Gambling Adversarial Nets for Hard Sample Mining and Structured Prediction: Application in Ultrasound Thyroid Nodule Segmentation

Masoumeh Bakhtiariziabari[1,2], and Mohsen Ghafoorian[3]

[1] University of Amsterdam, The Netherlands
[2] 3DUniversum, Amsterdam, The Netherlands
mbakhtiariz@gmail.com
[3] TomTom, Amsterdam, The Netherlands

Abstract. Most real-world datasets are characterized by long-tail distributions over classes or, more generally, over underlying visual representations. Consequently, not all samples contribute equally to the training of a model and therefore, methods properly evaluating the importance/difficulty of the samples can considerably improve the training efficiency and effectivity. Moreover, preserving certain inter-pixel/voxel structural qualities and consistencies in the dense predictions of semantic segmentation models is often highly desirable; accordingly, a recent trend of using adversarial training is clearly observable in the literature that aims for achieving higher-level structural qualities. However, as we argue and show, the common formulation of adversarial training for semantic segmentation is ill-posed, sub-optimal, and may result in side-effects, such as the disability to express uncertainties.
In this paper, we suggest using recently introduced Gambling Adversarial Networks that revise the conventional adversarial training for semantic segmentation, by reformulating the fake/real discrimination task into a correct/wrong distinction. This forms then a more effective training strategy that simultaneously serves for both hard sample mining as well as structured prediction. Applying the gambling networks to the ultrasound thyroid nodule segmentation task, the new adversarial training dynamics consistently improve the qualities of the predictions shown over different state-of-the-art semantic segmentation architectures and various metrics.

Keywords: adversarial training · hard sample mining · structured prediction · ultrasound thyroid nodule segmentation.

1 Introduction

Semantic segmentation is arguably among the longest-standing and most important problems in medical image analysis. Despite the significant improvements in semantic segmentation, due to much better representation learning capabilities of modern deep learning architectures [1], the task is still facing inherent

challenges that remain not fully resolved. In the following, we briefly discuss two such important challenges, namely the hard sample mining and structured prediction.

Hard sample mining Most of the real-world datasets exhibit long-tail distributions as certain categories are observed much more frequently than others; This issue becomes even more pronounced in the world of medical image analysis, where the pathological observations are generally much less abundant compared to the normal. Arguing beyond the class-imbalance problem, even within the same category of a frequent or an infrequent class, not all samples are of the same difficulty and therefore are not equally informative for the model to attend while being trained [2]. Thus, identically treating different samples is potentially inefficient and should be avoided. While some methods modify the sampling distribution of the training data [3], many others, including [4–10], approximate the sampling process by re-weighting the sample contributions to the objective function to be minimized. For instance, class-based re-weighting, among the most simple and popular approaches, uses the class infrequency as a proxy metric for the sample difficulty/importance. Some other approaches [4, 5, 11, 12] directly use the corresponding errors as a measure of sample importance. Such methods, however, may suffer from their strategy when dealing with noisy labels [13].

Structured prediction Semantic segmentation models not only do need to optimize for per-pixel accuracy, but also certain inter-pixel/voxel structural qualities should be preserved in many cases. For instance, the predictions should be smooth, preserve certain shapes, geometries, and be semantically consistent. Besides using graphical models such as dense CRFs [14], adversarial semantic segmentation [15, 16, 7, 17–21] has been widely employed recently, where a discriminator is incorporated to provide higher-level feedback by learning to distinguish model predictions (fake) from the GT annotations (real). However, GAN models are notoriously difficult to train and are very sensitive to the hyperparameters. More importantly, a major issue when applying adversarial training to semantic segmentation is an inherent triviality of the distinction between real and fake predictions using only low-level cues. An obvious example of this is using value-based clues to contrast the soft values of the model predictions with the bimodal sharp GT labels. This not only hinders learning to improve on high-level structural qualities but also pushes the model to be overconfident, even in the presence of high uncertainties, to close this low-level unimportant gap.

