CRISP - Reliable Uncertainty Estimation for Medical Image Segmentation

Thierry Judge, Olivier Bernard, Mihaela Porumb, Agisilaos Chartsias, Arian Beqiri, Pierre-Marc Jodoin

CRISP - Reliable Uncertainty Estimation for Medical Image Segmentation

Thierry Judge[1⋆], Olivier Bernard[3], Mihaela Porumb[2], Agisilaos Chartsias[2], Arian Beqiri[2], and Pierre-Marc Jodoin[1]

[1] Department of Computer Science, University of Sherbrooke, Canada
[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. Accurate uncertainty estimation is a critical need for the medical imaging community. A variety of methods have been proposed, all direct extensions of classification uncertainty estimations techniques. The independent pixel-wise uncertainty estimates, often based on the probabilistic interpretation of neural networks, do not take into account anatomical prior knowledge and consequently provide sub-optimal results to many segmentation tasks. For this reason, we propose *CRISP* a ContRastive Image Segmentation for uncertainty Prediction method. At its core, *CRISP* implements a contrastive method to learn a joint latent space which encodes a distribution of valid segmentations and their corresponding images. We use this joint latent space to compare predictions to thousands of latent vectors and provide anatomically consistent uncertainty maps. Comprehensive studies performed on four medical image databases involving different modalities and organs underlines the superiority of our method compared to state-of-the-art approaches. Code is available at: `https://github.com/ThierryJudge/CRISP-uncertainty`.

Keywords: Medical imaging, Segmentation, Uncertainty, Deep learning

1 Introduction

Deep neural networks are the *de facto* solution to most segmentation, classification and clinical metric estimation. However, they provide no anatomical guarantees nor any safeguards on their predictions. Error detection and uncertainty estimation methods are therefore paramount before automatic medical image segmentation systems can be effectively deployed in clinical settings.

In this work, we present a novel uncertainty estimation method based on joint representations between images and segmentations trained with contrastive learning. Our method, *CRISP* (ContRastive Image Segmentation for uncertainty Prediction), uses this representation to overcome the limitations of state-of-the-art (SOTA) methods which heavily rely on probabilistic interpretations of neural networks as is described below.

⋆ Corresponding author

Uncertainty is often estimated assuming a probabilistic output function by neural networks. However, directly exploiting the maximum class probability of the *Softmax* or *Sigmoid* usually leads to suboptimal solutions [7]. Some improvements can be made by considering the entire output distribution through the use of entropy [22] or by using other strategies such as temperature scaling [7].

Uncertainty may also come from Bayesian neural networks, which learn a distribution over each parameter using a variational inference formalism [10]. This enables weight sampling, which produces an output distribution that can model the prediction uncertainty. As Bayesian networks are difficult to train, they are often approximated by aggregating the entropy of many dropout forward runs [5,6]. Alternatively, a network ensemble trained with different hyperparameters can also estimate uncertainties through differences in predictions [14].

In addition to modeling weight uncertainty, referred to as epistemic uncertainty, uncertainty in the data itself (aleatoric) can also be predicted [11]. However, it has been shown that these methods are less effective for segmentation [9].

Other methods explicitly learn an uncertainty output during training. DeVries and Taylor [4] proposed Learning Confidence Estimates (LCE) by adding a confidence output to the network. The segmentation prediction is interpolated with the ground truth according to this confidence. This confidence can also be learned after training by adding a confidence branch and finetuning a pretrained network. This enables learning the True Class Probability which is a better confidence estimate than the maximum class probability [2].

Recent works have modeled the disagreement between labelers for ambiguous images [13,1]. Both these methods use a form of variational sampling to make their output stochastic. However, these methods require datasets with multiple labels per image to perform at their best. As these datasets are rarely available, we consider these methods out of scope for this paper.

With the exception of methods modeling disagreement, all other methods can be applied to classification and, by extension, to segmentation tasks with an uncertainty prediction at each pixel. In theory, uncertainty maps should identify areas in which the prediction is erroneous. However, as these methods produce per-pixel uncertainties, they do not take into account higher-level medical information such as anatomical priors. Such priors have been used in segmentation [25,17], but are yet to be exploited in uncertainty estimation. For instance, Painchaud et al. [18] remove anatomical errors by designing a latent space dedicated to the analysis and correction of erroneous cardiac shapes. However, this approach does not guarantee that the corrected shape matches the input image.

