Appendix

We offer further insights into our study through various sections in the appendix. Firstly, we provide additional details regarding the datasets utilized in Section A and elaborate on the training recipe for the models, as discussed in Section B. In Section C, we delve deeper into the *white box* analysis, providing ablations across all datasets, examining pixel and frequency-based attacks at different perturbation budgets. Likewise, Section E expands on the frequency analysis initially presented in Section 4.3, covering different adversarial attacks. Lastly, in Section F, we present results on transfer-based *black box* attacks across all datasets, considering various perturbation strengths, and offer further insights derived from these findings. **Our well-documented code and pretrained weights will be made publicly available.**

A Datasets

Medical imaging data sets encompass a range of imaging techniques, including Positron Emission Tomography (PET), Computed Tomography (CT), and Magnetic Resonance Imaging (MRI). PET scans effectively show the metabolic or biochemical activity within the body, and CT imaging offers high-resolution images of the body's internal structure. Similarly, MRIs effectively differentiate between soft tissues without the use of ionizing radiation. These imaging modalities acquire comprehensive and complementary information about the body's organs, functions, and tumors. Thus, to conduct a comprehensive benchmarking analysis of the model's robustness and susceptibility to adversarial attacks, we utilize four different segmentation datasets: `BTCV`, `ACDC`, `Hecktor`, and `Abdomen-CT`, which consist of medical images from CT and MRI modalities encompassing different tumor and organ segmentations.

BTCV: The `BTCV` dataset consists of 30 abdominal CT scans from metastatic liver cancer patients acquired from a single medical center. Each CT scan is manually annotated for 13 abdominal organs (Spleen, Right Kidney, Left Kideny, Gallbladder, Esophagus, Liver, Stomach, Aorta, IVC, Portal and Splenic Veins, Pancreas, Right adrenal gland, and Left adrenal gland). The CT scan size is 512×512 pixels, the number of slices ranges from 80 to 225, and the slice thickness ranges from 1 to 6 mm.

ACDC: The Automated Cardiac Diagnosis Challenge (`ACDC`) dataset consists of 150 MRI images from patients with cardiac abnormalities acquired from a single medical center. Each MRI scan is manually annotated for different heart organs, such as the left ventricle (LV), right ventricle (RV), and myocardium (MYO). The number of MRI slices ranges from 28 to 40, and the slice thickness ranges from 5 to 8 mm. The spatial resolution goes from 1.37 to 1.68 mm^2/pixel.

HECKTOR: The `Hecktor` dataset consists of 524 CT/PET scans of head and neck cancer patients collected from seven medical centers. It was manually annotated for primary gross tumor volumes (GTVp) and nodal gross tumor volumes (GTVn). The CT scan size ranges from 128×128 to 512×512, the number of slices ranges from 67 to 736, and the slice thickness ranges from 1 to 2.8 mm.

AbdomenCT-1k: The AbdomenCT-1k dataset consists of 1112 abdominal CT scans from 12 medical centers, including multi-phase, multi-vendor, and multi-disease cases. All the scans' annotations for the liver, kidney, spleen, and pancreas are provided. The CT scan size has resolutions of 512×512pixels with varying voxel sizes and slice thicknesses between 1.25 to 5 mm.

B Training Details

For BTCV, Abdomen-CT, and ACDC all models trained for 5000 epochs, while for Hecktor we use only 500 epochs. A batch size of 3 and a learning rate (lr) of $1e-4$ with the warmup_cosine scheduler is used. During training, images are normalized to the range of $[0,1]$, and a 3D random crop of size $96 \times 96 \times 96$ is selected as an input to the segmentation model. Augmentations include RandomFlip (for all three spatial dimensions), RandomRotate90, RandomScaleIntensity, and RandomShiftIntensity. During inference, we employed a sliding window approach, dividing the input volume of arbitrary size into 3D sliding windows of size $96 \times 96 \times 96$ with a 50% overlap. The predictions of overlapping voxels were combined using a Gaussian kernel.

C Robustness against White-box Attacks

In Figure 5, we report robustness of the volumetric segmentation models on *white box* attacks across BTCV, Abdomen-CT, Hecktor, and ACDC datasets. For pixel-based attacks we craft adversarial examples at l_∞ perturbation budget $\varepsilon \in \{\frac{4}{255}, \frac{8}{255}\}$ for FGSM, PGD, and CosPGD. For frequency-based attack VAFA we craft adversarial examples with $q_{max} \in \{10, 30\}$. We report DSC and LPIPS score on generated adversarial examples. Similar to our results in Table 1 in the main paper, we observe that VAFA causes the most drop in DSC score of models on BTCV and Abdomen-CT dataset, while iterative pixel-based attacks PGD and CosPGD cause the most drop on Hecktor and ACDC dataset. Furthermore, we provide ablation across VAFA attack with varying $q_{max} \in \{10, 20, 30\}$ in Figure 6. With an increase in q_{max}, we observe a decrease in the DSC score across all the datasets. Consequently, we also observe a drop in the LPIPS score of the generated adversarial examples.

D Adversarial Examples

In Figure 7, adversarial example using VAFA is crafted on UNet model trained on Abdomen-CT dataset. Segmentation prediction of all the volumetric segmentation models is shown for the clean sample and the adversarial one. We can clearly observe that the adversarial example causes the predictions to change across all models.

E Frequency Analysis of White-Box Attacks

In this section, we delve deeper into the frequency analysis of adversarial attacks to study which frequency components lead to drop in performance of models. Following [44], we implement an adversarial attack incorporating a frequency filter M, restricting perturbations to specific frequency domains. The filter operation is defined as $x'_{\text{freq}} = \text{IDCT}(\text{DCT}(x' - x) \odot M) + x$, where DCT and IDCT denote Discrete Cosine Transform and its inverse, respectively. Similar to [44], using the filter M, we extract 3D cubes of varying size n from the top left corner as part of the low-frequency components $(0, n)$ where $n \in \{8, 16, 32\}$. Similarly, mid-frequency

