MICCAI 2023 Vancouver CANADA

Frequency Domain Adversarial Training for Robust Volumetric Medical Segmentation

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Asif Hanif, Muzammal Naseer, Salman Khan, Mubarak Shah, Fahad Shahbaz Khan
MBZUAI (UAE), UCF (USA), Linköping University (Sweden)
{asif.hanif, muzammal.naseer, salman.khan, fahad.khan}@mbzuai.ac.ae shah@crcv.ucf.edu

Key Contribution

Introduction of a 3D frequency domain adversarial attack for volumetric medical image segmentation models and demonstration of its advantages over conventional input or voxel domain attacks. Using the proposed attack, a novel frequency domain adversarial training approach is presented for optimizing a robust model against voxel and frequency domain attacks.

Introduction

- Semantic segmentation of organs, or anomalies in medical images (e.g. CT scans) is one of the fundamental tasks in medical image analysis.
- With recent advances in deep learning, performance of volumetric medical image segmentation has also improved. Despite this progress, the real-world deployment of medical image models is not straightforward due to their vulnerabilities towards adversarial attacks.
- 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.
- The adversarial robustness of the medical imaging models is still an open and under-explored area.

Method

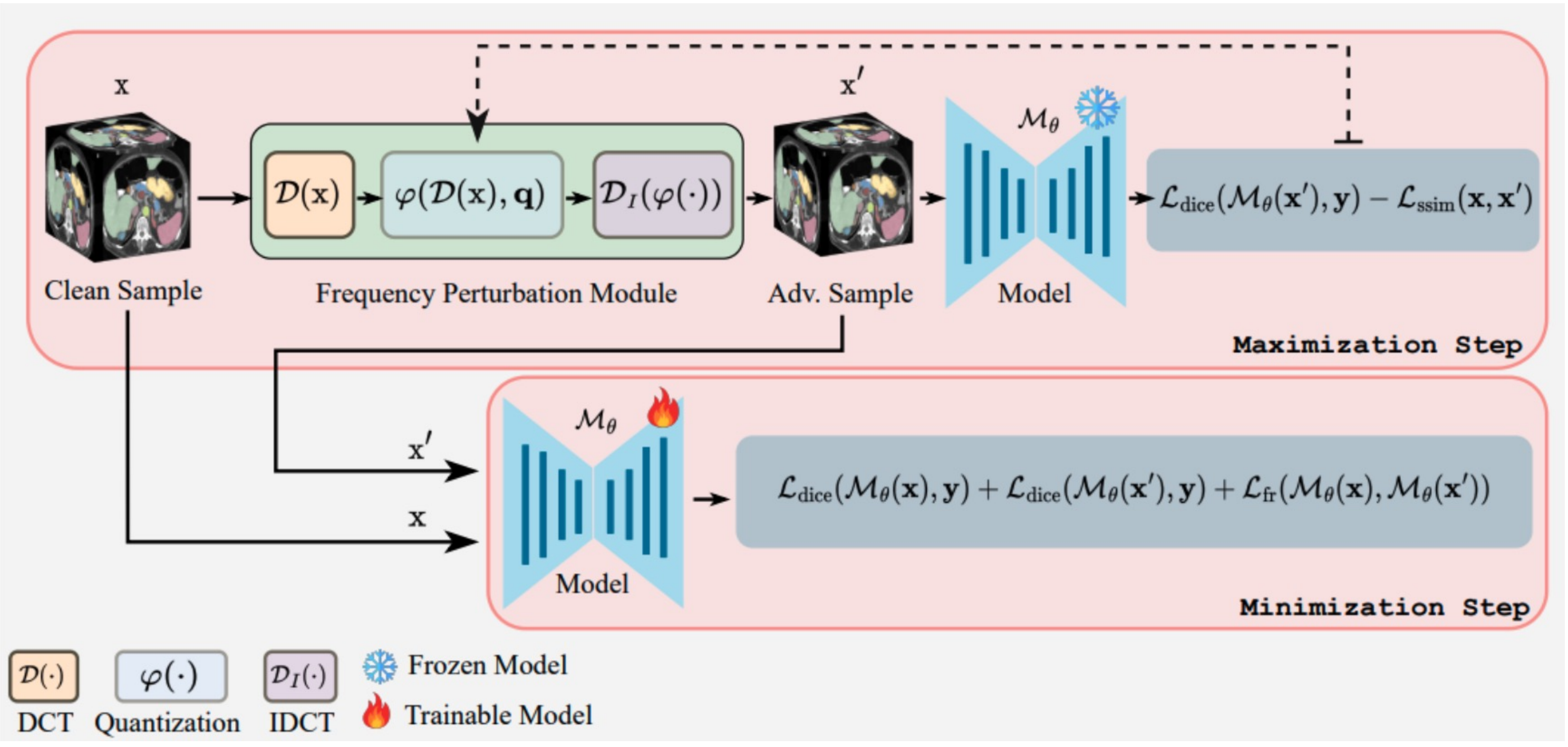
x
x'
Clean Sample
$\mathcal{D}(\mathbf{x})$
$\varphi(\mathcal{D}(\mathbf{x}),\mathbf{q})$
$\mathcal{D}_I(\varphi(\cdot))$
Frequency Perturbation Module
Adv. Sample
$\mathcal{M}_\theta$
Model
$\mathcal{L}_{\text{dice}}(\mathcal{M}_\theta(\mathbf{x}'),\mathbf{y}) - \mathcal{L}_{\text{ssim}}(\mathbf{x},\mathbf{x}')$
Maximization Step
x'
x
$\mathcal{M}_\theta$
Model
$\mathcal{L}_{\text{dice}}(\mathcal{M}_\theta(\mathbf{x}),\mathbf{y}) + \mathcal{L}_{\text{dice}}(\mathcal{M}_\theta(\mathbf{x}'),\mathbf{y}) + \mathcal{L}_{\text{fr}}(\mathcal{M}_\theta(\mathbf{x}),\mathcal{M}_\theta(\mathbf{x}'))$
Minimization Step
$\mathcal{D}(\cdot)$ DCT
$\varphi(\cdot)$ Quantization
$\mathcal{D}_I(\cdot)$ IDCT
Frozen Model
Trainable Model