In this paper, we propose to use Gambling adversarial networks [6], a recent method that tackles the aforementioned shortcomings and simultaneously serves as an adversarial hard sample mining and structured semantic segmentation strategy. We demonstrate that applying the proposed method to the task of ultrasound thyroid nodule segmentation, consistently improves results compared to the state-of-the-art hard sample mining and structured prediction approaches, measured over various segmentation architectures and metrics.

2 Methods

Consider $\mathcal{D} = \{(x^i, y^i)\}_{i\in\{1..N\}}$, a dataset of N supervised samples in which $x^i \in \mathbb{R}^{W\times H}$ and $y^i \in \{0,1\}^{W\times H\times C}$ represent the i-th image and the corresponding pixel-wise one-hot labels of size $W \times H$ on a C-class problem. A standard approach to solve such a problem is to train a deep neural network F_θ by minimizing a (weighted) aggregation of pixel-wise loss terms, e.g., categorical cross-entropy:

$$\mathcal{L}_{\text{ce}}(x, y; \theta) = -\frac{1}{W \times H \times N} \sum_{i,j,k} \sum_{c} w(x^i, y^i)_{j,k}\ y^i_{j,k,c} \ln(F_\theta(x^i)_{j,k,c}), \tag{1}$$

where $w(x^i, y^i)$ is a sample weighting function. The weighting function w is, in a vast majority of cases, a uniform weighting of 1 for each sample or is set based on class frequencies as a heuristic, agnostic to the structure and difficulty of the input sample x^i. On the other hand, focal loss [4] sets this weighting based on the sample error, but still ignores the structure in samples x^i:

$$w_{\text{focal}}(x^i, y^i)_{j,k} = \sum_{c} y^i_{j,k,c}(1 - F_\theta(x^i)_{j,k,c})^\gamma, \tag{2}$$

Here γ is a hyperparameter factor that controls the extent to which the faultier sample predictions would contribute more to the final loss. Adversarial confidence learning [7], a recently proposed method, suggests to use sample confidences taken from a discriminator D_ϕ, trained to distinguish GT labels (real) from the network predictions (fake) to extract such sample weighting function:

$$w_{\text{conf}}(x^i, y^i)_{j,k} = (1 - D_\phi(x^i, F_\theta(x^i))_{j,k})^\gamma. \tag{3}$$

Meanwhile, a multitude of methods further add adversarial loss terms by incorporating discriminators distinguishing real and fake predictions, to improve the structural qualities of the predictions:

$$\mathcal{L}(x, y; \theta, \phi) = \mathcal{L}_{\text{ce}}(x, y; \theta) + \alpha\mathcal{L}_{\text{adv}}(x, y; \theta, \phi), \tag{4}$$

where the adversarial loss is either computed with the standard non-saturated cross-entropy loss [15] as $\mathcal{L}_{\text{adv}}(x, y; \theta, \phi) = \mathbb{E}_i \ln(D_\phi(x^i, F_\theta(x^i)))$, or as a distance to be minimized in the embedding space [16, 21] mapped by the discriminator: $\mathcal{L}_{\text{adv}}(x, y; \theta, \phi) = \mathbb{E}_i\, ||D_\phi(x^i, F_\theta(x^i)) - D_\phi(x^i, y^i)||$. However, as discussed, there is an inherent ill-posedness in the real/fake distinction task for semantic segmentation. Therefore, we propose to use gambling adversarial nets [6], reformulating the real/fake discrimination task into a *correct/wrong* distinction, to overcome this shortcoming.

This is achieved by replacing a conventional discriminator with a gambler model (G_φ) generating dense $W\times H$ betting/investment maps, where $G_\varphi(x^i, F_\theta(x^i))_{j,k}$, i.e. the betting map at position (j, k), aims to predict how likely the prediction $F_\theta(x^i)$ is *wrong* (rather than fake). The gambler attempts to maximize the

weighted cross-entropy loss for the segmenter network as in Equation (1), where its betting map forms the weight terms (w_{gam}). This translates into the following loss function for the gambler:

$$\mathcal{L}_{\text{g}}(x, y; \varphi, \theta) = \frac{1}{W \times H \times N} \sum_{i,j,k} \sum_{c} w_{\text{gam}}(x^i)_{j,k} \; y^i_{j,k,c} \; \ln(F_\theta(x^i)_{j,k,c}), \tag{5}$$