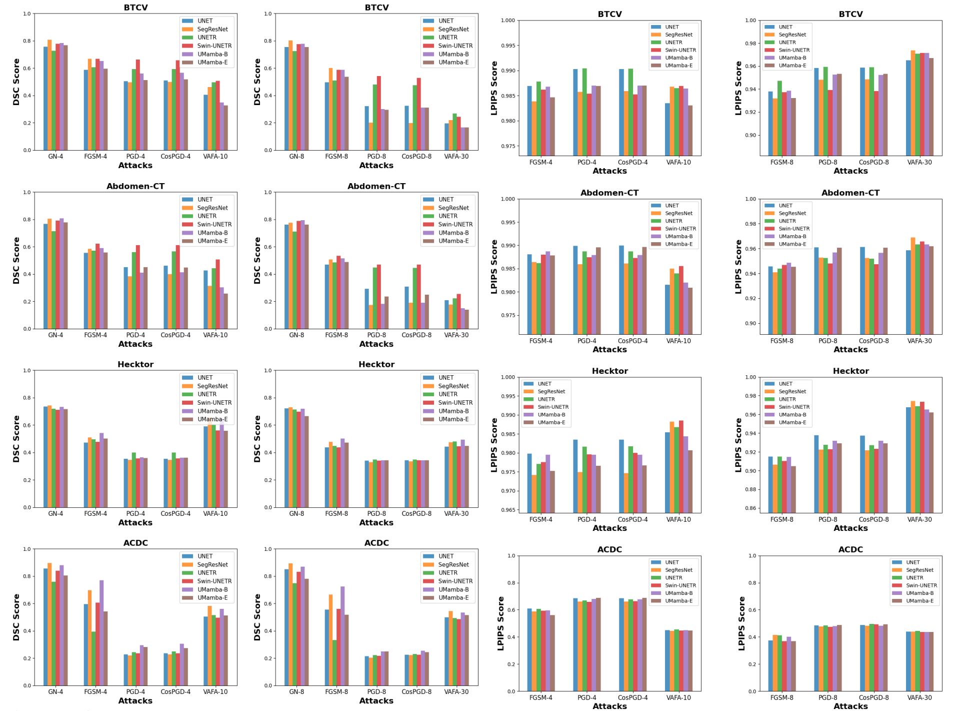

Figure 5: **White Box Attack Ablation:** Evaluating robustness of volumetric segmentation models on *white box* attacks. For pixel-based attacks results are reported for $\varepsilon = \frac{4}{255}$ and $\varepsilon = \frac{8}{255}$ indicated by attack names followed by the suffixes -4 or -8, respectively. Regarding frequency-based attack `VAFA`, the results are reported with a constraint on $q_{\max}$ set to 10 and 30, denoted as `VAFA-10` and `VAFA-30`, respectively. DSC score *(lower is better)* and LPIPS score *(higher is better)* are reported on the generated adversarial examples.

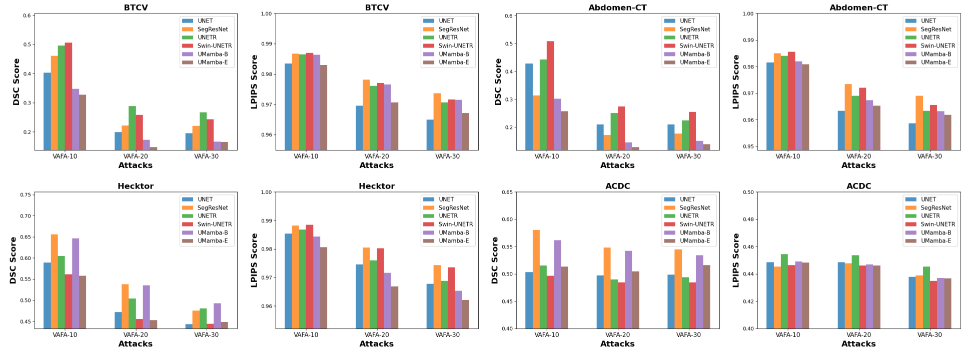

Figure 6: **Frequency Attack Ablation:** Examining the robustness of volumetric segmentation models `VAFA` based *white box* attack. Adversarial examples are generated at $q_{max} \in \{10, 20, 30\}$, represented as `VAFA-10`, `VAFA-20`, and `VAFA-30`, respectively. DSC score *(lower is better)* and LPIPS score *(higher is better)* are reported on the generated adversarial examples.

$(16-48)$ and high-frequency $(16-96)$ components are also extracted. See Figure 8 for the design of filters. While in Figure 3 of the main paper, we provide frequency analysis on `VAFA`, which shows the best transferability across target models. Figure 9 expands this anal-

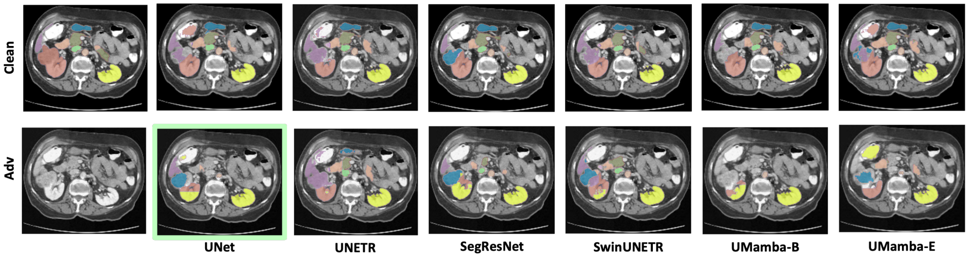

Figure 7: Comparing multi-organ segmentation across various models under transfer-based *black box* attacks, where adversarial examples are generated on UNet and transferred to other unseen models.

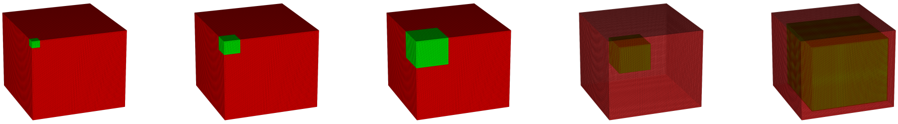

Figure 8: **Frequency Analysis Filters:** The frequencies associated with the red section are eliminated, while those linked to the green section are allowed to pass. These filters are labeled as $(0-8)$, $(0-16)$, $(0-32)$, $(16-48)$, and $(16-96)$ *(from right to left)*.

ysis to encompass all the adversarial attacks employed in our experiments. For pixel-based attacks, we report results at $\varepsilon = \frac{8}{255}$ and for VAFA at $q_{max} = 30$. In the case of the VAFA attack, which demonstrates significant *transferability* to the target models in *black box* setting, we note that the low-frequency components of the adversarial examples predominantly contribute to the performance decline across surrogate volumetric segmentation models. While a similar trend is observed with pixel-based attacks, it is not as pronounced as with VAFA. For instance, when analyzing the Abdomen-CT and ACDC datasets, we find that the high-frequency components of adversarial examples generated by pixel-based attacks also result in a noticeable performance decrease across models. However, as discussed in Section 4.2, these adversarial examples produced by pixel-based attacks exhibit very limited transferability.

F Robustness against Black-box Attacks

In Tables 6, 7, 8, and 9, we report robustness of the volumetric segmentation models on *black box* attacks across BTCV, Abdomen-CT, Hecktor, and ACDC datasets. For pixel-based attacks we craft adversarial examples at l_∞ perturbation budget $\varepsilon = \frac{4}{255}$ for FGSM, PGD, and CosPGD. For frequency-based attack VAFA we craft adversarial examples with $q_{max} = 20$. We report DSC and LPIPS score on generated adversarial examples. Similar to our observations for $\varepsilon = \frac{8}{255}$ and $q_max = 30$ in Section 4.2, we observe frequency-based attack VAFA results in significant transferability of adversarial examples to target models at $\varepsilon = \frac{4}{255}$ and $q_max = 20$ as well. Further, we also report the DSC and HD95 score for results reported in Section 4.2 of the main paper in Table 10, 11, 12, and 13. FInally, in Table 14, we report performance of SAM-Med3D on adversarial examples crafted on surrogate models trained on BTCV, Abdomen-CT, Hecktor, and ACDC datasets.