To this end, we propose *CRISP*, a method which does not take into account the probabilistic nature of neural networks, but rather uses a joint latent representation of anatomical shapes and their associated image. This paper will describe the *CRISP* method and propose a rigorous evaluation comparing *CRISP* to SOTA methods using four datasets.

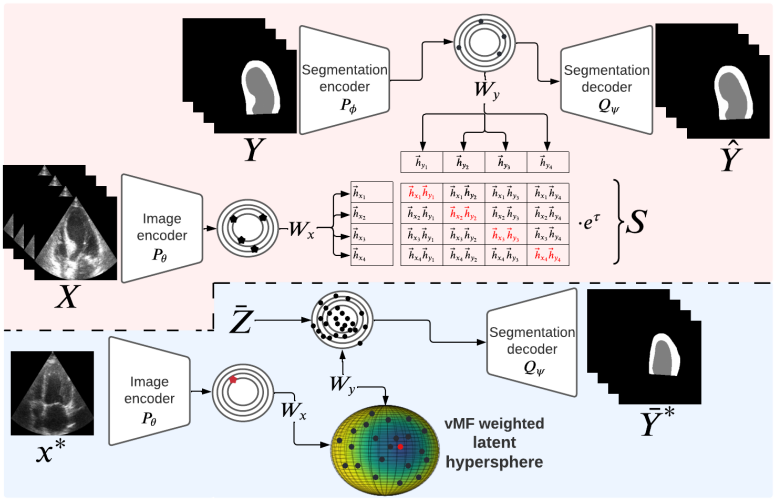

Fig. 1. Schematic representation of our method. Top depicts the training phase and bottom illustrate the uncertainty estimation on an input-prediction pair (x^*, y^*).

2 *CRISP*

The overarching objective of our method is to learn a joint latent space, in which the latent vector of an input image lies in the vicinity of its corresponding segmentation map's latent vector in a similar fashion as the "CLIP" method does for images and text [21]. As such, a test image x whose latent vector does not lie close to that of its segmentation map y is an indication of a potentially erroneous segmentation. Further details are given below.

Training. As shown in Fig. 1, at train time, *CRISP* is composed of two encoders: the image encoder P_θ and the segmentation encoder P_ϕ. They respectively encode an image x_i and its associated segmentation groundtruth y_i into latent vectors $\vec{z}_{x_i} \in \Re^{D_x}$ and $\vec{z}_{y_i} \in \Re^{D_y}$ $\forall i$. Two weight matrices $W_x \in \Re^{D_h \times D_x}$ and $W_y \in \Re^{D_h \times D_y}$ linearly project the latent vectors into a joint D_h-dimensional latent space where samples are normalized and thus projected onto a hyper-sphere. As such, the image latent vector $\vec{z}_{x_i}$ is projected onto a vector $\vec{h}_{x_i} = \frac{W_x \cdot \vec{z}_{x_i}}{||W_x \cdot \vec{z}_{x_i}||}$ and similarly for $\vec{z}_{y_i}$. A successful training should lead to a joint representation for which $\vec{h}_{x_i} \approx \vec{h}_{y_i}$.

During training, images and groundtruth maps are combined into batches of B elements, $\boldsymbol{X} = [x_1 x_2 ... x_B] \in \Re^{B \times C \times H \times W}$ and $\boldsymbol{Y} = [y_1 y_2 ... y_B] \in \{0,1\}^{B \times K \times H \times W}$ for images with C channels and K segmentation classes. As mentioned before, these batches are encoded by P_θ and P_ϕ into sets of latent vectors Z_X and Z_Y and then projected and normalized into sets of joint latent vectors H_X and H_Y.

At this point, a set of $2 \times B$ samples lie on the surface of a unit hypersphere of the joint latent space. Much like CLIP [21], the pair-wise distance

between these joint latent vectors is computed with a cosine similarity that we scale by a learned temperature factor τ to control the scale of the logits. This computation is done by taking a weighted product between H_X and H_Y which leads to the following square matrix: $S = (H_X \cdot H_Y^T)e^\tau \in \Re^{B \times B}$. As shown in Fig. 1, the diagonal of S corresponds to the cosine similarity of the latent image vectors with their corresponding latent groundtruth vector while the off-diagonal elements are cosine similarity of unrelated vectors.

