Towards Robust General Medical Image Segmentation

Laura Daza[1(✉)], Juan C. Pérez[1,2], and Pablo Arbeláez[1]

[1] Universidad de los Andes, Bogotá, Colombia
la.daza10@uniandes.edu.co
[2] King Abdullah University of Science and Technology (KAUST),
Thuwal, Saudi Arabia

Abstract. The reliability of Deep Learning systems depends on their accuracy but also on their robustness against adversarial perturbations to the input data. Several attacks and defenses have been proposed to improve the performance of Deep Neural Networks under the presence of adversarial noise in the natural image domain. However, robustness in computer-aided diagnosis for volumetric data has only been explored for specific tasks and with limited attacks. We propose a new framework to assess the robustness of general medical image segmentation systems. Our contributions are two-fold: *(i)* we propose a new benchmark to evaluate robustness in the context of the Medical Segmentation Decathlon (MSD) by extending the recent *AutoAttack* natural image classification framework to the domain of volumetric data segmentation, and *(ii)* we present a novel lattice architecture for RObust Generic medical image segmentation (ROG). Our results show that ROG is capable of generalizing across different tasks of the MSD and largely surpasses the state-of-the-art under sophisticated adversarial attacks.

Keywords: Robustness assessment · Adversarial training · Adversarial attacks · General medical segmentation

1 Introduction

The observation that imperceptible changes in the input can mislead Deep Neural Networks (DNN) [6,29] has attracted great interest in the deep learning community. This behaviour has been studied for various tasks, such as classification [19,29], object detection [34] and semantic segmentation [2,7,32], highlighting the importance of *reliably* assessing robustness. To improve this dimension, adversarial robustness has been studied both from the side of the attacks [3,9,14,22,24] and the defenses [17,23,27,35], obtaining sizable progress towards DNN models resistant to adversarial perturbations; however, still much is left to advance. In particular, recognition tasks in medical domains are of
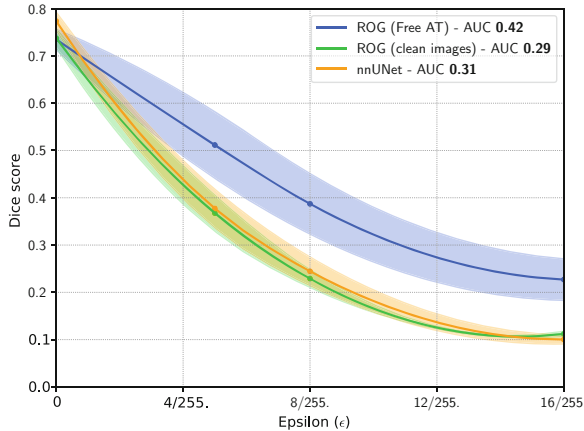
Fig. 1. Medical segmentation robustness benchmark. We introduce a new benchmark for studying adversarial robustness on the MSD [28]. We propose a novel method for robust general segmentation that significantly outperforms the state-of-the-art in our benchmark [13], providing a strong baseline for future reference.

utmost importance for robustness, as these tasks aim at safety-critical applications in which brittleness could have dramatic consequences.

Most semantic segmentation methods for computer aided-diagnosis are designed for specific anatomical structures [12,20,30,38]. This level of specialization has been attributed to the large variability in structures, acquisition protocols, and image modalities, the limited annotated datasets, and the computational cost of processing 3D data. The Medical Segmentation Decathlon (MSD) [28], an experimental framework that combines ten diagnostic tasks, aims at addressing this over-specialization. Top-performing techniques in the MSD solve each task by combining multiple models and automatically adjusting task-specific parameters [13], or conducting Neural Architecture Search (NAS) [33,37]. The MSD, thus, is a promising benchmark to develop general methods for medical image segmentation. However, adversarial robustness, a critical dimension of deep learning systems, remains uncharted territory in the MSD.

In this paper, we propose a new experimental framework to evaluate the adversarial robustness of medical segmentation methods. With this aim, we build on the AutoAttack benchmark [9], an ensemble of adversarial attacks to evaluate recognition models in natural images. We extend AutoAttack from the image classification domain to volumetric multi-channel semantic segmentation. Our results suggest that the adversarial vulnerability of methods in the medical domain, while previously evidenced [16,18], has most likely been *underestimated*. We find that subjecting segmentation methods to our extension of AutoAttack exposes their brittleness against adversarial perturbations.

