FREQUENCY DOMAIN ADVERSARIAL TRAINING FOR ROBUST VOLUMETRIC MEDICAL SEGMENTATION

Asif Hanif, Muzammal Naseer, Salman Khan, Mubarak Shah, Fahad Shahbaz Khan

Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), UAE

University of Central Florida (UCF), USA

Linköping University, Sweden

MICCAI2023 Vancouver CANADA

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.
- Deep learning and volumetric medical image segmentation
- Adversarial vulnerability of volumetric medical image segmentation
- Adversarial robustness of medical image models
- Understanding the vulnerability and robustness of medical image models

Contributions

- Volumetric Adversarial Frequency Attack (VAFA)
- Volumetric Adversarial Frequency Training (VAFT)

Voxel-domain Attacks vs. Frequency-domain Attacks

- Voxel-domain Attacks: Directly perturb input space
- Frequency-domain Attacks: Preturb the frequency-domain representation of an image e.g. DCT

Method

- **V**olumetric **A**dversarial **F**requency domain **A**ttack (**VAFA**)
- **V**olumetric **A**dversarial **F**requency domain **T**raining (**VAFT**)

Volumetric Adversarial Frequency Domain Attack (VAFA)

Volumetric Adversarial Frequency Domain Attack (VAFA)

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

Volumetric Adversarial Frequency Domain Attack (VAFA)

$$\varphi(\mathcal{D}(\mathbf{x}), \mathbf{q}) := \left\lfloor \frac{\mathcal{D}(\mathbf{x})}{\mathbf{q}} \right\rfloor \odot \mathbf{q}$$

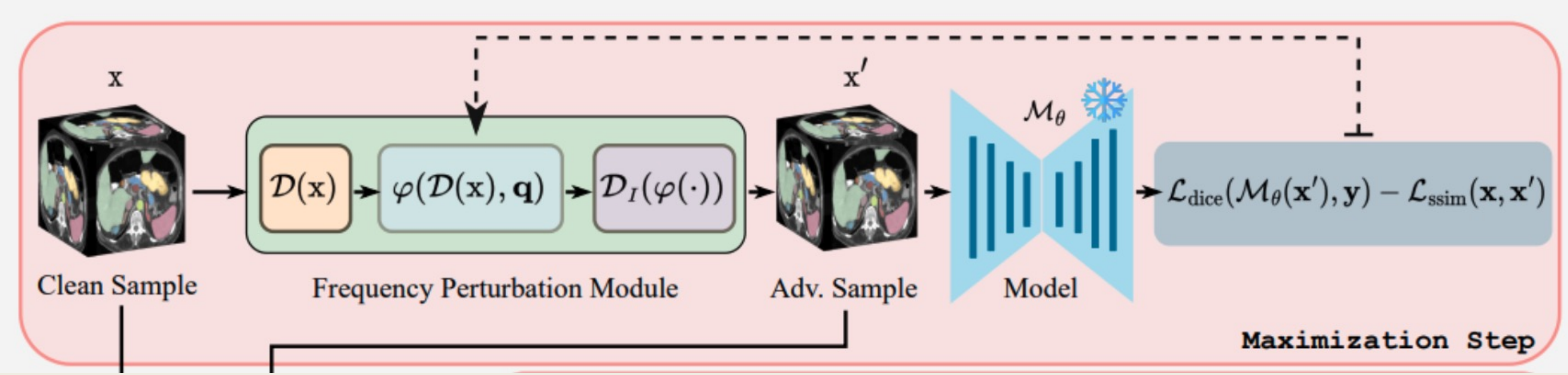
$$\mathbf{q} \in \mathbb{Z}^{h \times w \times d}$$