The goal during training is to push S towards an identity matrix, such that the latent vectors $\vec{h}_{x_i}$ and $\vec{h}_{y_j}$ lie on the same spot in the joint latent space when $i = j$ and are orthogonal when $i \neq j$. This would lead to similarities close to 1 on the diagonal and close to 0 outside of it. To enforce this, a cross-entropy loss on the rows and columns of S is used as a contrastive loss [21],

$$\mathcal{L}_{cont} = -\frac{1}{2}\left(\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B} I_{ij}\log S_{ij} + \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B} I_{ji}\log S_{ji}\right). \tag{1}$$

CRISP also has a segmentation decoder Q_ψ to reconstruct segmentation latent vectors, a critical feature for estimating uncertainty. This decoder is trained with a reconstruction loss $\mathcal{L}_{rec}$ which is a weighted sum of the Dice coefficient and cross-entropy loss. The model is trained end-to-end to minimize $\mathcal{L} = \mathcal{L}_{cont} + \mathcal{L}_{rec}$.

Uncertainty prediction. Once training is over, the groundtruth segmentation maps $\boldsymbol{Y}$ are projected one last time into the Z and H latent spaces. This leads to a set of N latent vectors $\bar{Z} \in \Re^{N \times D_y}$ and $\bar{H} \in \Re^{N \times D_h}$ which can be seen as latent anatomical prior distributions that will be used to estimate uncertainty.

Now let x^* be a non-training image and y^* its associated segmentation map computed with a predetermined segmentation method (be it a deep neural network or not). To estimate an uncertainty map, x^* is projected into the joint latent space to get its latent vector $\vec{h}_{x^*} \in \Re^{D_h}$. We then compute a weighted dot product between $\vec{h}_{x^*}$ and each row of $\bar{H}$ to get $\bar{S} \in \Re^N$, a vector of similarity measures between $\vec{h}_{x^*}$ and every groundtruth latent vector. Interestingly enough, the way *CRISP* was trained makes $\bar{S}$ a similarity vector highlighting how each groundtruth map fits the input image x^*.

Then, the M samples of $\bar{Z}$ with the highest values in $\vec{S}$ are selected. These samples are decoded to obtain $\bar{Y}^*$, *i.e.* various anatomically valid segmentation maps whose shapes are all roughly aligned on x^*. To obtain an uncertainty map, we compare these samples to the initial prediction y^*. We compute the average of the pixel-wise difference between y^* and $\bar{Y}^*$ to obtain an uncertainty map U.

$$U = \frac{1}{M}\sum_{i=1}^{M} w_i(\bar{y}_i^* - y^*) \tag{2}$$

As not all samples equally correspond to x^*, we add a coefficient w_i which corresponds to how close a groundtruth map y_i is from x^*. Since the joint latent space is a unit hyper-sphere, we use a *von Mises-Fisher distribution* (vMF) [16] centered on $\vec{h}_{x^*}$ as a kernel to weigh its distance to $\vec{h}_{y_i}$. We use Taylor's

method [24] to define the kernel bandwidth b (more details are available in the supplementary materials). We define the kernel as:

$$w_i = e^{\frac{1}{b}\vec{h_i}^T\vec{h}_{x^*}} / e^{\frac{1}{b}\vec{h}_{x^*}^T\vec{h}_{x^*}} = e^{\frac{1}{b}(\vec{h_i}^T\vec{h}_{x^*}-1)}. \quad (3)$$

3 Experimental setup

3.1 Uncertainty metrics

Correlation. Correlation is a straightforward method for evaluating the quality of uncertainty estimates for a full dataset. The absolute value of the Pearson correlation score is computed between the sample uncertainty and the Dice score. In this paper, sample uncertainty is obtained by dividing the sum of the uncertainty for all pixels by the number of foreground pixels. Ideally, the higher the Dice is, the lower the sample uncertainty should be. Therefore, higher correlation values indicate more representative uncertainty maps.
Calibration. A classifier is calibrated if its confidence is equal to the probability of being correct. Calibration is expressed with Expected Calibration Error (ECE) computed by splitting all n samples into m bins and computing the mean difference between the accuracy and average confidence for each bin. Please refer to the following paper for more details [19].
Uncertainty-error mutual information. Previous studies have computed Uncertainty-error overlap by obtaining the Dice score between the thresholded uncertainty map and a pixel-wise error map between the prediction and the ground-truth segmentation map [9]. As the uncertainty error overlap requires the uncertainty map to be thresholded, much of the uncertainty information is lost. We therefore propose computing the mutual information between the raw uncertainty map and the pixel-wise error map. We report the average over the test set weighted by the sum of erroneous pixels in the image.