As a strong baseline for RObust General (ROG) medical segmentation, we propose an efficient lattice architecture that segments organs and lesions on MRI and CT scans. Further, we leverage Adversarial Training (AT) [10,19] to protect our models against adversarial perturbations. Since AT's great computational

demands can render training of 3D models prohibitive, we equip ROG with "Free" AT [27], an efficient version of AT. Figure 1 shows the variation in Dice score as the attack's strength, ϵ, increases, comparing ROG trained with clean images, trained with Free AT and nnUnet [13], the winner of the MSD challenge.

Our main contributions can be summarized as follows: *(i)* we extend AutoAttack to the domain of 3D segmentation to introduce an experimental framework for evaluating adversarial robustness in medical segmentation tasks, and *(ii)* we introduce ROG, the first generic medical segmentation model that is robust against adversarial perturbations. We will make our code publicly available to ensure reproducibility and to encourage robustness assessments as a fundamental dimension in the evaluation of 3D DNNs in safety-critical applications[1].

2 Methodology

2.1 Adversarial Robustness

Extending *AutoAttack* to 3D Segmentation. Assessing adversarial robustness has proven to be a complex task [5], hence, many approaches have been proposed [6,22]. In this work, we build on AutoAttack [9], a parameter-free ensemble of attacks that has shown remarkable capacity for breaking adversarial defenses and has emerged as the leading benchmark to assess adversarial robustness. AutoAttack is composed of: *(i)* AutoPGD-CE and *(ii)* AutoPGD-DLR, two Projected Gradient Descent-based (PGD) adversaries that optimize Cross Entropy and Difference-of-Logits Ratio [9] respectively, *(iii)* targeted Fast Adaptive Boundary (FAB-T) [8] which minimizes a decision-changing perturbation, and *(iv)* Square Attack [1], a score-based black-box attack.

Image classification is the most studied setting for adversarial robustness [19,21]. Thus, AutoAttack was designed to fool models that assign a probability distribution to *each image*. However, we are interested in attacking models that assign a probability distribution to *each spatial location*. Therefore, we generalize the definition of the attacks to consider both input and output as four-dimensional tensors: three spatial dimensions plus the channels and the spatial dimensions plus the probability distribution over the classes. In addition, every attack requires a notion of "success", *i.e.* the state in which the attack has fooled the model. While the success of an attack is readily defined for classification, it is not clearly defined for segmentation. Thus, we re-define this notion as a drop in the Dice score below $\mu_i/2$, where μ_i is the average performance in task i over clean images. With this definition, we modify each attack to compute voxel-wise functions across all spatial locations. Finally, we preserve the definition of "Robust Accuracy" as the per-instance worst case across the four attacks.

Adversarial Training. Given the critical applications of medical image segmentation, it is crucial to develop methods that are both accurate and robust to perturbations. To improve ROG's robustness, we apply "Free" Adversarial

---

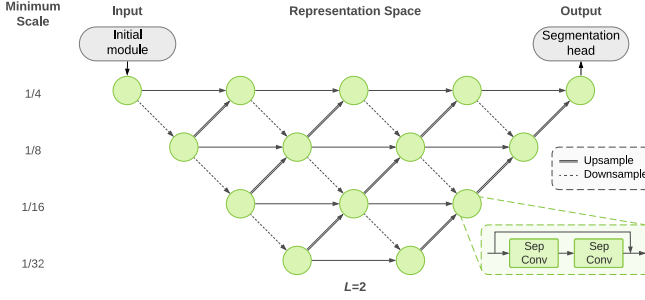[1] https://github.com/BCV-Uniandes/ROG.

Fig. 2. Overview of ROG. Our lattice architecture for generic medical segmentation preserves high resolution features while also exploiting multiple image scales. The initial module adjusts the input by reducing its scale to a minimum of $1/4$ of the original size, while the segmentation head adapts to produce a mask of the same size as the input.

Training [27] in our task domain to fine-tune our models, following the spirit of [4,11]. For Free AT, we set ϵ, the strength of the attack, to $8/255$, and m, the number of times an image is replayed for gradient computation, to 5.