We believe our empirical results showing higher transferability obtained from frequency

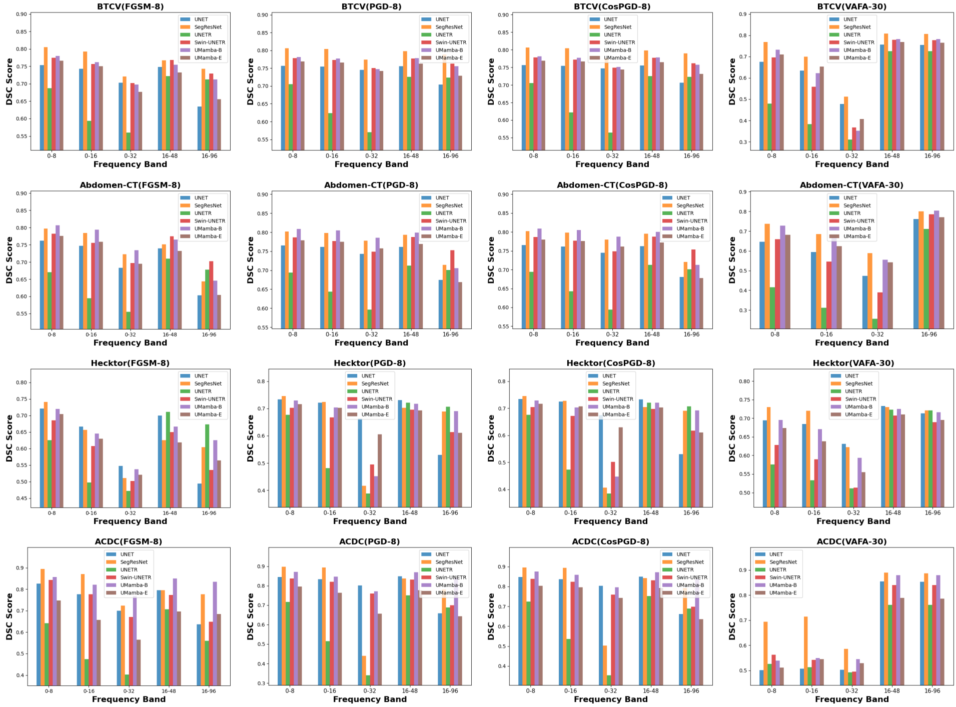

Figure 9: **Frequency Analysis on FGSM, PGD, CosPGD, and VAFA**: We report the performance drop on models in *white box* settings across adversarial attacks. We restrict the adversarial perturbations to be added to the image within different frequency ranges. For pixel-based attacks results are reported for $\varepsilon = \frac{8}{255}$ indicated by attack names followed by the suffixes -8, respectively. Regarding frequency-based attack VAFA, the results are reported with a constraint on $q_{\max}$ set to 30, denoted as VAFA−30. DSC score *(lower is better)* is reported on the generated adversarial examples.

domain attack can be attributed to the ability of frequency domain perturbations to affect a wider range of model architectures and training datasets compared to spatial domain perturbations. One possible reason for the higher transferability obtained from frequency domain adversarial attacks is that frequency domain perturbations can capture more abstract and general features of the input data. These perturbations may affect the underlying patterns and structures that are common across different models and datasets, making them more transferable across various scenarios. Additionally, frequency domain transformations can be less sensitive to small changes in pixel values, leading to more robust and consistent adversarial examples across different scenarios.

Surrogate	Attack	UNet		SegResNet		UNETR		SwinUNETR		UMamba-B		UMamba-E	
		DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑
	Clean	75.72	9.15	80.84	8.14	72.53	15.08	78.07	10.01	78.37	8.12	77.06	11.11
	GN	75.65	9.69	80.72	8.37	72.66	14.72	77.88	9.98	78.24	8.18	76.68	16.04
UNet	FGSM-4	58.77	37.71	79.6	8.36	71.72	14.72	76.7	10.39	76.85	9.4	75.18	15.44
	PGD-4	50.53	53.9	80.42	8.51	72.42	14.68	77.57	10.21	77.7	8.69	76.22	14.28
	CosPGD-4	50.97	48.99	80.4	8.49	72.37	14.78	77.57	10.18	77.66	8.57	76.26	14.53
	VAFA-20	56.13	39.73	69.69	18.68	53.19	31.66	66.8	28.34	72.53	14.38	69.63	15.5
SegResNet	FGSM-4	74.0	10.75	66.68	32.41	71.45	15.21	75.27	10.69	73.33	11.1	72.95	15.6
	PGD-4	74.57	13.03	49.49	65.31	71.83	15.25	76.67	10.45	74.96	11.63	73.97	16.1
	CosPGD-4	74.51	13.02	49.89	60.42	71.78	15.45	76.64	10.47	74.99	11.41	73.85	16.77
	VAFA-20	61.59	20.73	51.44	37.74	43.33	41.75	58.62	28.87	68.92	13.88	66.03	17.45
UNETR	FGSM-4	73.68	9.95	79.08	8.72	60.52	32.3	75.12	10.79	76.02	8.82	74.91	12.88
	PGD-4	74.34	11.04	79.8	8.54	59.27	36.6	76.27	10.48	76.83	9.08	75.47	11.66
	CosPGD-4	74.29	10.62	79.79	8.43	59.17	38.39	76.29	10.48	76.83	9.05	75.35	12.85
	VAFA-20	55.31	26.62	57.5	23.07	35.09	45.42	50.44	30.85	63.53	17.54	59.1	21.81
SwinUNETR	FGSM-4	73.91	11.86	78.36	9.34	70.89	15.55	66.8	24.71	75.41	9.41	74.37	13.14
	PGD-4	74.22	12.24	79.47	9.24	71.45	15.96	66.37	32.63	76.36	10.24	75.21	16.84
	CosPGD-4	74.24	12.22	79.4	9.24	71.43	15.96	65.82	32.86	76.37	10.22	75.09	19.19
	VAFA-20	59.36	24.77	61.67	26.59	47.41	40.06	54.22	45.64	65.13	17.85	62.44	21.08
UMamba-B	FGSM-4	73.81	13.17	76.57	11.88	71.52	14.8	75.39	10.84	65.17	24.69	71.63	20.33
	PGD-4	74.69	11.6	78.78	10.44	71.92	15.03	76.88	10.58	56.03	47.41	73.43	19.04
	CosPGD-4	74.56	11.52	78.7	11.24	71.88	15.13	76.85	10.54	56.53	45.15	73.18	19.13
	VAFA-20	73.3	14.16	76.82	10.92	67.21	16.77	74.83	14.62	68.22	19.06	73.58	15.0
UMamba-E	FGSM-4	74.34	10.36	78.49	10.11	71.69	15.08	76.37	10.56	75.04	11.29	59.47	36.77
	PGD-4	75.2	10.54	80.34	8.6	72.31	14.89	77.71	10.29	77.41	10.21	51.3	57.83
	CosPGD-4	75.23	12.35	80.28	9.36	72.31	14.74	77.62	10.26	77.33	10.1	51.67	54.16
	VAFA-20	73.74	13.94	78.67	10.6	68.15	13.88	75.36	16.89	77.09	8.56	62.15	38.97