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 use a frequency consistency loss to improve the model performance.

Frequency Perturbation:

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

Algorithm

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

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

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

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

Results

Models → Attacks ↓	UNETR DSC↓	 HD95↑	 LPIPS↑	UNETR++ 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 1: Comparison of VAFA with other voxel-domain attacks (Synapse dataset)

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

Table 2: Performance (Dice Score) of different attacks on adversarially trained (robust) models

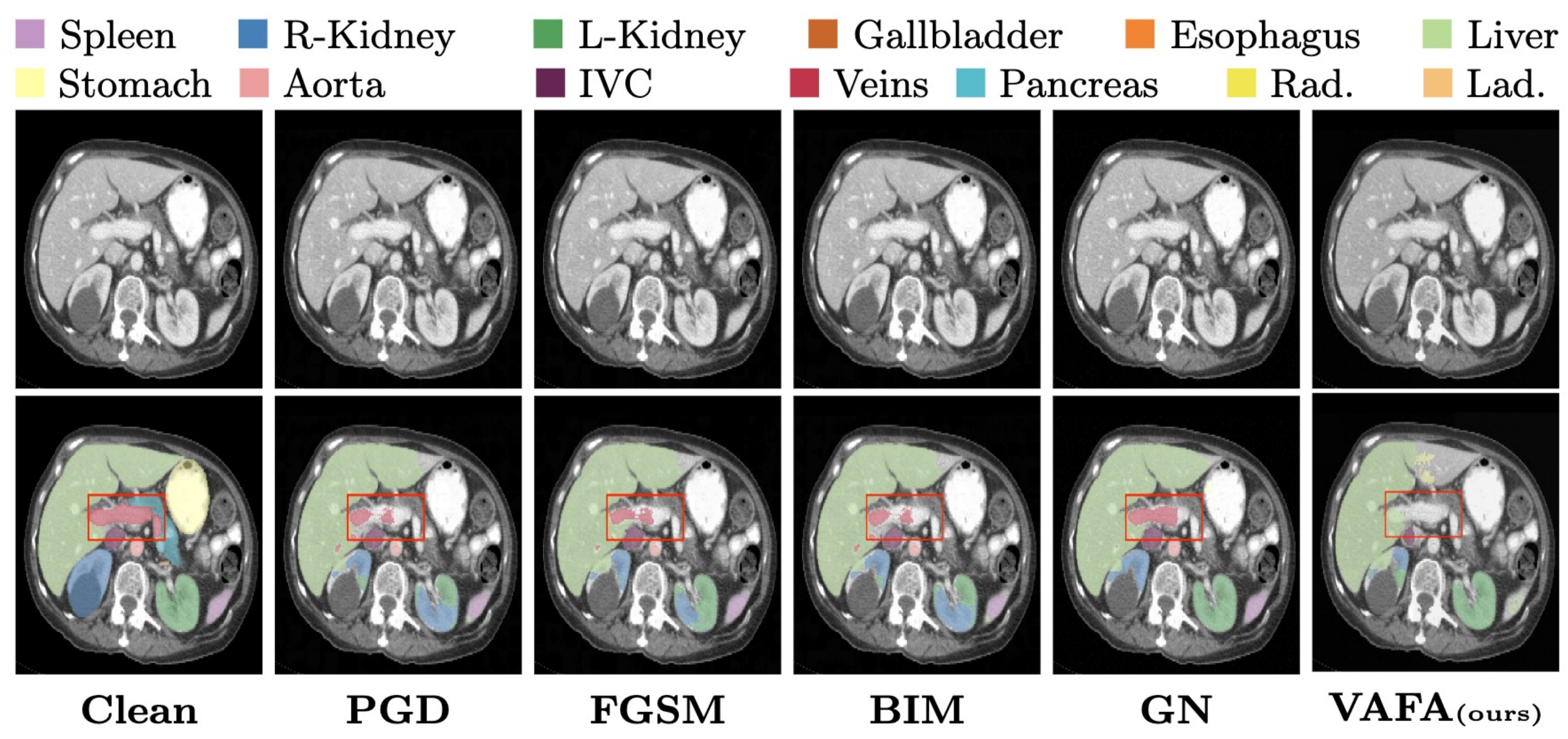

Fig. 2: Qualitative multi-organ segmentation comparison under different attacks on the normally trained UNETR model with Synapse dataset. Top row shows example images and bottom row shows the corresponding segmentation masks predicted by the model under different attacks. Compared to different voxel-domain attacks (PGD, FGSM, BIM and GN), our attack (VAFA) achieves higher fooling rate (highlighted in red bounding box) while maintaining comparable perceptual similarity.