3.2 Data

CAMUS. The CAMUS dataset [15] consists of cardiac ultrasound clinical exams performed on 500 patients. Each exam contains the 2D apical four-chamber (A4C) and two-chamber view (A2C) sequences. Manual delineation of the endocardium and epicardium borders of the left ventricle (LV) and atrium were made by a cardiologist for the end-diastolic (ED) and end-systolic (ES) frames. The dataset is split into training, validation and testing sets of 400, 50 and 50 patients respectively.
HMC-QU. The HMC-QU dataset [3] is composed of 162 A4C and 130 A2C view recordings. 93 A4C and 68 A2C sequences correspond to patients with scarring from myocardial infarction. The myocardium (MYO) of 109 A4C (72 with myocardial infarction/37 without) recordings was manually labeled for the full cardiac cycle. These sequences were split into training, validation and testing sets of 72, 9 and 28 patients.

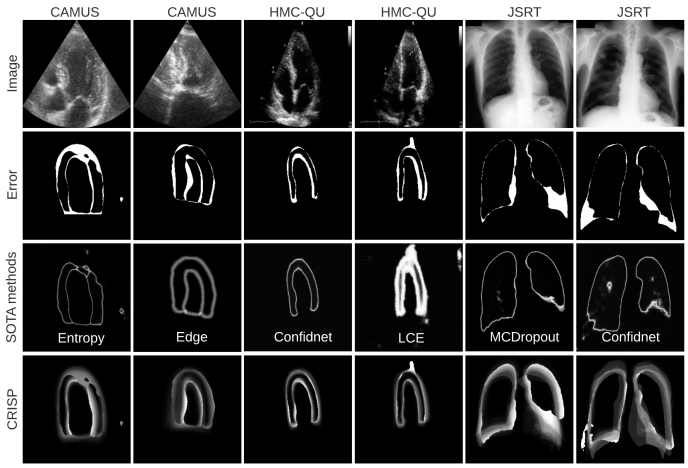

Fig. 2. From top to bottom: raw images, corresponding error maps, uncertainty estimation of SOTA methods and CRISP uncertainty. White indicates erroneous pixels in the error maps [row 2] and high uncertainty in the uncertainty maps [rows 3 and 4].

Shenzen. The Shenzen dataset [8] is a lung X-ray dataset acquired for pulmonary tuberculosis detection. The dataset contains 566 postero-anterior chest radiographs and corresponding manually segmented masks to identify the lungs. The dataset was split into training and validation sets of 394 and 172 patients.
JSRT. We use the Japanese Society of Radiological Technology (JSRT) [23] lung dataset which contains images and segmentation maps for 154 radiographs with lung nodules, and corresponding segmentation masks.

3.3 Implementation details

CRISP was compared to several SOTA methods mentioned before. To make comparison fair, every method use the same segmentation network (an Enet [20] in our case). All methods were trained with a batch size of 32 and the Adam optimizer [12] with a learning rate of 0.001 and weight decay of 1e-4. We added early stopping and selected the weights with the lowest validation loss. The **Entropy** method was tested using the baseline network. We tested **MC Dropout** by increasing the baseline dropout value from 10% to 25% and 50% (we report best results with respect to the Dice score) and computing the average of 10 forward passes. For **LCE**, we duplicated the last bottleneck of the Enet to output confidence. The **Confidnet** method was trained on the baseline Enet pre-trained network. The full decoder was duplicated to predict the True Class Probability. For methods or metrics that require converting pixel-wise confidence (c) to uncertainty (u), we define the relationship between the two as $u = 1 - c$ as all methods produce values in the range $[0, 1]$.

To highlight some limitations of SOTA methods, we also added a naïve method for computing uncertainty which we referred to as **Edge**. The uncer-

Training data	CAMUS			CAMUS			Shenzen		
Testing data	CAMUS			HMC-QU			JSRT		
Method	Corr. ↑	ECE ↓	MI ↑	Corr. ↑	ECE ↓	MI ↑	Corr. ↑	ECE ↓	MI ↑
Entropy	0.66	0.12	0.02	0.34	0.27	0.02	**0.89**	0.08	0.02
ConfidNet [1]	0.34	0.08	0.04	0.36	0.17	0.04	0.69	0.09	0.01
CRISP	**0.71**	0.09	**0.20**	**0.41**	0.14	**0.06**	0.83	0.19	**0.11**
McDropout [3]	0.67	0.13	0.03	0.26	0.26	0.02	**0.82**	**0.06**	0.03
CRISP-MC	**0.78**	**0.11**	**0.26**	**0.29**	**0.14**	**0.06**	**0.82**	0.21	**0.08**
LCE [2]	0.58	0.44	0.08	**0.35**	0.37	**0.07**	**0.87**	0.37	0.06
CRISP-LCE	**0.59**	**0.08**	**0.15**	0.34	**0.13**	**0.07**	0.85	**0.18**	**0.11**

Table 1. Uncertainty estimation results (average over 3 random seeds) for different methods. Bold values indicate best results.

tainty map for *Edge* amounts to a trivial edge detector applied to baseline predicted segmentation maps. The resulting borders have a width of 5 pixels.

