Frequency Domain Adversarial Training for Robust Volumetric Medical Segmentation

Asif Hanif[1(✉)], Muzammal Naseer[1], Salman Khan[1], Mubarak Shah[2], and Fahad Shahbaz Khan[1,3]

[1] Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE
{asif.hanif,muzammal.naseer,salman.khan,fahad.khan}@mbzuai.ac.ae
[2] University of Central Florida (UCF), Orlando, USA
shah@crcv.ucf.edu
[3] Linköping University, Linköping, Sweden

Abstract. It is imperative to ensure the robustness of deep learning models in critical applications such as, healthcare. While recent advances in deep learning have improved the performance of volumetric medical image segmentation models, these models cannot be deployed for real-world applications immediately due to their vulnerability to adversarial attacks. We present a 3D frequency domain adversarial attack for volumetric medical image segmentation models and demonstrate its advantages over conventional input or voxel domain attacks. Using our proposed attack, we introduce a novel frequency domain adversarial training approach for optimizing a robust model against voxel and frequency domain attacks. Moreover, we propose frequency consistency loss to regulate our frequency domain adversarial training that achieves a better tradeoff between model's performance on clean and adversarial samples. Code is available at https://github.com/asif-hanif/vafa.
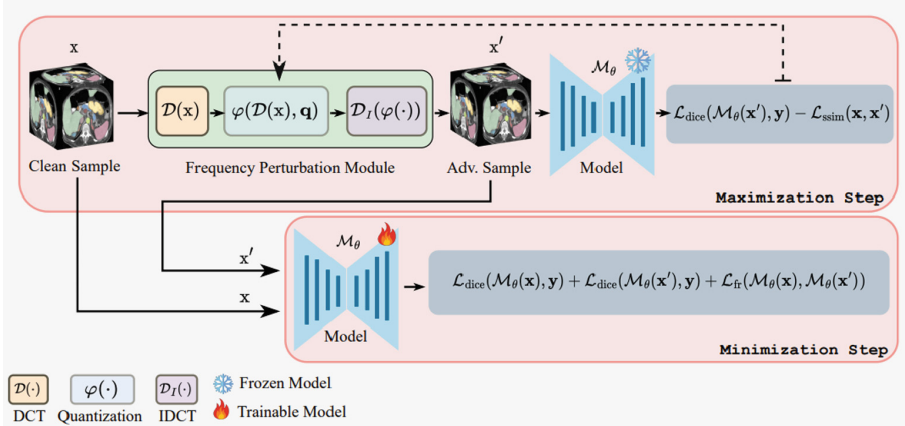
Keywords: Adversarial attack · Adversarial training · Frequency domain attack · Volumetric medical segmentation

1 Introduction

Semantic segmentation of organs, anatomical structures, or anomalies in medical images (e.g. CT or MRI scans) remains one of the fundamental tasks in medical image analysis. Volumetric medical image segmentation (MIS) helps healthcare professionals to diagnose conditions more accurately, plan medical treatments, and perform image-guided procedures. Although deep neural networks (DNNs) have shown remarkable improvements in performance for different vision tasks,

Fig. 1. Overview of Adversarial Frequency Attack and Training: A model trained on voxel-domain adversarial attacks is vulnerable to frequency-domain adversarial attacks. In our proposed adversarial training method, we generate adversarial samples by perturbing their frequency-domain representation using a novel module named "Frequency Perturbation". The model is then updated while minimizing the dice loss on clean and adversarially perturbed images. Furthermore, we propose a frequency consistency loss to improve the model performance.

including volumetric MIS, their real-world deployment is not straightforward particularly due to the vulnerabilities towards adversarial attacks [26]. An adversary can deliberately manipulate input data by crafting and adding perturbations to the input that are imperceptible to the human eye but cause the DNN to produce incorrect outputs [10]. Adversarial attacks pose a serious security threat to DNNs [1], as they can be used to cause DNNs to make incorrect predictions in a wide range of applications, including DNN-based medical imaging systems. To mitigate these threats, various techniques have been explored, including adversarial training, input data transformations, randomization, denoising auto-encoders, feature squeezing, and robust architectural changes [1]. Although significant progress has been made in adversarial defenses, however, this area is still evolving due to the development of attacks over time [3].

