Using Soft Labels to Model Uncertainty in Medical Image Segmentation

João Lourenço-Silva
Arlindo L. Oliveira

Instituto Superior Técnico / INESC-ID
{joao.lourenco.silva,arlindo.oliveira}@tecnico.ulisboa.pt

Abstract. Medical image segmentation is inherently uncertain. For a given image, there may be multiple plausible segmentation hypotheses, and physicians will often disagree on lesion and organ boundaries. To be suited to real-world application, automatic segmentation systems must be able to capture this uncertainty and variability. Thus far, this has been addressed by building deep learning models that, through dropout, multiple heads, or variational inference, can produce a set - infinite, in some cases - of plausible segmentation hypotheses for any given image. However, in clinical practice, it may not be practical to browse all hypotheses. Furthermore, recent work shows that segmentation variability plateaus after a certain number of independent annotations, suggesting that a large enough group of physicians may be able to represent the whole space of possible segmentations. Inspired by this, we propose a simple method to obtain soft labels from the annotations of multiple physicians and train models that, for each image, produce a single well-calibrated output that can be thresholded at multiple confidence levels, according to each application's precision-recall requirements. We evaluate our method on the MICCAI 2021 QUBIQ challenge, showing that it performs well across multiple medical image segmentation tasks, produces well-calibrated predictions, and, on average, performs better at matching physicians' predictions than other physicians.

Keywords: Uncertainty Estimation · Medical Image Segmentation · Soft Labels.

1 Introduction

Accurate segmentation of medical images is crucial in diagnosing and planning the treatment of multiple pathologies. Nevertheless, it is also very laborious and time-consuming, spurring great interest in the development of automatic segmentation mechanisms.

In the last few years, deep learning systems have achieved high performance in the segmentation of several organs and anatomical structures [30]. However, most methods do not account for the uncertainty inherent to these tasks. For a given image, there may be multiple plausible segmentations, and physicians will

often disagree on the zones of interest and their contours. Thus, models should be able to capture uncertainty and express it in their predictions. Otherwise, they risk biasing physicians, which may lead to misdiagnosis and sub-optimal treatment.

To date, most work on uncertainty estimation in medical image segmentation focuses on being able to produce multiple plausible outputs for a given image [1,11,22,23,33]. However, in clinical practice, it may impractical to browse all hypotheses. Furthermore, recent research [15] shows that, even though segmentation variability increases with the number of annotators, it plateaus after a certain data and task-dependent number of independent annotations, implying that, although multiple plausible segmentations exist for a given input, they can be encompassed by the annotations of a sufficiently large group of physicians.

In this work, we follow a trend orthogonal to that of previous work. Rather than aiming to build probabilistic models that can produce various plausible hypotheses, we propose to train deterministic models on soft labels built from the annotations of multiple physicians. We evaluated our method on datasets from the MICCAI 2021 QUBIQ challenge. The results showed that it performs well compared to alternative approaches and produces well-calibrated outputs across a range of medical image segmentation tasks and imaging modalities.

2 Related Work

Monte Carlo Dropout is a technique used by early approaches for uncertainty estimation in image segmentation, which use dropout [42] over spatial features to induce probability distributions over the models' outputs [16,17], allowing the drawing of multiple samples at test-time. However, these methods quantify uncertainty pixel-wise, leading them to produce spatially inconsistent segmentation hypotheses.

Ensembles [26,28] and *Multi-Head Neural Networks* [13,29,38] are simple methods to produce plausible and consistent output hypotheses. While they may not be able to capture diversity and learn rare variants when ensemble members and network heads are trained independently, that can be circumvented by joint training on an oracle loss [7], which only accounts for the lowest-error prediction. The main disadvantages of these approaches are their poor scaling with the number of hypotheses and the latter's requirement to be set at training time.

