Exploring the Relationship Between Segmentation Uncertainty, Segmentation Performance and Inter-observer Variability with Probabilistic Networks

Elisa Chotzoglou(✉) and Bernhard Kainz

Department of Computing, BioMedIA, Imperial College London, London, UK
{e.chotzoglou16,b.kainz}@imperial.ac.uk

Abstract. Medical image segmentation is an essential tool for clinical decision making and treatment planning. Automation of this process led to significant improvements in diagnostics and patient care, especially after recent breakthroughs that have been triggered by deep learning. However, when integrating automatic tools into patient care, it is crucial to understand their limitations and to have means to assess their confidence for individual cases. Aleatoric and epistemic uncertainties have been subject of recent research. Methods have been developed to calculate these quantities automatically during segmentation inference. However, it is still unclear how much human factors affect these metrics. Varying image quality and different levels of human annotator expertise are an integral part of aleatoric uncertainty. It is unknown how much this variability affects uncertainty in the final segmentation. Thus, in this work we explore potential links between deep network segmentation uncertainties with inter-observer variance and segmentation performance. We show how the area of disagreement between different ground-truth annotators can be developed into model confidence metrics and evaluate them on the LIDC-IDRI dataset, which contains multiple expert annotations for each subject. Our results indicate that a probabilistic 3D U-Net and a 3D U-Net using Monte-Carlo dropout during inference both show a similar correlation between our segmentation uncertainty metrics, segmentation performance and human expert variability.

1 Introduction

Segmentation, *i.e.*, delineation of anatomical structures in 2D/3D, is a core necessity in medical imaging analysis. In most cases, segmentation is carried out manually by an expert. It is well known that manual segmentation suffers from inter-observer variability and that segmentation quality is influenced by factors such as fatigue, different domain knowledge, level of expertise, and image resolution. As a result, manual segmentations contain aleatoric uncertainty and can thus be ambiguous for diagnosis or confusing for supervised learning methods. Nevertheless, annotator confidence can be an important source of information for

L. Zhou et al. (Eds.): LABELS 2019/HAL-MICCAI 2019/CuRIOUS 2019, LNCS 11851, pp. 51–60, 2019.
https://doi.org/10.1007/978-3-030-33642-4_6

clinical decision making. Varying annotator confidence can be a trigger for additional imaging tests and an indicator for quality control and treatment options. Confidence is an important factor to weigh individual test result but it is only qualitatively assessed in the clinical practice. Recent successes of deep learning for image segmentation [2,8] promise to reduce clinical annotation workload. Currently, the majority of these methods lack the ability to communicate annotator confidence.

Quantitative assessment of uncertainties is key to guarantee quality of care, increases trust and can have great impact on therapeutic decisions. Thus, in this work we explore whether human inter-observer variability can be correlated with the distribution of two different probabilistic neural networks and investigate the impact of this variability on the estimation of segmentation uncertainty and segmentation performance. To achieve this, we

(1) discuss an extension of a probabilistic U-Net [10] to 3D,
(2) compare the properties of a 3D probabilistic U-Net with a Monte-Carlo dropout extension of a standard 3D U-Net [2] on the proof-of-concept task of lung nodule segmentation,
(3) examine and present both qualitatively and quantitatively at which extent automatically predicted confidence and uncertainty metrics, disagreement aware metric (which is proposed) and segmentation performance metrics are correlated.

