IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 15, 2020, accepted September 28, 2020, date of publication October 12, 2020,
date of current version November 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3030235

Adversarial Perturbation on MRI Modalities in Brain Tumor Segmentation

GUOHUA CHENG[1] AND HONGLI JI[2]

[1]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Institute of Science and Technology for Brain-Inspired Intelligence, Ministry of Education, Fudan University, Shanghai 200433, China
[2]Jianpei Technology Company Ltd., Hangzhou 310000, China

Corresponding author: Guohua Cheng (17110850005@fudan.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1000400 and Grant 2019YFA0709502, in part by the Scientific Research Foundation of National Health and Family Planning Commission under Grant WKJ-ZJ-1814, in part by the Key Research and Development Program of Zhejiang Province under Grant 2019C03002, in part by the Hangzhou Major Science and Technology Innovation Project under Grant 20172011A038, in part by the 111 Project under Grant B18015, in part by the Key Project of Shanghai Science and Technology under Grant 16JC1420402, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2018SHZDZX01, in part by the ZJLab, in part by the National Key Research and Development Program of China under Grant 2018YFC1312900, and in part by the National Natural Science Foundation of China (NSFC) under Grant 91630314.

ABSTRACT Convolutional neural networks (CNNs) have been widely used by biomedical image segmentation applications. U-net, as a semantic segmentation method, has become a mainstream approach to brain tumor segmentation. However, the intrinsic vulnerability of CNNs also brings potential risks to all CNN-based applications, including semantic segmentation applications. In this paper, we create a universal adversarial perturbation and apply it on every modality in order to investigate how the adversarial perturbation affects each Magnetic Resonance Imaging(MRI) modality and the MRI images overall. We evaluate the performance when all four modalities are attacked and when one modality is attacked. The results show the following: 1) The adversarial perturbation affects the accuracy performance greatly, regardless of the size of the perturbation; 2) When only one modality is attacked, the network structure and the other three modalities provide some resistance to the adversarial perturbation; and 3) There are performance differences in different modalities, which are strongly related to the intensity distribution. T2 is least affected by the adversarial perturbation, while T1 and T1ce are more affected by the adversarial perturbation.

INDEX TERMS Adversarial perturbation, semantic segmentation, adversarial training.

I. INTRODUCTION

In recent years, deep learning networks and especially convolutional neural networks (CNNs) have achieved remarkable success in many computer vision areas, including object recognition [1], [2], semantic segmentation [3], [4] [5], depth estimation [6], object detection [7], [8], etc. Instead of crafting features by humans, CNNs use convolutional layers to automatically extract the features from the input images and provide end-to-end solutions to the perceptional task. Because CNN-based solutions have more accurate feature extraction and are more convenient in the processing pipeline, convolutional neural networks have also been widely used for biomedical segmentation tasks including lung segmentation [9], [10], brain tumor segmentation [11], etc.

The goal of brain tumor segmentation is to detect and localize tumor regions by comparing the tested brain tissue images to the normal brain tissue images [12]. The ground truth brain tissue images are labeled by a medical professional on special medical images such as X-ray or MRI images. Automatic brain tumor segmentation has been used to help medical personnel, reducing the time necessary for identification of abnormal regions. Automatic segmentation methods will be critical for early tumor prescreening, especially when doctors need to examine a large number of biomedical images.

Currently, magnetic resonance imaging (MRI) is a major type of biomedical imagery in brain tumor analysis, monitoring and surgery planning. Compared to the methods using handcrafted features, methods based on CNNs use a set of convolutional filters that can extract the convolutional features directly from the input data, providing end-to-end solutions to MRI images. Many state-of-the-art automatic brain tumor segmentation methods have been developed in the recent years, such as U-net [13] and V-net [12], [14]. Although both X-rays and CT images can be used for

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Shen.