Variational Bayesian Inference methods, like the sPU-Net [23], HPU-Net [22] and PHiSeg [1] combine conditional variational autoencoders [19,20,36,41] with U-Net-based [37] networks to model the distribution of segmentations given an input image. Input images are encoded into multivariate normal latent spaces that the decoder samples at test-time to produce arbitrarily complex and diverse segmentation hypotheses. However, this approach requires a training-only posterior network, and the placement of the latent variables within the model entails a partial forward pass for each output hypothesis. Recent work addresses these issues with a more constrained low-rank multivariate normal distribution over

the logit space, which avoids the use of a posterior network and allows efficient sampling without compromising performance [33]. Other work [11] extends the sPU-Net using variational dropout [21] to predict epistemic uncertainty, and intergrader variability as a target for supervised aleatoric uncertainty estimation.

3 Method

3.1 Motivation

For many years, due to optimization difficulties and lack of computing power, it was very difficult to train deep neural networks. Recently, however, better hardware and new architectural components, such as batch normalization [14] and residual connections [9], have enabled training increasingly deeper and wider networks [9,12,25,40,43,44], which achieve high performance in a wide range of tasks. Nevertheless, unlike their shallower and less accurate counterparts from the past, like the LeNet [27], modern neural networks are poorly calibrated, leading to a situation where the probabilities they assign to classes do not reflect their real likelihoods. Though a set of factors such as model capacity, batch normalization and lack of regularization have been put forward as possible causes for miscalibration, the use of hard labels is probably one of the causes at the heart of the problem. When training neural networks to make their predictions match a set of hard labels, which are often the only available ones, it is unreasonable to interpret them as probabilistic models and expect them to output well-calibrated confidence values.

A simple approach to address this issue would be to use soft labels conveying information about real class likelihood. These would not only allow modeling the uncertainty inherent to the data, but would also be likely to enable faster and more data-efficient training. As noted in seminal work on knowledge distillation by Hinton et al. [10], compared to hard targets, high entropy soft targets provide much more information per training case and much less variance in the gradient across samples, allowing models to be trained on much fewer data and with significantly higher learning rates. In fact, even noisy soft labels can be of great value, as showcased by recent research on semi-supervised learning [35,45].

3.2 Proposed Method

We propose to use soft labels built from multiple annotations to model uncertainty and address network calibration. Given a set of labels for an image, we average them to produce probabilistic ground-truth masks. Beyond expressing real physicians' uncertainty about zones of interest and their contours, these high entropy soft labels enable our models to enjoy the advantages pointed out by Hinton et al. [10]. Additionally, note that, for binary variables, the variance can be obtained from the mean[1]. Therefore, although it can be used as an auxiliary supervision signal, it does not need to be predicted directly by the models.

[1] $X = X^2 \implies \mathbf{V}(X) = \mathbf{E}(X^2) - \mathbf{E}(X)^2 = \mathbf{E}(X) - \mathbf{E}(X)^2$.

Depending on the number of annotations per image, the set S of possible ground-truth probabilities will be more or less granular. Formally, let A be the number of annotations, then $S = \{\frac{i}{A} : i \in \mathbb{N} \wedge i \leq A\}$. More annotations per image result in smoother and less noisy ground-truth, which, intuitively, should allow better segmentation performance and uncertainty modeling.

Following Hinton et al. [10], we minimize the cross-entropy between the probabilities predicted by the model, p, and the ground-truth soft targets, g. Note that the dice loss (DL) function, commonly used in segmentation tasks, is not suitable to be used with soft targets. Let C be the number of classes and N the total number of pixels. DL can be defined as

$$\mathrm{DL}(p, g) = 1 - 2\frac{\sum_{c=1}^{C}\sum_{i=1}^{N}[g_{ci}p_{ci}]}{\sum_{c=1}^{C}\sum_{i=1}^{N}[g_{ci} + p_{ci}]}. \tag{1}$$

Without loss of generalization, consider the binary classification of a single-pixel image. For $g > 0$, DL is a monotonically decreasing function of p. Hence, for $p \in [0, 1]$, the minimum DL will be obtained for $p = 1$. Consequently, the model is encouraged to binarize its outputs and does not learn to predict uncertainty. This problem could be mitigated by measuring DL at multiple confidence thresholds. However, it would require defining the optimal number of thresholds and their values. Thus, we opt for the more principled cross-entropy loss. Researching overlap-based loss functions that, unlike DL, can be used to match soft targets, is a possible direction for future work.

