Types	Attacks	No defense		FP			NAD			I-BAU			FT-SAM			CL-Guard(OURS)		
		ASR	ACC	ASR	ACC	SEI	ASR	ACC	SEI	ASR	ACC	SEI	ASR	ACC	SEI	ASR	ACC	SEI
Pixel Space	Badnets	94.80	96.67	35.25	96.41	62.54	76.78	95.81	18.10	0.11	95.72	98.88	21.85	96.69	76.95	0.29	96.91	99.69
	Blended	99.41	95.54	82.40	94.96	16.52	98.43	95.11	0.55	70.41	94.68	28.30	57.24	96.56	42.42	18.26	96.53	81.63
	TrojanNet	99.94	95.95	47.94	95.93	52.01	99.35	96.36	0.59	3.58	95.51	95.97	1.03	95.55	98.56	0.85	93.96	97.09
	SIG	97.51	95.68	71.03	95.33	26.79	69.43	95.30	28.40	0.38	88.32	92.06	15.67	96.03	83.92	0.06	94.99	99.23
	Dynamic	96.93	95.33	5.85	95.89	93.96	40.22	95.67	58.50	13.02	94.92	86.14	6.08	96.59	93.72	4.83	96.75	95.02
	ISSBA	99.83	95.83	13.89	96.15	86.08	49.62	95.91	50.29	60.32	93.04	36.78	0.10	88.74	92.79	2.29	96.60	97.71
	WaNet	96.30	95.82	0.15	97.56	99.84	6.05	97.83	93.71	0.07	97.22	99.92	0.02	97.68	99.97	0.23	97.47	99.76
	BPPA	95.79	96.14	5.75	97.41	93.99	20.48	97.39	78.61	0.06	97.45	99.93	0.18	97.03	99.81	0.54	97.39	99.44
Feature Space	Refool	87.02	93.01	27.78	95.58	68.07	40.12	95.15	53.89	81.74	94.25	6.07	36.17	94.37	0.00	0.65	94.70	99.25
	FBA	92.02	93.01	12.58	96.52	86.33	44.45	95.74	51.69	50.74	94.33	44.86	12.17	95.37	86.77	2.36	96.72	97.43
Avg. one the above attacks		-	-	32.22	96.13	66.65	55.60	96.05	42.52	25.52	94.56	71.56	15.05	86.46	73.15	3.03	96.19	96.63

Table 1: Comparison with state-of-the-art mitigations on GTSRB with 10% benign data on VGG-16 (%).

Appendix

Experimental Setup

All experiments were conducted on a system running Linux 5.13.0, equipped with a 3.1 GHz 16-core Intel Core i9 processor, 128 GB of RAM, and an NVIDIA GeForce RTX 4090 GPU. The software environment includes Python 3.8, PyTorch 2.0.0+cu118, Torchvision 0.15.1+cu118, SciPy 1.10.1, NumPy 1.23.5, and scikit-image 0.21.0.

Experimental Supplement

Table 1 presents the experimental results of VGG-16 on the GTSRB dataset. For pixel-space attacks, although the proposed method yields slightly higher ASR values than FT-SAM under ISSBA, WaNet, and BPPA by approximately $\langle 2.19, 0.12, 0.36 \rangle$ percentage points, it demonstrates superior average performance overall. Compared to FP, NAD, I-BAU, and FT-SAM, the proposed method achieves average improvements of approximately $\langle 29.37, 0.19, 29.73 \rangle$, $\langle 54.13, 0.15, 55.11 \rangle$, $\langle 15.08, 1.72, 16.45 \rangle$, and $\langle 9.36, 0.72, 10.24 \rangle$ percentage points in ACC, ASR, and SEI, respectively. For feature-space attacks, the proposed method consistently performs best, with average SEI improvements of $\langle 20.14, 45.55, 72.88, 76.67 \rangle$ percentage points over FP, NAD, I-BAU, and FT-SAM, respectively. Across all aforementioned attack types, the proposed method still exhibits a significant overall advantage, outperforming the best values of ACC, ASR, and SEI by approximately $\langle 12.26, 0.01, 11.07 \rangle$ percentage points.