A THE EFFECT OF REMOVING MILD GRADIENTS AND EXTREME GRADIENTS RESPECTIVELY ON THE CAIT-S/24 MODEL

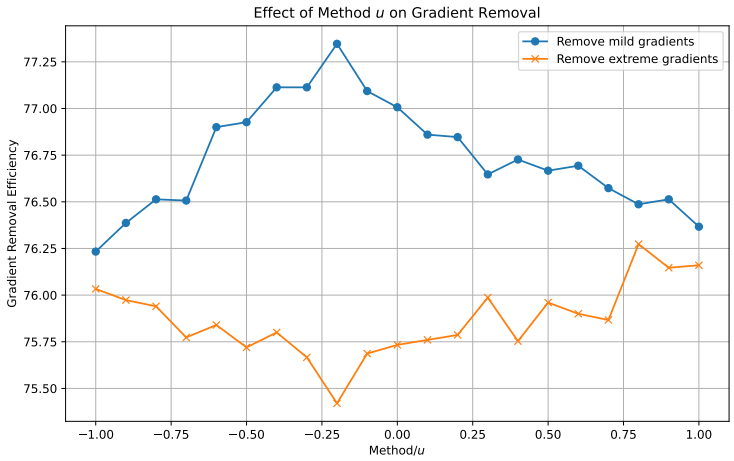

Figure 5: Removing mild gradients and extreme gradients respectively on the CaiT-S/24 model

As shown in Fig. 5, we constructed the adversarial samples on the CaiT-S/24 model by removing mild gradients and extreme gradients respectively, and calculated the average attack success rate after these adversarial samples were transferred to attack other ViT and CNN models. Different hyperparameters in the table adjust the division range of mild gradients. In transferable attacks, the transferability of adversarial samples can be largely affected by overfitting during local training, thus showing different attack success rates on the target models. Therefore, we believe that a higher average attack success rate represents a lower possibility of overfitting.

B THE IMPACT OF GNS AND HFA ON SAMPLE TRANSFERABILITY

Table 5: Ablation study of GNS only, HFA only, both used and none used

	ViT								CNN							
Method	LeViT-256	PiT-B	DeiT-B	ViT-B/16	TNT-S	ConViT-B	Visformer-S	CaiT-S/24	Inc-v3	Inc-v4	IncRes-v2	ResNet-101	Inc-v3-adv-3	Inc-v3-adv-4	IncRes-v2-adv	Average
NONE USED	34.10%	34.00%	62.80%	100.00%	50.60%	64.80%	37.10%	64.70%	32.30%	30.60%	26.30%	32.30%	23.30%	21.00%	19.70%	42.24%
GNS_ONLY	64.40%	58.60%	88.90%	99.90%	78.30%	89.40%	62.60%	87.50%	54.80%	51.00%	42.90%	49.00%	38.10%	38.80%	33.20%	62.49%
HFA_ONLY	58.40%	55.40%	84.00%	100.00%	74.50%	84.10%	57.19%	84.80%	55.19%	52.79%	49.00%	51.99%	45.10%	43.79%	39.80%	62.40%
BOTH	76.30%	70.10%	93.50%	99.90%	88.00%	92.90%	73.70%	92.80%	68.00%	63.40%	59.10%	63.00%	54.90%	54.50%	47.90%	73.20%

We conducted ablation experiments for GNS only or HFA only in Tab. 5 with the same parameters in **Sec.5.1** of the main paper. The experimental results indicate that, compared to attack methods without either GNS or HFA (average success rate of 42.24%), both GNS and HFA play nearly equally crucial roles in enhancing the transferability of adversarial samples (averaging 62.49% and 62.40%, respectively). The strategy to combine GNS and HFA together yields the best algorithm performance (average 73.20%), demonstrating that gradient normalization and scaling for 'mild gradients', coupled with frequency-domain exploration, effectively improve the transferability of adversarial samples. The results of the ablation experiments align with our assumptions regarding the roles played by GNS and HFA.