3.3 Model Architecture

We conducted our experiments using a U-Net decoder [37] with 16-channel feature maps at the highest resolution level. As encoder, we used an EfficientNet-B0 [44], a good-performing feature extractor that makes for segmentation models whose compute scales better with input image size than the default U-Net encoder and other popular architectures [39], which can be of great value when segmenting high pixel-count medical images.

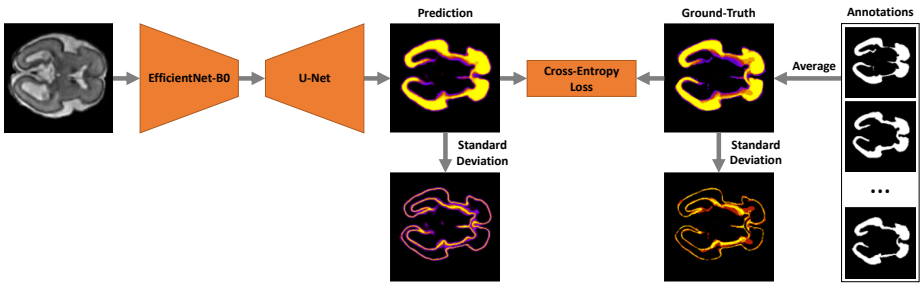


Fig. 1: Pipeline of the proposed method and segmentation model architecture. Our method allows modeling annotations without overfitting them, resulting in predictions with smooth probability and standard deviation variations.

4 Experimental Setup

4.1 Datasets and Data Augmentation

We evaluated our method on datasets from the MICCAI 2021 QUBIQ challenge[2], composed of CTs and MRIs with multiple annotations per case. Except for the four-channel brain tumor MRIs, all images are single-channel. Due to architectural constraints, the images were cropped during training to ensure their dimensions were multiples of 32. Crop sizes were set empirically to balance segmentation performance and training time. Table 1 summarizes the details of each dataset.

Table 1: Summary of the MICCAI 2021 QUBIQ challenge datasets used.

Dataset	Modality	Tasks	Annotators	Size		Cases	
				Slice	Crop	Train	Validation
Brain Growth	2D MRI	1	7	256^2	256^2	34	5
Brain Tumor	2D MRI	3	3	240^2	224^2	28	4
Kidney	2D CT	1	3	497^2	320^2	20	4
Prostate	2D MRI	2	6†	640^2 640×960	480^2	48	7

Data augmentation is performed online and consists of the following sequentially applied random transformations: 1) -10% to 10% horizontal and vertical translation; 2) $-15°$ to $15°$ rotation; 3) -10% to 10% zoom; 3) horizontal flip with 50% probability; 4) vertical flip with 0% probability for the kidney dataset and 50% for the remaining. Following Fort et al. [4], we draw multiple augmentation samples per image in a growing batch regime. However, unlike them, we keep the original images in the batch, since we observed this improves performance slightly. Specifically, we augment each batch with three transformations of each image.

4.2 Evaluation Metrics

Segmentation is a spatially structured prediction task. Therefore, segmentation models' calibration must be assessed using metrics that take spatial structure into account. Hence, instead of common pixel-wise calibration metrics [2,5,34], we measured overlap and surface distance at multiple confidence thresholds. Additionally, following previous work [22,23,33], we used the generalized energy distance to measure the statistical distance between the ground-truth masks and our models' predictions at the 50% confidence threshold. Below, we briefly de-

[2] Challenge information and datasets available at this https url.

† Except for three cases, which only have 5 annotations.

scribe the more domain-specific metrics, which may be unknown to some readers. In addition to those, we also measured precision and recall.

Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) are measures of the overlap between two segmentations. DSC is the ratio between the area of overlap and the sum of the two areas, and is equal to $1 - \mathrm{DL}$ (see Eq. 1). IoU is the ratio between the overlap and union areas. When both segmentation masks are empty, we set DSC and IoU to 1.