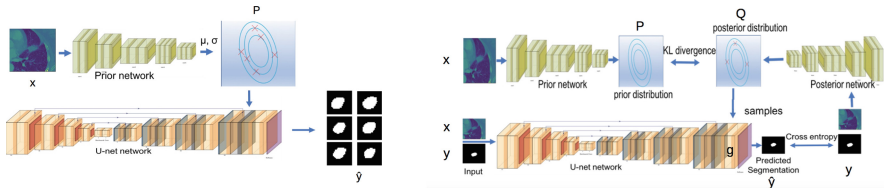
Related Work: Estimation of uncertainty in the medical imaging domain has been attempted in works such as [11,12,14]. In [14] authors use Monte Carlo samples from the posterior distribution of a Bayesian fully Convolutional neural network which are derived using dropout at test phase. Based on these samples, they compute structure-wise and voxel-wise uncertainties metrics, which as they prove, are highly correlated with segmentation accuracy. Application field is infant brain segmentation. In another work [12] Monte Carlo dropout is used for uncertainty estimation in Multiple Sclerosis lesion detection and segmentation. Four different voxel-wise uncertainties were utilised including prediction variance, Monte Carlo sample variance, predictive Entropy and Mutual Information. As it was proved by the results, filtering based on uncertainty leads to improvement on the lesion detection accuracy. In [11] authors propose a framework to approximate Bayesian inference in deep neural networks by imposing Bernoulli distribution directly on the weights of the deep model. Then Monte Carlo samples from posterior distribution are utilised to compute Mutual Information as metric for uncertainty in CT-organ semantic segmentation. Furthermore, the effect of inter-observer variability for estimation of uncertainty in segmentation is studied in [7]. Authors, in MRI images from brain tumors, explore the impact of different label fusion techniques (e.g. no fusion, staples, union, intersection, majority) in estimation of segmentation uncertainty. As it is proved, there is a link between uncertainty estimation and inter-observer variability. Monte Carlo dropout is also used in this work for estimation of uncertainty (entropy). Finally, an alternative way to produce plausible segmentation hypotheses is proposed in [10] where authors use generative segmentation model, a combination

of U-Net and conditional variational autoencoder, in order to produce plausible segmentation hypotheses (diverse samples) for lung abnormalities segmentation task.

2 Background

Two different probabilistic networks are utilised in our work: a $3D$ probabilistic U-Net (**PUNet**) and a $3D$ U-Net using Monte Carlo Dropout during inference (**DUNet**).

PUNet: We extend a $2D$ probabilistic U-Net [10], which is a combination of a U-Net [2,13] and a conditional variational autoencoder [17] to 3D. The whole architecture consists of three networks, which is shown in Fig. 1.



(a) Training (b) Inference

Fig. 1. PUNet [10] as we use it for our method.

Let x be an input volume, M the segmentation map, $\hat{y}$ the predicted segmentation, y the ground truth segmentation as it is produced by several experts ($n = 4$ for LIDC), C the number of classes and N number of voxels per volume similar as proposed by [10]. The Prior net is conditioned on the input volume x. It computes the distribution over the (low-dimensional) latent space R^K. At inference stage samples that are produced by this distribution are concatenated with the last layer's feature maps of the segmentation network, which produces a segmentation map for each sample. More precisely the prior probability distribution P is modelled as an axis-aligned Gaussian distribution with mean $\mu_{prior}(x; w_{prior}) \in R^K$ and variance $\sigma_{prior}(x; w_{prior}) \in R^K$. To sample T segmentations we apply the network T times to the same input volume. In each iteration a sample $z_t, t = \{1, 2, ..., T\}$ is drawn from the distribution:

$$z \sim P(.|x) = \mathcal{N}(\mu_{prior}(x; w_{prior}), diag(\sigma_{prior}(x; w_{prior}))) \tag{1}$$

Each sample is reshaped to a K-channel feature map with the same shape as the segmentation map. This feature map is concatenated to the last activation map of a U-Net. Then, a segmentation map, which corresponds to sample t, is produced by $M_t = f(g(x, w), z_i, \psi)$ where w is the U-Net parameters and ψ weights of the last layer of U-Net.

The posterior net is conditioned on the volume x as well as the ground truth y. It learns to recognize (embeds) segmentation variants $\mu_{post}(x, y; \nu) \in R^K$ with some uncertainty $\sigma_{post}(x, y; \nu) \in R^K$ in the low dimensional latent space. The output is denoted as posterior distribution Q. A sample z from this distribution

$$z \sim Q(.|x, y) = \mathcal{N}(\mu_{post}(x, y; \nu), diag(\sigma_{post}(x, y; \nu)) \tag{2}$$

combined with the activation map of the U-Net will result in a predicted segmentation $\hat{y}$.

