

Epistemic and Heteroscedastic Uncertainty Estimation in Retinal Blood Vessel Segmentation

Pedro Costa¹, Asim Smailagic², Jaime S. Cardoso³, Aurélio Campilho⁴




¹Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal (up201000588@edu.fe.up.pt) ORCID 0000-0002-9528-1292; ²Institute for Complex Engineered Systems, Carnegie Mellon University, Pittsburgh, PA 15213, USA (asim@cs.cmu.edu) ORCID 0000-0001-8524-997X; ³INESC TEC-Institute for Systems and Computer Engineering, Technology and Science, Faculty of Engineering campus, Rua Dr. Roberto Frias, Building I, 4200-465 Porto, Portugal; Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal (jaime.cardoso@inesctec.pt) ORCID 0000-0002-3760-2473; ⁴Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal (campilho@fe.up.pt) ORCID 0000-0002-5317-6275

Abstract

Current state-of-the-art medical image segmentation methods require high quality datasets to obtain good performance. However, medical specialists often disagree on diagnosis, hence, datasets contain contradictory annotations. This, in turn, leads to difficulties in the optimization process of Deep Learning models and hinder performance. We propose a method to estimate uncertainty in Convolutional Neural Network (CNN) segmentation models, that makes the training of CNNs more robust to contradictory annotations. In this work, we model two types of uncertainty, heteroscedastic and epistemic, without adding any additional supervisory signal other than the ground-truth segmentation mask. As expected, the uncertainty is higher closer to vessel boundaries, and on top of thinner and less visible vessels where it is more likely for medical specialists to disagree. Therefore, our method is more suitable to learn from datasets created with heterogeneous annotators. We show that there is a correlation between the uncertainty estimated by our method and the disagreement in the segmentation provided by two different medical specialists. Furthermore, by explicitly modeling the uncertainty, the Intersection over Union of the segmentation network improves 5.7 percentage points.

Author Keywords. Diabetic Retinopathy, Blood Vessel Segmentation, Uncertainty Estimation, Deep Learning, Convolutional Neural Networks.

Type: Research Article

 Open Access  Peer Reviewed  CC BY

1. Introduction

Retinal vasculature provides information about many conditions including vision threatening diseases, such as Diabetic Retinopathy, and cardiovascular diseases, such as coronary artery disease (Nguyen and Wong 2009). A commonly used biomarker to diagnose these diseases is the Ratio between Arteriolar and Venular diameters (AVR) (Dashtbozorg, Mendonça, and Campilho 2013). Therefore, the task of segmenting the blood vessels in retinal images is an important first step towards automatically diagnosing these diseases.

Current state-of-the-art blood vessel segmentation methods rely on Convolutional Neural Networks (CNNs) (Imran et al. 2019; Meyer et al. 2017; Meyer et al. 2018) which typically require large high-quality datasets to achieve best performance. The best performing methods (Meyer et al. 2017; Meyer et al. 2018) typically use a U-Net style architecture (Ronneberger, Fischer, and Brox 2015) to segment the input images. The U-Net consists of an encoder-decoder CNN, with skip connections between the encoder and the decoder layers, to

help preserve fine details from the input image in the output segmentation mask, such as the edges of the object of interest.