Ensuring the adversarial robustness of the models involved in safety-critical applications such as, medical imaging and healthcare is of paramount importance because a misdiagnosis or incorrect decision can result in life-threatening implications. Moreover, the weak robustness of deep learning-based medical imaging models will create a trust deficit among clinicians, making them reluctant to rely on the model predictions. The adversarial robustness of the medical imaging models is still an open and under-explored area [6,20]. Furthermore, most adversarial attacks and defenses have been designed for 2D natural images and little effort has been made to secure volumetric (3D) medical data [20].

In the context of 2D natural images, it has been recently observed that frequency-domain based adversarial attacks are more effective against the

defenses that are primarily designed to "undo" the impact of pixel-domain adversarial noise in natural images [7]. Motivated by this observation in 2D natural images, here we explore the effectiveness of frequency-domain based adversarial attacks in the regime of volumetric medical image segmentation and aim to obtain a volumetric MIS model that is robust against adversarial attacks. To achieve this goal, we propose a *min-max* objective for adversarial training of *volumetric MIS model in frequency-domain*. For *maximization* step, we introduce **V**olumetric **A**dversarial **F**requency **A**ttack - **VAFA** (Fig. 1, Sect. 2.1) which operates in the frequency-domain of the data (unlike other prevalent voxel-domain attacks) and explicitly takes into account the 3D nature of the volumetric medical data to achieve higher fooling rate. For *minimization* step, we propose **V**olumetric **A**dversarial **F**requency-domain **T**raining - **VAFT** (Fig. 1, Sect. 2.2) to obtain a model that is robust to adversarial attacks. In VAFT, we update model parameters on clean and adversarial (obtained via VAFA) samples and further introduce a novel *frequency consistency loss* to keep frequency representation of the logits of clean and adversarial samples close to each other for a better accuracy tradeoff. In summary, our contributions are as follows:

– We propose an approach with a min-max objective for adversarial training of volumetric MIS model in the frequency domain. In the maximization step, we introduce a volumetric adversarial frequency attack (VAFA) that is specifically designed for volumetric medical data to achieve higher fooling rate. Further, we introduce a volumetric adversarial frequency-domain training (VAFT) based on a frequency consistency loss in the minimization step to produce a model that is robust to adversarial attacks.
– We conduct experiments with two different hybrid CNN-transformers based volumetric medical segmentation methods for multi-organ segmentation.

Related Work: There are three main types of popular volumetric MIS model architectures: CNN [23], Transformer [13] and hybrid [12,24]. Research has shown that medical machine learning models can be manipulated in various ways by an attacker, such as adding imperceptible perturbation to the image, rotating the image, or modifying medical text [8]. Adversarial attack studies on medical data have primarily focused on classification problems and voxel-domain adversaries. For example, Ma *et al.* [20] have used four types of pixel-domain attacks [4,10,16,21] on two-class and multi-class medical datasets. Li *et al.* [19] and Daza *et al.* [6] have focused on single-step and iterative adversarial attacks [5,10,15] on the volumetric MIS. In constant to voxel-domain adversarial attacks, our approach works in the frequency-domain.