As our **CRISP** method can be used to evaluate any image-segmentation pair, regardless of the segmentation method, we tested it on all the segmentation methods that produce different results (baseline, MC Dropout, LCE). This allows for a more robust evaluation as the evaluation of uncertainty metrics is directly influenced by the quality of the segmentation maps [9]. The value of M was determined empirically and kept proportional to the size of $\bar{Z}$. It can be noted, that the vMF weighting in the latent space attenuates the influence of M.

3.4 Experimental setup

We report results on both binary and multi-class segmentation tasks. As our datasets are relatively large and homogeneous, Dice scores are consistently high. This can skew results as methods can simply predict uncertainty around the prediction edges. Thus, as mentioned below, we tested on different datasets or simulated domain shift through data augmentation.

Tests were conducted on the CAMUS dataset for LV and MYO segmentation. We simulated a domain shift by adding brightness and contrast augmentations (factor=0.2) and Gaussian noise ($\sigma^2 = 0.0001$) with probability of 0.5 for all test images. We used the 1800 samples from the training and validation sets to make up the $\bar{Z}$ set and used $M = 50$ samples to compute the uncertainty map.

We also tested all methods trained on the CAMUS dataset on the HMC-QU dataset for myocardium segmentation. We added brightness and contrast augmentations (factor=0.2) and RandomGamma (0.95 to 1.05) augmentations during training and normalized the HMC-QU samples using the mean and variance of the CAMUS dataset. We used the A4C samples from the CAMUS dataset (along with interpolated samples between ES and ED instants) to create the set

of latent vectors $\bar{Z}$. This corresponds to 8976 samples, of which $M = 150$ were selected to compute the uncertainty map.

Finally, we tested our method on a different modality and organ by using the lung X-ray dataset. We trained all the methods on the Shenzen dataset and tested on the JSRT dataset. We normalized JSRT samples with the mean and variance of the Shenzen dataset. We used the 566 samples from the Shenzen dataset to form the $\bar{Z}$ set and used $M = 25$ samples to compute the uncertainty.

4 Results

Uncertainty maps are presented in Fig. 2 for samples on 3 datasets. As can be seen, *Entropy, ConfidNet*, and *MCDropout* have a tendency to work as an edge detector, much like the naive *Edge* method. As seen in Table 1, different methods perform to different degrees on each of the datasets. However, CRISP is consistently the best or competitive for all datasets for the correlation and MI metrics. ECE results for *CRISP* are also competitive but not the best. Interestingly, the trivial *Edge* method often reports the best ECE results. This is probably due to the fact that errors are more likely to occur near the prediction boundary and the probability of error decreases with distance. These results might encourage the community to reconsider the value of ECE for specific types of segmentation tasks.

Fig. 3 shows the distribution of pixel confidence according to the well-classified and misclassified pixels. This figure allows for a better understanding of the different shortcomings of each method. It clearly shows that both MC Dropout and *Confidnet* methods produce over-confident results. On the other hand, LCE appears to produce slightly under-confident predictions which explains the higher mutual information value. Finally, *CRISP* is the only method that can clearly separate certain and uncertain pixels. These results are consistent with what is observed in Fig. 2 as both MC Dropout and *Confidnet* produce very thin uncertainty and LCE predicts large areas of uncertainty around the border. Only *CRISP* produces varying degrees of uncertainty according to the error.

It is apparent that there is a slight decrease in performance for *CRISP* on the JSRT dataset. This is most likely caused by the fact that the latent space is not densely populated during uncertainty estimation. Indeed, the 566 samples in $\bar{Z}$ might not be enough to produce optimal uncertainty maps. This is apparent in Fig. 2 where the uncertainty maps for the JSRT samples are less smooth than the other datasets that have more latent vectors. Different techniques such as data augmentation or latent space rejection sampling [18] are plausible solutions.