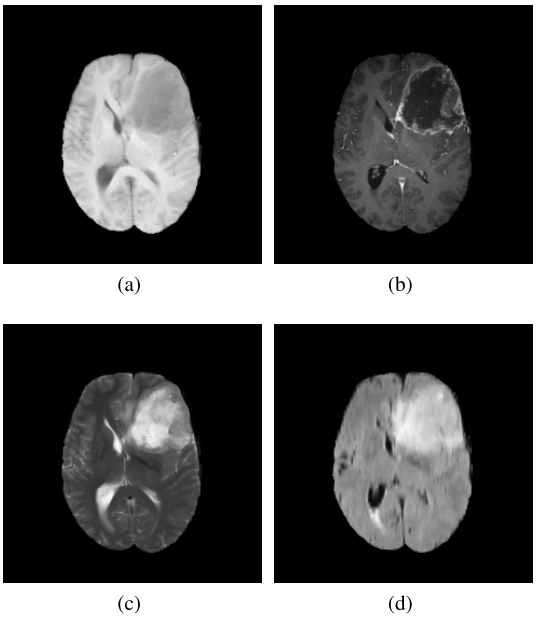


FIGURE 1. Image examples in MICCAI BraTS 2019. Four images belonging to a subject. a) T1 mode, b) T1ce mode, c) T2 mode, d) FLAIR mode.

biomedical research, MRI images are the major data source type for these neural networks.

Research on adversarial attacks has shown that typical convolution neural networks have a universal vulnerability to these attacks. Similar to other CNN-based applications, semantic segmentation applications have also been proven to be vulnerable to adversarial sample attack [15], [16]. However, compared to other semantic segmentation applications, brain tumor segmentation has several unique characteristics. Because brain tumors have different sizes, shapes, and locations in different patients, doctors and other medical personnel always use several modalities of MRI images to help tumor region segmentation and labeling. Different modalities have differences in the pixel intensity and contain different information. The comprehensive consideration of multiple modalities provides tumor tissues at multiple intensity levels for analysis by doctors. There are four types of modalities of MRI images: T1 (spin-lattice relaxation), contrast-enhanced T1 (T1ce), T2 (spin-spin relaxation) and FlAIR (fluid-attenuated inversion recovery). Each modality corresponds to grayscale images that highlight different kinds of tissue. In current brain tumor segmentation methods, all four modalities are used jointly for computation during the model training process.

As a result, it is beneficial to investigate how adversarial perturbation affects the brain tumor segmentation methods and to elucidate the impact of adversarial perturbation in different modalities.

The research described in this paper can be useful for the mitigation of many threats: 1) As the images are acquired from MRI equipment, system failures and human error may cause imperfect MRI images, possibly leading to the errors in the segmentation results. 2) As MRI images are valuable personal information, they are vulnerable to cyber attack and can be deliberately altered by adversaries.

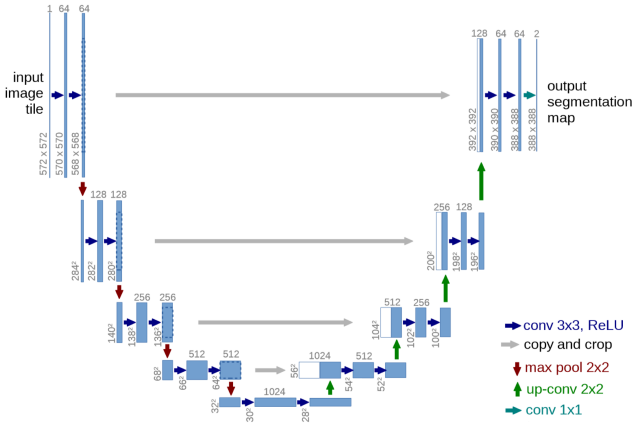


FIGURE 2. Layout of the U-net network structure.

In this paper, we first evaluate the adversarial perturbation effect on the current automatic brain tumor segmentation methods in terms of segmentation accuracy and then investigate the performance reduction for the adversarial attacks on each modality. This paper also presents an explanation for the accuracy differences among the four modalities. The adversarial training recommendation will be given in the end of the paper.

The rest of the paper is organized as follows: Section II introduces the background of U-net-like neural networks, and the background of adversarial sample attack. Section III introduces the dataset used for this research. Section IV describes the method for the generation of the perturbations for the brain tumor MRI images. Section V shows a typical U-net-like segmentation target model that serves as the attack target. The experimental configuration and results are discussed in Section VI. The conclusion and future work are given at the end of the paper.

