Asymmetric Contour Uncertainty Estimation for Medical Image Segmentation

Thierry Judge[1(✉)], Olivier Bernard[3], Woo-Jin Cho Kim[2], Alberto Gomez[2], Agisilaos Chartsias[2], and Pierre-Marc Jodoin[1]

[1] Department of Computer Science, University of Sherbrooke, Sherbrooke, Canada
thierry.judge@usherbrooke.ca
[2] Ultromics Ltd., Oxford OX4 2SU, UK
[3] University of Lyon, CREATIS, CNRS UMR5220, Inserm U1294, INSA-Lyon, University of Lyon 1, Villeurbanne, France

Abstract. Aleatoric uncertainty estimation is a critical step in medical image segmentation. Most techniques for estimating aleatoric uncertainty for segmentation purposes assume a Gaussian distribution over the neural network's logit value modeling the uncertainty in the predicted class. However, in many cases, such as image segmentation, there is no uncertainty about the presence of a specific structure, but rather about the precise outline of that structure. For this reason, we explicitly model the location uncertainty by redefining the conventional per-pixel segmentation task as a contour regression problem. This allows for modeling the uncertainty of contour points using a more appropriate multivariate distribution. Additionally, as contour uncertainty may be asymmetric, we use a multivariate skewed Gaussian distribution. In addition to being directly interpretable, our uncertainty estimation method outperforms previous methods on three datasets using two different image modalities. Code is available at: https://github.com/ThierryJudge/contouring-uncertainty.

Keywords: Uncertainty estimation · Image segmentation

1 Introduction

Segmentation is key in medical image analysis and is primarily achieved with pixel-wise classification neural networks [4,14,20]. Recently, methods that use contours defined by points [10,13] have been shown more suitable for organs with a regular shape (e.g. lungs, heart) while predicting the organ outline similarly to how experts label data [10,13]. While various uncertainty methods have been investigated for both pixel-wise image segmentation [6,16,29] and landmark regression [25,27], few uncertainty methods for point-defined contours in the context of segmentation exists to date.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43898-1_21.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 210–220, 2023.
https://doi.org/10.1007/978-3-031-43898-1_21

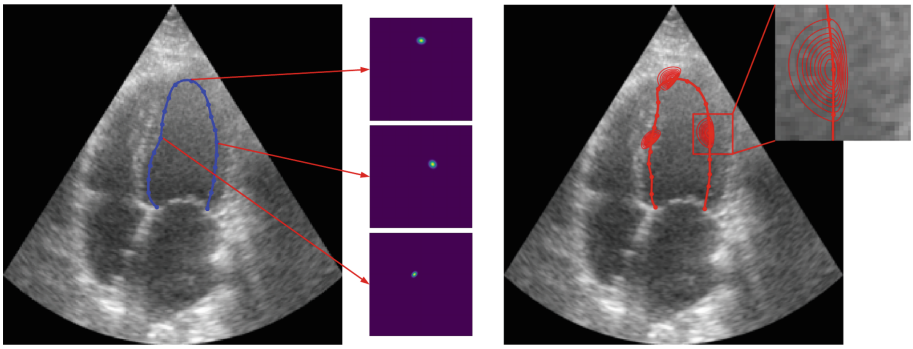

Fig. 1. Overview of our method. Left to right: predicted left ventricle contour landmarks; heatmaps associated to three points; and skewed-normal uncertainty estimation for these three points.

Uncertainty can be *epistemic* or *aleatoric* by nature [16]. Epistemic uncertainty models the network uncertainty by defining the network weights as a probabilistic distribution instead of a single value, with methods such as Bayesian networks [5], MC dropout [11,12] and ensembles [19]. Aleatoric uncertainty is the uncertainty in the data. Most pixel-wise segmentation methods estimate per-pixel aleatoric uncertainty by modeling a normal distribution over each output logit [16]. For regression, it is common practice to assume that each predicted output is independent and identically distributed, and follows an univariate normal distribution. In that case, the mean and variance distribution parameters μ and σ are learned with a loss function that maximizes their log-likelihood [16].