2.2   Generic Medical Segmentation

We aim to study the unexplored dimension of adversarial robustness in general medical image segmentation. However, existing methods lack publicly available resources [33] or require training multiple models for each task [13]. Combined with the high computational demands of adversarial training, these limitations prevent their adoption for our primary goal. Hence, we propose ROG, an efficient single-architecture model for *RObust Generic* segmentation that exploits increasingly larger receptive fields while preserving high-resolution features. ROG comprises an initial module with four convolutions, the main lattice of processing nodes, and a segmentation head with two convolutions. We organize the nodes in a triangular lattice with four scales, as shown in Fig. 2. Unlike UNet++ [36], we connect each node with both upper and lower level nodes, and we remove the dense connections. As a result, ROG has $10\times$ fewer parameters ($2.6M$ vs. $26.2M$) than [36]. We find that our topology is significantly more efficient than a squared lattice since it reduces the number of layers while preserving the advantages of simultaneous multi-scale processing. We also observe that varying the length L of the lattice, *i.e.* the number of nodes in the last scale, benefits different types of objects. In particular, lower lengths tend to benefit small objects but are detrimental for large objects, while longer lengths favor the opposite. We set $L = 2$ as a trade-off between accuracy and model size.

Within each node, we use two separable convolutions, instance normalization layers, and residual connections. We also use swish activations [26], given the benefit of its smooth gradient for adversarial training [31]. To connect nodes at different levels, we use $1 \times 1 \times 1$ convolutions to adjust the number of feature maps and change the spatial resolution by either using a stride of 2 or adding a trilinear interpolation. Finally, we automatically adjust ROG for each task

using the average size of the volumes in the dataset to select the input patch size, strides in the initial module, and upsampling factors in the segmentation head. This selection is driven by two main factors: the target objects' spatial dimensions and the model's memory consumption. Also, we make sure that the minimum possible spatial resolution is $1/32$ of the input resolution.

3 Experiments

3.1 Adversarial Robustness Assessment

We train two variants of ROG for each task: using only clean images and using Free AT. We then use our extended version of AutoAttack to attack the models. For the AutoPGD and FAB attacks, we fix the number of iterations to 5 and vary ϵ between $5/255$ and $16/255$ to assess the adversarial robustness of each model to perturbations of different magnitudes. However, since our experimental framework differs significantly from natural image classification, we also explore the attacks using between 5 and 20 iterations with $\epsilon = 8/255$ and report these results in the supplemental material. For the Square attack, we vary the number of queries between 500 and 5000. We train our models on 80% of the data and validate on the remaining 20%. To reduce bias during the evaluation, all design choices are made based on the results over the tasks of pancreas and prostate segmentation of the MSD. For reference, we evaluate the robustness of nnU-Net, the highest performing method on the MSD with publicly available resources.

3.2 General Medical Segmentation

We evaluate the generalization capacity of ROG on the MSD. Each task has a target that can be an organ or a tumor and source data that can be CT or MRI with one or multiple modalities. The number of patients varies between 30 and 750 for the tasks, with a total of 2,633 volumes [28]. We use Adam [15] with a weight decay of 10^{-5}, and an initial learning rate of 10^{-3} that is reduced by a factor of 0.5 when the validation loss has stagnated for the previous 50 epochs. We optimize a combination of the Dice loss and the Cross-Entropy loss. For data augmentation, we use random rotations, scaling, mirroring, and gamma correction. We sample cubic patches with a 50% probability of being centered at a foreground category to address data imbalance. For pre-processing, we re-sample the volumes to ensure a homogeneous voxel spacing. Next, we follow [13] and clip the intensities to the [0.5, 99.5] percentiles of the foreground values and perform z-score normalization.

For inference, we uniformly sample patches and combine the predictions to generate the volume. We combine the predictions of the same voxel in multiple patches via weights inversely proportional to the distance to the patch's center. To obtain the final predictions for the MSD test set, we perform simple noise reduction: for organs, we preserve only the largest element; for tumors, we fuse all elements smaller than a threshold with the surrounding category. The threshold is defined based on the average object size in the dataset.
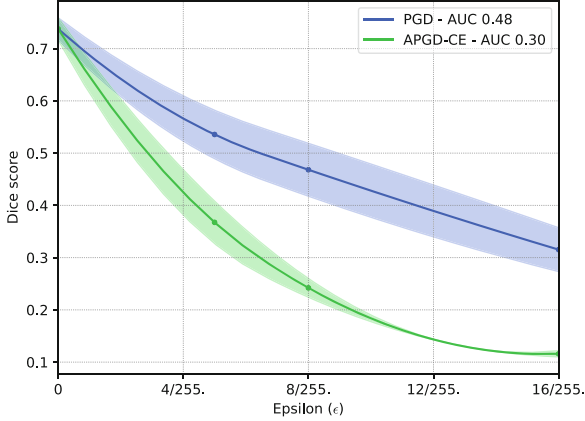
Fig. 3. PGD vs. AutoPGD-CE. We compare the performance of PGD and AutoPGD-CE by attacking ROG trained on clean images averaging the performance over the 10 tasks in the MSD. The results demonstrate that the AutoPGD-CE attack is significantly stronger than regular PGD.