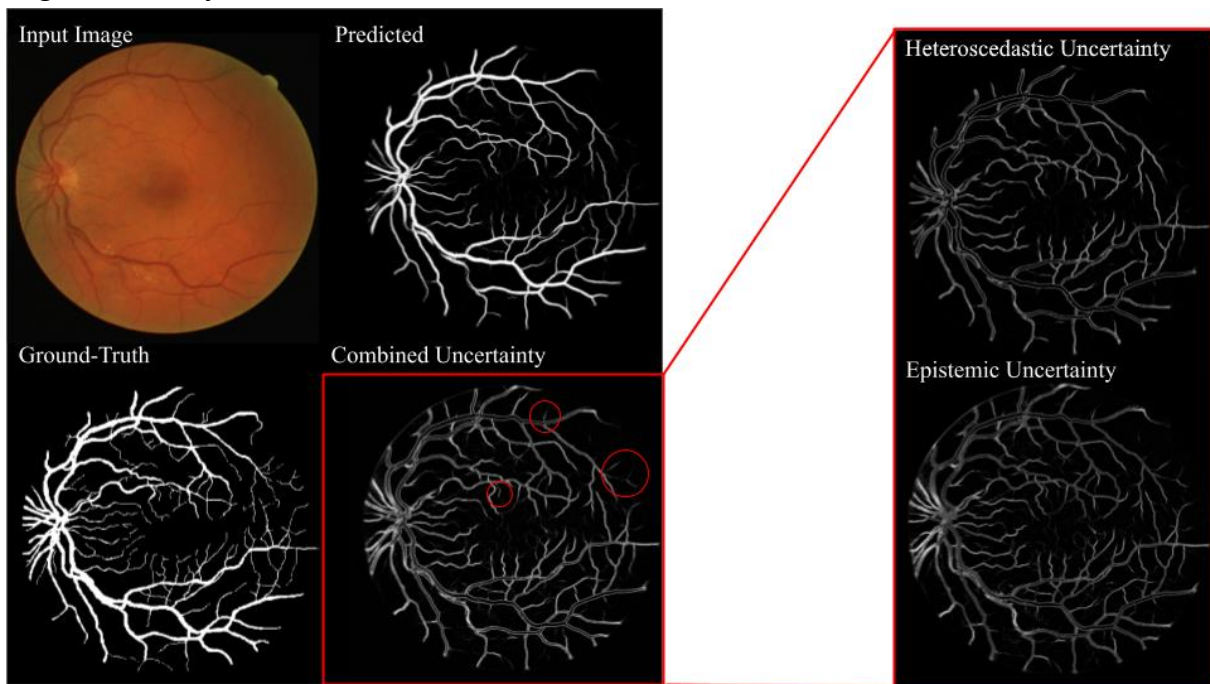


Figure 1: We propose a system that segments blood vessels in eye fundus images. Additionally, the system models two types of uncertainty in the prediction: heteroscedastic and epistemic uncertainty

However, medical doctors often disagree on diagnosis ([Krause et al. 2018](#); [Wanderley et al. 2019](#)) leading to inconsistent annotations that may hinder performance. For instance, in the case of Diabetic Retinopathy grading, specialists agree in only 71% of the images on one dataset with 406 eye fundus images. This problem may be even bigger for segmentation tasks, where each pixel of the image is annotated. For the case of a blood vessel segmentation task, 2 different annotators agree in only 60% of the pixels that were labeled as containing blood vessels by any annotator ([Lampert, Stumpf, and Gançarski 2016](#)). This number can reduce drastically as the number of different annotators increases. In the task of segmenting fissures in high resolution images acquired by an unmanned aerial vehicle, 13 different annotators only agreed in 0.6979% of the pixels marked as fissures by any annotator ([Lampert, Stumpf, and Gançarski 2016](#)).

To solve this issue, it is common to have images annotated by multiple doctors and then have a committee reach a consensus for each image, but this reduces the total size of the dataset, hence, data variability. One possible solution to this problem is to estimate the uncertainty in the model's predictions ([Kendall and Gal 2017](#)). Recently, uncertainty estimation in medical imaging has attracted much interest ([Awate, Garg, and Jena 2019](#); [Galdran et al. 2019](#); [Garifullin, Lensu, and Uusitalo 2020](#); [Tanno et al. 2017](#); [Wang et al. 2019](#)).

These methods can be divided in two main approaches: domain knowledge and Bayesian approaches. Some methods pre-process the segmentation masks to include an "uncertain" class using domain knowledge. For instance, for the task of segmenting arteries and veins in retinal images, crossings in the vasculature and thin blood vessels can be labeled as uncertain ([Galdran et al. 2019](#)). However, most existing methods aim to estimate the uncertainty directly from data without any additional domain knowledge information. Some works model epistemic uncertainty through an approximate Bayesian inference by means of variational dropout ([Wang et al. 2019](#)). Other works model the heteroscedastic uncertainty by adding

noise to the predictions with an estimated diagonal covariance representing the intrinsic uncertainty (Tanno et al. 2017).