Other methods estimate the aleatoric uncertainty from multiple forward passes of test-time augmentation [1,29]. Some methods do not explicitly model epistemic nor aleatoric uncertainty, but rather use custom uncertainty losses [9] or add an auxiliary confidence network [8]. Other works predict uncertainty based on an encoded prior [15] or by sampling a latent representation space [3,18]. The latter however requires a dataset containing multiple annotations per image to obtain optimal results.

Previous methods provide pixel-wise uncertainty estimates. These estimates are beneficial when segmenting abnormal structures that may or may not be present. However, they are less suited for measuring uncertainty on organ delineation because their presence in the image are not uncertain.

In this work, we propose a novel method to estimate aleatoric uncertainty of point-wise defined contours, independent on the model's architecture, without compromising the contour estimation performance. We extend state-of-the-art point-regression networks [21] by modeling point coordinates with Gaussian and skewed-Gaussian distributions, a novel solution to predict asymmetrical uncertainty. Conversely, we demonstrate that the benefits of point-based contouring also extend to uncertainty estimation with highly interpretable results.

2 Method

Let's consider a dataset made of N pairs $\{x_i, y_i^k\}_{i=1}^N$, each pair consisting of an image $x_i \in R^{H \times W}$ of height H and width W, and a series of K ordered points y_i^k, drawn by an expert. Each point series defines the contour of one or more organs depending on the task. A simple way of predicting these K points is to regress $2K$ values (x-y coordinates) with a CNN, but doing so is sub-optimal due to the loss of spatial coherence in the output flatten layer [21]. As an alternative, Nabili et al. proposed the DSNT network (*differentiable spatial to numerical transform*) designed to extract numerical coordinates of K points from the prediction of K heatmaps [21] (c.f. the middle plots in Fig. 1 for an illustration).

Inspired by this work, our method extends to the notion of heatmaps to regress univariate, bivariate, and skew-bivariate uncertainty models.

2.1 Contouring Uncertainty

Univariate Model - In this approach, a neural network $f_\theta(\cdot)$ is trained to generate K heatmaps $Z^k \in R^{H \times W}$ which are normalized by a softmax function so that their content represents the probability of presence of the center c^k of each landmark point. Two coordinate maps $\mathbf{I} \in R^{H \times W}$ and $\mathbf{J} \in R^{H \times W}$, where $\mathbf{I}_{i,j} = \frac{2j-(W+1)}{W}$ and $\mathbf{J}_{i,j} = \frac{2i-(H+1)}{H}$, are then combined to these heatmaps to regress the final position μ^k and the corresponding variance (σ_x^k, σ_y^k) of each landmark point through the following two equations:

$$\mu^k = E[c^k] = \left[\langle \hat{Z^k}, \mathbf{I} \rangle_F, \langle \hat{Z^k}, \mathbf{J} \rangle_F\right] \in R^2, \tag{1}$$

$$\begin{aligned} Var[c^{k_x}] = (\sigma^{k_x})^2 &= E[(c^{k_x} - E[c^{k_x}])^2) \\ &= \langle \hat{Z}^k, (\mathbf{I} - \mu^{k_x}) \odot (\mathbf{I} - \mu^{k_x}) \rangle_F \in R, \end{aligned} \tag{2}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product, $\odot$ corresponds to the Hadamard product, and $(\sigma^{k_y})^2$ is computed similarly. Thus, for each image x_i, the neural network $f_\theta(x_i)$ predicts a tuple (μ_i, σ_i) with $\mu_i \in R^{2K}$ and $\sigma_i \in R^{2K}$ through the generation of K heatmaps. The network is finally trained using the following univariate aleatoric loss adapted from [16]

$$\mathcal{L}_{\mathcal{N}_1} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \log\left(\sigma_i^{k_x} \sigma_i^{k_y}\right) + \frac{1}{2} \frac{(\mu_i^{k_x} - y_i^{k_x})^2}{(\sigma_i^{k_x})^2} + \frac{(\mu_i^{k_y} - y_i^{k_y})^2}{(\sigma_i^{k_y})^2}, \tag{3}$$

where y_i^k is the k^{th} reference landmark point of image x_i.