The loss function is composed by two terms. The first is the cross entropy loss $E_{z\sim Q(.|y,x)}[-\log P_c(y|M(x, z))]$, which penalizes the difference between the ground truth and the segmentation map. The second one is the Kullback-Leibler (KL) divergence $D_{KL}(Q(z|y, x)||P(z|x))$ which penalizes differences between the posterior distribution Q and the prior distribution P. Both terms are combined as a weighted sum with a weighting factor β as proposed by [10]. Thus, the total loss function is defined as:

$$L(y, x) = E_{z\sim Q(.|y,x)}[-\log P_c(y|M(x, z))] + \beta * D_{KL}(Q(z|y, x)||P(z|x)) \tag{3}$$

In our experiments we use $\beta = 0.2$. Differences between training and inference are outlined in Fig. 1.

DUNet: We utilise a U-Net where dropout layers are activated during inference. During test phase, dropout is similar to Bayesian approximation [4]. In this way, we can take Monte Carlo samples over the posterior distribution $p(w|x, y)$ of the models' weight w and volume x and labels y. Cross entropy between ground truth and predicted segmentation is utilised as loss function.

3 Method

To produce plausible segmentation samples, we utilise PUNet and DUNet. In order to exploit volumetric information, $3D$ versions of the above models are trained using $3D$ convolutions. The U-Nets consist of 3 layers. Each layer consists of $3D$ convolution blocks followed by Rectified Linear Unit (ReLU) activation, batch normalization and max pooling. Filter size is $3\times3\times3$. We start the number of feature maps at 32 and double it after each block. For the prior net as well as for the posterior net in the PUNet, we utilize the encoder part of the U-Net. We train the networks using exponential decay learning rate and the Adam optimizer. For the DUNet, dropout is used after each layer in the encoding part of U-Net. We use a dropout probability of 0.2. We generate an equal number of samples T for both $3D$ networks. All networks are implemented in Python using Tensorflow, on a workstation with NVIDIA Titan X GPU.

In order to estimate model uncertainty we compute two uncertainty scores: Z_{var} and Z_S using variance [9,16] and predictive entropy [5] of samples respectively.

We define mean variance across all classes C as:

$$\sigma^2(x^*) = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{T}\sum_{t=1}^{T}(p_t(y=c|x^*,w) - \hat{p}(y=c|x^*,w))^2, \tag{4}$$

where $\hat{p}(y|x^*,w)$ is the average of softmax probabilities of T samples for each $c \in [1,...,C]$ and p_t the output of the network for sample t. Subsequently we define Z_{var} as

$$Z_{var} = \frac{1}{N}\sum_{v=1}^{N}\sigma^2(x^*(v)), \tag{5}$$

and predictive entropy S as

$$S(x^*) = -\sum_{c=1}^{C}\hat{p}(y=c|x^*,w) \times \log(\hat{p}(y=c|x^*,w)). \tag{6}$$

Thus, for each subject x^*, Z_S is computed as:

$$Z_S = \frac{1}{N}\sum_{v=1}^{N}S(x^*(v)) \tag{7}$$

We utilise the Sørensen–Dice coefficient (Dice score) to characterise segmentation performance. To examine possible linear correlation between segmentation performance and model uncertainty, we compute the Pearson correlation coefficient (ρ) between Z_S and Z_{var} and the Dice score.

To investigate the relationship between Z_S and Z_{var} and the variability among human experts we define the area of human disagreement (Γ) as an XOR ($\oplus$) of the different annotations for each subject. For each voxel the $\oplus$ operation will result 1 indicating that at least one annotator disagrees (disagreement) while 0 is used where all annotators agree (agreement). For a fair comparison with Z_S and Z_{var} we utilize the same schemes for deriving quantitative uncertainty: predictive Entropy(S) Eq. 6 and variance σ^2 Eq. 4. For qualitative analysis we imply Mutual Information (MI) and a map of softmax output probabilities for the predominant class (Softmax). MI is defined using Eq. 6 as the entropy of the average of samples minus the mean of the sum of the entropy of each sample, *i.e.*,