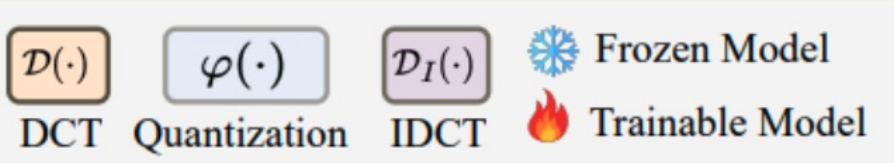
Volumetric Adversarial Frequency Domain Attack (VAFA)

$$\underset{\mathbf{q}}{\text{maximize}} \quad \mathcal{L}_{\text{dice}}(\mathcal{M}_\theta(\mathrm{X}'), \mathrm{Y}) - \mathcal{L}_{\text{ssim}}(\mathrm{X}, \mathrm{X}')$$

$$\text{s.t.} \quad \|\mathbf{q}\|_\infty \leq q_{\max},$$

Volumetric Adversarial Frequency Domain Attack (VAFA)





Volumetric Adversarial Frequency Domain Attack (VAFA)

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

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

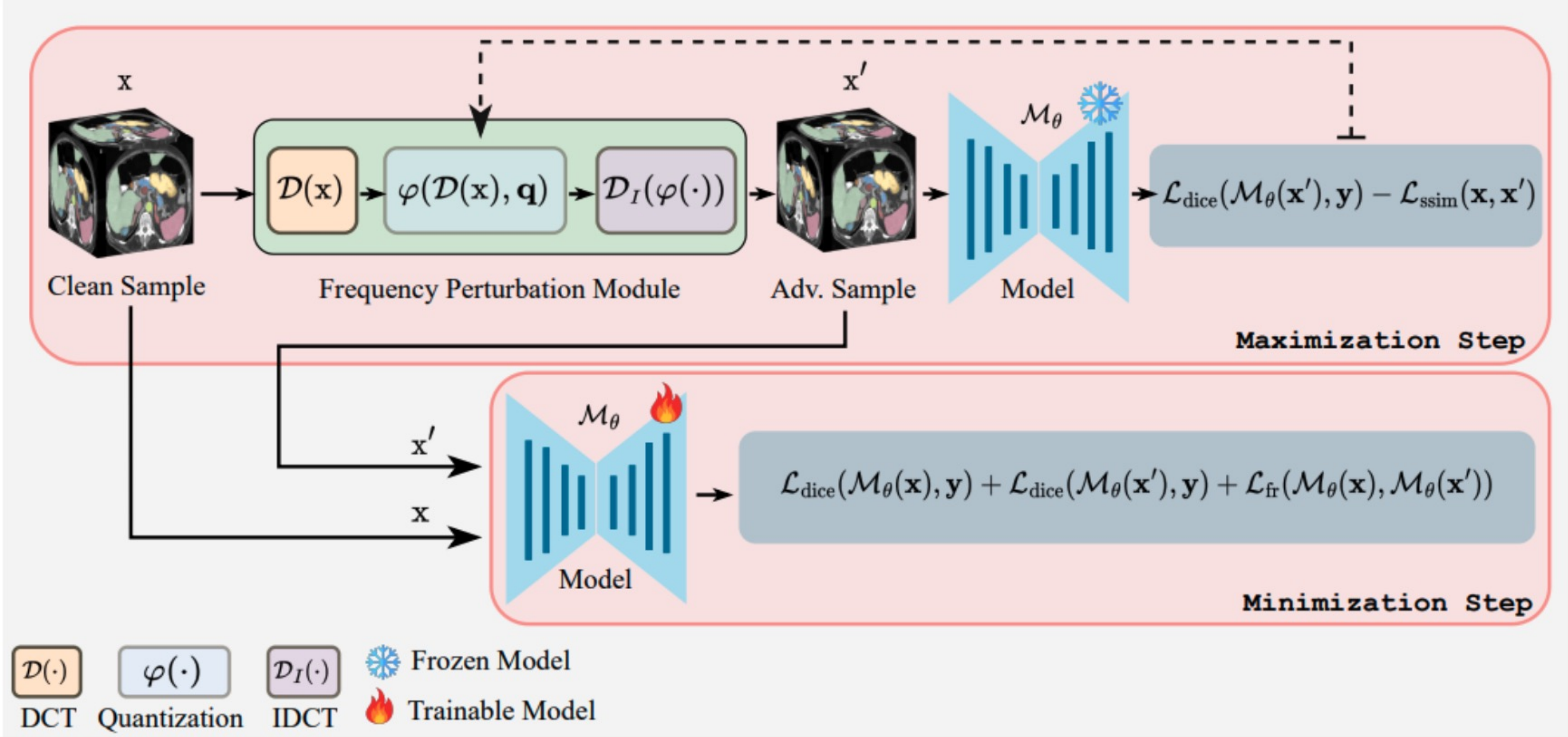
Volumetric Adversarial Frequency Domain Training (VAFT)