Table 6: Performance of models against transfer-based *black box* attacks on BTCV dataset. For pixel-based attacks results are reported for $\varepsilon = \frac{4}{255}$ indicated by attack names followed by the suffixes -4, respectively. Regarding frequency-based attack VAFA, the results are reported with a constraint on $q_{\max}$ set to 20, denoted as VAFA-20. DSC score *(lower is better)* is reported on the generated adversarial examples.

Surrogate	Attack	**UNet**		**SegResNet**		**UNETR**		**SwinUNETR**		**UMamba-B**		**UMamba-E**	
		DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑
	Clean	76.79	19.72	80.89	13.30	71.35	27.73	79.33	25.79	81.08	15.51	78.05	18.31
	GN	-	-	-	-	-	-	-	-	-	-	-	-
`UNet`	`FGSM-4`	55.49	49.07	77.63	14.51	70.19	31.9	77.31	28.89	77.51	18.02	74.69	18.88
	`PGD-4`	45.11	60.73	78.86	14.2	70.78	32.96	78.49	27.38	79.26	18.67	76.37	17.3
	`CosPGD-4`	46.15	56.43	78.92	13.65	70.78	32.49	78.48	27.06	79.38	16.8	76.49	17.09
	`VAFA-20`	20.98	79.82	25.59	64.99	39.1	65.7	38.01	60.38	22.87	69.81	17.97	72.59
`SegResNet`	`FGSM-4`	73.99	21.52	58.49	40.26	70.12	32.88	76.92	26.18	73.01	22.27	71.63	19.58
	`PGD-4`	75.32	21.71	38.38	76.59	70.74	32.11	78.1	27.63	76.2	18.0	74.31	19.26
	`CosPGD-4`	75.28	21.9	39.99	71.41	70.73	31.93	78.09	28.11	76.09	20.27	74.31	18.12
	`VAFA-20`	37.16	48.6	17.07	86.27	45.66	48.14	43.38	50.99	22.17	66.08	17.25	72.17
`UNETR`	`FGSM-4`	74.07	21.03	77.61	14.12	57.17	47.62	75.8	27.04	77.75	19.74	74.62	18.03
	`PGD-4`	74.88	21.39	78.53	14.16	56.09	54.17	77.06	27.34	78.7	18.36	75.46	17.04
	`CosPGD-4`	74.9	21.2	78.62	14.25	56.47	51.35	77.14	26.5	78.76	17.46	75.59	16.86
	`VAFA-20`	41.79	51.35	33.3	54.93	25.07	81.29	39.73	54.54	29.44	55.98	27.03	55.56
`SwinUNETR`	`FGSM-4`	73.67	21.14	76.4	15.09	69.15	32.35	62.36	53.75	76.46	18.36	73.79	19.52
	`PGD-4`	74.7	22.26	77.75	14.97	69.93	35.14	61.12	62.08	78.21	17.08	75.29	18.76
	`CosPGD-4`	74.74	21.24	77.81	15.17	69.95	33.62	61.11	59.67	78.14	16.96	75.3	18.39
	`VAFA-20`	31.55	58.31	27.55	59.83	35.04	61.92	27.41	72.71	24.01	65.52	21.14	68.78
`UMamba-B`	`FGSM-4`	73.88	23.47	73.57	17.24	70.06	33.54	76.92	27.25	59.06	45.82	70.01	25.71
	`PGD-4`	75.23	21.66	76.18	16.31	70.71	32.94	78.15	27.58	41.0	72.28	72.81	21.73
	`CosPGD-4`	75.14	22.07	76.19	16.21	70.71	32.76	78.13	26.97	41.31	73.11	72.87	21.47
	`VAFA-20`	37.29	55.72	22.86	65.32	45.02	57.95	43.66	52.85	14.45	91.51	17.12	76.75
`UMamba-E`	`FGSM-4`	74.1	21.44	75.43	15.89	70.08	33.24	77.17	27.36	73.23	21.73	55.75	46.17
	`PGD-4`	75.58	21.91	77.85	15.16	70.86	32.12	78.45	27.14	76.53	18.44	45.08	64.08
	`CosPGD-4`	75.54	21.46	77.77	15.09	70.82	32.04	78.44	27.76	76.72	18.25	44.7	59.92
	`VAFA-20`	38.54	54.64	25.12	64.86	45.39	60.19	45.1	53.93	22.68	72.02	12.76	93.07

Table 7: Performance of models against transfer-based *black box* attacks on `Abdomen-CT` dataset. For pixel-based attacks results are reported for $\varepsilon = \frac{4}{255}$ indicated by attack names followed by the suffixes -4, respectively. Regarding frequency-based attack `VAFA`, the results are reported with a constraint on $q_{\max}$ set to 20, denoted as `VAFA-20`. DSC score *(lower is better)* is reported on the generated adversarial examples.