$$MI(x^*) = S(x^*) + \sum_{c=1}^{C}\frac{1}{T}\sum_{t=1}^{T}p_t(y=c|x^*,w)\log(p_t(y=c|x^*,w)), \tag{8}$$

where $p_t(y=c|x^*,w)$ is the softmax output of the network for each sample. We then characterise a voxel v of a new sample x^* as certain/uncertain using

$$x^*(v) = \begin{cases}\text{uncertain, if } S(x^*(v)) >= \theta,\\ \text{certain, otherwise,}\end{cases} \tag{9}$$

where θ is a threshold and v a voxel. Alternatively, we can replace $S(x^*)$ in Eq. 9 with variance σ^2 for estimation of uncertainty. We use a threshold since we assume that human perception of uncertainty is more accurate when interpreted binary than continuous. Evidence for this is given in behavioural sciences literature, *e.g.* [3,6].

We compute the ROC curve between True Positive Rate (TPR) and False Positive Rate (FPR) for the binary case. This allows us to correlate model uncertainty with aleatoric expert uncertainty. With Γ and Eq. 9 we define TPR and FPR as

$$\text{TPR} = p(\text{uncertain}|\text{disagreement}) = \frac{p(\text{uncertain}, \text{disagreement})}{p(\text{disagreement})}, \tag{10}$$

and

$$\text{FPR} = p(\text{uncertain}|\text{agreement}) = \frac{p(\text{uncertain}, \text{agreement})}{p(\text{agreement})}. \tag{11}$$

To evaluate segmentation uncertainty with respect to Γ we use disagreement accuracy ($DisAcc$) as metric [11,15]. $DisAcc$ correlates positively with expert variability. It requires the definition of true invalid predictions, TI, as the voxels that are uncertain within in the area of disagreement (uncertain and disagreement) and true non-invalid predicitons, TU, as the voxels that are certain in the area of agreement (certain and agreement). Similarly to conventional accuracy, $DisAcc$ can be written as $DisAcc = \frac{TI+TU}{N}$, normalised by the total number of voxels N.

4 Evaluation and Results

Data. We use the LIDC-IDRI [1,18] dataset for training and testing. This CT dataset contains images of lung nodules and their delineations from four independent expert observers. We resample data to an isotropic volume resolution of $1 \times 1 \times 1\,\text{mm}^3$. We use 700 patients as a training dataset and 175 patients as a test set for performance evaluation. We crop each volume at the center of the nodule position and produce volumes of $128 \times 128 \times 128$. For the evaluation of the method we use the Dice score as a metric of volume overlap.

Correlation of Z_S and Z_{var} and Dice Score: We analyze correlation between Z_{var} and Z_S (Sect. 3) and the actual Dice score in Fig. 2. Dice score is computed between the absolute ground truth (average of 4 annotators) and the predicted segmentation. In Fig. 2(a, b, e, f) we observe linear negative correlation ($p < 0.001$) between Z_{var}, Z_S and the segmentation performance for both networks. Higher negative correlation is observed for DUNet between Z_{var} and Dice score (Fig. 2e, $\rho = -0.75$) and between Z_S and Dice score (Fig. 2f, $\rho = -0.67$).

There are some cases (10 cases) in both methods that produce uncertainty scores that are not representative for the segmentation quality. Although these nodules do not have any special visual characteristics, the model produces high Dice scores with high uncertainty scores. In Fig. 2c, d and g, h the distributions of uncertainty scores are plotted for two different groups of segmentations.

Successful segmentations have been empirically defined as those where the Dice score is $\geq$0.80 and unsuccessful segmentations with Dice scores $\leq$0.65. Thus, a threshold for the uncertainty score, which divides the two groups of segmentations can be defined as the intersection of the two distributions, which is close to 0.25.

