFlow Net (Planar) model in the *Single Rater* setting shows a large improvement when compared to Prob. U-Net model. This is also demonstrated in Fig. 3 seen as more realistic and diverse samples generated only by training only on a single (the first) rater. There is a small reduction in perfor-

mance of all the conditional generative models when compared to the Det. U-Net model in dice accuracy.

The significant improvements in d^2_{GED} for the cFlow Net models reported in Table 1 are also reflected qualitatively in the samples shown in Fig. 3. Samples from cFlow Net (row 2) are not only able to capture the variations amongst all four raters (row 1) but the remainder samples appear plausible. When trained with a single rater (row 3), the cFlow Net model is still able to capture a richer diversity of segmentations. As annotations are available from only a single rater in majority of applications, this behaviour of the cFlow Net of being able to capture diverse segmentations from single rater is desirable. This is in contrast with the samples from Prob. U-Net even when trained with all raters (row 4), where the samples appear blurry and are unable to reflect the diversity of the four raters. This lack of diversity becomes more pronounced when trained with a single rater, as the Prob. U-Net model outputs almost identical looking samples (row 5).

In the last column of Fig. 3 we also show the mean prediction obtained from samples of each model (brighter regions have higher probability). The mean predictions from the cFlow Net model trained on a single rater could be more informative than the mean prediction from Prob. U-Net trained on a single rater. This further strengthens our argument that improving the approximation to the latent posterior distribution with conditional normalizing flows helps capture meaningful uncertainty with the possibility of sampling unlimited number of diverse segmentations.

A similar trend is also observed with the Retinal Vessel dataset. This is a far more challenging dataset as the images are acquired differently and the quality of annotations vary between the six annotators. This is captured as higher d^2_{GED} and -CLL across all models. Even within this setting, the cFlow Net models fare better than the Prob. U-Net model in both the single and multiple rater experiments. There was no significant difference in dice accuracy between any of the methods indicating the stochastic generative components of the proposed models do not affect segmentation accuracy.

5 Conclusion

We proposed a novel conditional generative model based on conditional normalizing flows to quantify uncertainty in segmentations. The use of cFlow steps improved the approximation of the latent posterior distribution, captured in the smaller negative conditional log likelihood values and also manifested in the diversity of samples. The primary contribution in this work is the incorporation of conditional normalizing flows for handling high dimensional data such as medical images. The *flow transformation* block is modular and can be easily replaced with any suitable normalizing flow providing access to a rich class of improved conditional generative models [17]. We demonstrated this feature of cFlow Net with two types of normalizing flow transformations: Planar [19] and Glow [10] with promising performance.

Acknowledgements. We thank Oswin Krause and the Medical Image Analysis group at DIKU for fruitful discussions and valuable feedback.

References

1. Anthony, L.F.W., Kanding, B., Selvan, R.: Carbontracker: tracking and predicting the carbon footprint of training deep learning models. In: ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems (2020). https://arxiv.org/abs/2007.03051
2. Armato III, S.G., et al.: Lung image database consortium: developing a resource for the medical imaging research community. Radiology **232**(3), 739–748 (2004)
3. Baumgartner, C.F., et al.: PHiSeg: capturing uncertainty in medical image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 119–127. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_14
4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
5. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059 (2016)
6. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Trans. Med. imaging **19**(3), 203–210 (2000)
7. Jensen, M.H., Jørgensen, D.R., Jalaboi, R., Hansen, M.E., Olsen, M.A.: Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 540–548. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_59
8. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems, pp. 5574–5584 (2017)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Kingma, D.P., Dhariwal, P.: Glow: generative flow with invertible 1x1 convolutions. In: Advances in Neural Information Processing Systems, pp. 10215–10224 (2018)
11. Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., Welling, M.: Improved variational inference with inverse autoregressive flow. In: Advances in Neural Information Processing Systems, pp. 4743–4751 (2016)
12. Knegt, S.: A Probabilistic U-Net for segmentation of ambiguous images implemented in PyTorch (2018). https://github.com/stefanknegt/Probabilistic-Unet-Pytorch
13. Kohl, S., et al.: A probabilistic u-net for segmentation of ambiguous images. In: Advances in Neural Information Processing Systems, pp. 6965–6975 (2018)
14. Kohl, S.A., et al.: A hierarchical probabilistic u-net for modeling multi-scale ambiguities. In: Workshop on Medical Imaging Meets NeurIPS (2019)
15. Lu, Y., Huang, B.: Structured output learning with conditional generative flows. In: ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models (2019)

16. Owen, C.G., et al.: Retinal arteriolar tortuosity and cardiovascular risk factors in a multi-ethnic population study of 10-year-old children; the child heart and health study in england (chase). Arteriosclerosis Thrombosis Vasc. Biol. **31**(8), 1933–1938 (2011)
17. Papamakarios, G., Nalisnick, E., Rezende, D.J., Mohamed, S., Lakshminarayanan, B.: Normalizing flows for probabilistic modeling and inference. arXiv preprint arXiv:1912.02762 (2019)
18. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp. 8024–8035 (2019)
19. Rezende, D.J., Mohamed, S.: Variational inference with normalizing flows. arXiv preprint arXiv:1505.05770 (2015)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Rupprecht, C., et al.: Learning in an uncertain world: representing ambiguity through multiple hypotheses. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3591–3600 (2017)
22. Selvan, R., Faye, F., Middleton, J., Pai, A.: Uncertainty quantification in medical image segmentation with Normalizing Flows (Supplementary material). arXiv preprint arXiv:2006.02683 (2020)
23. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems, pp. 3483–3491 (2015)
24. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging **23**(4), 501–509 (2004)
25. Wilson, R., Spann, M.: Image Segmentation and Uncertainty. Wiley, Hoboken (1988)
26. Zhou, T., Ruan, S., Canu, S.: A review: deep learning for medical image segmentation using multi-modality fusion. Array 100004 (2019)