II. RELATED WORK

A. U-NET STYLE NEURAL NETWORKS

Since the first application of U-net [13] in the biomedical segmentation area, a group of U-net style neural networks has been invented and applied to brain tumor segmentation [12], [14]. Depending on the dimension of the convolutional kernels, U-nets can be divided into two categories: 2D U-nets and 3D U-net. Similar to the 2D U-net, the 3D U-net also consider a few images before or after the current image, which is the so-called depth, using 3D covolutional kernels. Moreover, the modalities in MRI images are treated as channels of the input images. As a result, the memory capacity is always the bottleneck for 3D U-net training.V-net [14] is a variant of 3D U-net, in that case in this paper, we focus on the general U-net style neural network.

Figure 2 shows a typical U-net network structure. The right-hand part of Figure 2 is similar to the VGG model. Multiple convolution layers extract the convolutional features from the images. Both 2D and 3D versions of U-net have convolutional layers for feature extraction. When the adversarial perturbation is added to either some images or all images, it will compromise the feature map and extensively affect the

results of the downpooling layers. After computation in the subsequent layers, the final result will be affected.

B. ADVERSARIAL SAMPLE ATTACK

Research on adversarial sample attacks started in 2013. Szegedy *et al.* first found the adversarial sample phenomenon in the early neural networks: LeNet [1], [17] and Alexnet [2]. Studies then started to craft small and imperceptible perturbations and added them into images in order to lead the classifier to make mistakes. In subsequent research, researchers found that this vulnerability may be caused by the linear characteristic in neural networks, and it was proved to affect most neural network applications, including object recognition [18], object segmentation [15], [16], [19], depth estimation [20], etc.

Depending on whether the structure and the parameters of models are known to the adversary, adversarial sample attack can be categorized into two types: white-box attack [21], [22] and black-box attack [23], [24]. In a white-box attack, adversarial samples are generated based on the knowledge about the model such as the layers, parameters and loss functions.

Depending on whether the adversary has a specific targeted class, the adversarial sample attack can be categorized into types: targeted attack or nontargeted attack. A targeted attack seeks to make the classifier misclassify one class to another class or segment one image into a particular map that is crafted by the adversary. A nontargeted attack does not have a specific target. Instead it disturbs the classifier and makes the output of neural networks models unpredictable. In this paper, we focus on a nontargeted attack. The perturbations were added to the input images in order to disturb the classifier and evaluate the reliability of the current segmentation model.

III. THE DATASET

The dataset that we use in this paper is MICCAI BraTS 2019 [25], [26]. It is the largest publicly available dataset with MRI images of brain tumors. It contains 259 high-grade Glimoas (HGG) cases and 76 low-grade Glimoas cases in the training set. The validation set contains 125 cases. The training set contains four modality images, namely, T1, T1ce, T2, and FLAIR, and the ground truth images. The validation set contains only the four modality images. The users only can upload the segmentation files to the website and compare to the ground truth images online. All of the modality images were collected from different MRI scanners and different institutions. In this case, certain normalization techniques must be applied prior to using the images for training or validating.

IV. ADVERSARIAL PERTURBATION

A. UNIVERSAL RANDOM PERTURBATION

The method for generating the adversarial perturbation in this paper is different from that used in the work by Moosavi-Dezfooli *et al.* [27]. Similar to [27], we search for a universal perturbation that does not depend on the model that we are targeting. To randomly generate the adversarial perturbation, the adversarial noise consists of two parts: a max norm $vec1$ and an L_2 norm $vec2$. $vec1$ can be considered a random vector inside a unit hypercube, whereas $vec2$ can be considered a random vector inside an unit hypersphere. Additionally, the two parameters ϵ and rad must be set during the configuration stage.

$$pert = \epsilon * sgn(vec1) + rad * vec2 \tag{1}$$

Equation 1 describes the generation of the adversarial perturbation. Here, $vec1$ and $vec2$ are generated randomly according to a Gaussian distribution. ϵ and rad are set by the users.

B. PERTURBATION IN MRI IMAGES

Figure 3 shows the comparison between the original four MRI modality images and the adversarial images with $\epsilon = 15$ and $rad = 15$. Upon examination of the adversarial images, a slight variation in the intensity can be observed. Here, two features must be noticed. First, because the intensities of the MRI images are often larger than 1000, to present the image in grayscale for this paper, the images were normalized to be $0 \sim 255$, as shown in Figure 3. In our experiment, we still generate and add the perturbation to the images without grayscale normalization. Second, the values of ϵ and rad should be between 0 and 255. Larger values lead to an increase in the magnitude of the noise.

