A. IGAM Hyperparameters

The IGAM hyperparameters are fined through grid search through the same range of hyperparameter values within each transfer task. We report the values of the IGAM models whose results are reported in this paper for reproducibility.

A.1. CIFAR-10 Target Task

IGAM-MNIST $\lambda_{\text{adv}} = 1$, $\lambda_{\text{diff}} = 100$, f_{disc} : 5 CNN layers (16-32-64-128-256 output channels) and updated once for every 10 classifier update steps

IGAM-TranposeConv $\lambda_{\text{adv}} = 1$, $\lambda_{\text{diff}} = 10$, f_{disc} : 5 CNN layers (8-16-32-64 output channels) and updated once for every 5 classifier update steps

IGAM-RandomPad $\lambda_{\text{adv}} = 1$, $\lambda_{\text{diff}} = 10$, f_{disc} : 5 CNN layers (8-16-32-64 output channels) and updated once for every 5 classifier update steps

IGAM-Pad $\lambda_{\text{adv}} = 2$, $\lambda_{\text{diff}} = 20$, f_{disc} : 5 CNN layers (8-16-32-64 output channels) and updated once for every 5 classifier update steps

IGAM-Upsize $\lambda_{\text{adv}} = 5$, $\lambda_{\text{diff}} = 10$, f_{disc} : 5 CNN layers (8-16-32-64 output channels) and updated once for every 5 classifier update steps

A.2. CIFAR-100 Target Task

IGAM-MNIST $\lambda_{\text{adv}} = 0.1$, $\lambda_{\text{diff}} = 200$, f_{disc} : 5 CNN layers (16-32-64-128-256 output channels) and updated once for every 5 classifier update steps

IGAM-CIFAR10 $\lambda_{\text{adv}} = 2$, $\lambda_{\text{diff}} = 10$, f_{disc} : 5 CNN layers (16-32-64-128-256 output channels) and updated once for every 10 classifier update steps

A.3. Tiny-ImageNet Target Task

IGAM-CIFAR10-Resize $\lambda_{\text{adv}} = 0.1$, $\lambda_{\text{diff}} = 200$, f_{disc} : 4 CNN layers (8-16-32-64 output channels) and updated once for every 5 classifier update steps

IGAM-CIFAR10-Crop $\lambda_{\text{adv}} = 2$, $\lambda_{\text{diff}} = 50$, f_{disc} : 4 CNN layers (8-16-32-64 output channels) and updated once for every 5 classifier update steps

IGAM-CIFAR100-Resize $\lambda_{\text{adv}} = 0.1$, $\lambda_{\text{diff}} = 200$, f_{disc} : 4 CNN layers (8-16-32-64 output channels) and updated once for every 5 classifier update steps

IGAM-CIFAR100-Crop $\lambda_{\text{adv}} = 0.5$, $\lambda_{\text{diff}} = 200$, f_{disc} : 4 CNN layers (8-16-32-64 output channels) and updated once for every 5 classifier update steps

B. Proof

Theorem B.1. *The global minimum of L_{adv} is achieved when $\mathrm{J}_s = \mathrm{J}_t$.*

Proof. From [7], the optimal discriminator is

$$f^*_{\text{disc}}(\mathrm{J}) = \frac{p_{\text{teacher}}(\mathrm{J})}{p_{\text{teacher}}(\mathrm{J}) + p_{\text{student}}(\mathrm{J})} \tag{19}$$

We can include the optimal discriminator into Equation (10) to get

$$\begin{aligned} L_{\text{adv}} &= \mathbb{E}_{\mathrm{J}\sim p_{\text{teacher}}}[\log f^*_{\text{disc}}(\mathrm{J})] + \mathbb{E}_{\mathrm{J}\sim p_{\text{student}}}[\log(1 - f^*_{\text{disc}}(\mathrm{J}))] \\ &= \mathbb{E}_{\mathrm{J}\sim p_{\text{teacher}}}\left[\log \frac{p_{\text{teacher}}(\mathrm{J})}{p_{\text{teacher}}(\mathrm{J}) + p_{\text{student}}(\mathrm{J})}\right] \\ &\quad + \mathbb{E}_{\mathrm{J}\sim p_{\text{student}}}\left[\log \frac{p_{\text{student}}(\mathrm{J})}{p_{\text{teacher}}(\mathrm{J}) + p_{\text{student}}(\mathrm{J})}\right] \\ &= KL\left(p_{\text{teacher}} \,\middle\|\, \frac{p_{\text{teacher}} + p_{\text{student}}}{2}\right) \\ &\quad + KL\left(p_{\text{student}} \,\middle\|\, \frac{p_{\text{teacher}} + p_{\text{student}}}{2}\right) - \log 4 \\ &= 2 \cdot JS(p_{\text{teacher}}||p_{\text{student}}) - \log 4 \end{aligned} \tag{20}$$

where KL and JS are the Kullback-Leibler and Jensen-Shannon divergence respectively. Since the Jensen-Shannon divergence is always non-negative, $L_{\text{adv}}(G)$ reaches its global minimum value of $-\log 4$ when $JS(p_{\text{teacher}}||p_{\text{student}}) = 0$. When $\mathrm{J_s} = \mathrm{J_t}$, we get $p_{\text{teacher}} = p_{\text{student}}$ and consequently $JS(p_{\text{teacher}}||p_{\text{student}}) = 0$, thus completing the proof.