5 Discussion and conclusion

While empirical results indicate that all methods perform to a certain degree, qualitative results in Fig. 2 show that most SOTA methods predict uncertainty around the prediction edges. While this may constitute a viable uncertainty prediction when the predicted segmentation map is close to the groundtruth,

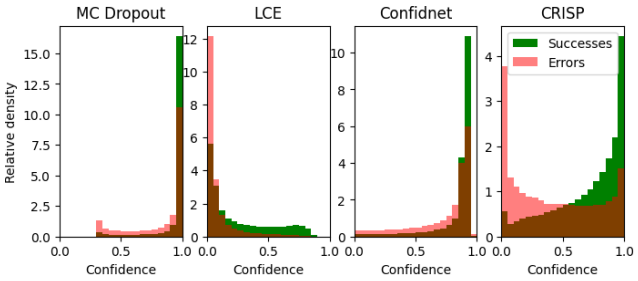

Fig. 3. Histograms of well classified pixels (Successes) and mis-classified pixels (Errors) for different methods on the HMC-QU dataset.

these uncertainty estimates are useless for samples with large errors. Whereas in other datasets and modalities, the uncertainty represents the probability of a structure being in an image and at a given position, lung X-Ray and cardiac ultrasound structures are always present and are of regular shape and position. This makes the task of learning uncertainty during training challenging as few images in the training set produce meaningful errors. Compared to other approaches, *CRISP* leverages the information contained in the dataset to a greater degree and accurately predicts uncertainty in even the worst predictions.

To conclude, we have presented a method to identify uncertainty in segmentation by exploiting a joint latent space trained using contrastive learning. We have shown that SOTA methods produce sub-optimal results due to the lack of variability in segmentation quality during training when segmenting regular shapes. We also highlighted this with the naïve *Edge* method. However, due to its reliance on anatomical priors, *CRISP* can identify uncertainty in a wide range of segmentation predictions.

References

1. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E.: Phiseg: Capturing uncertainty in medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 119–127. Springer International Publishing, Cham (2019)
2. 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 32, pp. 2902–2913. Curran Associates, Inc. (2019)
3. Degerli, A., Zabihi, M., Kiranyaz, S., Hamid, T., Mazhar, R., Hamila, R., Gabbouj, M.: Early detection of myocardial infarction in low-quality echocardiography. IEEE Access **9**, 34442–34453 (2021)
4. DeVries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks. arXiv preprint arXiv:1802.04865 (2018)
5. Gal, Y., Ghahramani, Z.: Bayesian convolutional neural networks with bernoulli approximate variational inference. ArXiv **abs/1506.02158** (2015)
6. 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 - Volume 48. p. 1050–1059. ICML'16, JMLR.org (2016)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330. PMLR (06–11 Aug 2017)
8. Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X., Thoma, G.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative Imaging in Medicine and Surgery **4**, 475 (2014)
9. Jungo, A., Reyes, M.: Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 48–56. Springer International Publishing, Cham (2019)
10. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30, pp. 5574–5584. Curran Associates, Inc. (2017)
11. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
12. 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, May 7-9, 2015, Conference Track Proceedings (2015)
13. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S.M.A., Jimenez Rezende, D., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. In: Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)

14. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
15. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., Lartizien, C., D'hooge, J., Lovstakken, L., Bernard, O.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019)
16. Mardia, K.V., Jupp, P.E.: Directional Statistics. Wiley (1999)
17. Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., de Marvao, A., Dawes, T., O'Regan, D.P., Kainz, B., Glocker, B., Rueckert, D.: Anatomically constrained neural networks (acnns): Application to cardiac image enhancement and segmentation. IEEE Transactions on Medical Imaging **37**(2), 384–395 (2018)
18. Painchaud, N., Skandarani, Y., Judge, T., Bernard, O., Lalande, A., Jodoin, P.M.: Cardiac segmentation with strong anatomical guarantees. IEEE Transactions on Medical Imaging **39**(11), 3703–3713 (2020)
19. Pakdaman Naeini, M., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. Proceedings of the AAAI Conference on Artificial Intelligence **29**(1) (Feb 2015)
20. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: Enet: A deep neural network architecture for real-time semantic segmentation. CoRR **abs/1606.02147** (2016)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. CoRR **abs/2103.00020** (2021)
22. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
23. 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. American Journal of Roentgenology **174**, 71–74 (2000)
24. Taylor, C.C.: Automatic bandwidth selection for circular density estimation. Computational Statistics and Data Analysis **52**(7), 3493–3500 (2008)
25. Zotti, C., Humbert, O., Lalande, A., Jodoin, P.M.: Gridnet with automatic shape prior registration for automatic mri cardiac segmentation. MICCAI - ACDC Challenge (2017)