A trivial solution now for the gambler to minimize its loss is to infinitely bet on every single location. To prevent this, we make the analogy to a real-world gambler complete, by giving the gambler a limited budget, so that the gambler needs to learn the mistake patterns to efficiently distribute its limited budget. This is obtained by spatially normalizing the betting maps, in the following form:

$$w_{\text{gam}}(x^i)_{j,k} = \frac{e^{G_\varphi(x^i, F_\theta(x^i))_{j,k} + \beta}}{\sum_{m,n} e^{G_\varphi(x^i, F_\theta(x^i))_{m,n} + \beta}}, \tag{6}$$

where β is a regularizing smoothing factor, ensuring that the weights are smoothly distributed over the different samples. The segmenter network F_θ is involved in a minimax game:

$$\mathcal{L}_{\text{f}}(x, y; \varphi, \theta) = -\mathcal{L}_{\text{g}}(x, y; \varphi, \theta). \tag{7}$$

Note that with such an adversarial game between the semantic segmentation model and the gambler network, the proposed method implements two strategies, hard-sample mining as well as improving structural qualities. Note that using this formulation, gambling nets do not suffer from the inherent problems of real/fake distinction, e.g. value-based discrimination. Structural inconsistencies are reliable investments and often easy-to-find clues for the gambler to bet on; therefore the resulting gradients will encourage the segmenter network to avoid structurally wrong predictions. An overview of the gambling adversarial nets is presented in Figure 1.

Notice that in contrast to all the aforementioned adversarial methods, our critic, i.e. the gambler, never perceives real ground-truth images as input and therefore learns the mistake patterns only through the structure of the model predictions, in combination with the input image. We argue that this an important aspect in not letting the critic learn to misuse some of the inherent and sometimes even desirable discrepancies between the GT labels and model predictions, such as soft values and uncertainties.

3 Experimental Setup

We evaluate and compare the proposed method with focal loss [4] and adversarial confidence learning [7, 8], as well as adversarial training [15] and SegAN [16], representing the hard sample mining and structured prediction literature respectively. To ensure fair comparisons, all the common hyperparameters are kept the same and the models only differ in the corresponding loss formulation, and the specific hyperparameters are tuned separately. To show the consistency of the

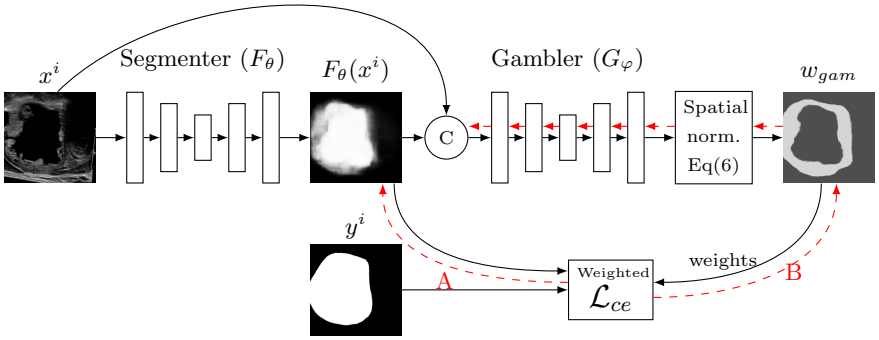


Fig. 1: A schematic of the gambling adversarial networks for semantic segmentation, where the node marked with C represents concatenation operation. Note the two gradient flows specified with the red dashed lines, A and B. While the gradients on path A support hard sample reweighting, the gradients on path B help improving structural qualities as containing information on why the gambler has picked certain regions to upweight.

comparison beyond a single network architecture, all the models are trained and evaluated with three state-of-the-art semantic segmentation network architectures, namely U-net [22], PSPNet [23] with ResNet-101 [24] backbone, and DeepLabV3+ [25] with an Xception [26] backbone network. The details of the training process and the hyperparameters are available in the supplementary materials. Please note that all reported metrics are averaged over three runs to suppress possible fluctuations due to random initialization.

