Table 1: Evaluation of proposed and baseline methods across multiple robustness benchmarks. The best results are highlighted in bold font and the second best is highlighted in underlines. All backbone models are ViT-B with 86M parameters. In most situations, our proposed method achieves better performance over baselines. We exclude FourierFormer as it only implements the tiny version of ViT at this moment.

Dataset	ImageNet-C	ImageNet-A	ImageNet-O	ImageNet-R
Metric	mCE↓	Top-1 Acc↑	AUPR↑	Top-1 Err Rate↓
ViT	55.54	26.62	21.43	73.60
RVT	46.80	**28.50**	25.87	51.30
ViT-RKDE (Huber)	46.72	26.68	28.33	50.26
ViT-RKDE (Hampel)	46.75	26.75	28.65	50.17
ViT-SPKDE	**46.34**	27.86	**30.37**	**49.54**
ViT-MoM	49.36	23.34	29.53	55.62

Table 2: The Top-1 accuracy (%) was measured on ImageNet's clean data as well as under adversarial attacks. All underlying models used are ViT-B with 86M parameters. The proposed methods demonstrate enhanced resilience against various adversarial attacks and simultaneously maintain competitive performance on the original ImageNet.

Method	Clean Data	FGSM	PGD	SPSA
ViT	79.96	63.14	48.74	54.29
RVT	82.70	67.48	51.66	58.77
ViT-RKDE (Huber)	82.24	69.13	52.43	58.24
ViT-RKDE (Hampel)	82.33	69.16	52.54	58.31
ViT-SPKDE	**82.95**	**70.08**	**52.89**	**58.96**
ViT-MoM	81.16	67.22	50.65	57.43

Table 3: We measured the computation time (in seconds per batch) during the training phase for both the proposed and baseline Transformers. These results were obtained using the WikiText-103 dataset for the language modeling task.

	Iterations of KIRWLS				Transformer	SPKDE	MoM	MGK
	1	2	3	5				
Time (s/batch)	0.32	0.43	0.56	0.72	0.17	0.86	0.18	0.45