4   Results

4.1   Adversarial Robustness

Adversarial Attacks. Previous assessments of adversarial robustness in segmentation tasks rely on the well-known PGD attack [23]. To elucidate the need for stronger adversaries, we compare the performances of AutoPGD-CE and PGD. We run both attacks on our model trained with clean images in all the datasets, vary the attacks' strength (ϵ), and report the average results in Fig. 3. In line with [9], we find that AutoPGD-CE is dramatically stronger for *every* ϵ. In particular, the AUC for PGD is 0.48, while it drops to 0.30 for AutoPGD-CE.

We test the individual attacks *within* AutoAttack following the same configuration and report the average Dice scores in Table 1. We find that both PGD-based adversaries are the strongest attacks. Further, we note that FAB-T's performance is deficient: FAB-T is unable to find adversaries. While studies on FAB [8,9] often report lower performance for FAB when compared to PGD-based adversaries, our results present a stark contrast. While this observation calls for further experimentation, our initial hypothesis is that FAB's formulation, based on decision-boundary approximation through linearization near the input, may be ill-suited for segmentation. Thus, a re-definition of FAB's inner workings to account for the particularities of segmentation methods may be called for. Regarding the Square attack, despite the simplicity of the random search-based optimization, it consistently finds adversaries. However, since the search space for variable-sized cubes for volumetric inputs is rather large, the computational demands are notably higher than the other attacks.

Table 1. Comparison of individuals attacks of AutoAttack. We report the average performance over all datasets in the MSD with the four attacks in AutoAttack.

	Clean	APGD-CE	APGD-DLR	FAB-T	Square
ROG	0.7375	0.1222	0.3581	0.7375	0.5791
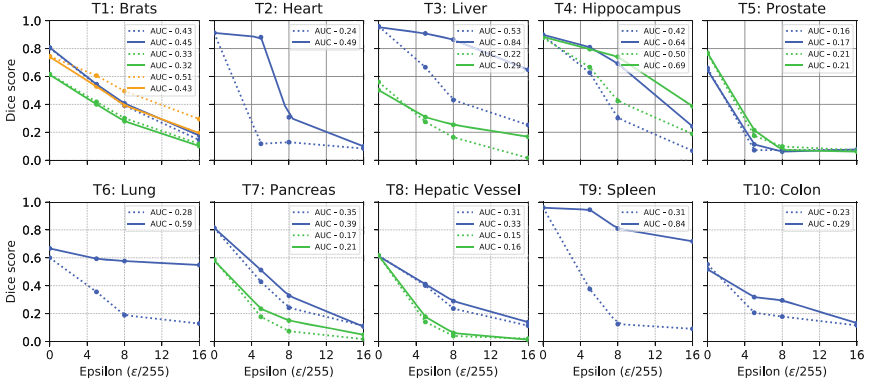
Fig. 4. Adversarial robustness for the 10 tasks in the MSD. We train ROG on clean images (dotted lines) and with Free AT (solid lines), and evaluate their adversarial robustness across various ϵ. Every color represents a different category within each task. In general, training with Free AT improves adversarial robustness against attacks.

Adversarial Robustness. We conduct attacks on every task of the MSD, vary the attack strength (ϵ), and report the Dice score in Fig. 4. Dotted lines represent the performance of ROG with regular training, solid lines show ROG with Free AT, and categories within each task are color-coded. The effect on the Dice score with regular training demonstrates the susceptibility of segmentation models to adversarial attacks. While the ranges for Dice score vary greatly across tasks and categories, our results suggest a coherent finding consistent with the robustness literature: training with Free AT is a tractable approach to improving the adversarial robustness of DNNs. In addition, in most tasks, Free AT enhances robustness *without* major costs to performance on *clean* samples. We identify that the robustness improvement is greater for large structures, such as the liver, and for definite structures like the left atrium of the heart.