Surrogate	Attack	**UNet**		**SegResNet**		**UNETR**		**SwinUNETR**		**UMamba-B**		**UMamba-E**	
		DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑
	Clean	73.91	11.36	74.73	11.08	72.36	14.61	71.61	22.07	73.50	10.89	72.19	13.29
	GN	-	-	-	-	-	-	-	-	-	-	-	-
`UNet`	`FGSM-4`	47.09	34.93	66.43	14.93	69.56	15.36	64.99	27.04	66.4	13.95	64.54	15.78
	`PGD-4`	35.45	74.93	70.95	12.63	71.06	15.5	68.1	25.57	69.56	12.38	68.93	14.33
	`CosPGD-4`	35.35	75.02	70.94	12.44	71.12	15.48	68.26	24.87	69.76	12.26	69.11	15.95
	`VAFA-20`	47.18	37.67	71.81	13.39	65.29	18.38	66.4	19.96	68.94	14.23	62.6	19.23
`SegResNet`	`FGSM-4`	65.95	19.65	51.08	29.89	70.11	16.15	63.28	29.1	60.19	19.86	61.2	21.39
	`PGD-4`	67.57	16.92	34.53	76.44	70.65	14.89	64.93	27.82	60.05	25.77	62.3	20.78
	`CosPGD-4`	68.18	16.59	34.44	76.67	70.78	14.84	65.21	27.15	60.25	24.63	62.28	18.18
	`VAFA-20`	70.06	17.43	53.74	26.07	66.36	17.73	65.13	23.35	67.99	14.94	63.06	21.2
`UNETR`	`FGSM-4`	65.86	16.76	67.22	13.6	49.69	28.51	62.87	28.55	66.27	13.45	64.41	15.66
	`PGD-4`	67.24	18.42	68.89	14.18	39.98	57.26	64.09	28.24	68.12	14.02	66.81	14.71
	`CosPGD-4`	67.62	18.51	68.9	14.62	39.97	58.22	64.41	27.93	68.09	14.1	66.77	15.9
	`VAFA-20`	66.46	16.01	71.08	13.66	50.4	30.36	58.58	26.82	66.81	15.32	59.47	22.95
`SwinUNETR`	`FGSM-4`	64.77	17.07	62.78	15.66	68.79	16.25	47.62	50.52	61.94	17.02	61.6	19.59
	`PGD-4`	64.77	20.54	63.71	17.34	69.26	17.86	35.71	71.66	63.03	19.62	62.32	19.43
	`CosPGD-4`	64.89	20.28	64.0	17.41	69.3	17.63	35.67	71.17	63.62	18.24	62.63	19.32
	`VAFA-20`	67.49	19.81	70.16	14.05	63.04	20.19	45.55	39.63	68.31	16.17	58.99	21.62
`UMamba-B`	`FGSM-4`	65.9	15.64	60.75	17.1	70.0	15.18	62.54	29.27	54.27	25.87	60.26	19.26
	`PGD-4`	68.38	15.01	57.51	27.59	71.01	15.12	64.41	28.32	36.54	73.08	59.32	23.49
	`CosPGD-4`	68.57	15.83	57.11	29.09	71.06	15.1	64.85	28.94	36.2	72.71	60.1	24.75
	`VAFA-20`	67.38	19.59	68.45	14.79	65.08	17.12	65.41	22.1	53.55	31.11	60.21	21.25
`UMamba-E`	`FGSM-4`	67.51	15.97	64.71	14.8	70.57	15.6	65.7	27.37	63.32	17.41	50.06	31.45
	`PGD-4`	71.99	12.47	71.35	12.6	71.57	14.44	69.5	23.35	70.11	13.39	35.94	74.19
	`CosPGD-4`	71.84	13.13	71.61	12.14	71.58	14.44	69.47	23.63	70.48	14.36	36.07	73.97
	`VAFA-20`	68.52	19.79	71.56	13.43	64.93	18.85	66.36	22.05	68.19	21.2	45.28	31.42

Table 8: Performance of models against transfer-based *black box* attacks on `Hecktor` dataset. For pixel-based attacks results are reported for $\varepsilon = \frac{4}{255}$ indicated by attack names followed by the suffixes -4, respectively. Regarding frequency-based attack `VAFA`, the results are reported with a constraint on $q_{\max}$ set to 20, denoted as `VAFA-20`. DSC score *(lower is better)* is reported on the generated adversarial examples.

Surrogate	Attack	**UNet**		**SegResNet**		**UNETR**		**SwinUNETR**		**UMamba-B**		**UMamba-E**	
		DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑
	Clean	85.52	5.75	89.65	2.56	76.37	16.31	84.19	7.93	88.22	6.01	80.91	8.48
	GN	-	-	-	-	-	-	-	-	-	-	-	-
UNet	FGSM-4	59.57	18.93	88.54	3.74	75.15	17.32	82.86	8.27	86.93	7.05	79.17	8.92
	PGD-4	22.75	38.22	89.38	3.18	75.73	17.29	83.75	8.63	87.72	6.8	80.02	8.73
	CosPGD-4	23.52	35.54	89.24	3.29	75.46	17.34	83.5	8.48	87.72	6.4	79.95	8.55
	VAFA-20	49.69	27.47	74.0	17.66	55.71	22.52	58.27	22.96	60.81	24.13	52.7	25.48
SegResNet	FGSM-4	84.67	5.92	69.85	11.22	75.08	17.52	82.49	8.16	85.2	10.12	78.95	8.57
	PGD-4	85.04	6.09	21.96	38.44	75.62	17.08	83.62	8.24	87.42	6.84	80.03	8.77
	CosPGD-4	85.08	5.85	22.68	38.34	75.67	17.15	83.62	8.16	87.31	6.77	79.3	8.73
	VAFA-20	51.79	27.04	54.8	23.06	54.41	22.32	56.64	22.84	59.04	24.27	51.82	26.04
UNETR	FGSM-4	83.39	8.43	87.74	4.14	39.45	26.49	80.67	10.34	85.09	12.68	78.52	9.45
	PGD-4	85.01	7.95	88.31	4.25	24.3	33.77	82.93	9.62	86.46	11.63	78.72	9.19
	CosPGD-4	85.09	7.79	88.16	4.37	24.79	33.27	82.63	10.53	86.22	12.23	78.15	9.53
	VAFA-20	52.38	26.55	73.19	16.56	48.99	23.31	55.29	24.52	57.86	24.69	50.38	26.23
SwinUNETR	FGSM-4	84.44	6.38	85.64	4.41	74.37	16.69	60.68	19.26	84.64	8.6	78.99	8.58
	PGD-4	84.95	5.85	87.83	4.6	74.53	17.5	23.47	34.53	86.65	8.74	77.56	9.95
	CosPGD-4	85.0	6.39	87.92	3.96	74.46	17.59	23.54	34.97	86.78	8.84	77.86	9.68
	VAFA-20	51.47	26.55	70.59	16.1	53.55	22.66	48.42	25.2	57.92	24.72	51.46	26.11
UMamba-B	FGSM-4	83.68	8.65	86.22	4.41	74.71	17.35	82.2	8.95	76.9	14.19	77.72	8.86
	PGD-4	84.22	8.86	86.58	5.62	74.53	17.49	82.45	9.33	29.55	31.7	74.43	10.37
	CosPGD-4	84.26	8.03	86.54	5.24	74.66	17.27	82.6	9.51	30.5	29.28	73.78	11.24
	VAFA-20	50.83	27.17	69.36	17.54	53.71	22.46	55.54	23.19	54.22	24.92	50.69	26.11
UMamba-E	FGSM-4	84.28	7.03	87.74	3.43	74.41	17.51	83.5	8.09	85.73	8.73	54.11	19.96
	PGD-4	85.13	6.02	89.17	3.39	75.56	17.04	83.65	8.09	87.36	6.35	28.09	29.77
	CosPGD-4	85.24	5.88	88.91	3.09	75.54	17.16	83.72	8.2	87.39	6.67	27.24	30.24
	VAFA-20	54.27	26.67	75.58	15.64	57.17	22.19	59.84	22.92	61.11	24.42	50.47	25.7

Table 9: Performance of models against transfer-based *black box* attacks on ACDC dataset. For pixel-based attacks results are reported for $\epsilon = \frac{4}{255}$ indicated by attack names followed by the suffixes -4, respectively. Regarding frequency-based attack VAFA, the results are reported with a constraint on $q_{\max}$ set to 20, denoted as VAFA-20. DSC score *(lower is better)* is reported on the generated adversarial examples.