95% Hausdorff Distance (95% HD). Given two sets of points - two segmentation masks, in this context -, the Hausdorff distance is the maximum distance from a point in one set to the closest point in the other set. The 95% HD disregards the 5% most distant pairs of points, ignoring outliers, but still providing a measure of the longest distance between the two sets of points.

Generalized Energy Distance (D_{GED}) measures the statistical distance between probability distributions. As long as the function $d(\cdot, \cdot)$ is a metric, so is D_{GED}. Following previous work [22,23,33] we define $d(x, y) = 1 - \mathrm{IoU}(x, y)$, which has been proven to be a metric [24,31]. Given the distributions of ground-truth segmentations, p, and predicted segmentations, $\hat{p}$, D^2_{GED} is defined by

$$D^2_{\mathrm{GED}}(p, \hat{p}) = 2\mathbb{E}_{y\sim p, \hat{y}\sim\hat{p}}[d(y, \hat{y})] - \mathbb{E}_{y,y'\sim p}[d(y, y')] - \mathbb{E}_{\hat{y},\hat{y}'\sim\hat{p}}[d(\hat{y}, \hat{y}')]. \quad (2)$$

Since our models are deterministic, the formula above can be simplified to

$$D^2_{\mathrm{GED}}(p, \hat{p}) = 2\mathbb{E}_{y\sim p}[d(y, \hat{y})] - \mathbb{E}_{y,y'\sim p}[d(y, y')], \quad (3)$$

where $\hat{y}$ is the predicted segmentation mask. The first term of Eq. 3 is the average distance between predicted and ground-truth annotations, and the second can be interpreted as a measure of ground-truth segmentation diversity.

4.3 Implementation and Training Details

We used encoders pre-trained on ImageNet [3]. Decoder hidden and output layers were initialized using Kaiming [8] and Xavier initialization [6], respectively. Models were trained for 150 epochs - 180 for the third brain tumor task -, using batches of 8 images. As optimizer, we used Adam [18], with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and no weight decay. Learning rates were initialized at 10^{-2} and decreased to 10^{-4} using a cosine annealing schedule [32]. To ensure reproducibility, we trained and tested each model three times, obtaining similar results across all runs. All experiments were conducted using public PyTorch implementations [46] under an MIT license.

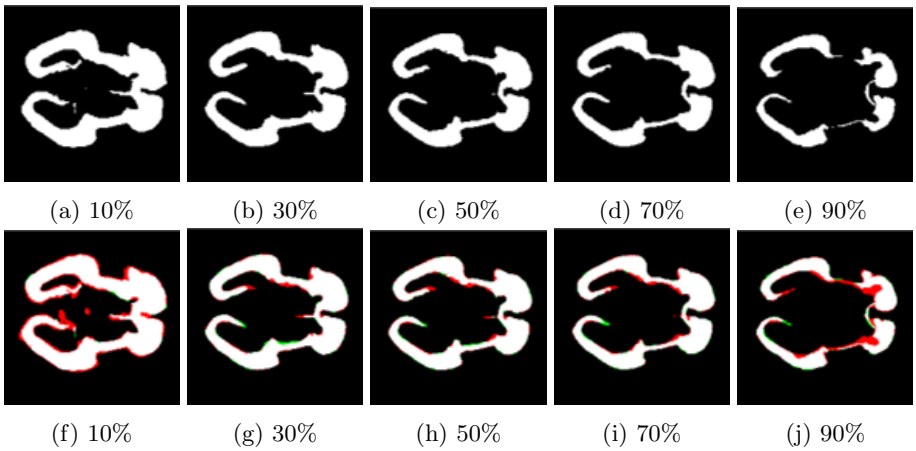
5 Results

To assess calibration, we started by measuring DSC, precision, recall and 95% HD between model predictions and averaged ground-truth masks at multiple

confidence thresholds. Specifically, we used thresholds ranging from 10% to 90% confidence, with a step size of 10%.

