GUARDIAN: Guarding Against Uncertainty and Adversarial Risks in Robot-Assisted Surgeries

Supplementary Material

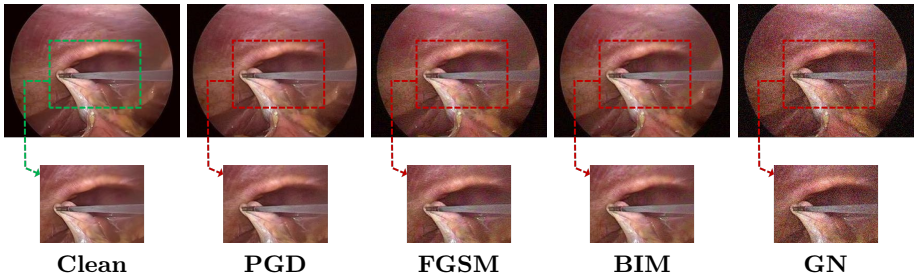

Fig. A. Visualizing clean versus adversarial perturbed images with alpha and epsilon values set to 8 and a standard deviation of 0.1.

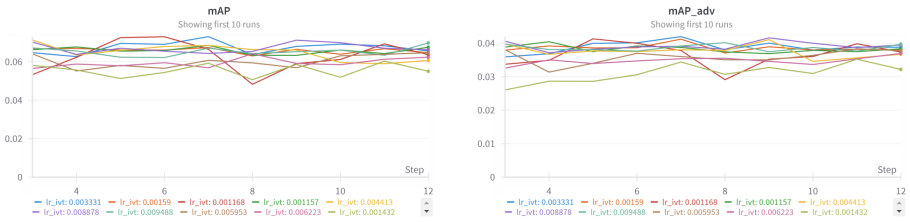

Fig. B. We conducted an ablation study to obtain optimal learning rates for the RDV model against the mAP for the triplet recognition across the initial 12 epochs, focusing on the volatility of learning rate trends to establish a robust starting point.

Table A. Impact of epsilon on image quality.

Attack	Epsilon	PSNR	SSIM	LPIPS
PGD	4	**38.74**	**0.878**	**0.901**
	8	33.68	0.732	0.833
	16	28.70	0.517	0.769
	32	23.34	0.285	0.684
FGSM	4	**36.59**	**0.843**	**0.876**
	8	30.59	0.629	0.782
	16	24.64	0.364	0.669
	32	18.81	0.169	0.521
BIM	4	**38.08**	**0.869**	**0.893**
	8	33.06	0.716	0.822
	16	28.45	0.529	0.745
	32	26.68	0.460	0.710

Table B. Impact of alpha on image quality.

Attack	Alpha	PSNR	SSIM	LPIPS
PGD	2	**33.68**	**0.732**	**0.833**
	4	33.05	0.710	0.819
	8	32.04	0.675	0.803
	16	30.59	0.616	0.776
BIM	2	**34.31**	**0.762**	**0.840**
	4	33.06	0.716	0.823
	8	32.11	0.687	0.805
	16	30.59	0.632	0.778

Table C. Impact of steps on image quality.

Attack	Num Steps	PSNR	SSIM	LPIPS
PGD	5	**34.14**	**0.746**	**0.847**
	10	33.68	0.732	0.833
	15	33.49	0.726	0.827
	20	33.40	0.724	0.824
BIM	5	**33.52**	**0.732**	**0.829**
	10	33.06	0.716	0.822
	15	33.16	0.720	0.823
	20	33.10	0.718	0.822

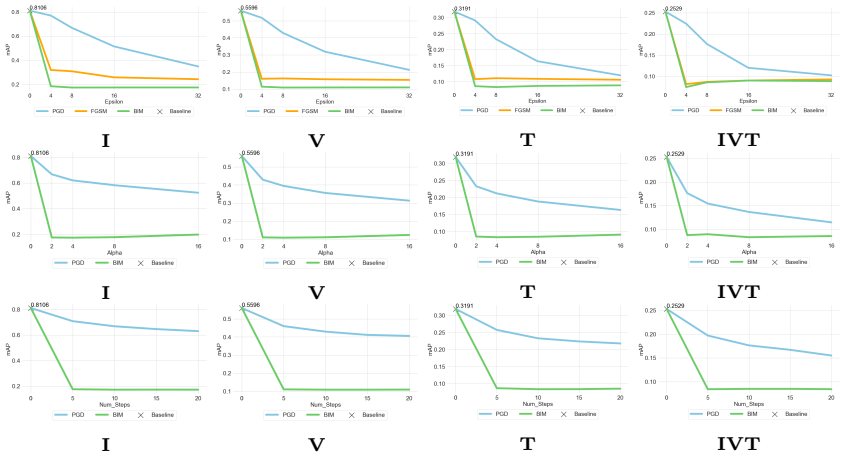

Fig. C. Visualization of the impact of hyperparameters from PGD, FGSM, and BIM attacks on the RDV model's mAP: varying epsilon (row one), alpha (row two), and steps (row three), with other parameters held constant in each row.