Surrogate	Attack	**UNet**		**SegResNet**		**UNETR**		**SwinUNETR**		**UMamba-B**		**UMamba-E**	
		DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑
	Clean	75.72	9.15	80.84	8.14	72.53	15.08	78.07	10.01	78.37	8.12	77.06	11.11
	GN	75.40	10.31	80.34	8.44	72.33	14.45	77.44	9.73	77.69	9.44	75.34	21.80
UNet	FGSM-8	49.49	48.23	77.98	9.49	70.51	15.99	74.97	12.26	74.64	10.03	71.84	25.95
	PGD-8	32.06	89.45	79.78	8.73	72.02	15.28	76.89	10.69	76.92	9.02	74.74	13.86
	CosPGD-8	32.51	75.89	79.78	8.71	71.99	14.94	76.87	11.02	76.83	10.36	74.79	16.50
	VAFA-30	19.49	86.75	34.39	53.31	45.49	34.68	41.48	37.00	29.00	57.25	19.38	71.95
SegResNet	FGSM-8	71.70	12.01	59.97	37.36	69.93	15.34	72.48	11.81	68.73	14.77	67.95	22.74
	PGD-8	73.34	13.94	20.11	100.13	70.74	15.15	74.95	12.29	70.79	14.25	70.08	21.67
	CosPGD-8	73.29	13.76	19.68	96.74	70.75	15.12	75.04	11.75	70.50	14.16	69.79	21.31
	VAFA-30	36.43	44.43	22.00	72.16	50.16	28.19	45.54	36.83	28.39	55.87	19.53	69.45
UNETR	FGSM-8	70.89	12.07	76.13	10.51	50.98	44.86	71.48	12.97	72.47	10.35	71.21	15.05
	PGD-8	72.24	13.49	78.46	9.57	48.01	53.78	74.64	12.62	74.62	10.29	72.73	15.29
	CosPGD-8	72.31	13.61	78.73	9.29	47.46	51.14	74.83	13.44	74.78	9.80	72.81	15.43
	VAFA-30	40.27	35.24	42.55	34.84	26.68	58.83	41.90	39.54	36.16	43.97	31.47	45.45
SwinUNETR	FGSM-8	71.47	12.58	75.09	10.86	68.94	15.74	58.80	40.84	71.88	11.46	70.15	19.12
	PGD-8	72.28	13.68	77.92	9.61	70.17	17.17	54.09	64.43	74.62	11.56	72.19	20.70
	CosPGD-8	72.40	14.59	77.87	9.71	70.20	18.13	52.93	58.46	74.72	12.03	72.19	21.12
	VAFA-30	29.22	54.11	30.30	44.20	38.69	38.22	24.30	62.38	25.69	49.59	21.96	56.46
UMamba-B	FGSM-8	71.39	14.14	71.96	13.29	70.13	15.97	72.63	12.73	58.63	35.45	65.75	23.13
	PGD-8	73.27	12.02	76.11	12.59	71.06	15.61	75.59	12.31	30.08	98.04	67.72	23.63
	CosPGD-8	73.30	12.92	76.41	12.78	71.11	14.97	75.74	11.72	30.96	82.75	68.26	25.44
	VAFA-30	35.18	42.78	30.45	49.72	46.93	34.42	42.89	38.19	16.65	79.04	17.01	73.73
UMamba-E	FGSM-8	72.59	12.33	75.69	11.32	70.49	16.27	74.48	11.02	71.36	14.58	53.59	45.54
	PGD-8	74.39	13.66	79.09	9.55	71.69	15.04	76.91	11.23	75.58	10.74	29.51	104.51
	CosPGD-8	74.41	13.73	79.14	9.94	71.75	15.38	76.95	11.24	75.83	11.52	31.10	79.69
	VAFA-30	44.08	36.78	42.47	36.74	53.61	25.57	54.41	25.99	32.90	47.13	16.45	91.35

Table 10: Performance of models against transfer-based *black box* attacks on BTCV dataset. For pixel-based attacks results are reported for $\varepsilon = \frac{8}{255}$ indicated by attack names followed by the suffixes -8, respectively. Regarding frequency-based attack VAFA, the results are reported with a constraint on $q_{\max}$ set to 30, denoted as VAFA-30. DSC score *(lower is better)* is reported on the generated adversarial examples.

Surrogate	Attack	**UNet**		**SegResNet**		**UNETR**		**SwinUNETR**		**UMamba-B**		**UMamba-E**	
		DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑
	Clean	76.79	19.72	80.89	13.30	71.35	27.73	79.33	25.79	81.08	15.51	78.05	18.31
	GN	76.13	21.97	77.59	18.16	71.13	34.37	78.95	29.09	79.33	16.78	76.27	18.16
`UNet`	`FGSM-8`	46.85	60.19	71.90	18.50	68.90	35.00	74.99	31.74	72.55	22.68	70.24	24.43
	`PGD-8`	29.14	90.85	76.35	15.56	70.19	32.92	77.73	28.77	77.22	19.29	74.47	20.15
	`CosPGD-8`	30.80	77.47	76.44	15.66	70.24	33.87	77.86	28.52	77.43	18.27	74.58	20.25
	`VAFA-30`	20.92	88.19	26.62	66.39	38.36	63.57	38.75	57.28	22.36	71.25	18.89	75.11
`SegResNet`	`FGSM-8`	70.50	23.59	50.73	49.07	68.71	34.66	74.29	27.96	64.91	29.14	63.94	26.72
	`PGD-8`	73.76	22.77	17.42	106.73	69.90	32.58	76.90	28.83	70.87	23.57	69.81	24.80
	`CosPGD-8`	74.07	24.07	18.88	102.25	69.93	32.26	77.00	29.17	71.59	23.13	70.36	23.56
	`VAFA-30`	38.47	51.36	17.68	84.91	43.33	54.59	45.00	47.85	22.43	63.84	19.75	64.63
`UNETR`	`FGSM-8`	70.38	23.55	71.68	16.66	48.48	58.85	71.54	31.48	71.70	25.19	69.28	21.42
	`PGD-8`	72.57	22.99	74.85	17.19	44.83	76.08	75.28	30.96	74.82	18.84	70.82	23.81
	`CosPGD-8`	72.67	22.73	75.11	17.05	44.64	74.34	75.46	30.61	74.82	18.47	71.00	23.05
	`VAFA-30`	39.68	48.60	32.63	56.05	22.34	88.08	38.11	54.49	29.60	54.23	26.77	58.62
`SwinUNETR`	`FGSM-8`	69.99	24.44	69.39	18.48	66.81	35.05	53.46	71.03	70.21	23.45	68.12	23.85
	`PGD-8`	72.67	24.14	73.19	18.20	68.77	38.35	46.95	88.88	74.39	23.43	71.40	24.23
	`CosPGD-8`	72.88	23.67	73.56	18.34	68.85	38.09	46.79	82.92	74.59	22.30	71.67	23.91
	`VAFA-30`	30.47	60.63	27.51	60.51	33.10	64.36	25.43	76.41	24.11	65.63	22.06	65.39
`UMamba-B`	`FGSM-8`	70.60	24.91	65.06	23.98	68.69	35.94	74.25	31.40	51.36	53.89	62.19	30.01
	`PGD-8`	73.49	22.17	69.75	21.84	69.93	32.71	76.89	28.43	18.20	111.79	65.72	28.54
	`CosPGD-8`	73.46	22.99	70.35	20.94	69.94	33.08	76.92	28.69	18.99	101.89	66.36	30.04
	`VAFA-30`	38.49	55.47	24.20	65.44	43.80	58.37	44.61	52.18	14.99	92.09	18.46	69.64
`UMamba-E`	`FGSM-8`	70.93	23.48	68.27	20.87	68.69	35.07	74.85	29.52	65.66	30.35	48.81	52.13
	`PGD-8`	74.20	21.71	73.82	17.58	70.25	32.12	77.55	28.89	71.18	23.11	23.59	96.84
	`CosPGD-8`	74.41	21.16	74.15	17.06	70.30	32.56	77.65	28.78	71.86	22.90	24.99	82.01
	`VAFA-30`	39.41	53.15	27.39	61.14	43.32	58.63	46.53	50.97	23.31	72.51	13.88	92.91