In this work, we model both epistemic and heteroscedastic uncertainty, as shown in Figure 1, without adding any additional supervisory signal other than the ground-truth segmentation masks. The ground-truth is produced by a heterogeneous group of annotators and we show that our combined uncertainty correlates with the disagreement between annotators. Moreover, simply by explicitly modeling the uncertainty during training, we are able to improve the segmentation Intersection over Union (IoU) results by 5.7 percentage points. In summary, our contributions are as follow:

- **Accuracy:** Segmentation results are improved by estimating uncertainty.
- **Second Opinion:** We show that the combined uncertainty is correlated with the disagreement between doctors.
- **Explainable:** The method estimated higher uncertainty near blood vessel edges and on top of thinner vessels.

2. Method

In this section, we will show how we segment blood vessels in retinal fundus images and how we estimate uncertainty. We will describe how to estimate two types of uncertainty, the epistemic and heteroscedastic uncertainties and, in the end, how to combine them, as shown in Figure 2.

2.1. Blood vessel segmentation

A U-Net (Ronneberger, Fischer, and Brox 2015) was used, which consists of an encoder-decoder architecture with skip connections between the encoder and the decoder. The encoder contains 8 Convolutional Layers followed by the ReLU activation function and BatchNorm (Ioffe and Szegedy 2015). Max-Pool is used after every two of these Conv-ReLU-BatchNorm blocks. For the decoder, we use 3 Conv-ReLU-BatchNorm blocks, with bilinear upsampling before each of them, followed by a Convolutional Layer with a single output unit corresponding to the predicted segmentation mask. The model was trained using Adam optimizer (Kingma and Ba 2014) with default parameters.

The per-pixel binary cross-entropy loss is used to train the segmentation model f :

$$L_i = -y_i \log(f(x_i)) - (1 - y_i) \log(1 - f(x_i)), \quad (1)$$

where x is the input image, y is the ground-truth segmentation mask and i is the pixel index. We then minimize the mean of the per-pixel loss $\frac{1}{N} \sum_i^N L_i$, where N is the number of pixels in image x .

2.2. Epistemic uncertainty

Epistemic uncertainty, also referred to as model uncertainty, accounts for the uncertainty in the model parameters. This type of uncertainty is related to the limited amount of information provided to the model and can be explained away given enough data.

We use dropout variational inference (Gal and Ghahramani 2015) to approximate epistemic uncertainty. Dropout with $p=10\%$ is added after each ReLU and then, at test time, dropout is also applied in T stochastic forward passes. The epistemic uncertainty can be defined as the predictive variance:

$$u_i = \frac{1}{T} \sum \left(f(x_i; \hat{\theta}) - E \left(f(x_i; \hat{\theta}) \right) \right)^2, \quad (2)$$