Qualitative Evaluation. We show qualitative results in Fig. 5, where the task is to segment liver and liver tumors. We compare ROG with nnUNet over clean and adversarial images computed through APGD-CE with $\epsilon = 8/255$ and $m = 5$. For more results, please refer to the supplemental material. Our results show that, on clean images, ROG provides similar segmentations to those of nnUNet. However, on adversarial images, both methods strongly diverge in their performance. In particular, ROG preserves the segmentations rather untouched, while nnUNet's segmentations are shattered, and the tumor completely disappears.
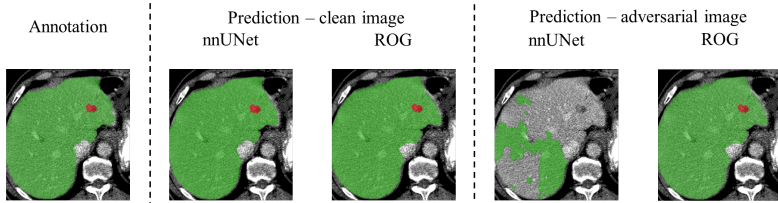
Fig. 5. Qualitative results. We compare the results of ROG and nnUNet on liver images. The adversarial examples are obtained using APGD-CE with $\epsilon = 8/255$ and $m = 5$. Both methods have comparable performance on clean images, while ROG is vastly superior in the presence of adversarial noise.

Table 2. MSD test set results. Comparison of Dice score with top performing methods of the MSD. The results were retrieved from the MSD official leaderboard.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Mean	Std
MPUNet [25]	0.6	0.89	0.75	0.89	0.77	0.59	0.48	0.48	0.95	0.28	0.66	0.22
nnUNet [13]	0.61	0.93	0.84	0.89	0.83	0.69	0.66	0.66	0.96	0.56	0.75	0.15
C2FNAS [33]	0.62	0.92	0.84	0.88	0.82	0.70	0.67	0.67	0.96	0.59	0.75	0.14
ROG (Ours)	0.56	0.91	0.75	0.89	0.79	0.36	0.66	0.62	0.88	0.46	0.69	0.18

We highlight a visually appealing phenomenon: nnUNet's segmentations appear to follow a pattern of segmenting elements whose boundaries usually correspond with high contrast boundaries. These patterns suggest that nnUNet, and perhaps other segmentation methods, are subject to strong inductive bias that may come from the training data or the convolutional layers with small kernels.

All in all, we observe that nnUNet's vulnerability to adversarial noise is high, failing both on the side of false positives and false negatives. In real-life clinical scenarios, the consequences of such mistakes, while largely unpredictable, could be appalling. Since ROG suffers from these effects to a minor degree, it appears as a better alternative for reliable medical image segmentation.

4.2   Results on the Medical Segmentation Decathlon

To verify the generalization capability of ROG in a clean setup, we evaluate our results in the test server of the MSD. Table 2 summarizes our results compared against all published methods that are evaluated over the entire MSD. ROG surpasses the performance of MPUNet in most tasks with nearly $24\times$ fewer parameters ($2.6M$ vs. $62M$). When we compare to more complex models, we see a trade-off between performance and efficiency. While nnUNet uses an ensemble of models, ROG relies on a *single* architecture. In contrast, C2FNAS uses a single fixed model for most tasks, but requires $6.5\times$ more parameters and almost $5\times$ more GFLOPs (150.78 *vs.* 30.55). Note that, since ROG adapts to each dataset, we compute the total number of FLOPs by averaging the results of all methods

using a maximum input size of $96 \times 96 \times 96$, as in [33]. Under this scheme, our most computationally expensive method has only 79.32 GFLOPs. Considering the large computational demands of studying adversarial robustness, specially in the domain of 3D segmentation, these results demonstrate the advantage of adopting ROG as a strong baseline for the analysis of this dimension in the task of generic medical segmentation.

5   Conclusion

We introduce a benchmark for evaluating the adversarial robustness of volumetric segmentation systems in the medical domain. In particular, we extend the state-of-the-art AutoAttack approach from the image-recognition domain to the volumetric semantic segmentation domain. Further, we propose ROG, a robust and efficient segmentation method for the MSD. We find that ROG provides competitive results over clean images in most tasks of the MSD, while vastly outperforming one of the top-performing methods in this benchmark under adversarial attacks. Our study provides further evidence that the adversarial robustness dimension has been profoundly under-explored in medical image segmentation. Hence, we hope that our work can serve as a tool for developing accurate and robust segmentation models in the clinical domain.