Volumetric Adversarial Frequency Domain Training (VAFT)

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

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

Volumetric Adversarial Frequency Domain Training (VAFT)



Volumetric Adversarial Frequency Domain Training (VAFT)

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

Train Dataset: $\mathcal{X} = \{(\mathrm{X}_i, \mathrm{Y}_i)\}_{i=1}^{N}, \ \ \mathrm{X}_i \in \mathbb{R}^{H \times W \times D}, \ \ \mathrm{Y}_i \in \{0,1\}^{\mathrm{NumClass} \times H \times W \times D}$

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

        $\mathrm{X}' \leftarrow \mathbf{VAFA}(\mathrm{X}, \mathrm{Y}) \ \ \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B}$ ▷ Adv. Freq. Attack on clean images.

        $\mathcal{L} = \mathcal{L}_{\mathrm{dice}}(\mathcal{M}_\theta(\mathrm{X}), \mathrm{Y}) + \mathcal{L}_{\mathrm{dice}}(\mathcal{M}_\theta(\mathrm{X}'), \mathrm{Y}) + \mathcal{L}_{\mathrm{fr}}(\mathcal{M}_\theta(\mathrm{X}), \mathcal{M}_\theta(\mathrm{X}'))$

        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 $\mathcal{M}_{\text{🛡}}$

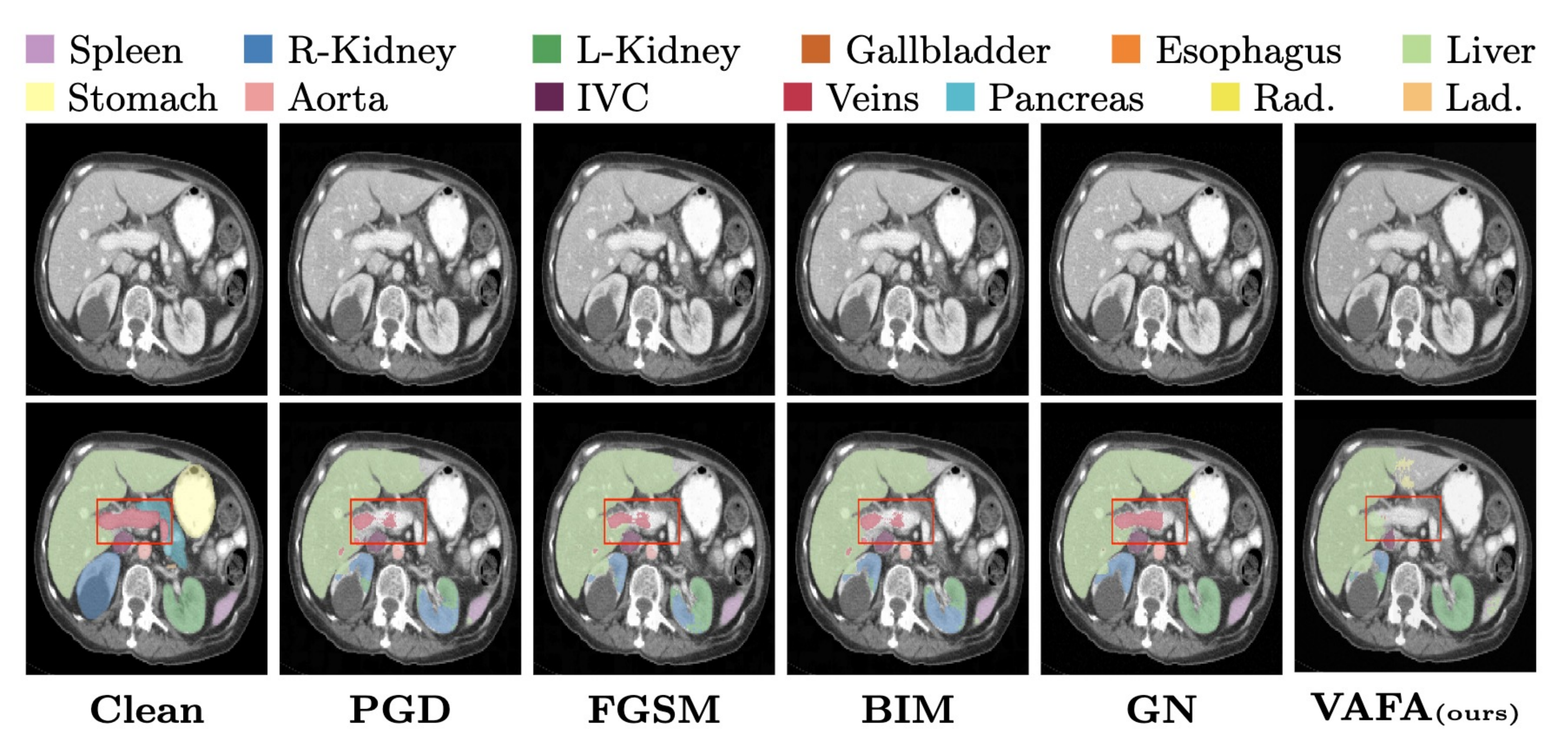
Results

- **Segmentation Models:** UNETR and UNETR++
- **Datasets:** Synapse and ACDC
- **Baseline Attacks:** PGD, FGSM, BIM, GN
- **Evaluation Metrics:** Dice Score, HD95 Distance, LPIPS
- **Programming Framework:** Pytorch

Volumetric Adversarial Frequency Domain Attack (VAFA)

