Transferability of Vision Transformer on Biomedical Image Segmentation

Eugene Yong Apurva Dilip Kokate Vy Bui Opeyemi Ajibuwa
Oregon State University
{yonge, kokatea, buivy, ajibuwao}@oregonstate.edu

1. Introduction

In recent years, the field of biomedical imaging has witnessed remarkable advancements with the increasing adoption of deep learning (DL) techniques, particularly in the area of image segmentation [7,10,24,25]. However, the vulnerability of DL models to adversarial attacks poses significant challenges to their reliability and performance in real-world applications [11, 18]. Adversarial attacks involve the deliberate manipulation of input data to deceive the model, leading to incorrect predictions and potential harm to patients.

While there has been extensive research on the robustness of convolutional neural networks (CNNs) to transfer-based attacks, the efficacy of transformer-based architectures, specifically in the context of image segmentation tasks in the biomedical field, remains relatively unexplored. Previous studies, such as [16,20,23] have demonstrated that adversarial examples do not readily transfer between CNNs and transformers in general computer vision tasks. However, it is yet to be determined whether this observation holds true for image segmentation, which has its unique set of challenges and characteristics.

The motivation behind this research lies in the critical importance of ensuring the robustness and performance of DL models in medical imaging. As the reliance on DL for medical inference grows, the potential consequences of adversarial attacks become increasingly significant. Hence, it becomes imperative to thoroughly investigate the transferability of adversarial examples and the robustness of vision transformers specifically in the context of biomedical image segmentation.

This paper aims to address several research gaps in the current understanding of robustness evaluation and defense techniques for segmentation models in biomedical research. Firstly, existing testing methods to evaluate the robustness of segmentation models in the biomedical domain are insufficient and need improvement [2, 13]. Secondly, there is a limited understanding of how different model architectures and hyperparameters influence their susceptibility to adversarial attacks [1,8,12]. Thirdly, there has been limited exploration of synergistic defense techniques that combine adversarial training with other regularization methods, such as dropout and weight decay, specifically tailored for segmentation tasks [4, 14].

To fill these gaps, we propose an investigation into the transferability of adversarial examples crafted on models with different architectures and their impact on the performance of the Swin-Unet [5], a transformer-based architecture designed for image segmentation in the biomedical domain. By conducting a series of experiments and evaluating the robustness of the Swin-Unet against adversarial attacks, we aim to shed light on the transferability of adversarial examples in the context of biomedical image segmentation. Furthermore, we will explore the effectiveness of synergistic defense techniques that combine adversarial training with other regularization methods to enhance the robustness of vision transformers. Through this research, we aim to contribute to the development of more reliable and secure DL models for biomedical image segmentation, thereby advancing the adoption of DL in the medical field while mitigating the risks associated with adversarial attacks. To summarize, the specific contributions made by this paper are the following:

- Finds out that adversarial examples don't transfer between different model architectures (CNN, CNN-Transformer, Transformer) in the field of biomedical image segmentation.
- Observed that UNet (CNN) is more resilient to white-box FGSM and transfer attack than TransUNet (CNN-Transformer) and Swin-Unet (Transformer)

The remainder of this paper is structured as follows. Section 2 briefly introduces prior work related to our research. Section 3 describes the methodology of our experiments for generating and testing the adversarial perturbations on different architectures. The results of the experiments are discussed in section 4, 5 and section 6 concludes the paper.

2. Related Works

This section reviews related works, briefly highlighting their main contributions and how our works differ.

[26] presented adversarial examples for complex tasks such as semantic segmentation and object detection. The authors propose the Dense Adversary Generation (DAG) algorithm for generating adversarial perturbations and demonstrate their transferability across networks with different architectures and training data. In [9], the authors demonstrates the transferability of adversarial attackers to semantic segmentation tasks, allowing for the creation of imperceptible perturbations that lead to misclassifications in specific classes while minimally affecting predictions outside those classes. Our proposed work diverges by focusing on the transferability and robustness of vision transformers, particularly the Swin-Unet, specifically in the context of biomedical image segmentation.