Table 11: Performance of models against transfer-based *black box* attacks on `Abdomen-CT` dataset. For pixel-based attacks results are reported for $\varepsilon = \frac{8}{255}$ indicated by attack names followed by the suffixes -8, respectively. Regarding frequency-based attack `VAFA`, the results are reported with a constraint on $q_{\max}$ set to 30, denoted as `VAFA-30`. DSC score *(lower is better)* is reported on the generated adversarial examples.

Surrogate	Attack	UNet		SegResNet		UNETR		SwinUNETR		UMamba-B		UMamba-E	
		DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑
	Clean	73.91	11.36	74.73	11.08	72.36	14.61	71.61	22.07	73.50	10.89	72.19	13.29
	GN	72.11	12.97	73.26	10.08	71.28	14.51	69.61	23.65	72.12	13.96	66.58	13.45
UNet	FGSM-8	43.78	37.79	59.11	18.23	66.49	17.23	58.76	30.61	59.26	18.63	57.47	19.08
	PGD-8	34.07	78.61	68.24	13.80	69.87	17.54	64.99	27.33	67.38	14.62	65.55	17.80
	CosPGD-8	34.22	78.62	69.48	12.92	70.02	16.83	66.11	26.00	68.58	13.11	66.21	16.10
	VAFA-30	44.30	42.19	68.81	14.80	63.10	18.37	65.10	22.05	64.99	17.93	60.37	21.85
SegResNet	FGSM-8	60.62	21.43	47.70	32.24	67.51	17.17	57.01	30.20	53.92	27.12	55.13	21.08
	PGD-8	61.85	20.89	33.02	81.17	68.59	17.13	58.63	29.07	51.21	31.48	54.53	25.55
	CosPGD-8	61.99	19.89	33.41	81.28	68.81	16.63	58.74	29.19	50.25	32.25	54.15	23.28
	VAFA-30	69.07	17.77	47.55	33.08	66.24	17.71	65.55	21.85	65.23	17.79	62.22	23.54
UNETR	FGSM-8	58.16	19.27	59.01	15.60	44.89	35.08	54.29	31.27	57.88	17.11	56.95	19.62
	PGD-8	58.33	20.35	61.02	16.89	34.93	74.58	55.55	32.75	60.38	19.40	58.98	17.85
	CosPGD-8	59.02	21.18	61.97	17.29	34.89	73.84	56.13	32.29	61.55	19.64	59.95	20.11
	VAFA-30	64.73	19.36	69.68	13.81	48.01	33.82	57.85	27.53	65.38	19.19	58.59	21.28
SwinUNETR	FGSM-8	57.96	21.82	55.59	18.97	64.66	19.39	43.82	52.95	54.54	24.71	53.39	21.48
	PGD-8	55.25	25.87	56.09	20.16	66.29	19.83	34.12	78.85	55.14	23.72	54.18	23.60
	CosPGD-8	55.83	23.88	56.61	21.35	66.53	19.43	34.21	77.62	55.37	25.18	54.70	23.08
	VAFA-30	65.58	19.76	68.38	15.46	61.97	20.25	44.44	42.14	65.93	16.53	58.56	20.68
UMamba-B	FGSM-8	60.34	18.63	55.50	18.78	67.23	16.73	56.29	33.02	50.16	30.12	54.51	19.86
	PGD-8	61.83	19.34	49.65	39.99	69.07	16.88	58.12	31.30	34.31	78.39	51.03	30.70
	CosPGD-8	61.87	20.34	48.57	40.81	69.12	17.61	59.05	31.02	34.41	78.44	51.39	31.16
	VAFA-30	66.61	20.86	64.99	16.32	64.11	18.50	64.53	21.53	49.27	35.99	58.52	22.13
UMamba-E	FGSM-8	62.64	16.67	59.29	16.55	68.43	17.10	60.34	29.29	57.49	22.28	47.15	31.65
	PGD-8	69.30	14.95	68.77	12.22	70.59	15.53	66.92	25.14	66.75	16.73	34.29	78.25
	CosPGD-8	69.55	13.86	69.49	12.88	70.67	15.07	67.34	26.04	67.43	16.99	34.37	77.85
	VAFA-30	68.09	20.51	69.84	13.90	64.26	19.12	65.86	24.17	66.03	20.12	44.81	36.84

Table 12: Performance of models against transfer-based *black box* attacks on Hecktor dataset. For pixel-based attacks results are reported for $\varepsilon = \frac{8}{255}$ indicated by attack names followed by the suffixes -8, respectively. Regarding frequency-based attack VAFA, the results are reported with a constraint on $q_{\max}$ set to 30, denoted as VAFA-30. DSC score *(lower is better)* is reported on the generated adversarial examples.