The averages of these metrics across thresholds are reported in Table 2. Note that the results should be interpreted taking into account image sizes, reported in Table 1, and ground-truth area to image size ratios. For example, the prostate tasks' regions of interest are relatively large, making them easy to overlap and leading to a high DSC. However, the large size of the images - 640×640 and 640×960 - and structures leads to apparently high 95% HDs, compared to those of other tasks. On the other hand, the tiny structures in the third brain tumor segmentation task are challenging to detect and segment, hence the relatively low DSC and recall. Nevertheless, the small image size and object make low 95% HDs relatively easy to achieve.

Apart from the second brain tumor segmentation task, discussed in more detail below, our method achieves high segmentation performance in all other tasks. Furthermore, the low standard deviations indicate consistent performance across confidence thresholds and, therefore, good model calibration. This can be visualized in Figure 2, where despite performing better at confidence values near 50%, our model does well across a range of thresholds, allowing the latter to be selected according to each application's precision and recall requirements.



(a) 10% (b) 30% (c) 50% (d) 70% (e) 90%

(f) 10% (g) 30% (h) 50% (i) 70% (j) 90%

Fig. 2: Annotation average (top row) and model prediction (bottom row) thresholded at multiple confidence thresholds. In the bottom row, the zones in red and green correspond to false positives and false negatives, respectively.

To further assess calibration and uncertainty modeling, we followed previous work [22,23,33] and measured the generalized energy distance (D_{GED}) between the multiple ground-truth annotations and the models' predictions at the 50% confidence threshold. Additionally, we measured the expected value of the DSC between predictions at 50% confidence and each physician's annotations - which is not equivalent to measuring the DSC between model predictions and average ground-truth masks at the same threshold. Results are reported in Table 3.

Table 2: Dice score, precision, recall, 95% Hausdorff distance and ground-truth area to image size ratio, averaged across confidence thresholds ranging from 10% to 90%, with a step size of 10%. Results presented as mean ± standard deviation.

Task	Dice Score [%]	Precision [%]	Recall [%]	95% HD [pixels]	$\frac{\text{Ground-Truth Area}}{\text{Image Size}}$ [%]
Brain Growth	93.19 ± 1.83	93.33 ± 2.55	93.16 ± 1.84	3.53 ± 0.73	8.33 ± 1.57
Brain Tumor 1	92.90 ± 1.09	93.79 ± 1.01	89.04 ± 2.30	6.84 ± 2.37	3.50 ± 0.31
Brain Tumor 2	65.06 ± 20.27	67.18 ± 21.75	63.74 ± 19.57	15.15 ± 13.18	1.33 ± 1.47
Brain Tumor 3	85.73 ± 4.88	97.67 ± 1.78	78.34 ± 6.67	2.27 ± 2.15	0.29 ± 0.04
Kidney	96.04 ± 1.43	96.57 ± 1.64	95.65 ± 2.53	7.67 ± 3.65	1.90 ± 0.12
Prostate 1	95.64 ± 0.04	94.13 ± 1.05	97.43 ± 0.78	16.18 ± 7.58	8.87 ± 0.68
Prostate 2	93.56 ± 4.58	91.78 ± 4.09	95.70 ± 5.32	12.48 ± 5.46	5.49 ± 0.58

Except for the second brain tumor segmentation task, the remaining tasks' D^2_{GED} is very low, meaning the models' predictions closely match the distributions of ground-truth annotations. In fact, in most cases, the expected value of the IoU distance between model predictions and ground-truth annotations is lower than that of the IoU distance between annotations by different physicians, indicating that, on average, our models do a better task at matching a physician's annotations than other physicians do, which is remarkable, especially considering the small dimension of the datasets, composed of 20 to 48 samples.

The overall worse performances are registered for the second and third brain tumor segmentation tasks. For the latter, the lower performance is largely justified by the difficulty of segmenting its tiny structures. However, in the former case, the difficulty lies in the high variability between ground-truth masks. Even though three annotators may not be enough to represent all the segmentation hypotheses in this task, we suspect the annotations from one of the physicians to be incorrect, as their average IoU distance to the others is 94.15%, and the distance between the other two physicians' annotations is only 19.82%.