The work in [15] conducted a comprehensive study on the transferability of adversarial examples, considering both non-targeted and targeted attacks. The paper introduced ensemble-based approaches for generating transferable adversarial examples, which significantly improve the success rate of targeted attacks and demonstrate their efficacy in attacking black-box image classification systems. Authors in [19] studied the adversarial feature space of Vision Transformers (ViTs) and observed the limited transferability of conventional attacks across different models. They proposed two novel strategies, namely Self-Ensemble and Token Refinement, specifically designed for ViT architectures to enhance attack transferability by leveraging class-specific information and refining tokens within an ensemble of classifiers.

Authors in [3] presented an evaluation of the robustness of semantic segmentation models to adversarial attacks. The paper analyzed the impact of network architectures, model capacity, and multiscale processing on the robustness of semantic segmentation models. They highlighted the differences between classification and semantic segmentation tasks and provide insights into which segmentation models exhibit inherent robustness, guiding the selection of models for safety-critical applications. In the work of [16], the security of ViTs under white-box and black-box attacks were analyzed, revealing their vulnerability to adversarial examples. The study specifically investigates the transferability of adversarial examples between CNNs and transformers, finding limited transferability. Additionally, the paper introduces a novel attack, the self-attention blended gradient attack, to evaluate the security of an ensemble defense.

The novel network architecture proposed by [22] is based on ViT model and incorporate modifications such as splitting features into multiple scales and utilizing skip connections to capture long-range dependencies. The paper evaluates the proposed network on four diverse biomedical image segmentation datasets, demonstrating its superior performance compared to state-of-the-art methods across various images. While [22] emphasizes the architectural modifications and performance evaluation on biomedical datasets, our research investigates the susceptibility of vision transformers to adversarial attacks and their robustness in the biomedical domain.

3. Methodology

In this section, we describe our experimental set-up including the datasets, DL model, adversarial attacks and evaluation metrics. To measure the transferability of adversarial pertubations between different model architectures, we crafted adversarial pertubations on one model and used them to attack other models. Our code is published at `https://shorturl.at/oDJX2` for reproducibility.

Models. We use three different model architectures, U-Net (a CNN) [21], TransUNet (a Hybrid of transformers and UNET) [6], and Swin-Unet (a pure transformer) [5]. We retrained these models and obtained performances comparable to the reported results in the original papers.

Datasets. We use two popular medical image segmentation datasets, Synapse multi-organ segmentation dataset and OASIS-1 [17]. Synapse multi-organ dataset consists of 90 axial abdominal clinical CT images delineating multiple organs: the esophagus, stomach, gallbladder, spleen, left kidney, liver, pancreas and duodenum. OASIS-1 contains cross-sectional brain scans of 416 subjects aged 18 to 96.

Adversarial Attacks. We employed two adversarial attacks, namely Fast Gradient Sign Method (FGSM) adapted for image segmentation [3] and DAG [26]. (TODO: Add hyperparameters for these two attacks here)

Metrics. We use Dice-Similarity coefficient (DSC) and Hausdorff Distance (HD), the primary metrics for evaluating image segmentation models. We computed DSC and HD for each region of interest and took the average of them as the final metric.

4. Observe Transferability

For the primary experiment, we trained UNet, TransUNet, and Swin-Unet on the Synapse training dataset. After that, we use FGSM ($\epsilon = 0.01$) to craft adversarial examples using the Synapse test dataset with each of the trained model, resulting in three adversarial examples crafted. Each of the models was then tested against the clean test dataset and the three crafted adversarial examples. If the adversarial examples generated were able to cause performance drop in models other than the model used to generate it, we say that the adversarial examples are transferable. Otherwise, we say that the adversarial examples don't transfer between different model architectures.

Results shown in Table 1 and Table 2. Both DSC and HD results show the same trend. The model performance drops more significantly when a model is tested against adversarial examples crafted with itself (i.e UNet

Table 1. Models against clean and adversarial examples. Each (row, col) in the table representing the result of model row tested with clean examples/adversarial examples crafted with model col. Dice similarity coefficient (DSC).