Bivariate Model - One of the limitations of the univariate model is that it assumes no x-y covariance on the regressed uncertainty. This does not hold true in many cases, because the uncertainty can be oblique and thus involve a non-zero x-y covariance. To address this, one can model the uncertainty of each point

with a 2×2 covariance matrix, Σ, where the variances are expressed with Eq. 2 and the covariance is computed as follows:

$$\begin{aligned} cov[c^k] &= E[(c^{k_x} - E[c^{k_x}])(c^{k_y} - E[c^{k_y}]) \\ &= \langle \hat{Z}, (\mathbf{I} - \mu^{k_x}) \odot (\mathbf{J} - \mu^{k_y}) \rangle_F. \end{aligned} \tag{4}$$

The network $f_\theta(x_i)$ thus predicts a tuple (μ_i, Σ_i) for each image x_i, with $\mu_i \in R^{K \times 2}$ and $\Sigma_i \in R^{K \times 2 \times 2}$. We propose to train f_θ using a new loss function $\mathcal{L}_{\mathcal{N}_2}$:

$$\mathcal{L}_{\mathcal{N}_2} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{1}{2} \log |\Sigma_i^k| + \frac{1}{2} (\mu_i^k - y_i^k)^T (\Sigma_i^k)^{-1} (\mu_i^k - y_i^k). \tag{5}$$

Asymmetric Model - One limitation of the bivariate method is that it models a symmetric uncertainty, an assumption that may not hold in some cases as illustrated on the right side of Fig. 2. Therefore we developed a third approach based on a bivariate *skew-normal* distribution [2]:

$$\mathcal{SN}_n(y|\mu, \Sigma, \alpha) = 2\phi_n(y|\mu, \Sigma)\Phi_1\Big(\alpha^T \omega^{-1}(y - \mu)\Big), \tag{6}$$

where ϕ_n is a multivariate normal, Φ_1 is the cumulative distribution function of a unit normal, $\Sigma = \omega \bar{\Sigma} \omega$ and $\alpha \in R^n$ is the skewness parameter. Note that this is a direct extension of the multivariate normal as the skew-normal distribution is equal to the normal distribution when $\alpha = 0$.

The corresponding network predicts a tuple (μ, Σ, α) with $\mu \in R^{K \times 2}$, $\Sigma \in R^{K \times 2 \times 2}$ and $\alpha \in R^{K \times 2}$. The skewness output α is predicted using a sub-network whose input is the latent space of the main network (refer to the supplementary material for an illustration). This model is trained using a new loss function derived from the maximum likelihood estimate of the *skew-normal* distribution:

$$\begin{aligned} \mathcal{L}_{\mathcal{SN}_2} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{1}{2} \log |\Sigma_i^k| + \frac{1}{2} (\mu_i^k - y_i^k)^T (\Sigma_i^k)^{-1} (\mu_i^k - y_i^k) \\ + \log \Phi_1\Big((\alpha_i^k)^T (\omega_i^k)^{-1} (y_i^k - \mu_i^k)\Big). \end{aligned} \tag{7}$$

2.2 Visualization of Uncertainty

As shown in Fig. 2, the predicted uncertainty can be pictured in two ways: (i) either by per-point covariance ellipses [left] and skewed-covariance profiles [right] or (ii) by an uncertainty map to express the probability of wrongly classifying pixels which is highest at the border between 2 classes. In our formalism, the probability of the presence of a contour (and thus the separation between 2 classes) can be represented by the component of the uncertainty that is perpendicular to the contour. We consider the perpendicular normalized marginal distribution at each point (illustrated by the green line). This distribution also