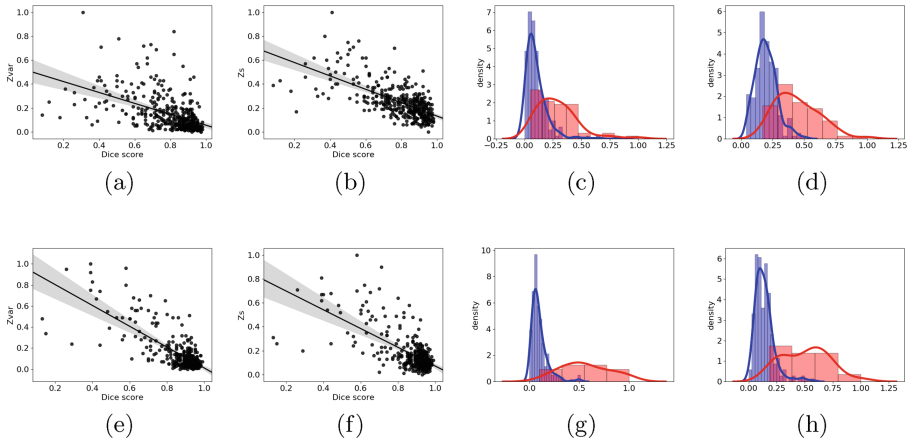
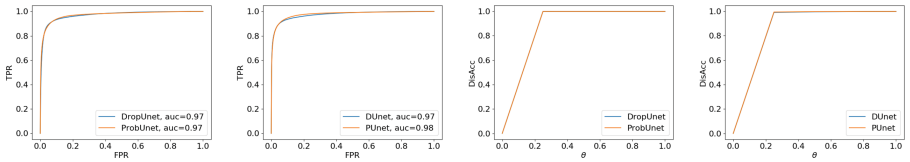


Fig. 2. Scatter plots of correlation between Dice score and uncertainty scores and probability density function (pdf) plots for both networks. Top row: PUNet. Bottom row: DUNet. Correlation between Dice score and Z_{var} and Z_S respectively: (a–b) PUNet and (e–f) for DUNet. Probability density function (PDF) for values of Z_{var} (and Z_S) of samples whose Dice scores is between 0.80 and 0.95 (blue) and the samples that their Dice scores is lower than 0.65 (red). (c–d) for PUNet and (g–h) for DUNet. (Color figure online)

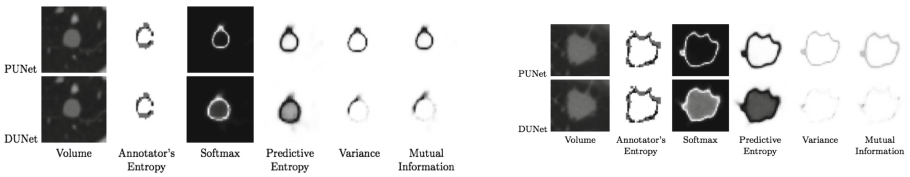
Inter-observer Variability vs. Segmentation Uncertainty: As a naïve baseline we evaluate a convolutional regressor network to predict the annotator variance directly from the volumes. The regressor consists of 5 (convolution-max pooling) layers which are followed by a global average pooling (GAP) layer to predict $Z_{var} \in [0, 1]$ directly. Mean square error between prediction and ground truth of variance among annotators is minimised during training. The performance of this approach is limited with a mean square error of 0.22 ± 0.0012. To evaluate TPR (Eq. 10) and FPR (Eq. 11) we compute ROC curves for each network as shown in Fig. 3 and evaluate *DisAcc* for a range of thresholds. The ROC curves of FPR and TPR (a–b) of both networks are quite similar with the best result for predictive entropy (Eq. 6) as uncertainty metric and PUNet with $AUC = 0.98$. Comparing *DisAcc* (c–d) for 5 different thresholds $\theta \in [0, 1]$ for both networks and *DisAcc* reaches 0.99 for $\theta = 0.2$ and then remains stable.

Qualitative Analysis of Inter-observer Variability and Segmentation Uncertainty. Finally, we present a qualitative comparison between segmentation uncertainty and human uncertainty. Here, human uncertainty is expressed as human annotator entropy. We present and compare the result among all the different uncertainty metrics for both networks in Fig. 4.