References

1. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12368, pp. 484–501. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58592-1_29
2. Arnab, A., Miksik, O., Torr, P.H.: On the robustness of semantic segmentation models to adversarial attacks. In: CVPR (2018)
3. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: reliable attacks against black-box machine learning models. In: ICLR (2018)
4. Cai, Q.Z., Liu, C., Song, D.: Curriculum adversarial training. In: International Joint Conference on Artificial Intelligence (IJCAI) (2018)
5. Carlini, N., et al.: On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705 (2019)
6. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP) (2017)
7. Cisse, M., Adi, Y., Neverova, N., Keshet, J.: Houdini: fooling deep structured prediction models. arXiv preprint arXiv:1707.05373 (2017)
8. Croce, F., Hein, M.: Minimally distorted adversarial examples with a fast adaptive boundary attack. In: International conference on machine learning (ICML) (2020)
9. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International Conference on Machine Learning (ICML) (2020)

10. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
11. Gupta, S., Dube, P., Verma, A.: Improving the affordability of robustness training for DNNs. In: CVPRW (2020)
12. Isensee, F., Jaeger, P.F., Full, P.M., Vollmuth, P., Maier-Hein, K.H.: nnU-net for brain tumor segmentation. arXiv preprint arXiv:2011.00848 (2020)
13. Isensee, F., et al.: nnU-net: self-adapting framework for u-net-based medical image segmentation. CoRR abs/1809.10486 (2018)
14. Joel, M.Z., et al.: Adversarial attack vulnerability of deep learning models for oncologic images. medRxiv (2021). https://doi.org/10.1101/2021.01.17.21249704
15. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
16. Li, Y., et al.: Volumetric medical image segmentation: a 3D deep coarse-to-fine framework and its adversarial examples. In: Lu, L., Wang, X., Carneiro, G., Yang, L. (eds.) Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics. ACVPR, pp. 69–91. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-13969-8_4
17. Liu, Q., et al.: Defending deep learning-based biomedical image segmentation from adversarial attacks: a low-cost frequency refinement approach. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 342–351. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_34
18. Ma, X., et al.: Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recogn. **110**, 107332 (2020)
19. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018)
20. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D vision (3DV) (2016)
21. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: CVPR (2017)
22. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: DeepFool: a simple and accurate method to fool deep neural networks. In: CVPR (2016)
23. Mummadi, C.K., Brox, T., Metzen, J.H.: Defending against universal perturbations with shared adversarial training. In: ICCV (2019)
24. Ozbulak, U., Van Messem, A., De Neve, W.: Impact of adversarial examples on deep learning models for biomedical image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 300–308. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_34
25. Perslev, M., Dam, E.B., Pai, A., Igel, C.: One network to segment them all: a general, lightweight system for accurate 3D medical image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 30–38. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_4
26. Ramachandran, P., Zoph, B., Le, Q.: Searching for activation functions. In: ICLR (2018)
27. Shafahi, A., et al.: Adversarial training for free! In: NeurIPS (2019)
28. Simpson, A.L., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. CoRR abs/1902.09063 (2019)
29. Szegedy, C., et al.: Intriguing properties of neural networks. In: ICLR (2014)
30. Tang, H., Zhang, C., Xie, X.: NoduleNet: decoupled false positive reduction for pulmonary nodule detection and segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 266–274. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_30

31. Xie, C., Tan, M., Gong, B., Yuille, A.L., Le, Q.V.: Smooth adversarial training. CoRR abs/2006.14536 (2020)
32. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: ICCV (2017)
33. Yu, Q., et al.: C2FNAS: coarse-to-fine neural architecture search for 3D medical image segmentation. In: CVPR (2020)
34. Zhang, H., Wang, J.: Towards adversarially robust object detection. In: ICCV (2019)
35. Zhang, H., Yu, Y., Jiao, J., Xing, E.P., Ghaoui, L.E., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning (ICML) (2019)
36. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: redesigning skip connections to exploit multiscale features in image segmentation. Trans. Med. Imaging (2019)
37. Zhu, Z., Liu, C., Yang, D., Yuille, A., Xu, D.: V-NAS: neural architecture search for volumetric medical image segmentation. In: 2019 International Conference on 3D Vision (3DV) (2019)
38. Zhu, Z., Xia, Y., Shen, W., Fishman, E., Yuille, A.: A 3D coarse-to-fine framework for volumetric medical image segmentation. In: International Conference on 3D Vision (3DV) (2018)