3.1 Dataset

The dataset used in this study is obtained from the TN-SCUI challenge [27] on Thyroid nodule segmentation and classification on ultrasound images. The dataset contains 3644 images of various resolutions each provided with a binary mask of the corresponding thyroid nodule, annotated by experienced doctors. We divide the data into training, validation, and test sets of 80%, 10%, and 10%. We tune the models on the validation set, and then use the full training and validation samples to train the final models to be evaluated on the test set.

3.2 Metrics

In addition to using Dice similarity score, and Jaccard index, widely used for semantic segmentation, we further assess our trained models with BF-score [28], Chamfer distance and Hausdorff distance that are commonly used in the structured semantic segmentation literature. These metrics mainly deal with the quality of prediction boundaries that, compared to pixel-wise metrics, better correlate

Table 1: Comparison of the different methods on the U-net [22] architecture. The metrics are averaged over three runs.

Loss formulation	Dice ↑	Jaccard ↑	BF-score ↑	Chamfer↓	Hausdorff↓
Cross-entropy	83.9	72.2	55.8	9.8	31.9
Cross-entropy + adv.	83.3	71.4	53.6	9.4	32.4
SegAN	80.4	67.2	41.8	14.5	39.3
Focal loss	83.5	71.7	54.0	10.2	32.3
Confidence adv.	83.6	71.9	54.8	9.9	31.8
Gambling adv.	**84.8**	**73.6**	**59.8**	**9.0**	**30.2**

with structural qualities. These metrics are computed as follows:

$$\mathrm{d}_{\mathrm{Chamfer}}(X,Y) = \frac{1}{2}\sum\left\{\frac{1}{|X|}\sum_{x\in X}\min_{y\in Y} d(x,y), \frac{1}{|Y|}\sum_{y\in Y}\min_{x\in X} d(x,y)\right\}, \tag{8}$$

$$d_\tau(X,Y) = \frac{1}{|X|}\sum_{x\in X}[\min_{y\in Y} d(x,y) < \tau], \tag{9}$$

$$BF(X,Y) = \frac{2d_\tau(X,Y)d_\tau(Y,X)}{d_\tau(X,Y)+d_\tau(Y,X)}, \tag{10}$$

where X and Y are the boundaries of the corresponding classes for the predictions and the ground-truth and τ represents a max tolerable distance. Note that the Hausdorff distance is quite similar to the Chamfer distance with the difference that the maximum of surface distances is computed rather than the average, which makes it more sensitive to the outliers. To assess the ability of the model to express uncertainties, we also report the mean maximum class likelihood (MMCL), representing the average likelihood of the most likely class, where the max likelihoods are averaged across the spatial dimensions and different samples.

4 Experimental Results and Discussion

Tables 1, 2 and 3 show the Dice, Jaccard, BF-score, Chamfer and Hausdorff distance metrics on the test set, on models trained with U-net, PSPNet and DeepLabV3+ architectures, respectively. Table 4 demonstrates and compares the MMCL values for the different methods. Extensive qualitative comparisons are available in Figures 1, 2, and 3 in the supplementary materials. As observed in the reported empirical results, models using gambling adversarial nets as the loss formulation consistently outperform the other adversarial and hard sample mining methods over different network architectures and metrics. In the following, we present a brief analysis for comparing the gambling nets to each of the other methods.

Table 2: Comparison of the different methods on the PSPNet [23] architecture, with Resnet101 backbone. The metrics are averaged over three runs.

Loss formulation	Dice ↑	Jaccard ↑	BF-score ↑	Chamfer↓	Hausdorff↓
Cross-entropy	86.6	76.4	65.4	6.9	23.8
Cross-entropy + adv.	85.7	74.9	64.4	7.6	26.0
SegAN	85.7	75.1	62.9	7.1	24.2
Focal loss	86.8	76.7	64.8	7.1	23.6
Confidence adv.	85.9	75.3	63.8	7.6	25.1
Gambling adv.	**87.4**	**77.6**	**68.5**	**6.7**	**22.1**