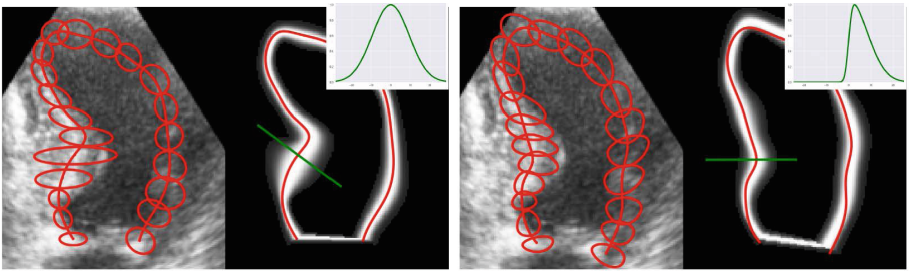

Fig. 2. Two uncertainty visualizations: a per-landmark representation and a pixel-wise uncertainty map. [Left] bi-variate normal model and [right] the skewed-normal distribution. In both case, the uncertainty map has been obtained by interpolating the landmark uncertainly along the contour.

happens to be a univariate normal [left] or skew-normal [right] distribution [2]. From these distributions, we draw isolines of equal uncertainty on the inside and outside of the predicted contour. By aggregating multiple isolines, we construct a smooth uncertainty map along the contours (illustrated by the white-shaded areas). Please refer to the supp. material for further details on this procedure.

3 Experimental Setup

3.1 Data

CAMUS. The CAMUS dataset [20] contains cardiac ultrasounds from 500 patients, for which two-chamber and four-chamber sequences were acquired. Manual annotations for the endocardium and epicardium borders of the left ventricle (LV) and the left atrium were obtained from a cardiologist for the end-diastolic (ED) and end-systolic (ES) frames. The dataset is split into 400 training patients, 50 validation patients, and 50 testing patients. Contour points were extracted by finding the basal points of the endocardium and epicardium and then the apex as the farthest points along the edge. Each contour contains 21 points.

Private Cardiac US. This is a proprietary multi-site multi-vendor dataset containing 2D echocardiograms of apical two and four chambers from 890 patients. Data comes from patients diagnosed with coronary artery disease, COVID, or healthy volunteers. The dataset is split into a training/validation set (80/20) and an independent test set from different sites, comprised of 994 echocardiograms from 684 patients and 368 echocardiograms from 206 patients, respectively. The endocardium contour was labeled by experts who labeled a minimum of 7 points based on anatomical landmarks and add as many other points as necessary to define the contour. We resampled 21 points equally along the contour.

JSRT. The Japanese Society of Radiological Technology (JSRT) dataset consists of 247 chest X-Rays [26]. We used the 120 points for the lungs and heart

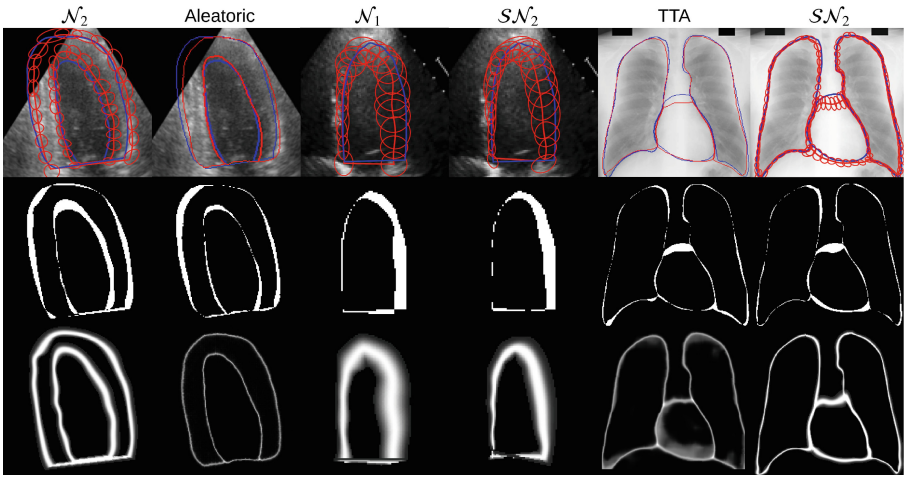

Fig. 3. Results from various methods on CAMUS, Private and JSRT datasets. [row 1] images with predicted (red) and groundtruth (blue) contours. Confidence intervals are shown in red around each point. [row 2] Error maps where white pixels indicate a prediction error. [row 3] Uncertainty maps where high uncertainty is shown in white. (Color figure online)