2 Frequency Domain Adversarial Attack and Training

We aim to train a model for volumetric medical segmentation that is robust against adversarial attacks. Existing adversarial training (AT) approaches rely on min-max optimization [10,16,21] and operate in the input space. They find adversaries by adding the adversarial perturbation to the input samples by
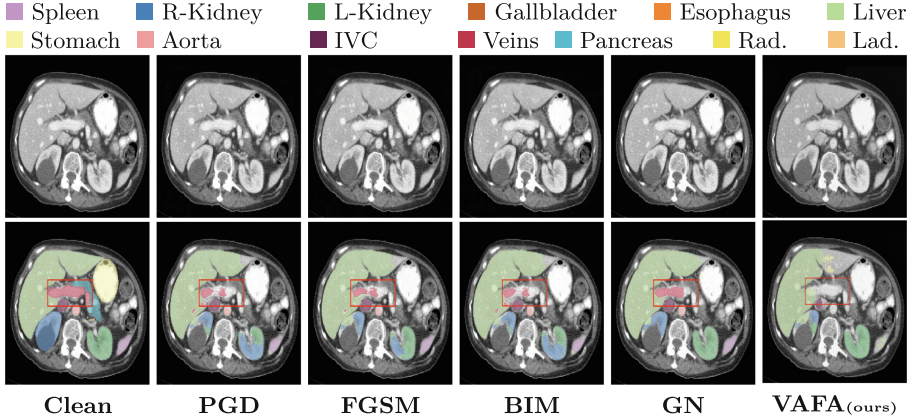
Fig. 2. Qualitative multi-organ segmentation comparison under different attacks on the UNETR [12] model. Top row shows example images and bottom row shows the corresponding segmentation masks predicted by the model under different attacks. Compared to different voxel-domain attacks (PGD [21], FGSM [10], BIM [16] and GN [14]), our attack (VAFA) achieves higher fooling rate (highlighted in red bounding box) while maintaining comparable perceptual similarity. Best viewed zoomed in.

maximizing the model loss (e.g., dice loss in segmentation). The loss function is then minimized on such adversaries to update the model parameters. In this work, we propose a frequency-domain adversarial attack that takes into account the 3D nature of the volumetric medical data and performs significantly better than the other voxel-domain as well as 2D frequency domain attacks (Table 1). Based on our attack, we then introduce a novel frequency-domain adversarial training to make the model resilient to adversarial attacks. Additionally, we observe that our approach improves/retains the performance of the robust model on clean samples when compared to the non-robust model. Our approach optimizes adversarial samples by perturbing the 3D-DCT coefficients within the frequency domain using our frequency perturbation module (Fig. 1) and adversarial guidance from the segmentation loss (Sect. 2.1). We find adversarial samples with high perceptual quality by maximizing the structural similarity between clean and adversarial samples. Using clean and adversarial samples, we propose updating the model parameters by simultaneously minimizing the segmentation loss (i.e. Dice loss) and the frequency consistency loss (Eq. 4) between the clean and adversarial outputs of the segmentation model.

3D Medical Segmentation Framework: Deep learning-based 3D medical segmentation generally uses encoder-decoder architectures [18]. The encoder produces a latent representation of the input sample. A segmentation map of the input sample is generated by the decoder using the latent feature representation. The decoder usually incorporates skip connections from the encoder to preserve spatial information [12]. Next, we describe our proposed volumetric frequency-domain adversarial attack in Sect. 2.1 and then training in Sect. 2.2.

2.1   Volumetric Adversarial Frequency Attack (VAFA)

Generally, adversarial attacks operate in the voxel domain by adding an imperceptible perturbation to the input data. In contrast, our attack perturbs the 3D-DCT coefficient to launch a frequency-domain attack for 3D medical image segmentation. Our *Frequency Perturbation Module* (FPM) transforms voxel-domain data into frequency-domain by using discrete cosine transforms (DCTs) and perturbs the DCT coefficients using a *learnable quantization*. It then takes an inverse DCT of the perturbed frequency-domain data and returns voxel-domain image. We keep the model in a "frozen" state while maximizing the dice loss [25] for segmentation and minimizing structural similarity loss [27] for perceptual quality. We represent a 3D (volumetric) single channel clean sample by $X \in \mathbb{R}^{1 \times H \times W \times D}$ and its ground-truth binary segmentation mask by $Y \in \{0, 1\}^{\text{NumClass} \times H \times W \times D}$, where "NumClass" is the number of classes. We split X into n 3D patches i.e. $X \mapsto \{\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^{h \times w \times d}$ and $h \leq H, w \leq W, d \leq D, h = w = d$. We apply our frequency perturbation module to each of these patches.