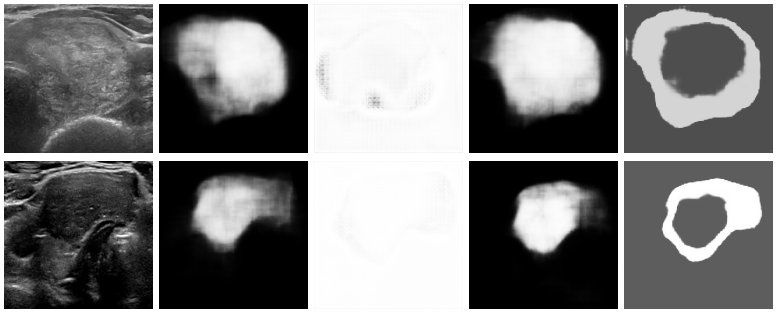
Table 3: Comparison of the different methods on the DeeplabV3+ [25] architecture with an Xception [26] backbone. The metrics are averaged over three runs.

Loss formulation	Dice ↑	Jaccard ↑	BF-score ↑	Chamfer↓	Hausdorff↓
Cross-entropy	86.6	76.4	64.8	7.3	24.7
Cross-entropy + adv.	87.2	77.4	66.6	8.0	43.2
SegAN	86.9	76.8	66.5	6.8	23.1
Focal loss	86.3	76.0	64.6	7.3	24.2
Confidence adv.	86.5	76.2	64.7	7.2	23.3
Gambling adv.	**87.5**	**77.7**	**69.3**	**6.5**	**21.6**

Table 4: The mean maximum class likelihoods from U-net reported for the different training methods.

Loss formulation	CE	Focal	CE + Adv.	SegAN	Gambling
MMCL	0.846	0.705	0.930	0.985	0.862

Focal loss Even though a normalized focal error map can be thought of as the minimization solution for the problem that the gambler deals with, training the segmenter network with the gambler sample weights has two clear advantages: Firstly, as illustrated in Figure 1, in addition to the plain up-weighting of the difficult samples, the gambler also provides structural information, representing why certain areas are considered investment-worthy. Secondly, focal loss can suffer from noisy labels, where a possibly correct prediction from the model is harshly penalized due to a noisy label. The gambling nets, on the other hand, suffer less from this, as the possibly incorrect labels only indirectly influence the resulting weights. Therefore as long as the gambler network is not overfitted to the noise patterns, the training framework is more resilient to label noise compared to the focal loss.



(a) (b) (c) (d) (e)

Fig. 2: Sample comparison of the obtained uncertainty maps from adv. confidence learning and adv. gambling nets. (a) input image, (b) adv. conf. learning prediction, (c) uncertainty map from adv. conf. learning (d) prediction from gambling nets, (e) betting map from the gambler.

Conventional adversarial training As suggested by the MMCL comparison in Table 4, the critic in both adversarial models push the model to fake to be more certain. However, we can also note that the segmentation models do not fully close the value-based gap, likely because otherwise, the pixel-wise loss would harshly penalize a confident wrong prediction. Therefore, with the remaining gap, the discriminator still has an easy job distinguishing the real and fake and thus will likely not go beyond such low-level remaining clues. This would obviously hinder learning higher-level structural qualities and consistencies. In contrast, a smooth likelihood, e.g. 0.8, is not a good investment for the gambler as long as the prediction is correct.

Adversarial confidence learning Even though adversarial confidence learning [7] is the closest work to the gambling networks, in the regard that it similarly aims to extract the samples' difficulty weights in a learnable fashion, there is still a major difference in how the critic is trained. This method still trains a discriminator to distinguish the real and fake labels. Apart from the argued ill-posedness of such formulation for semantic segmentation, we found it very difficult in practice to get meaningful confidence maps from the discriminator. As visible in Figure 2, the uncertainty map values were almost always very close to one. This can be attributed to the confidence model loss formulation [7] that forces the model to predict fake at 'every' spatial position in the network predictions; therefore finding any single clue, the confidence model is encouraged to propagate the fake prediction all over the spatial locations, no matter if the corresponding predictions were correct or wrong.

5 Conclusion and Future Work

In this paper, we showed that a simple but fundamental reformulation of the critic in adversarial training can consistently and effectively improve semantic segmentation results on the thyroid nodule segmentation task and the advantages were intuitively and empirically analyzed. We believe that not only using gambling nets as a 'learned' hard sample learning policy is potentially useful for (medical) image segmentation tasks, but also can be studied in combination with other image recognition tasks such as detection and classification; for instance in presence of controlled noise for comparison with other hard sample mining methods, which is left as future work.