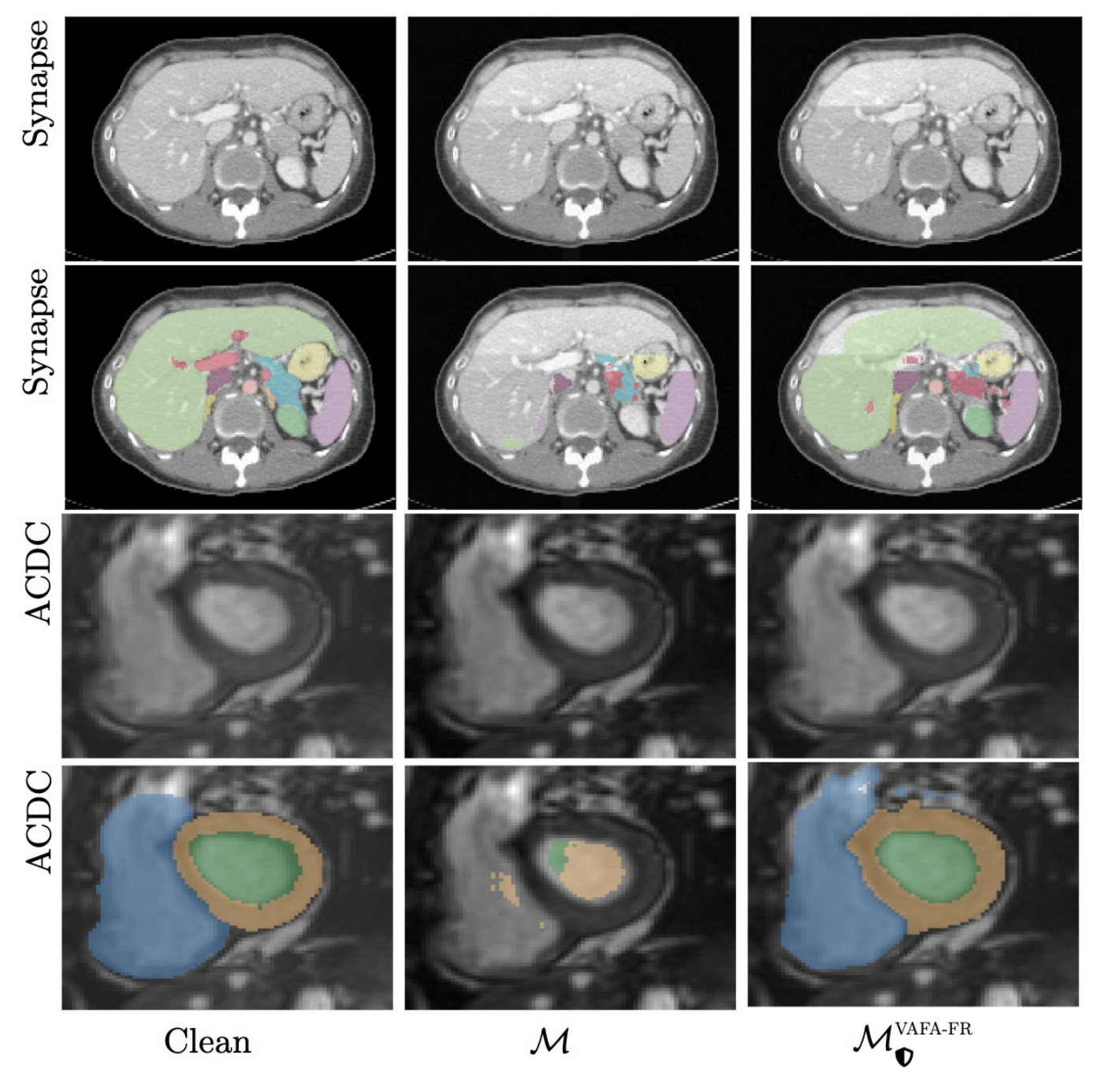

Fig. 3: Qualitative multi-organ segmentation comparison (before and after frequency-domain adversarial training) under VAFA attack on the UNETR model. First column shows clean images (along with their corresponding ground-truth segmentation masks). Second and third columns show adversarial images (and corresponding predicted segmentation masks) obtained by attacking the normally trained and adversarially trained UNETR model with VAFA attack respectively. Adversarially trained model shows robustness towards VAFA attack.

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.