Surrogate	Attack	UNet		SegResNet		UNETR		SwinUNETR		UMamba-B		UMamba-E	
		DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑	DSC↓	HD95↑
	Clean	85.52	5.75	89.65	2.56	76.37	16.31	84.19	7.93	88.22	6.01	80.91	8.48
	GN	85.01	5.67	89.26	2.89	74.88	18.49	83.06	10.69	86.95	6.68	78.11	11.67
UNet	FGSM-8	55.55	21.99	86.65	4.55	73.32	18.87	81.12	11.22	84.58	11.19	74.96	13.42
	PGD-8	21.42	39.14	88.05	5.21	74.67	18.96	82.57	10.59	86.32	10.08	77.82	10.82
	CosPGD-8	22.62	36.84	88.18	5.12	74.36	18.73	82.92	10.12	86.50	9.92	77.33	11.97
	VAFA-30	49.85	27.44	74.15	18.17	56.09	21.77	58.72	22.95	60.27	24.46	53.68	25.98
SegResNet	FGSM-8	83.63	6.32	66.58	11.88	73.03	19.72	80.41	12.75	81.15	19.95	74.57	14.46
	PGD-8	84.38	6.34	20.42	37.34	74.59	17.62	82.27	9.84	85.72	10.48	76.39	11.34
	CosPGD-8	84.54	6.08	22.33	38.61	74.79	18.05	82.44	9.58	85.69	9.58	76.41	11.32
	VAFA-30	51.83	27.22	54.47	23.38	54.96	21.60	57.00	24.20	59.24	24.27	53.12	25.38
UNETR	FGSM-8	80.15	12.78	83.33	6.48	33.16	29.77	75.79	16.74	77.48	25.93	72.88	14.93
	PGD-8	83.25	13.69	85.54	7.19	22.31	35.44	80.12	15.19	80.67	22.48	68.37	16.98
	CosPGD-8	82.98	13.48	86.08	7.09	23.08	34.80	80.12	15.41	81.37	21.11	70.52	17.51
	VAFA-30	52.67	26.66	74.18	15.78	49.39	22.63	55.90	24.47	58.27	24.85	52.16	25.09
SwinUNETR	FGSM-8	83.11	7.61	81.34	5.11	71.91	18.15	56.11	21.83	80.69	15.61	74.72	11.74
	PGD-8	84.04	7.49	84.37	6.75	71.73	19.22	21.69	36.99	82.42	15.23	72.15	12.77
	CosPGD-8	83.92	7.75	84.38	7.25	71.91	20.24	22.50	36.20	83.34	19.98	70.49	13.23
	VAFA-30	51.50	26.94	70.18	17.16	53.50	23.25	48.44	25.57	57.68	25.01	52.52	25.27
UMamba-B	FGSM-8	81.55	14.53	83.27	5.11	71.83	19.99	79.61	13.74	72.55	21.19	71.54	14.86
	PGD-8	82.82	12.81	82.49	6.18	72.14	19.13	79.95	12.33	24.95	34.66	69.66	15.29
	CosPGD-8	82.58	10.26	82.69	7.37	72.50	18.78	80.14	11.97	25.54	31.32	68.93	14.86
	VAFA-30	51.46	27.35	71.36	17.79	54.18	22.82	57.14	13.39	53.38	25.22	51.83	26.27
UMamba-E	FGSM-8	82.44	13.41	84.96	4.75	71.23	20.59	81.82	12.16	81.97	15.77	51.65	24.22
	PGD-8	84.66	6.11	87.71	3.43	74.51	18.16	82.95	9.17	86.12	7.46	25.01	32.43
	CosPGD-8	84.79	6.32	87.31	3.84	74.86	17.58	83.01	9.52	86.19	7.14	24.44	31.55
	VAFA-30	55.01	27.08	75.57	15.77	57.40	20.80	60.22	23.85	61.26	24.54	51.61	25.72

Table 13: Performance of models against transfer-based *black box* attacks on ACDC dataset. For pixel-based attacks results are reported for $\varepsilon = \frac{8}{255}$ indicated by attack names followed by the suffixes -8, respectively. Regarding frequency-based attack VAFA, the results are reported with a constraint on $q_{\max}$ set to 30, denoted as VAFA-30. DSC score *(lower is better)* is reported on the generated adversarial examples.

Surrogate	Attack	BTCV		ACDC		Hecktor		Abdomen-CT	
		DSC↓	IoU↓	DSC↓	IoU↓	DSC↓	IoU↓	DSC↓	IoU↓
UNet	FGSM-8	72.56	60.84	61.74	48.09	36.41	24.66	76.73	65.07
	PGD-8	73.23	61.39	60.44	46.54	36.15	24.55	77.15	65.53
	CosPGD-8	73.34	61.51	60.50	46.68	36.32	24.59	77.05	65.44
	VAFA-30	66.26	53.80	47.01	35.51	36.77	24.70	69.82	56.99
SegResNet	FGSM-8	73.16	61.34	63.15	49.86	36.64	24.73	76.82	65.21
	PGD-8	72.95	61.24	61.38	47.64	34.98	23.57	76.84	85.27
	CosPGD-8	73.05	61.31	61.99	48.35	34.76	23.38	76.92	65.32
	VAFA-30	66.28	53.76	45.13	33.64	36.27	24.39	71.57	58.92
UNETR	FGSM-8	71.59	59.82	61.18	47.65	37.52	25.49	75.49	63.66
	PGD-8	72.15	60.33	58.93	45.80	38.19	26.08	76.15	64.44
	CosPGD-8	72.34	60.46	59.74	46.55	38.32	26.37	76.16	64.43
	VAFA-30	65.31	52.77	45.16	33.43	40.00	27.24	68.35	55.07
SwinUNETR	FGSM-8	73.05	61.19	62.72	48.68	35.94	24.20	76.54	64.91
	PGD-8	73.06	61.23	61.59	47.82	36.35	24.81	76.67	64.99
	CosPGD-8	73.04	61.26	61.32	47.65	36.21	24.62	76.82	65.22
	VAFA-30	65.23	52.59	43.32	31.76	38.61	26.12	68.03	54.81
UMamba-B	FGSM-8	72.93	61.12	62.85	49.62	36.86	25.05	76.83	65.23
	PGD-8	72.94	61.16	60.10	46.45	35.56	24.09	77.01	65.38
	CosPGD-8	73.17	61.34	60.91	46.99	36.37	24.64	77.12	65.54
	VAFA-30	65.47	53.15	44.51	32.98	37.54	25.41	71.61	58.82
UMamba-E	FGSM-8	73.03	61.19	63.56	50.06	36.95	25.04	77.01	65.38
	PGD-8	73.04	61.28	60.56	46.78	35.72	24.14	77.15	65.53
	CosPGD-8	73.24	61.42	61.15	47.62	35.67	24.12	77.02	65.43
	VAFA-30	67.43	55.23	45.47	34.04	36.55	24.63	71.47	58.75

Table 14: Evaluating SAM-Med3D on adversarial examples crafted on surrogate models trained on BTCV, ACDC, Hecktor, and Abdomen-CT datasets.