References

1. Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
2. Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. *arXiv preprint arXiv:1803.00942*, 2018.
3. Mark JJP Van Grinsven, Bram van Ginneken, Carel B Hoyng, Thomas Theelen, and Clara I Sánchez. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *IEEE transactions on medical imaging*, 35(5):1273–1284, 2016.
4. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
5. Samuel Rota Bulo, Gerhard Neuhold, and Peter Kontschieder. Loss max-pooling for semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2126–2135, 2017.
6. Laurens Samson, Nanne van Noord, Olaf Booij, Michael Hofmann, Efstratios Gavves, and Mohsen Ghafoorian. I bet you are wrong: Gambling adversarial networks for structured semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.
7. Dong Nie, Li Wang, Lei Xiang, Sihang Zhou, Ehsan Adeli, and Dinggang Shen. Difficulty-aware attention network with confidence learning for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1085–1092, 2019.
8. Dong Nie and Dinggang Shen. Adversarial confidence learning for medical image segmentation and synthesis. *International Journal of Computer Vision*, pages 1–20, 2020.
9. Alireza Mehrtash, William M Wells III, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *arXiv preprint arXiv:1911.13273*, 2019.
10. Mohsen Ghafoorian, Jonas Teuwen, Rashindra Manniesing, Frank-Erik de Leeuw, Bram van Ginneken, Nico Karssemeijer, and Bram Platel. Student beats the teacher: deep neural networks for lateral ventricles segmentation in brain mr. In *Medical Imaging 2018: Image Processing*, volume 10574, page 105742U. International Society for Optics and Photonics, 2018.

11. Pei Wang and Albert CS Chung. Focal dice loss and image dilation for brain tumor segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 119–127. Springer, 2018.
12. S Mazdak Abulnaga and Jonathan Rubin. Ischemic stroke lesion segmentation in ct perfusion scans using pyramid pooling and focal loss. In *International MICCAI Brainlesion Workshop*, pages 352–363. Springer, 2018.
13. Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. *arXiv preprint arXiv:1803.09050*, 2018.
14. Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.
15. Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. In *NIPS Workshop on Adversarial Training*, 2016.
16. Yuan Xue, Tao Xu, Han Zhang, L Rodney Long, and Xiaolei Huang. Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation. *Neuroinformatics*, 16(3-4):383–392, 2018.
17. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
18. Mina Rezaei, Konstantin Harmuth, Willi Gierke, Thomas Kellermeier, Martin Fischer, Haojin Yang, and Christoph Meinel. A conditional adversarial network for semantic segmentation of brain tumor. In *International MICCAI Brainlesion Workshop*, pages 241–252. Springer, 2017.
19. Farhad Ghazvinian Zanjani, David Anssari Moin, Bas Verheij, Frank Claessen, Teo Cherici, Tao Tan, et al. Deep learning approach to semantic segmentation in 3d point cloud intra-oral scans of teeth. In *International Conference on Medical Imaging with Deep Learning*, pages 557–571, 2019.
20. Pim Moeskops, Mitko Veta, Maxime W Lafarge, Koen AJ Eppenhof, and Josien PW Pluim. Adversarial training and dilated convolutions for brain mri segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 56–64. Springer, 2017.
21. Mohsen Ghafoorian, Cedric Nugteren, Nóra Baka, Olaf Booij, and Michael Hofmann. El-gan: Embedding loss driven generative adversarial networks for lane detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
22. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.
23. Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.
24. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
25. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

26. François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.
27. Jianqiao Zhou, Xiaohong Jia, Dong Ni, Alison Noble, Ruobing Huang, Tao Tan, and Manh The Van. Thyroid Nodule Segmentation and Classification in Ultrasound Images, March 2020.
28. Gabriela Csurka, Diane Larlus, Florent Perronnin, and France Meylan. What is a good evaluation measure for semantic segmentation?. In *BMVC*, volume 27, page 2013, 2013.