Frequency Perturbation Module: We apply a 3D discrete cosine transform (DCT), represented as $\mathcal{D}(\cdot)$, to each patch $\mathbf{x}_i$. The resulting DCT coefficients are then processed through a function $\varphi(\cdot)$, which performs three operations: quantization, differentiable rounding (as described in [9]), and subsequent de-quantization. $\varphi(\cdot)$ utilizes a learnable quantization table $\mathbf{q} \in \mathbb{Z}^{h \times w \times d}$ to modify the DCT coefficients, setting some of them to zero. In particular, $\varphi(\mathcal{D}(\mathbf{x}), \mathbf{q}) := \lfloor \frac{\mathcal{D}(\mathbf{x})}{\mathbf{q}} \rfloor \odot \mathbf{q}$, where DCT coefficients of a patch (i.e. $\mathcal{D}(\mathbf{x})$) are element-wise divided by quantization table $\mathbf{q}$. After the division operation, the result undergoes rounding using a differentiable rounding operation [9], resulting in some values being rounded down to zero. The de-quantization step involves element-wise multiplication of $\lfloor \frac{\mathcal{D}(\mathbf{x})}{\mathbf{q}} \rfloor$ with the same quantization table $\mathbf{q}$. This step allows us to reconstruct the quantized DCT coefficients. Since quantization table is in the denominator of the division operation, therefore, higher quantization table values increase the possibility of more DCT coefficients being rounded down to zero. To control the number of DCT coefficients being set to zero, we can constrain the values of the quantization table to a maximum threshold (constraint in Eq. 2). In other words, $\varphi(\cdot)$ performs a 3D adversarial lossy compression on input through a learnable quantization table. Finally, a 3D inverse DCT (IDCT) is performed on the output of $\varphi(\cdot)$ in order to obtain an adversarially perturbed voxel-domain representation, denoted by $\mathbf{x}'$. We show our frequency perturbation module in Eq. 1 as follows:

$$\mathbf{x} \mapsto \mathcal{D}(\mathbf{x}) \mapsto \underbrace{\varphi(\mathcal{D}(\mathbf{x}), \mathbf{q})}_{\substack{\text{quantization,} \\ \text{rounding and} \\ \text{de-quantization}}} \mapsto \mathcal{D}_I(\varphi(\cdot)) \mapsto \mathbf{x}' \tag{1}$$

We repeat the above mentioned sequence of transformations for all patches and then merge $\{\mathbf{x}_i'\}_{i=1}^n$ to form adversarial image $X' \in \mathbb{R}^{H \times W \times D}$.

Quantization Constraint: We learn quantization table $\mathbf{q}$ by maximizing the $\mathcal{L}_{\text{dice}}$ while ensuring that $\|\mathbf{q}\|_\infty \leq q_{\max}$. Quantization threshold $q_{\max}$ controls the

---

Algorithm 1. Volumetric Adversarial Frequency Attack (**VAFA**)