annotation made available by [10]. The set of points contains specific anatomical points for each structure (4 for the right lung, 5 for the left lung, and 4 for the heart) and equally spaced points between each anatomical point. We reconstructed the segmentation map with 3 classes (background, lungs, heart) with these points and used the same train-val-test split of 70%–10%–20% as [10].

3.2 Implementation Details

We used a network based on ENet [24] for the ultrasound data and on DeepLabV3 [7] for the JSRT dataset to derive both the segmentation maps and regress the per-landmark heatmaps. Images were all reshaped to 256×256 and B-Splines were fit on the predicted landmarks to represent the contours. Training was carried out with the Adam optimizer [17] with a learning rate of 1×10^{-3} and with ample data augmentation (random rotation and translations, brightness and contrast changes, and gamma corrections). Models were trained with early stopping and the models with best validation loss were retained for testing.

3.3 Evaluation Metrics

To assess quality of the uncertainty estimates at image and pixel level we use:

Correlation. The correlation between image uncertainty and Dice was computed using the absolute value of the Pearson correlation score. We obtained

Table 1. Uncertainty estimation results for segmentation (top rows) and regression (bottom rows) methods. Best and second best results are highlighted in red and blue respectively.

Data	CAMUS			Private Card. US			JSRT		
Method	Corr. ↑	MCE ↓	MI ↑	Corr. ↑	MCE ↓	MI ↑	Corr. ↑	MCE ↓	MI ↑
Aleatoric [16]	.397	.327	.028	.101	.487	.010	.660	.294	.037
MC-Dropout [6]	.424	.349	.030	.276	.467	.011	.559	.346	.060
TTA [29]	.538	.340	.023	.261	.400	.009	.432	.422	.036
MC-Dropout	.271	.380	.021	.600	.378	.009	.453	.368	.007
$\mathcal{N}_1$	.403	.088	.052	.635	.103	.033	.713	.129	.047
$\mathcal{N}_2$	.386	.114	.049	.697	.173	.032	.595	.118	.050
$\mathcal{SN}_2$	.454	.104	.051	.562	.332	.025	.824	.152	.055

image uncertainty be taking the sum of the uncertainty map and dividing it by the number of foreground pixels.

Maximum Calibration Error (MCE). This common uncertainty metric represents the probability if a classifier (here a segmentation method) of being correct by computing the worst case difference between its predicted confidence and its actual accuracy [23].

Uncertainty Error Mutual-Information. As proposed in [15], uncertainty error mutual-information measures the degree of overlap between the unthresholded uncertainty map and the pixel-wise error map.

4 Results

We computed uncertainty estimates for both pixel-wise segmentation and contour regression methods to validate the hypothesis that uncertainty prediction is better suited to per-landmark segmentation than per-pixel segmentation methods. For a fair comparison, we made sure the segmentation models achieve similar segmentation performance, with the average Dice being $.90 \pm .02$ for CAMUS, $.86 \pm .02$ for Private US., and $.94 \pm .02$ for JSRT.

For the pixel-wise segmentations, we report results of a classical *aleatoric* uncertainty segmentation method [16] as well as a Test Time Augmentation (TTA) method [29]. For TTA, we used the same augmentations as the ones used during training. We also computed *epistemic* uncertainty with MC-Dropout [6] for which we selected the best results of 10%, 25%, and 50% dropout rates. The implementation of MC-Dropout for regression was trained with the DSNT layer [21] and mean squared error as a loss function.