	Clean	UNet	TransUNet	Swin-Unet
UNet	78.86	**67.30**	76.91	76.25
TransUNet	77.08	75.51	**62.93**	75.47
Swin-Unet	79.17	76.81	76.04	**63.61**

Table 2. Same format as Table 1. Hausdorff Distance (HD).

	Clean	UNet	TransUNet	Swin-Unet
UNet	33.57	**59.68**	36.86	37.64
TransUNet	30.87	37.31	**61.40**	39.09
Swin-Unet	22.19	25.70	25.72	**61.52**

against adversarial examples crafted on UNet). On the other hand, when models were tested against adversarial examples generated by other models, their performances barely drop compared to when tested against adversarial examples crafted with itself. From this, we can say that the adversarial examples generated by FGSM doesn't transfer between different model architectures.

In order to make sure the adversarial examples only not transferable between different model architectures but was transferable between same model architecture, we did another set of experiment. We trained another set of UNet, TransUNet and Swin-Unet model with different random seed than the previous ones. These models are then used to craft adversarial examples like the previous experiment. The same testing procedure applies to this experiment except we only measure the performances of models against the same model architecture.

From Table [3, 4, 5], we can see that the adversarial examples between same model architecture do transfer and make the models perform worse comparing to transfer between different model architectures. By this, we show that the FGSM attack can in fact transfer, but only to the same model architecture.

One interesting observation throughout the experiments was that UNet model is more robust against the FGSM attack and the transfer attack using FGSM. We were expecting the results to be the other way around since there are plenty of research pointing out that Transformer is a more robust model than CNN.

5. Investigate DAG Attack

We trained UNet models using a sample of the OASIS dataset (150 Training data points, 50 Test data points). We use Dense Adversarial Generation attack with gamma=0.5 to craft the adversarial samples. The attack success rate is measured using the Dice score and HD loss. A effective attack with reduce the dice score compared to clean predictions and increase HD.

1. Table 6 shows the attack success across different attack iterations. We observe that the attack success rate is dependent on the correct hyper parameter choice. We find 30 iterations to be optimal.

2. Tables 7 and 8 shows the transferability of the DAG adversarial examples. We observe that the attack is not transferable across dissimilar models but we can transfer between similar models.

Table 3. Same format as Table 1. Transfer attack between UNet models. Tuples representing (DSC, HD).

	Clean	UNet	UNet 2
UNet	(78.86, 33.57)	**(67.30, 59.68)**	(73.34, 47.85)
UNet 2	(78.26, 36.06)	(73.54, 42.26)	**(67.64, 56.78)**

Table 4. Same format as Table 1. Transfer attack between TransUNet models. Tuples representing (DSC, HD).

	Clean	TransUNet	TransUNet 2
TransUNet	(77.08, 30.87)	**(62.93, 61.40)**	(69.80, 50.19)
TransUNet 2	(75.86, 34.75)	(68.67, 50.44)	**(62.57, 60.40)**

Table 5. Same format as Table 1. Transfer attack between Swin-Unet models. Tuples representing (DSC, HD).

	Clean	Swin-Unet	Swin-Unet 2
Swin-Unet	(79.17, 22.19)	**(63.61, 61.52)**	(66.70, 48.75)
Swin-Unet 2	(78.04, 23.22)	(66.62, 50.15)	**(61.85, 61.45)**

Table 6. DAG Attack differing in number of iterations required for crafting attack. Target model is a UNet trained for 100 epochs

Attack Iters	DSC (Original)	DSC (Adversarial)	HD (Originall)	HD (Adversarial)
20	0.54026	0.60976	12.1645	10.3826
30	0.48732	0.61004	14.0706	10.3130
40	0.5035	0.6096	13.443	10.32823

6. Conclusion

There are plenty of room for future direction or improvement for the study of this topic. Mainly, a more concrete way to measure transferability for image segmentation task

Table 7. DAG Attack performance across Unet models differing only in initial seed

Attack Model	Target Model	DSC	HD
Model 1	Model 1	0.50763	13.48418
Model 2	Model 1	0.52175	13.40334
Model 1	Model2	0.50452	5.63809
Model 2	Model 2	0.52738	4.262631