1: Number of Steps: T, Quantization Threshold: $q_{\max}$
2: **Input:** $X \in \mathbb{R}^{H \times W \times D}$, $Y \in \{0,1\}^{\text{NumClass} \times H \times W \times D}$ **Output:** $X' \in \mathbb{R}^{H \times W \times D}$
3: **function VAFA**(X,Y)
4:     $\mathbf{q}_i \leftarrow \mathbf{1}$ $\forall\, i \in \{1,2,\ldots,n\}$ ▷ Initialize all *quantization tables* with ones.
5:     **for** $t \leftarrow 1$ to T **do**
6:         $\{\mathbf{x}_i\}_{i=1}^n \leftarrow \text{Split}(X)$ ▷ Split X into 3D patches of size $(h \times w \times d)$
7:         $\mathbf{x}_i' \leftarrow \mathcal{D}_I\big(\,\varphi(\mathcal{D}(\mathbf{x}_i), \mathbf{q}_i)\,\big)$ $\forall\, i \in \{1,2,\ldots,n\}$ ▷ Frequency Perturbation
8:         $X' \leftarrow \text{Merge}(\{\mathbf{x}_i'\}_{i=1}^n)$ ▷ Merge all adversarial patches to form X'
9:         $\mathcal{L}(X, X', Y) = \mathcal{L}_{\text{dice}}(\mathcal{M}_\theta(X'), Y) - \mathcal{L}_{\text{ssim}}(X, X')$
10:        $\mathbf{q}_i \leftarrow \mathbf{q}_i + \text{sign}(\nabla_{\mathbf{q}_i}\mathcal{L})$ $\forall\, i \in \{1,2,\ldots,n\}$
11:        $\mathbf{q}_i \leftarrow \text{clip}(\mathbf{q}_i,\ \min{=}1,\ \max{=}q_{\max})$ $\forall\, i \in \{1,2,\ldots,n\}$
12:    **end for**
13: **end function**
14: **Return** X'

Algorithm 2. Volumetric Adversarial Frequency Training (**VAFT**)

1: Train Dataset: $\mathcal{X} = \{(X_i, Y_i)\}_{i=1}^N$, $X_i \in \mathbb{R}^{H \times W \times D}$, $Y_i \in \{0,1\}^{\text{NumClass} \times H \times W \times D}$
2: NumSamples$=N$, BatchSize$=B$, Target Model: $\mathcal{M}_\theta$, AT Robust Model: $\mathcal{M}_{\boldsymbol{\varphi}}$
3: **for** $i \leftarrow 1$ to NumEpochs **do**
4:     **for** $j \leftarrow 1$ to $\lfloor N/B \rfloor$ **do**
5:         Sample a mini-batch $\mathcal{B} \subseteq \mathcal{X}$ of size B
6:         $X' \leftarrow$ **VAFA**(X, Y) $\forall (X, Y) \in \mathcal{B}$ ▷ Adv. Freq. Attack on clean images.
7:         $\mathcal{L} = \mathcal{L}_{\text{dice}}(\mathcal{M}_\theta(X), Y) + \mathcal{L}_{\text{dice}}(\mathcal{M}_\theta(X'), Y) + \mathcal{L}_{\text{fr}}(\mathcal{M}_\theta(X), \mathcal{M}_\theta(X'))$
8:         Backward pass and update $\mathcal{M}_\theta$
9:     **end for**
10: **end for**
11: $\mathcal{M}_{\boldsymbol{\varphi}} \leftarrow \mathcal{M}_\theta$ ▷ AT robust model after training completion.
12: **Return** $\mathcal{M}_{\boldsymbol{\varphi}}$

---

extent to which DCT coefficients are perturbed. The higher the value of $q_{\max}$, the more information is lost. The drop in perception quality of the adversarial sample and the accuracy of the model are directly proportional to the value of $q_{\max}$. To increase the perceptual quality of adversarial samples, we also minimize the structural similarity loss [27] between clean and adversarial samples, denoted by $\mathcal{L}_{\text{ssim}}(X, X')$, in optimization objective. Our attack optimizes the following objective to fool a target model $\mathcal{M}_\theta$:

$$\underset{\mathbf{q}}{\text{maximize}}\ \ \mathcal{L}_{\text{dice}}(\mathcal{M}_\theta(X'), Y) - \mathcal{L}_{\text{ssim}}(X, X') \tag{2}$$
$$\text{s.t.}\ \ \|\mathbf{q}\|_\infty \leq q_{\max},$$

where $\mathcal{L}_{\text{ssim}}(X, X') = 1 - \frac{1}{n}\sum_{i=1}^n \text{SSIM}(\mathbf{x}_i, \mathbf{x}_i')$ is structural similarity loss [27]. Algorithm 1 presents our volumetric adversarial frequency attack (VAFA). An overview of the attack can be found in *maximization* step of Fig. 1.