Finally, note that the expected value of the DSC between predictions at 50% confidence and each physician's annotations is generally high, meaning that beyond matching the averaged predictions of multiple physicians, the masks produced by our models also match individual physicians' annotations well.

Table 3: From the 2^{nd} to the 5^{th} column: squared generalized energy distance; expected IoU distance between predictions and ground-truth masks; ground-truth diversity; expected DSC between predictions at 50% confidence and each physician's annotations. Results presented as mean ± standard deviation.

Task	D^2_{GED}	$\mathbb{E}_{y\sim p}[1-\text{IoU}(y,\hat{y})]$	$\mathbb{E}_{y,y'\sim p}[1-\text{IoU}(y,y')]$	$\mathbb{E}_{y\sim p}[\text{DSC}(y,\hat{y})]$
Brain Growth	0.1323 ± 0.0077	0.1876 ± 0.0087	0.2429 ± 0.0124	89.63 ± 01.60
Brain Tumor 1	0.1455 ± 0.0622	0.1393 ± 0.0492	0.1330 ± 0.0497	92.39 ± 03.78
Brain Tumor 2	0.6731 ± 0.5631	0.6843 ± 0.3167	0.6955 ± 0.0835	34.27 ± 43.45
Brain Tumor 3	0.2515 ± 0.1928	0.2272 ± 0.1523	0.2030 ± 0.1306	86.25 ± 10.22
Kidney	0.0613 ± 0.0077	0.0814 ± 0.0105	0.1015 ± 0.0150	95.73 ± 01.59
Prostate 1	0.0950 ± 0.0692	0.1096 ± 0.0569	0.1242 ± 0.0478	94.07 ± 03.80
Prostate 2	0.0988 ± 0.0907	0.1431 ± 0.0679	0.1874 ± 0.0828	90.99 ± 15.25

6 Discussion

We proposed a new way of approaching uncertainty modeling in image segmentation. Instead of building models that learn independently from the annotations of multiple physicians and can produce multiple segmentation hypotheses for a given image, we train deliberately deterministic models on the joint predictions of physician ensembles, using the averages of their predictions as soft labels.

We evaluated our method on datasets from the MICCAI 2021 QUBIQ challenge, showing that it results in well-calibrated models that, on average, match physicians' predictions better than other physicians. The results show that our system exhibits a good performance on this task, competitive with other approaches.

A limitation of our method is that annotation averaging leads to loss of information about the spatial correlation between pixels, which could be valuable to train better and more accurate models. Additionally, by relying on annotation averaging to model uncertainty, our method's applicability is limited to datasets with more than one label per image, and the quality of its results is particularly dependent on the granularity of the ground-truth probabilistic targets.

In future work, we plan to test this technique on larger datasets and more challenging tasks with multiple classes, possibly including problems outside the scope of medical image segmentation. Additionally, we intend to investigate if the soft labels used in our method allow the more data-efficient training generic soft labels do [10]. Finally, given soft labels' role in recent teacher-student semi-supervised learning methods [45,35], we plan to assess if networks trained on soft labels, like ours, can be better teachers than those trained on hard labels.

Acknowledgments

This work was supported by national funds through Fundação para a Ciência e Tecnologia (FCT), under the project with reference UIDB/50021/2020 and the project PRELUNA, with the reference PTDC/CCI-INF/4703/2021.