(a) Variance (b) Entropy (c) Variance (d) Entropy

Fig. 3. ROC curves and *DisAcc* plots using predictive Entropy and σ^2 for both probabilistic networks



(a) (b)

Fig. 4. Uncertainty maps using maximum Softmax ($\max(M)$), Predictive Entropy (Eq. 6), Variance (Eq. 4) and Mutual Information (Eq. 8) using both networks (darker colour, larger value).

5 Discussion

A limitation of our evaluation is that for a few cases the evaluated uncertainty scores are not a representative metric for how good or bad is a segmentation and it is likely dependant on the used data set. Furthermore, the impact of added parameter capacity in the probabilistic U-Net architecture compared to the dropout-only architecture will need to be carefully investigated in future work. Also, uncertainty as perceived by humans might be fundamentally different from model confidence. Although there is evidence that model uncertainty could capture/include also human disagreement area, it is not clear yet at which extend this happens. Finally, the impact of the different label fusion techniques in the estimation of ground truth and segmentation uncertainty will need to be further examined.

6 Conclusion

Using probabilistic 3D segmentation networks, we examine the relationship between segmentation uncertainty and segmentation performance. We explore to which extent human expert inter-observer variability can effect and correlate with model segmentation uncertainty. Our results show that both, a U-Net using MC dropout during inference as well as a 3D probabilistic U-Net architecture can quantitatively correlate the posterior segmentation distribution with true uncertainties. We present results that show a relationship between segmentation

uncertainty and the area of annotator disagreement. Thus, in most cases model segmentation uncertainty indicates also likely human disagreement. The integration of the evaluated metrics into clinical quality control or for example into an active learning framework, where 'uncertain' parts of segmentations will be re-processed by a human, might show benefit in future work.

References

1. Armato III, S.G., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2), 915–931 (2011)
2. 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, Part II. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
3. Fox, C.R., Rottenstreich, Y.: Partition priming in judgment under uncertainty. Psychol. Sci. **14**(3), 195–200 (2003)
4. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning, ICML 2016 (2016)
5. Gal, Y., Islam, R., Ghahramani, Z.: Deep Bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, pp. 1183–1192 (2017)
6. Huang, L., Pashler, H.: Symmetry detection and visual attention: a "binary-map" hypothesis. Vis. Res. **42**(11), 1421–1430 (2002)
7. Jungo, A., et al.: On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018, Part I. LNCS, vol. 11070, pp. 682–690. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_77
8. Kamnitsas, K., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. **36**, 61–78 (2017)
9. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian SegNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In: British Machine Vision Conference 2017, BMVC 2017, London, UK, 4–7 September 2017 (2017)
10. Kohl, S., et al.: A probabilistic U-Net for segmentation of ambiguous images. In: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, Montréal, Canada, 3–8 December 2018, pp. 6965–6975 (2018)
11. Mobiny, A., Nguyen, H.V., Moulik, S., Garg, N., Wu, C.C.: DropConnect Is Effective in Modeling Uncertainty of Bayesian Deep Networks. arXiv preprint arXiv:1906.04569 (2019)
12. Nair, T., Precup, D., Arnold, D.L., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018, Part I. LNCS, vol. 11070, pp. 655–663. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_74

13. 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
14. Roy, A.G., Conjeti, S., Navab, N., Wachinger, C.: Inherent brain segmentation quality control from fully ConvNet Monte Carlo sampling. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018, Part I. LNCS, vol. 11070, pp. 664–672. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_75
15. Sakaridis, C., Dai, D., Van Gool, L.: Semantic nighttime image segmentation with synthetic stylized data, gradual adaptation and uncertainty-aware evaluation. arXiv preprint arXiv:1901.05946 (2019)
16. Smith, L., Gal, Y.: Understanding measures of uncertainty for adversarial example detection. In: Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, 6–10 August 2018, pp. 560–569 (2018)
17. 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)
18. Wang, S., et al.: Central focused convolutional neural networks: developing a data-driven model for lung nodule segmentation. Med. Image Anal. **40**, 172–183 (2017)