2.2   Volumetric Adversarial Frequency Training (VAFT)

The model parameters are then updated by minimizing the segmentation loss on both clean and adversarial samples (Eq. 3). Since our attack disrupts the frequency domain to find adversaries, we develop a novel frequency consistency loss (Eq. 4) to encourage frequency domain representation of the model's output (segmentation logits) for the clean sample close to the adversarial sample. Our frequency consistency loss not only boosts the robustness of the model against adversarial attacks but also improves/retains the performance of the robust model on clean images (Sect. 3). We present our volumetric adversarial frequency training (VAFT) in Algorithm 2.

$$\underset{\theta}{\text{minimize}}\ \mathcal{L}_{\text{dice}}(\mathcal{M}_\theta(X), Y) + \mathcal{L}_{\text{dice}}(\mathcal{M}_\theta(X'), Y) + \mathcal{L}_{\text{fr}}(\mathcal{M}_\theta(X), \mathcal{M}_\theta(X')), \quad (3)$$

$$\mathcal{L}_{\text{fr}}(\mathcal{M}_\theta(X), \mathcal{M}_\theta(X')) = \|\mathcal{D}(\mathcal{M}_\theta(X)) - \mathcal{D}(\mathcal{M}_\theta(X'))\|_1, \quad (4)$$

where $X' = \textbf{VAFA}(X, Y)$ and $\mathcal{D}(\cdot)$ is 3D DCT function. An overview of the adversarial training can be found in *minimization* step of Fig. 1.

Figure 2 presents a qualitative results of adversarial examples under different attacks on the standard UNETR model. We highlight areas by red bounding box in Fig. 2 to show the impact of each attack on the model performance, when compared with prediction on clean sample. Our attack (VAFA) achieves higher fooling rate as compared to other voxel-domain attacks, while maintaining comparable perceptual similarity.

3   Experiments and Results

Implementation Details: We demonstrate the effectiveness of our approach using two medical segmentation models: UNETR [12], UNETR++ [24] and two datasets: Synapse (18–12 split) [17], and ACDC [2]. Using pre-trained models from open-source Github repositories by the corresponding authors, we launch different adversarial attacks and conduct adversarial training with default parameters. We use the Pytorch framework and single NVIDIA A100-SXM4-40GB GPU for our experiments. For a pixel/voxel range [0, 255], we create l_∞ adversarial examples under perturbation budgets of $\epsilon \in \{4, 8\}$ for voxel-domain attacks following [7] and compare it with our attack **VAFA**. Unless otherwise specified, all attacks are run for a total of 20 optimization steps. More details about the parameters of the attacks used in different experiments can be found in Appendix. We use mean Dice Similarity Score (DSC), mean 95% Hausdorff Distance (HD95). We also report perceptual similarity between clean and adversarial sample (LPIPS) [28].

Results: For each evaluation metric, we take mean across all classes (including background) and test images. In each table (where applicable), green values show DSC and HD95 on clean images. Table 1 shows comparison of voxel-domain attacks (e.g. PGD [21], FGSM [10], BIM [16], GaussianNoise(GN) [14]) with VAFA-2D (2D DCT in FPM applied on each scan independently) and VAFA on

Table 1. Voxel vs. Freq. Attacks

Attack	DSC↓	LPIPS↑
-	74.31	-
PGD	62.67	**98.94**
FGSM	62.77	98.82
BIM	62.76	98.93
GN	74.19	97.71
VAFA-2D	61.66	98.43
VAFA	**52.54**	97.84