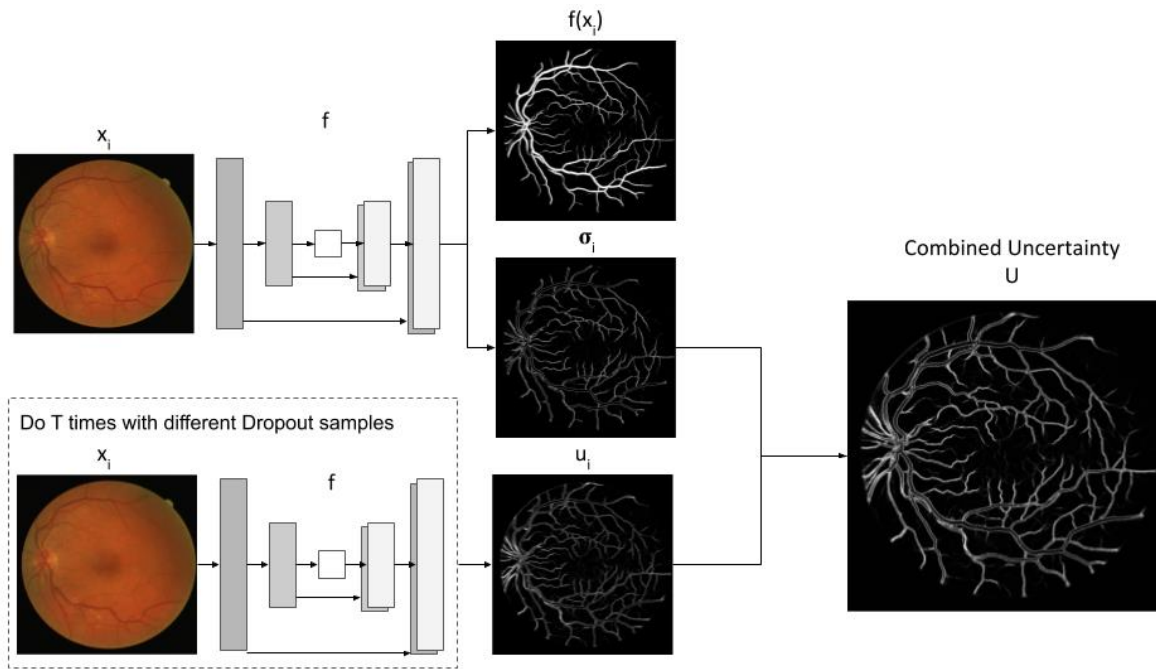


Figure 2: A U-net CNN f outputs a blood vessel segmentation mask $f(x_i)$ and heteroscedastic uncertainty σ_i . The same U-net f is run T times with dropout to compute the epistemic uncertainty u_i . Finally, σ_i and u_i are combined into a single uncertainty mask U

where $\hat{\theta}$ are the model's parameters sampled from the Dropout distribution and $E(f(x_i; \hat{\theta}))$ is the predictive mean. This uncertainty is reduced when all parameter samples $\hat{\theta}$ result in the same prediction.

2.3. Heteroscedastic uncertainty

Heteroscedastic uncertainty captures the observation noise in x_i and can not be reduced by gathering more data. For instance, for the task of vessel segmentation, the heteroscedastic uncertainty should be high near badly defined blood vessel edges. To capture this type of uncertainty, we make our model predict the log variance σ_i and modify the loss function:

$$L_U = \frac{1}{N} \sum_i e^{-\sigma_i} L_i + \sigma_i. \quad (3)$$

By multiplying the binary cross-entropy loss by $e^{-\sigma_i}$, the model is able to identify erroneous or ambiguous labels and ignore them. In order to avoid the degenerate solution of minimizing the loss by simply estimating high uncertainty in all pixels, we add the σ_i term. Therefore, the model is optimized to have low uncertainty in all predictions while, at the same time, to ignore labels where the model is likely to have high loss. We used the ELU activation function in the σ_i estimation to prevent the model from predicting very large negative values.

Finally, we combine the epistemic uncertainty and the heteroscedastic uncertainty. Before combining these two uncertainties we need to make sure they are in the same range, otherwise the uncertainty with larger range could have a bigger weight in the combined uncertainty. We normalize u and σ to have a minimum value of 0 and a maximum value of 1

in the training set: $\mu_{norm} = \frac{\mu^- \min_{\mu_i \in \mu_{train}} \mu_i}{\max_{\mu_i \in \mu_{train}} \mu_i}$, $\sigma_{norm} = \frac{\sigma^- \min_{\sigma_i \in \sigma_{train}} \sigma_i}{\max_{\sigma_i \in \sigma_{train}} \sigma_i}$, where μ_{train} and σ_{train} are

the sets containing all the epistemic and heteroscedastic uncertainties in the training set. Then, we average μ_{norm} and σ_{norm} to compute the final combined uncertainty $U = \frac{\mu_{norm} + \sigma_{norm}}{2}$.

3. Evaluation

3.1. Dataset

We evaluate our method on the publicly available DRIVE dataset (Staal et al. 2004). This dataset contains 40 images from different patients with 7 images containing signs of mild diabetic retinopathy. The dataset is equally divided into train and test sets, with 20 images in each set. Furthermore, the test set was annotated by two different observers, while the train set contains a single ground-truth annotation. This dataset is still one of the most widely used for the blood vessel segmentation task (Imran et al. 2019) and, additionally, as it contains two different annotations for each test set image, it allows us to compare the uncertainty estimation with the disagreement between observers. The images are resized to 512x512px and random translation, scale, rotation and flip operations are performed as dataset augmentation.

3.2. Segmentation results

In order to evaluate the effects of modeling the epistemic and heteroscedastic uncertainties on our U-Net model, we start by training a baseline. The baseline model consists of the U-Net as described in section 2.1 and trained with binary cross-entropy. All our models were developed using PyTorch (Paszke et al. 2019).