V. THE TARGETED MODEL

Although the adversarial perturbation that we generated is applicable to many semantic segmentation methods, we selected a pretrained model from the work of Lachinov *et al.* [28]. The goal of their research was to combine three U-net style neural networks and compare the performance between the ensemble model trained with the regular quantity of data and one single neural network trained with an additional quantity of data.

The ensemble neural networks consist of three U-net style neural networks. The first method is the typical U-net with a negative modification [29]. This method ranked second in the MICCAI BraTS 2018 challenge. The second method [12] in the ensemble is a U-net with residual connections [30]. This method combines the residual connections with an autoencoder branch that has a group batch regularization and reached first place in the MICCAI BraTS 2018 challenge. The third method [31] in the ensemble is Cascade U-net. A cascade of U-nets is used. Each U-net has multiple encoders that correspond to input modalities. This method also joined the MICCAI BraTS 2018 challenge.

VI. EXPERIMENT

All of the computational work is conducted on a virtual host provided by bitahub.com. The virtual host is equipped with an Nvidia GTX 1080ti GPU, 2vCPUs, and 16 GB RAM. Our evaluation experiment consists of two parts. The first part involves testing the adversarial perturbation effect on each modality. In this test, the adversarial samples are generated to attack the model trained with original images, in order to test whether the segmentation model can be tolerant to the adversarial samples. To do that, the images of one modality of

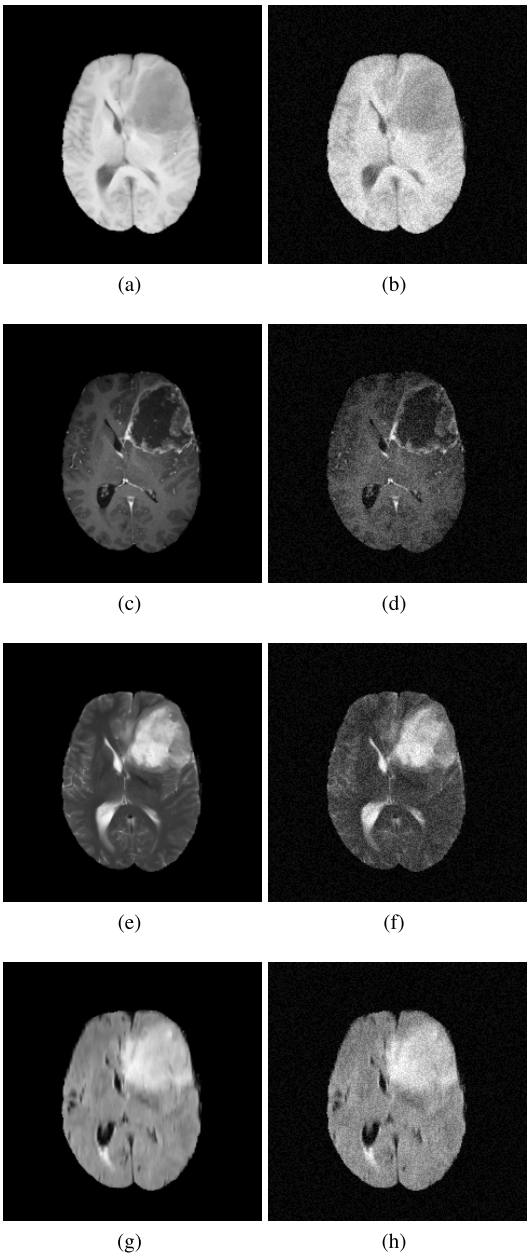


FIGURE 3. Adversarial samples generated with $\epsilon = 15$ and $rad = 15$ in Equation 1 in MICCAI 2019. All of the images belong to the same subject. The left-hand side images a), c), e), g) are T1, T1ce, T2, FLAIR modality images. The right-hand side of the images shows their corresponding adversarial samples.