Table 2. Impact of q_{max} on VAFA

q_{max}	DSC↓	LPIPS↑
-	74.31	-
10	65.95	**99.10**
20	56.24	98.70
30	50.96	98.33
40	49.58	97.90
60	48.83	96.60
80	**48.76**	94.50

Table 3. Impact of steps on VAFA

Steps	DSC↓	LPIPS↑
-	74.31	-
10	61.33	**98.85**
20	56.24	98.70
30	54.37	98.64
40	53.31	98.59
50	52.97	98.54
60	**52.25**	98.52

Table 4. Impact of patch size on VAFA

Size	DSC↓	LPIPS↑
-	74.31	-
4	63.48	**98.90**
8	56.24	98.70
16	41.30	98.14
32	32.40	97.49
48	28.19	97.16
96	**28.08**	96.47

Table 5. Comparison of VAFA with other voxel-domain attacks (Synapse dataset).

Models → Attacks ↓	UNETR			UNETR++		
	DSC↓	HD95↑	LPIPS↑	DSC↓	HD95↑	LPIPS↑
Clean Images	74.3	14.0	-	84.7	12.7	-
PGD ($\epsilon = 4/8$)	62.7/50.8	40.4/64.5	**98.9**/95.3	77.5/67.1	48.1/78.3	95.7/85.1
FGSM ($\epsilon = 4/8$)	62.8/53.9	34.8/48.7	98.8/94.7	73.1/67.1	37.3/43.2	94.7/82.2
BIM ($\epsilon = 4/8$)	62.8/50.7	39.9/**65.8**	98.8/95.3	77.3/66.8	46.6/78.1	**95.8**/85.3
GN ($\sigma = 4/8$)	74.2/73.9	17.0/15.4	97.7/91.1	84.7/84.3	12.3/13.4	93.3/78.2
VAFA ($q_{max} = 20/30$)	**32.2**/**29.8**	**57.6**/59.9	97.5/**96.9**	**45.3**/**39.3**	**73.9**/**85.2**	94.2/**94.7**

Table 6. Performance of different attacks on adversarially trained (robust) models.

	Attacks → Models ↓	UNETR					UNETR++				
		Clean	PGD	FGSM	BIM	VAFA	Clean	PGD	FGSM	BIM	VAFA
Synapse	$\mathcal{M}^{PGD}$	73.47	65.53	65.68	65.51	42.47	75.43	67.81	67.82	67.80	38.22
	$\mathcal{M}^{FGSM}$	72.44	64.80	66.31	64.76	39.02	81.06	73.84	74.76	73.77	37.48
	$\mathcal{M}^{BIM}$	75.12	67.78	68.32	67.78	45.97	74.80	67.58	67.46	67.57	35.72
	$\mathcal{M}^{GN}$	73.17	61.40	61.77	61.29	30.00	80.05	76.23	70.96	74.51	41.44
	$\mathcal{M}^{VAFA}$	74.67	64.83	65.49	64.73	66.31	81.88	69.09	65.40	68.90	76.47
	$\mathcal{M}^{VAFA\text{-}FR}$	**75.66**	65.90	66.79	65.83	66.33	**82.65**	70.61	67.00	70.41	78.19
ACDC	$\mathcal{M}^{VAFA}$	81.95	60.77	68.16	60.75	69.76	89.00	76.28	80.41	76.56	88.45
	$\mathcal{M}^{VAFA\text{-}FR}$	**83.44**	60.63	69.33	60.61	73.05	**91.36**	85.42	87.42	83.90	91.23