As for the landmark prediction, since no uncertainty estimation methods have been proposed in the literature, we adapted the MC-Dropout method to it. We also report results for our method using univariate, ($\mathcal{N}_1$), bivariate, ($\mathcal{N}_2$) and bivariate skew-normal distributions ($\mathcal{SN}_2$).

The uncertainty maps for TTA and MC-Dropout (i.e. those generating multiple samples) were constructed by computing the pixel-wise entropy of multiple

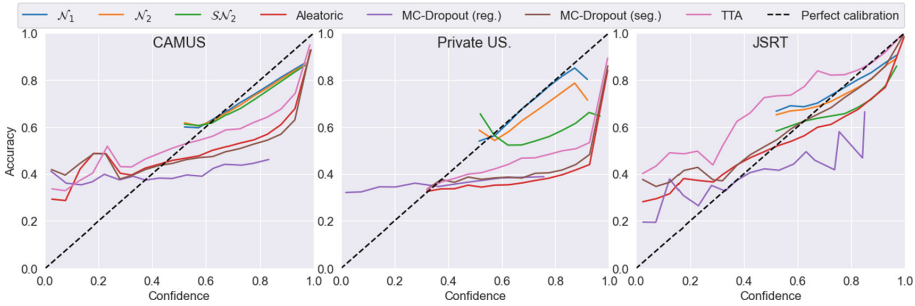

Fig. 4. Reliability diagrams [22] for the 3 datasets. For uncertainty (u) bounded by 0 and 1, confidence (c) is defined as $c = 1 - u$ [28]

forward passes. It was found that doing so for the aleatoric method produces better results than simply taking the variance. The uncertainty map for the landmark predictions was obtained with the method described in Sect. 2.2.

Quantitative results are presented in Table 1 and qualitative results are shown in Fig. 3. As can be seen, our uncertainty estimation method is globally better than the other approaches except for the correlation score on the CAMUS dataset which is slightly larger for TTA. Furthermore, our point-based aleatoric uncertainty better detects regions of uncertainty consistently, as reflected in the Mutual Information (MI) metric. The reliability diagrams in Fig. 4 show that our method is systematically better aligned to perfect calibration (dashed line) for all datasets, which explains why our method has a lower MCE. With the exception of the Private Cardiac US dataset, the skewed normal distribution model shows very similar or improved results for both correlation and mutual information compared to the univariate and bivariate models. It can be noted, however, that in specific instances, the asymmetric model performs better on Private Cardiac US dataset (c.f. column 2 and 3 in Fig. 3). This confirms that it is better capturing asymmetric errors over the region of every contour point.

5 Discussion and Conclusion

The results reported before reveal that approaching the problem of segmentation uncertainty prediction via a regression task, where the uncertainty is expressed in terms of landmark location, is globally better than via pixel-based segmentation methods. It also shows that our method ($\mathcal{N}_1$, $\mathcal{N}_2$ and $\mathcal{SN}_2$) is better than the commonly-used MC-Dropout. It can also be said that our method is more interpretable as is detailed in Sect. 2.2 and shown in Fig. 3.

The choice of distribution has an impact when considering the shape of the predicted contour. For instance, structures such as the left ventricle and the myocardium wall in the ultrasound datasets have large components of their contour oriented along the vertical direction which allows the univariate and bivariate models to perform as well, if not better, than the asymmetric model.

However, the lungs and heart in chest X-Rays have contours in more directions and therefore the uncertainty is better modeled with the asymmetric model.

Furthermore, it has been demonstrated that skewed uncertainty is more prevalent when tissue separation is clear, for instance, along the septum border (CAMUS) and along the lung contours (JSRT). The contrast between the left ventricle and myocardium in the images of the Private Cardiac US dataset is small, which explains why the simpler univariate and bivariate models perform well. This is why on very noisy and poorly contrasted data, the univariate or the bivariate model might be preferable to using the asymmetric model.