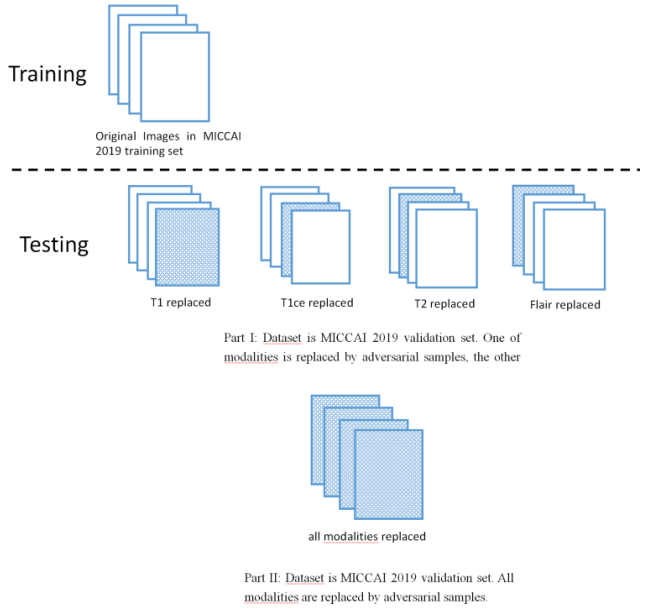


FIGURE 4. Configuration of our experiments. The model training used the original images from the MICCAI 2019 training set. The experiment for the reliability testing used the MICCAI 2019 validation set with some modalities altered by our adversarial samples.

TABLE 1. Original image performance baseline.

parameters	Dice_ET	Dice_WT	Dice_TC
Base Line with no perturbation	0.76	0.91	0.84

TABLE 2. Adversarial perturbations in Flair modality.

parameters	Dice_ET	Dice_WT	Dice_TC
$\epsilon = 5, rad = 5$	0.73	0.89	0.82
$\epsilon = 15, rad = 15$	0.73	0.89	0.82
$\epsilon = 30, rad = 30$	0.71	0.87	0.79

TABLE 3. Adversarial perturbations in T1 modality.

parameters	Dice_ET	Dice_WT	Dice_TC
$\epsilon = 5, rad = 5$	0.71	0.88	0.76
$\epsilon = 15, rad = 15$	0.71	0.88	0.76
$\epsilon = 30, rad = 30$	0.71	0.88	0.76

TABLE 4. Adversarial perturbations in T1ce modality.

parameters	Dice_ET	Dice_WT	Dice_TC
$\epsilon = 5, rad = 5$	0.72	0.87	0.78
$\epsilon = 15, rad = 15$	0.72	0.87	0.77
$\epsilon = 30, rad = 30$	**0.68**	0.87	0.75

the original four modalities are replaced by our crafted adversarial samples, and keep the other modalities unchanged. The second part involves testing the adversarial perturbation effect on all of the modalities. In this test, all of the modalities are replaced by our adversarial samples. Each part of our experiment imitates the potential risk in reality. The configuration of the experiment is shown in Figure 4

A. PART I. ADVERSARIAL PERTURBATION EFFECT ON EACH MODALITY

Since all four modalities participate in the training and testing phase as four channels of the input images, this experiment simulates the case where one modality image is contaminated by equipment failures or an adversary. For each subject in the dataset, one of the modalities among the original four modalities is replaced by adversarial samples, and the images of the other modalities are included with no change. The segmentation results are sent to the CBICA official website[1] for the evaluation. The results are presented in Tables 2, 3, 4 and 5.

Tables 2, 3, 4 and 5 compare the results obtained for the validation dataset. After testing different ϵ and *rad* values, the results are presented in terms of mean of Dice_ET,

[1] https://ipp.cbica.upenn.edu/

TABLE 5. Adversarial perturbations in T2 modality.

parameters	Dice_ET	Dice_WT	Dice_TC
$\epsilon = 5, rad = 5$	0.75	0.88	0.84
$\epsilon = 15, rad = 15$	0.75	0.89	0.84
$\epsilon = 30, rad = 30$	0.75	0.90	**0.84**

TABLE 6. Adversarial perturbations in all four modalities.

parameters	Dice_ET	Dice_WT	Dice_TC
$\epsilon = 1, rad = 1$	0.40	0.23	0.32
$\epsilon = 5, rad = 5$	0.39	0.23	0.31
$\epsilon = 15, rad = 15$	0.40	0.23	0.30
$\epsilon = 30, rad = 30$	0.41	0.26	0.31

Dice_WT and Dice_TC. An examination of the results shows that 1) T2 has the lowest performance degradation in an adversarial attack among the four modalities. Meanwhile, T1ce displays the greatest performance degradation in an adversarial attack; 2) by changing the size of the perturbation, the degradation shows a slight variation for different parameter settings but mostly remains stable within one modality.