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**

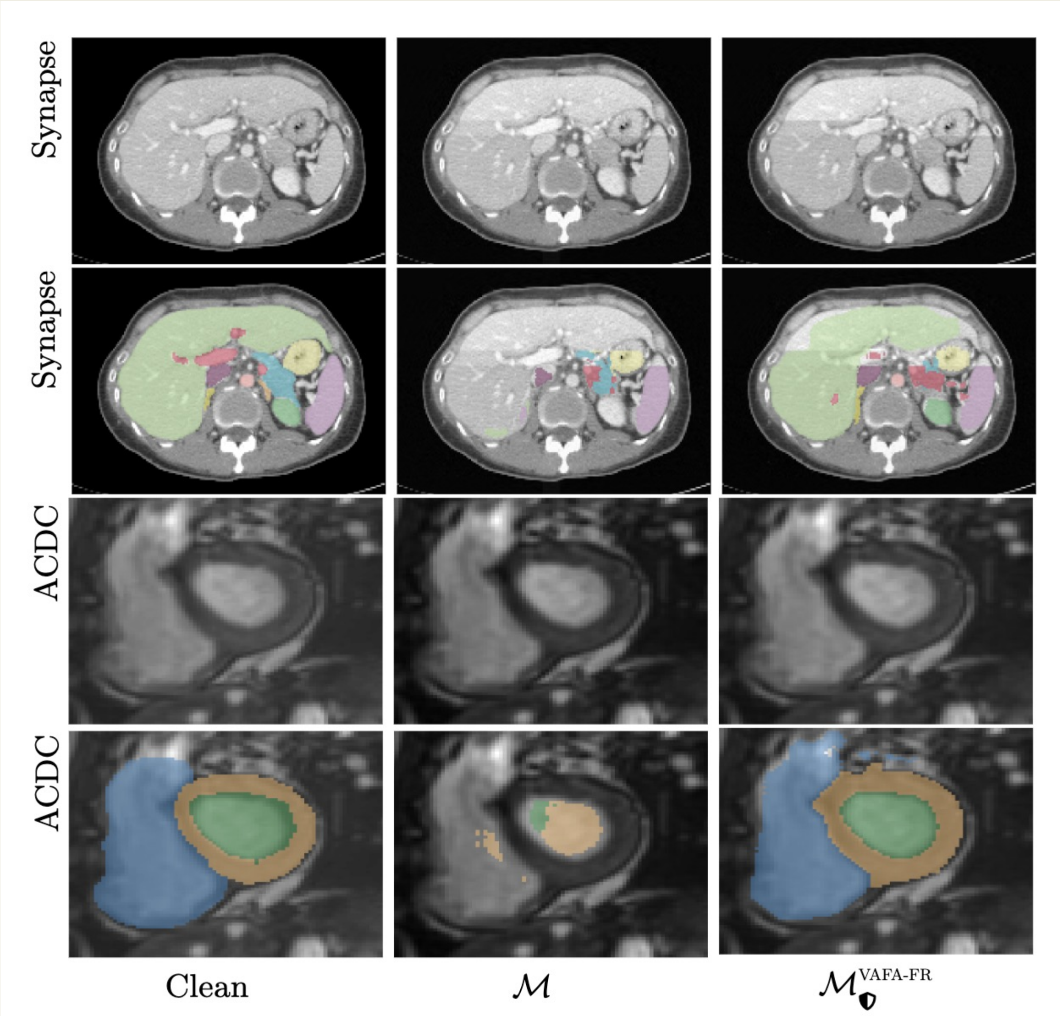
Volumetric Adversarial Frequency Domain Attack (VAFA)



Volumetric Adversarial Frequency Domain Training (VAFT)

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

Volumetric Adversarial Frequency Domain Training (VAFT)



Github Repository

https://github.com/asif-hanif/vafa

Thank you !

VAFA – Ablation Study

- Impact of quantization threshold
- Impact of steps
- Impact of patch size

VAFA – Ablation Study
Impact of Quantization Threshold: $q_{\max}$

$q_{\max}$	DSC↓	HD95↑	LPIPS↑
-	74.31	14.03	-
10	65.95	26.25	**99.10**
20	56.24	35.92	98.70
30	50.96	44.09	98.33
40	49.58	43.66	97.90
60	48.83	44.55	96.60
80	**48.76**	**45.30**	94.50

VAFA – Ablation Study
Impact of Steps

Steps	DSC↓	HD95↑	LPIPS↑
-	74.31	14.03	-
10	61.33	33.20	**98.85**
20	56.24	35.92	98.70
30	54.37	38.00	98.64
40	53.31	37.76	98.59
50	52.97	**39.23**	98.54
60	**52.25**	39.19	98.52

VAFA – Ablation Study
Impact of Patch-Size

Patch Size	DSC↓	HD95↑	LPIPS↑
-	74.31	14.03	-
$(4 \times 4 \times 4)$	63.48	32.63	**98.90**
$(8 \times 8 \times 8)$	56.24	35.92	98.70
$(16 \times 16 \times 16)$	41.30	45.98	98.14
$(32 \times 32 \times 32)$	32.40	56.64	97.49
$(48 \times 48 \times 48)$	28.19	**66.08**	97.16
$(96 \times 96 \times 96)$	**28.08**	59.09	96.47