We compare our results with the baseline in Table 1 using three different metrics: Area Under the ROC Curve (AUC), Dice Coefficient and Intersection over Union (IoU). The ROC curve plots the sensitivity and specificity of the model at all classification thresholds and is a standard metric for classification models. The Dice Coefficient and IoU are two standard metrics for segmentation models. The Dice Coefficient doubles the intersection of the predicted and ground-truth segmentation masks and divides by the sum of the areas of the predicted and ground-truth masks. The IoU, as the name implies, divides the intersection of the predicted and ground-truth masks by the union of the two.

By modeling both the epistemic and heteroscedastic uncertainties, we are able to improve the performance of the segmentation model in all 3 metrics. The performance improvement is more significant in the Dice Coefficient and IoU as they are more robust to class imbalance. Modeling both epistemic and heteroscedastic uncertainties is better than modeling each of them individually. However, both the epistemic and heteroscedastic versions perform better than the baseline in most metrics.

	AUC	Dice	IoU
U-Net	0.982	0.768	0.623
+ Epistemic	0.983	0.800	0.667
+ Heteroscedastic	0.977	0.796	0.661
+ Combined	0.984	0.809	0.680

Table 1: Blood Vessel Segmentation Results. Modeling epistemic and heteroscedastic uncertainty improves the segmentation performance on the Drive dataset for all metrics in evaluation

3.3. Uncertainty evaluation

In order to evaluate quantitatively our combined uncertainty, we compared our estimated uncertainty with the annotators disagreement. In this work we define disagreement between annotators as the absolute difference between the two annotations $d = |y' - y''|$.

We show that there is some correlation between the estimated uncertainty and the annotators' disagreement in Figure 3. The model tends to estimate high uncertainty close to the boundaries of the blood vessels and on top of thin vessels, which is similar to the places where the annotators disagree. Furthermore, we can see that, in some situations, there is high

uncertainty in places where the model did not predict to have blood vessels. These results indicate that it could be possible to extract clinically relevant retinal biomarkers with associated uncertainty that correlates with the disagreement between specialists.