B. PART II. ADVERSARIAL PERTUBATION EFFECT ON ALL MODALITIES

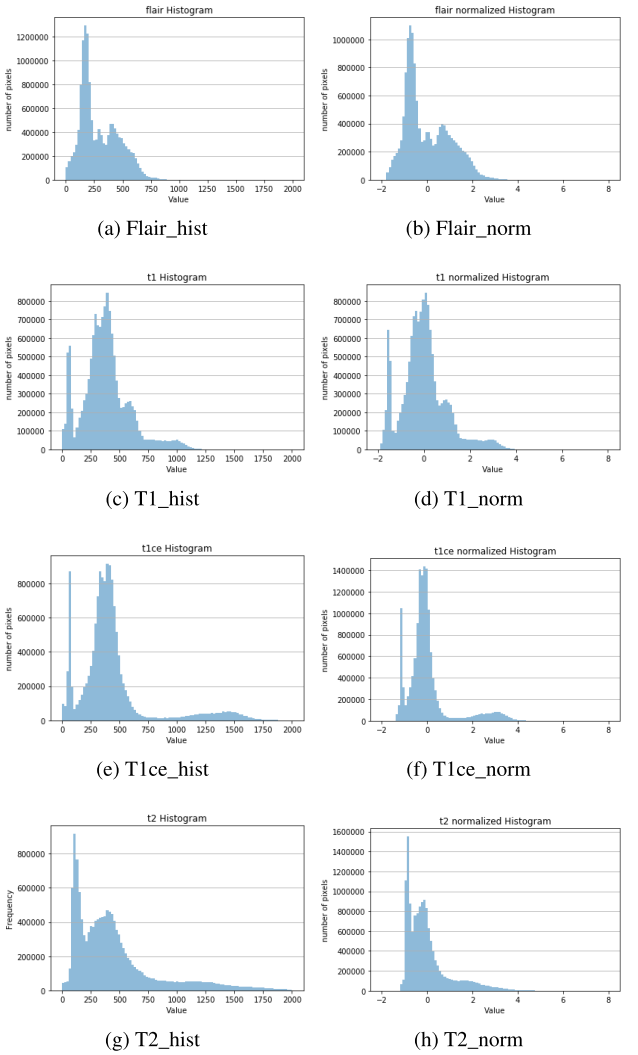
To test the adversarial perturbation effect on all of the modalities, we replaced the images for all four modalities with our adversarial samples.

Table 6 shows the obtained dice scores for different ϵ and *rad* values. The performance shows severe degradation when the images for all four modalities are replaced. Even when the perturbation size is very small ($\epsilon = 1, rad = 1$), the degradation is also severe. Similar to the results in Tables 2, 3, 4, and 5, changing the size of ϵ and *rad* does not appreciably change the performance.

Comparing to the results presented in Tables 2, 3, 4, and 5, one interesting observation is that although degradation is obtained when testing on each modality, the other three modalities show a certain tolerance to adversarial perturbation, which can prevent extreme performance degradation. This result implies that even if the images for one modality are compromised for a certain reason, medical personnel can still obtain usable results.

VII. DISCUSSION

The intensity difference is the major difference between the images of the different MRI modalities. We believe that the distribution of the intensity for each modality is strongly related to the different performance degradations obtained in the Part I experiments. To verify this hypothesis, we tried two sets of experiments. We first plotted the histograms for both the original images of the validation dataset and the images after normalization. Figure 5 compares the histogram distribution between the original four modality images and the images after normalization. Regardless of the normalization, the distribution of the intensity in the images does not change. As a result, the adversarial perturbation added to the images will not be affected by normalization. The accuracy degradation shoulde not related to any kind of normalization.