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: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 119–127. Springer (2019)
2. Brier, G.W., et al.: Verification of forecasts expressed in terms of probability. Monthly Weather Review **78**(1), 1–3 (1950)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)
4. Fort, S., Brock, A., Pascanu, R., De, S., Smith, S.L.: Drawing multiple augmentation samples per image during training efficiently decreases test error. arXiv preprint 2105.13343 (2021)

5. Friedman, J., Hastie, T., Tibshirani, R., et al.: The elements of statistical learning, vol. 1. Springer Series in Statistics New York (2001)
6. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
7. Guzman-Rivera, A., Batra, D., Kohli, P.: Multiple choice learning: learning to produce multiple structured outputs. In: Advances in Neural Information Processing Systems. vol. 1, p. 3 (2012)
8. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1026–1034. IEEE (2015)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint 1503.02531 (2015)
11. Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., Welling, M.: Supervised uncertainty quantification for segmentation with multiple annotations. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 137–145. Springer (2019)
12. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)
13. Ilg, E., Çiçek, Ö., Galesso, S., Klein, A., Makansi, O., Hutter, F., Brox, T.: Uncertainty estimates for optical flow with multi-hypotheses networks. arXiv preprint 1802.07095 p. 81 (2018)
14. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456 (2015)
15. Joskowicz, L., Cohen, D., Caplan, N., Sosna, J.: Inter-observer variability of manual contour delineation of structures in ct. European Radiology **29**(3), 1391–1399 (2019)
16. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint 1511.02680 (2015)
17. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? arXiv preprint 1703.04977 (2017)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint 1412.6980 (2014)
19. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: Advances in Neural Information Processing Systems. pp. 3581–3589 (2014)
20. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint 1312.6114 (2013)
21. Kingma, D.P., Salimans, T., Welling, M.: Variational dropout and the local reparameterization trick. Advances in Neural Information Processing Systems **28**, 2575–2583 (2015)
22. Kohl, S.A., Romera-Paredes, B., Maier-Hein, K.H., Rezende, D.J., Eslami, S., Kohli, P., Zisserman, A., Ronneberger, O.: A hierarchical probabilistic u-net for modeling multi-scale ambiguities. arXiv preprint 1905.13077 (2019)

23. Kohl, S.A., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K.H., Eslami, S., Rezende, D.J., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. arXiv preprint 1806.05034 (2018)
24. Kosub, S.: A note on the triangle inequality for the Jaccard distance. Pattern Recognition Letters **120**, 36–38 (2019)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems **25**, 1097–1105 (2012)
26. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. arXiv preprint 1612.01474 (2016)
27. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
28. Lee, S., Prakash, S.P.S., Cogswell, M., Ranjan, V., Crandall, D., Batra, D.: Stochastic multiple choice learning for training diverse deep ensembles. In: Advances in Neural Information Processing Systems. pp. 2119–2127 (2016)
29. Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., Batra, D.: Why M heads are better than one: training a diverse ensemble of deep networks. arXiv preprint 1511.06314 (2015)
30. Lei, T., Wang, R., Wan, Y., Zhang, B., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: a survey. arXiv preprint 2009.13120 (2020)
31. Lipkus, A.H.: A proof of the triangle inequality for the Tanimoto distance. Journal of Mathematical Chemistry **26**(1), 263–265 (1999)
32. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint 1608.03983 (2016)
33. Monteiro, M., Folgoc, L.L., de Castro, D.C., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., Glocker, B.: Stochastic segmentation networks: modelling spatially correlated aleatoric uncertainty. arXiv preprint 2006.06015 (2020)
34. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using Bayesian binning. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
35. Pham, H., Dai, Z., Xie, Q., Luong, M.T., Le, Q.V.: Meta pseudo labels (2021)
36. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: International Conference on Machine Learning. pp. 1278–1286. PMLR (2014)
37. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 234–241. Springer (2015)
38. Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., Hager, G.D.: Learning in an uncertain world: representing ambiguity through multiple hypotheses. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3591–3600 (2017)
39. Silva, J.L., Menezes, M.N., Rodrigues, T., Silva, B., Pinto, F.J., Oliveira, A.L.: Encoder-decoder architectures for clinically relevant coronary artery segmentation. arXiv preprint 2106.11447 (2021)
40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint 1409.1556 (2014)
41. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in Neural Information Processing Systems **28**, 3483–3491 (2015)

42. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**(1), 1929–1958 (2014)
43. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2015)
44. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114 (2019)
45. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10687–10698 (2020)
46. Yakubovskiy, P.: Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch (2020)