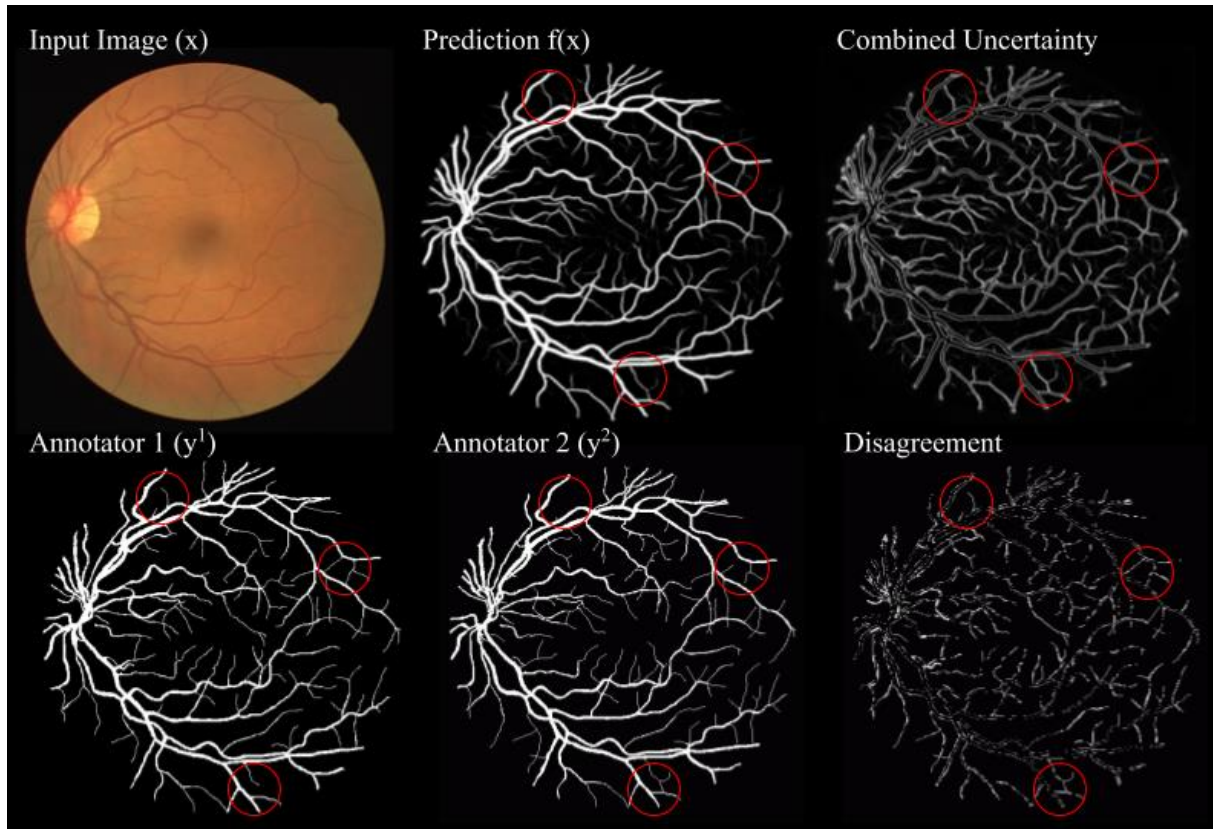


Figure 3: Comparing the disagreement between annotators and the estimated uncertainty. There is more disagreement between annotators close to the boundaries of the blood vessels and thinner vessels. The estimated uncertainty displays the same behavior. We highlight interesting regions where the uncertainty is similar to the disagreement

We evaluated quantitatively the similarity between the annotators' disagreement and the estimated uncertainty. For that, we treat the disagreement as ground-truth and compare each uncertainty map with the disagreement. The results are compiled in [Table 2](#) and show that there is some correlation between the estimated uncertainties and the disagreement.

	AUC	Dice	IoU
Epistemic	0.924	0.351	0.212
Heteroscedastic	0.927	0.378	0.233
Combined	0.929	0.377	0.232

Table 2: Comparing annotation disagreement with estimated uncertainty. We can see that there is some correlation between the annotators' disagreement and the estimated uncertainty

All estimated uncertainty maps are very similar to each other and, therefore, obtain very similar results when compared with the disagreement. The heteroscedastic uncertainty attributes higher uncertainty in the edges of the predicted blood vessels while the epistemic uncertainty may attribute higher uncertainty in regions where the segmentation model predicted to be background.

4. Conclusions

We proposed a method to estimate uncertainty in eye fundus blood vessel segmentation. We modeled both heteroscedastic and epistemic uncertainty and then combined them into a single uncertainty estimation map. The resulting uncertainty correlates with the disagreement in annotations from specialists, which indicates that our method may act as a second opinion. Moreover, this method learns from heterogeneous annotators as it predicts which pixel annotations are most likely to be annotated differently by medical doctors and includes that information in the loss function. Therefore, it may be possible to eliminate the need of having multiple annotators labeling the same images, and discussing to reach a consensus, allowing the creation of larger and more variable datasets without hindering performance.

In the future, we want to apply these ideas to multi-class segmentation problems, such as the artery-vein segmentation problem in eye fundus images. Additionally, we want to test the robustness of this method to different levels of noise.