(a) Flair_hist (b) Flair_norm

(c) T1_hist (d) T1_norm

(e) T1ce_hist (f) T1ce_norm

(g) T2_hist (h) T2_norm

FIGURE 5. Comparison between the histogram of the original MRI images and the histogram of original images after normalization.

The second set of experiments is plotting the quantile-quantile figure for each MRI modality and Gaussian distribution. The quantile-quantile (q-q) plot is a graphical technique for determining whether two data sets come from populations with a common distribution. In Figure 6, we compare the intensity in each modality to a Gaussian distribution. If the curve is close to the reference line, then it has a similar distribution to a Gaussian distribution. A smaller distance of the curve from the reference line implies a greater similarity of the curve to a Gaussian distribution. T2 shows the highest deviation from the reference line and is the least similar to the Gaussion distribution.

Table 7 shows the slope and intercept of the reference line generated by least-squares regression (best-fit). The slope represents the standard deviation of the intensity, and the intercept represents the mean of the intensity. It is observed that T2 has the highest standard deviation and mean, indicating that it is the least similar to the Gaussian distribution. The other modalities show lower standard deviation and mean values.

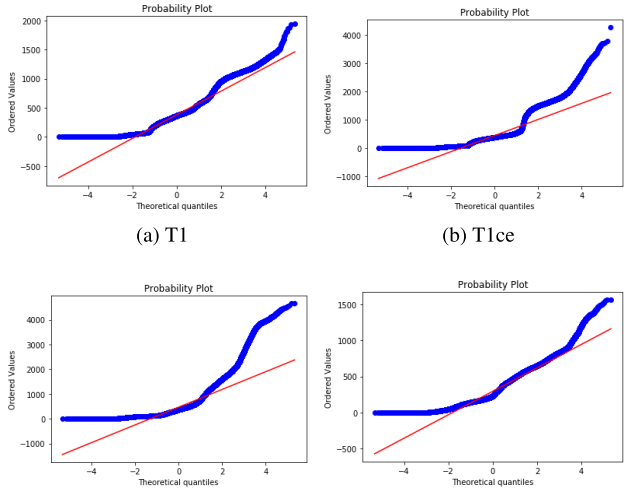


FIGURE 6. The probability plotting of the intensity distribution fitting to the Gaussian distribution. The plots are based on least-squares regression (best-fit). If the curve is close to the reference line, then the curve has a similar distribution to the reference line. The closer the curve is to the reference line, the more similar it is to the Gaussian distribution. T2 shows the highest deviation from the reference line and is the least similar to a Gaussian distribution. More quantitative information is shown in Table 7.

TABLE 7. Slope and intercept summary of probplot in Figure 6.

Modalities	slope	intercept
T1	203.136	379.927
T1ce	285.585	437.846
T2	**358.540**	**466.197**
Flair	162.865	294.221

The explanation for the differences in accuracy degradation is as follows. Because the perturbation that we used for the experiment uses the Gaussian distribution to randomly generate $vec1$ and $vec2$ in Equation 1, the intensity distribution of the modalities that is similar to the Gaussian distribution will be affected more than those in the other cases, and the intensity distributions of the modalities that are less like a Gaussian distribution will be less affected. It is observed from Figure 5 that T2 is the least similar to the Gaussian distribution, and is the least affected by perturbation. By contrast, T1 and T1ce basically follow a Gaussian distribution pattern and are more vulnerable to our adversarial perturbation.

VIII. CONCLUSION

In this paper, we generated a universal random perturbation for each modality in brain tumor segmentation. Several general conclusions are obtained by attacking the state-of-the-art segmentation model: 1) When four modalities are attacked or damaged, a severe performance degradation in accuracy will occur. However, when one of the modalities is attacked or damaged, the accuracy also drops. In both of our tests, the performance reduction is not related to the size of the perturbation. 2) If only one of the modality images is attacked or damaged, the performance does not decrease strongly. Due to the use of images from other channels, the existing U-net-style neural network models are robust if only one modality image is attacked or damaged. 3) For an individual modality, the modalities that have similar intensity distributions to the Gaussian distribution are more vulnerable to our adversarial perturbation and vice versa.