Table 8. DAG Attack performance across UNet models differing in Number of epochs. Model 1 trained for 100 epochs, Model 2 trained for 60 epochs

Attack Model	Target Model	DSC	HD
Model 1	Model 1	0.50482	13.8421
Model 3	Model 1	0.51585	12.4458
Model 1	Model 3	0.52173	5.47198
Model 3	Model 3	0.49895	5.92498

could be used. One idea we had but wasn't able to adapt is by comparing the dice similarity coefficient between the prediction of a model against its own adversarial examples and the prediction of another model against the same adversarial examples.

To conclude, the main observation we made through out the experiments is that for image segmentation, adversarial examples only transfer between the same model architecture but not to different model architectures. There may be some practical application that can be benefit using this observation. For example, in an image segmentation task which speed of inference is less important, we can have an ensemble model that consist of 3 different model architectures that vote against each other. This could make the model more robust to transfer adversarial attack crafted on only one model architecture.

References

[1] Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? *Advances in Neural Information Processing Systems*, 32, 2019. 1

[2] Kyriakos D Apostolidis and George A Papakostas. A survey on adversarial deep learning robustness in medical image analysis. *Electronics*, 10(17):2132, 2021. 1

[3] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 888–897, 2018. 2

[4] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021. 1

[5] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 205–218. Springer, 2023. 1, 2

[6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2

[7] Jianxu Chen, Lin Yang, Yizhe Zhang, Mark Alber, and Danny Z Chen. Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. *Advances in neural information processing systems*, 29, 2016. 1

[8] Evelyn Duesterwald, Anupama Murthi, Ganesh Venkataraman, Mathieu Sinn, and Deepak Vijaykeerthy. Exploring the hyperparameter landscape of adversarial robustness. *arXiv preprint arXiv:1905.03837*, 2019. 1

[9] Volker Fischer, Mummadi Chaithanya Kumar, Jan Hendrik Metzen, and Thomas Brox. Adversarial examples for semantic image segmentation. *arXiv preprint arXiv:1703.01101*, 2017. 2

[10] Lin Gu, Yinqiang Zheng, Ryoma Bise, Imari Sato, Nobuaki Imanishi, and Sadakazu Aiso. Semi-supervised learning for biomedical image segmentation via forest oriented super pixels (voxels). In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 702–710. Springer, 2017. 1

[11] Hokuto Hirano, Akinori Minagi, and Kazuhiro Takemoto. Universal adversarial attacks on deep neural networks for medical image classification. *BMC medical imaging*, 21:1–13, 2021. 1

[12] Sandesh Kamath, Amit Deshpande, and KV Subrahmanyam. How do sgd hyperparameters in natural training affect adversarial robustness? *arXiv preprint arXiv:2006.11604*, 2020. 1

[13] Sara Kaviani, Ki Jin Han, and Insoo Sohn. Adversarial attacks and defenses on ai in medical imaging informatics: A survey. *Expert Systems with Applications*, page 116815, 2022. 1

[14] Sheeba Lal, Saeed Ur Rehman, Jamal Hussain Shah, Talha Meraj, Hafiz Tayyab Rauf, Robertas Damaševičius, Mazin Abed Mohammed, and Karrar Hameed Abdulkareem. Adversarial attack and defence through adversarial training and feature fusion for diabetic retinopathy recognition. *Sensors*, 21(11):3922, 2021. 1

[15] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and blackbox attacks. *arXiv preprint arXiv:1611.02770*, 2016. 2

[16] Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial ex-

amples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7838–7847, 2021. 1, 2

[17] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007. 2

[18] Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks. In *CVPR Workshops*, volume 2, page 2, 2017. 1

[19] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. *arXiv preprint arXiv:2106.04169*, 2021. 2

[20] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022. 1

[21] 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: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[22] Abhinav Sagar. Vitbis: vision transformer for biomedical image segmentation. In *Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning: 10th Workshop, CLIP 2021, Second Workshop, DCL 2021, First Workshop, LL-COVID19 2021, and First Workshop and Tutorial, PPML 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 2*, pages 34–45. Springer, 2021. 2