UNETR model (Synapse). VAFA achieves a higher fooling rate as compared to other attacks with comparable LPIPS. We posit that VAFA-2D on volumetric MIS data is sub-optimal and it does not take into account the 3D nature of the data and model's reliance on the 3D neighborhood of a voxel to predict its class. Further details are provided in the supplementary material. We show impacts of different parameters of VAFA e.g. quantization threshold (q_{max}), steps, and patch size ($h \times w \times d$) on DSC and LPIPS in Table 2, 3 and 4 respectively. DSC and LPIPS decrease when these parameters values are increased. Table 5 shows a comparison of VAFA (patch size = $32 \times 32 \times 32$) with other voxel-domain attacks on UNETR and UNETR++ models. For adversarial training experiments, we use $q_{max} = 20$ (for Synapse), $q_{max} = 10$ (for ACDC) and patch-size of $32 \times 32 \times 32$ (chosen after considering the trade-off between DSC and

LPIPS from Table 4) for VAFA. For voxel-domain attacks, we use $\epsilon = 4$ (for Synapse) and $\epsilon = 2$ (for ACDC) by following the work of [11, 22]. Table 6 presents a comparison of the performance (DSC) of various adversarially trained models against different attacks. $\mathcal{M}_{\circlearrowleft}^{\text{VAFA-FR}}$, $\mathcal{M}_{\circlearrowleft}^{\text{VAFA}}$ denote our robust models which were adversarially trained with and without frequency consistency loss ($\mathcal{L}_{\text{fr}}$, Eq. 4) respectively. In contrast to other voxel-domain robust models, our approach demonstrated robustness against both voxel and frequency-based attacks.

4 Conclusion

We present a frequency-domain based adversarial attack and training for volumetric medical image segmentation. Our attack strategy is tailored to the 3D nature of medical imaging data, allowing for a higher fooling rate than voxel-based attacks while preserving comparable perceptual similarity of adversarial samples. Based upon our proposed attack, we introduce a frequency-domain adversarial training method that enhances the robustness of the volumetric segmentation model against both voxel and frequency-domain based attacks. Our training strategy is particularly important in medical image segmentation, where the accuracy and reliability of the model are crucial for clinical decision making.

References

1. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access **6**, 14410–14430 (2018)
2. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)
3. Carlini, N., Wagner, D.: Adversarial examples are not easily detected: bypassing ten detection methods. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 3–14 (2017)
4. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57. IEEE (2017)
5. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International Conference on Machine Learning, pp. 2206–2216. PMLR (2020)
6. Daza, L., Pérez, J.C., Arbeláez, P.: Towards robust general medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 3–13. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_1
7. Duan, R., Chen, Y., Niu, D., Yang, Y., Qin, A.K., He, Y.: Advdrop: adversarial attack to DNNs by dropping information. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7506–7515 (2021)
8. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019)
9. Gong, R., et al.: Differentiable soft quantization: bridging full-precision and low-bit neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4852–4861 (2019)

10. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
11. Guo, C., Rana, M., Cisse, M., Van Der Maaten, L.: Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117 (2017)
12. Hatamizadeh, A., et al.: UNETR: transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)
13. Karimi, D., Vasylechko, S.D., Gholipour, A.: Convolution-free medical image segmentation using transformers. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 78–88. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_8
14. Kim, H.: Torchattacks: a pytorch repository for adversarial attacks. arXiv preprint arXiv:2010.01950 (2020)
15. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236 (2016)
16. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: Artificial Intelligence Safety and Security, pp. 99–112. Chapman and Hall/CRC (2018)
17. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: MICCAI multi-atlas labeling beyond the cranial vault-workshop and challenge. In: Proceedings of MICCAI Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge, vol. 5, p. 12 (2015)
18. Lei, T., Wang, R., Wan, Y., Du, X., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: a survey. arXiv preprint arXiv:2009.13120 (2020)
19. 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
20. Ma, X., et al.: Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recogn. **110**, 107332 (2021)
21. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
22. Prakash, A., Moran, N., Garber, S., DiLillo, A., Storer, J.: Deflecting adversarial attacks with pixel deflection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8571–8580 (2018)
23. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
24. Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: UNETR++: delving into efficient and accurate 3D medical image segmentation. arXiv preprint arXiv:2212.04497 (2022)
25. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 240–248. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_28
26. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)

27. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
28. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)