□

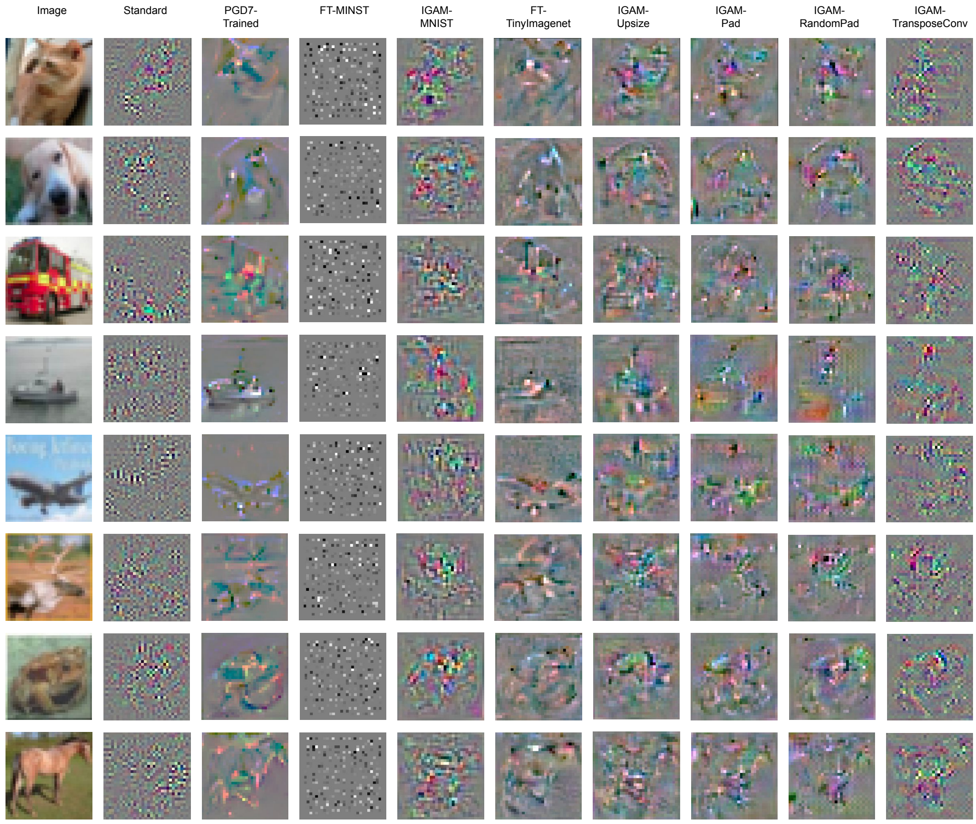

Figure 6: Input gradients of different models.

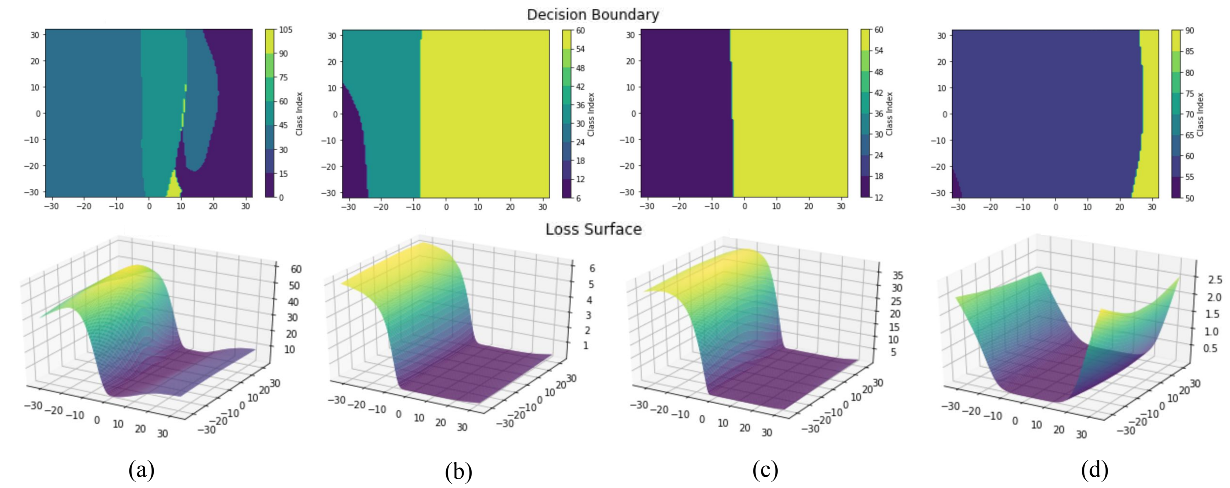

Figure 7: Decision boundaries and loss landscapes of (a) standard trained, (b) IGAM-CIFAR10 ($\lambda_{\text{adv}} = 2, \lambda_{\text{diff}} = 0$), (c) IGAM-CIFAR10 ($\lambda_{\text{adv}} = 0, \lambda_{\text{diff}} = 10$) and (d) IGAM-CIFAR10 ($\lambda_{\text{adv}} = 2, \lambda_{\text{diff}} = 10$) along the adversarial perturbation and a random direction. Correct class: #53.