While our method works well on the tasks presented, it is worth noting that it may not be applicable to all segmentation problems like tumour segmentation. Nevertheless, our approach is broad enough to cover many applications, especially related to segmentation that is later used for downstream tasks such as clinical metric estimation. Future work will look to expand this method to more general distributions, including bi-modal distributions, and combine the aleatoric and epistemic uncertainty to obtain the full predictive uncertainty.

References

1. Ayhan, M.S., Berens, P.: Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In: International Conference on Medical Imaging with Deep Learning (2018)
2. Azzalini, A.: Institute of Mathematical Statistics Monographs: The Skew-Normal and Related Families Series Number 3. Cambridge University Press, Cambridge (2013)
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. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)
5. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 1613–1622. PMLR, Lille, France, 07–09 July 2015
6. Camarasa, R., et al.: Quantitative comparison of Monte-Carlo dropout uncertainty measures for multi-class segmentation. In: Sudre, C.H., et al. (eds.) UNSURE/GRAIL -2020. LNCS, vol. 12443, pp. 32–41. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60365-6_4
7. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR abs/1706.05587 (2017)
8. Corbière, C., Thome, N., Bar-Hen, A., Cord, M., Pérez, P.: Addressing failure prediction by learning model confidence. In: Advances in Neural Information Processing Systems, vol. 32, pp. 2902–2913. Curran Associates, Inc. (2019)
9. DeVries, T., Taylor, G.W.: Leveraging uncertainty estimates for predicting segmentation quality. CoRR abs/1807.00502 (2018)
10. Gaggion, N., Mansilla, L., Mosquera, C., Milone, D.H., Ferrante, E.: Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis. IEEE Trans. Med. Imaging (2022)

11. Gal, Y., Ghahramani, Z.: Bayesian convolutional neural networks with Bernoulli approximate variational inference. arXiv abs/1506.02158 (2015)
12. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning. ICML'16, vol. 48, pp. 1050–1059. JMLR.org (2016)
13. Gomez, A., et al.: Left ventricle contouring of apical three-chamber views on 2d echocardiography. In: Aylward, S., Noble, J.A., Hu, Y., Lee, S.L., Baum, Z., Min, Z. (eds.) ASMUS 2022. LNCS, vol. 13565, pp. 96–105. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16902-1_10
14. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: NNU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
15. Judge, T., Bernard, O., Porumb, M., Chartsias, A., Beqiri, A., Jodoin, P.M.: Crisp - reliable uncertainty estimation for medical image segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022 MICCAI 2022. LNCS, vol. 13438, pp. 492–502. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1_47
16. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 5574–5584. Curran Associates, Inc. (2017)
17. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015)
18. Kohl, S., et al.: A probabilistic u-net for segmentation of ambiguous images. In: Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018)
19. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
20. Leclerc, S., et al.: Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Trans. Med. Imaging **38**(9), 2198–2210 (2019)
21. Nibali, A., He, Z., Morgan, S., Prendergast, L.: Numerical coordinate regression with convolutional neural networks. arXiv preprint arXiv:1801.07372 (2018)
22. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. ICML '05, pp. 625–632. Association for Computing Machinery, New York, NY, USA (2005)
23. Pakdaman Naeini, M., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using Bayesian binning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29, no. 1, February 2015
24. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: Enet: a deep neural network architecture for real-time semantic segmentation. CoRR abs/1606.02147 (2016)
25. Schobs, L.A., Swift, A.J., Lu, H.: Uncertainty estimation for heatmap-based landmark localization. IEEE Trans. Med. Imaging **42**(4), 1021–1034 (2023)
26. Shiraishi, J., et al.: Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. Am. J. Roentgenol. **174**, 71–74 (2000)
27. Thaler, F., Payer, C., Urschler, M., Štern, D.: Modeling annotation uncertainty with Gaussian heatmaps in landmark localization. Mach. Learn. Biomed. Imaging **1**, 1–27 (2021)

28. Tornetta, G.N.: Entropy methods for the confidence assessment of probabilistic classification models. Statistica (Bologna) **81**(4), 383–398 (2021)
29. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019)