REFERENCES

[1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[4] Y. Tian, G. Cheng, J. Gelernter, S. Yu, C. Song, and B. Yang, "Joint temporal context exploitation and active learning for video segmentation," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107158.
[5] Y. Tian, Y. Zhang, D. Zhou, G. Cheng, W.-G. Chen, and R. Wang, "Triple attention network for video segmentation," *Neurocomputing*, vol. 417, pp. 202–211, Dec. 2020.
[6] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5162–5170.
[7] Y. Tian, X. Wang, J. Wu, R. Wang, and B. Yang, "Multi-scale hierarchical residual network for dense captioning," *J. Artif. Intell. Res.*, vol. 64, pp. 181–196, Jan. 2019.
[8] Y. Tian, H. Wang, and X. Wang, "Object localization via evaluation multi-task learning," *Neurocomputing*, vol. 253, pp. 34–41, Aug. 2017.
[9] S. Hwang and S. Park, "Accurate lung segmentation via network-wise training of convolutional networks," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2017, pp. 92–99.
[10] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.
[11] A. Casamitjana, S. Puch, A. Aduriz, and V. Vilaplana, "3D convolutional neural networks for brain tumor Segmentation: A comparison of multi-resolution architectures," in *Proc. Int. Workshop Brainlesion, Glioma, Multiple Sclerosis, Stroke Traumatic Brain Injuries*. Cham, Switzerland: Springer, 2016, pp. 150–161.
[12] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *Proc. Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, 2018, pp. 311–320.
[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[14] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.
[15] V. Fischer, M. Chaithanya Kumar, J. Hendrik Metzen, and T. Brox, "Adversarial examples for semantic image segmentation," 2017, *arXiv:1703.01101*. [Online]. Available: http://arxiv.org/abs/1703.01101
[16] J. H. Metzen, M. C. Kumar, T. Brox, and V. Fischer, "Universal adversarial perturbations against semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2755–2764.
[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. *LeNet-5, Convolutional Neural Networks*. Accessed: Mar. 2020. [Online]. Available: http://yann.lecun.com/exdb/lenet
[18] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2016, *arXiv:1607.02533*. [Online]. Available: http://arxiv.org/abs/1607.02533
[19] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1369–1378.
[20] J. Hu and T. Okatani, "Analysis of deep networks for monocular depth estimation through adversarial attacks with proposal of a defense method," 2019, *arXiv:1911.08790*. [Online]. Available: http://arxiv.org/abs/1911.08790
[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*. [Online]. Available: http://arxiv.org/abs/1412.6572

[22] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.
[23] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, Apr. 2017, pp. 506–519.
[24] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," 2018, *arXiv:1804.08598*. [Online]. Available: http://arxiv.org/abs/1804.08598
[25] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, and L. Lanczi, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.
[26] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, and M. Prastawa, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018, *arXiv:1811.02629*. Accessed: Mar. 2020. [Online]. Available: http://arxiv.org/abs/1811.02629
[27] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1765–1773.
[28] D. Lachinov, E. Shipunova, and V. Turlapov, "Knowledge distillation for brain tumor segmentation," 2020, *arXiv:2002.03688*. [Online]. Available: http://arxiv.org/abs/2002.03688
[29] A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers*, vol. 10670. Cham, Switzerland: Springer, 2018.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[31] D. Lachinov, E. Vasiliev, and V. Turlapov, "Glioma segmentation with cascaded UNet," in *Proc. Int. MICCAI Brainlesion Workshop*. Springer, 2018, pp. 189–198.

GUOHUA CHENG received the master's degree from Nanyang Technological University, Singapore. He is currently pursuing the Ph.D. degree with Fudan University, Shanghai, China. He has been the CEO of Jianpei Technology Company Ltd., a 1000 Talents Plan Member of Zhejiang Province. He also works on medical image artificial intelligence. His current interests include machine learning and biomedical engineering.

HONGLI JI received the bachelor's degree from Shanghai Jiao Tong University (SJTU) and the Ph.D. degree from Nanyang Technological University. She, after returning to her country in 2012, has participated in the establishment of Jianpei Technology Company Ltd., in the capability of Chief Technology Officer (CTO) in an attempt to analyze the medical data mining and study the artificial intelligence technology (AIT).