[23] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021. 1

[24] Nikhil Kumar Tomar, Debesh Jha, Michael A Riegler, Håvard D Johansen, Dag Johansen, Jens Rittscher, Pål Halvorsen, and Sharib Ali. Fanet: A feedback attention network for improved biomedical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 1

[25] Jeya Maria Jose Valanarasu, Vishwanath A Sindagi, Ilker Hacihaliloglu, and Vishal M Patel. Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pages 363–373. Springer, 2020. 1

[26] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1369–1378, 2017. 2

Appendix

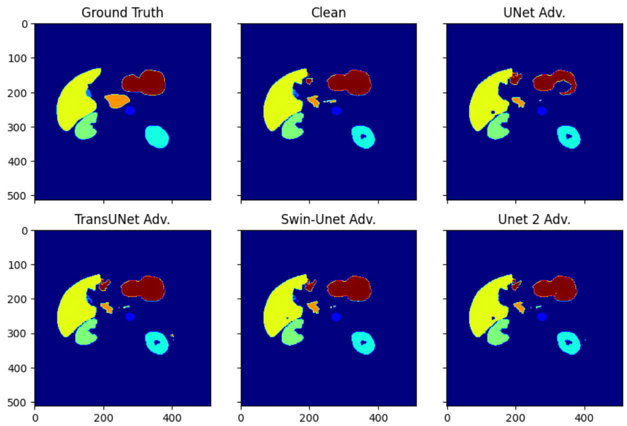

Figure 1. UNet against clean and adversarial examples

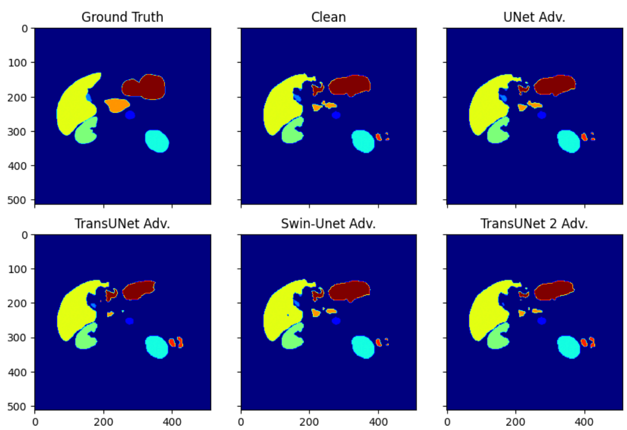

Figure 2. TransUNet against clean and adversarial examples

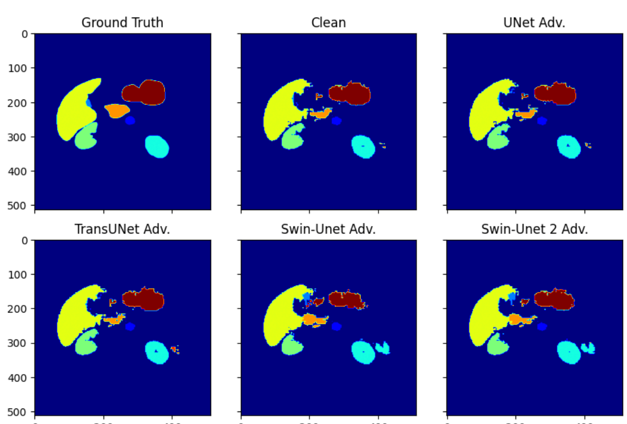

Figure 3. Swin-Unet against clean and adversarial examples

Looking at the predictions by eyes, it matches our expectation. From the tables in Section 4, the models against their own adversarial example should cause more change in the prediction than the others. As we noted previously, UNet seems to be more robust against transfer adversarial attack. We observed that the UNet 2 adversarial barely change the prediction of the UNet model. Whereas the adversarial example crafted on Swin-Unet 2 changes Swin-Unet prediction similar to the adversarial example crafted on Swin-Unet itself. These results match and reflect the scores in tables from Section 4.