References

- Awate, S. P., S. Garg, and R. Jena. 2019. "Estimating uncertainty in MRF-based image segmentation: A perfect-MCMC approach". *Medical Image Analysis* 55 (july): 181-96. <https://doi.org/10.1016/j.media.2019.04.014>.
- Dashtbozorg, B., A. M. Mendonça, and A. Campilho. 2013. "An automatic method for the estimation of Arteriolar-to-Venular Ratio in retinal images". In *Proceedings of CBMS 2013 - 26th IEEE International Symposium on Computer-Based Medical Systems*, 512-13. IEEE. <https://doi.org/10.1109/CBMS.2013.6627854>.
- Gal, Y., and Z. Ghahramani. 2015. "Bayesian convolutional neural networks with Bernoulli approximate variational inference". Preprint, submitted June 6, 2015. <https://arxiv.org/abs/1506.02158>.
- Galdran, A., M. Meyer, P. Costa, Mendonca, and A. Campilho. 2019. "Uncertainty-aware artery/vein classification on retinal images". In *Proceedings - 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 556-60. IEEE. <https://doi.org/10.1109/ISBI.2019.8759380>.
- Garifullin, A., L. Lensu, and H. Uusitalo. 2020. "On the uncertainty of retinal artery-vein classification with dense fully-convolutional neural networks". In *Advanced concepts for intelligent vision systems. ACIVS 2020*, 87-98. Lecture Notes in Computer Science. Springer International Publishing. https://doi.org/10.1007/978-3-030-40605-9_8.
- Imran, A., J. Li, Y. Pei, J. J. Yang, and Q. Wang. 2019. "Comparative analysis of vessel segmentation techniques in retinal images". *IEEE Access* 7: 114862-87. <https://doi.org/10.1109/ACCESS.2019.2935912>.
- Ioffe, S., and C. Szegedy. 2015. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". Preprint, submitted February 11, 2015. <https://arxiv.org/abs/1502.03167>.
- Kendall, A., and Y. Gal. 2017. "What uncertainties do we need in Bayesian deep learning for computer vision?". In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 5575-85.
- Kingma, D. P., and J. Ba. 2014. "Adam: A method for stochastic optimization". Preprint, submitted December 22, 2014. <https://arxiv.org/abs/1412.6980>.
- Krause, J., V. Gulshan, E. Rahimy, P. Karth, K. Widner, G. S. Corrado, L. Peng, and D. R. Webster. 2018. "Grader variability and the importance of reference standards for evaluating machine

- learning models for diabetic retinopathy". *Ophthalmology* 125, no. 8 (august): 1264-72. <https://doi.org/10.1016/j.ophtha.2018.01.034>.
- Lampert, T. A., A. Stumpf, and P. Gañçarski. 2016. "An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation". *IEEE Transactions on Image Processing* 25, no. 6 (june): 2557-72. <https://doi.org/10.1109/TIP.2016.2544703>.
- Meyer, M. I., P. Costa, A. Galdran, A. M. Mendonça, and A. Campilho. 2017. "A deep neural network for vessel segmentation of Scanning Laser Ophthalmoscopy images". In *Image Analysis and Recognition. ICIAR 2017*, 507-15. Lecture Notes in Computer Science, vol. 10317. https://doi.org/10.1007/978-3-319-59876-5_56.
- Meyer, M. I., A. Galdran, P. Costa, A. M. Mendonça, and A. Campilho. 2018. "Deep convolutional artery/vein classification of retinal vessels". In *Image Analysis and Recognition. ICIAR 2018*, 622-30. Lecture Notes in Computer Science, vol. 10882. Springer International Publishing. https://doi.org/10.1007/978-3-319-93000-8_71.
- Nguyen, T. T., and T. Y. Wong. 2009. "Retinal vascular changes and diabetic retinopathy". *Current Diabetes Reports* 9, no. 4: 277-83. <https://doi.org/10.1007/s11892-009-0043-4>.
- Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, et al. 2019. "PyTorch: An imperative style, high-performance deep learning library". In *Advances in Neural Information Processing Systems*, 8024-35. <https://papers.nips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html>.
- Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-Net: Convolutional networks for biomedical image segmentation". In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234-41. Lecture Notes in Computer Science, vol. 9351. Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28.
- Staal, J., M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken. 2004. "Ridge-based vessel segmentation in color images of the retina". *IEEE Transactions on Medical Imaging* 23, no. 4 (april): 501-09. <https://doi.org/10.1109/TMI.2004.825627>.
- Tanno, R., D. E. Worrall, A. Ghosh, E. Kaden, S. N. Sotiropoulos, A. Criminisi, and D. C. Alexander. 2017. "Bayesian image quality transfer with CNNs: Exploring uncertainty in dMRI super-resolution". In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, 611-19. Springer International Publishing. https://doi.org/10.1007/978-3-319-66182-7_70.
- Wanderley, D. S., T. Araújo, C. B. Carvalho, C. Maia, S. Penas, Carneiro Â, A. M. Mendonça, and A. Campilho. 2019. "Analysis of the performance of specialists and an automatic algorithm in retinal image quality assessment". In *6th ENBENG IEEE Portuguese Meeting on Bioengineering - Proceedings Book*, 1-4. <https://doi.org/10.1109/ENBENG.2019.8692506>.
- Wang, G., W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren. 2019. "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks". *Neurocomputing* 338 (april): 34-45. <https://doi.org/10.1016/j.neucom.2019.01.103>.