Supplementary Material

Jiayang Liu[1], Daniel Ts'o[2], Yiming Bu[1], Qinru Qiu[1]
[1]Department of Electrical Engineering and Computer Science, Syracuse University
[2] SUNY Upstate Medical University
{jliu206, ybu104, qiqiu}@syr.edu, {tsod}@upstate.edu

The code is availale at github

A. Details for predictive reconstruction model

In the experiment, the Predictive Reconstruction Encoder consists of 4 Convolutional-LSTM layers and the Predictive Reconstruction Decoder have 3 convolutional layers. As shown in figure 1, a foveal-peripheral view ,V_t,is received and processed by the predictive reconstruction model in each time step t based on the following equations:

$$h^1_{1:t} = Convlstm^1(V_t, h^1_{1:t-1}) \quad (1)$$

$$h^2_{1:t} = Convlstm^2(h^1_{1:t}, h^2_{1:t-1}) \quad (2)$$

$$\mathrm{h}^3_{1:t} = Convlstm^3(h^2_{1:t}, h^3_{1:t-1}) \quad (3)$$

$$\mathrm{s}_{1:t} = Convlstm^4(h^3_{1:t}, s_{1:t-1}) \quad (4)$$

$$G_t = Conv(s_{1:t}) \quad (5)$$

The variables $h^1_{1:t}$, $h^2_{1:t}$, $h^3_{1:t}$and $s_{1:t}$ are the hidden states of the ConvLSTM layers, while G_t is the predictive reconstruction outcome at time t. In our model, the initial three layers of the ConvLSTM gradually increase the depth of feature maps from 3, to 8, 16 and 24, while the last layer compresses all the generated feature maps into channel dimension 3.

B. How peripheral sampling proportion affect system robustness

To determine the optimal sampling ratio for the peripheral region, we evaluate the ViT-Base model against white-box APGD and MIFGSM attacks under varying attack strengths using ImageNet dataset. As shown in Figure 1, a 6% sampling ratio consistently achieves the highest robustness across all tested conditions. When the sampling ratio is too sparse (e.g., below 3%), insufficient visual information is captured, making it difficult to reconstruct semantically coherent images. In contrast, when the sampling ratio is too dense (e.g., above 7%), adversarial perturbations dominate the input, reducing the effectiveness of the reconstruction and weakening the defense. A 6% ratio strikes the optimal balance by preserving enough informative content while suppressing adversarial noise.

C. How number of glimpses affect system robustness

To determine the optimal number of glimpses, each corresponding to a new foveal-peripheral sample from the environment, we evaluate our system under APGD and MIFGSM attacks using the ViT-Base model and ImageNet dataset. Both attacks are configured with an $\epsilon = 16/255$ and run for 25 steps and 10 steps respectively, where the undefended ViT Base model will fail completely.

As shown in the Figure 2, we observe that with 1 or 2 glimpses, the system underperforms due to insufficient visual information. In contrast, 4 or 5 glimpses introduce excessive adversarial perturbations, which degrade the quality of predictive reconstruction.

This highlights an important insight: more information is not always beneficial and there exists a trade-off between information gain and robustness. Empirically, we find that 3 glimpses is the best balance under this setup, and we adopt this setting in all subsequent experiments.

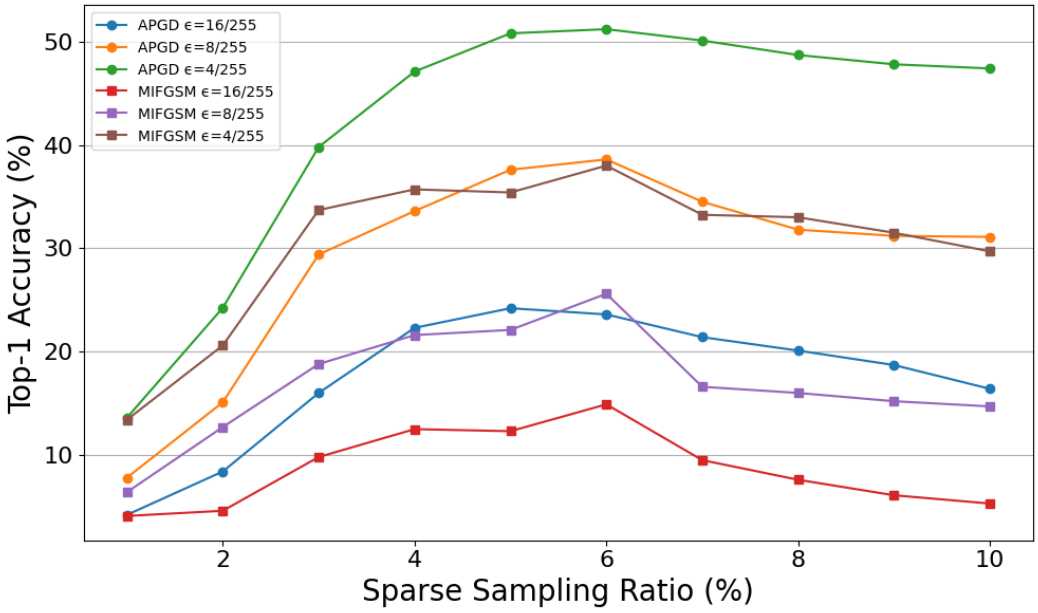

Figure 1. Accuracy comparison under different ϵ values.

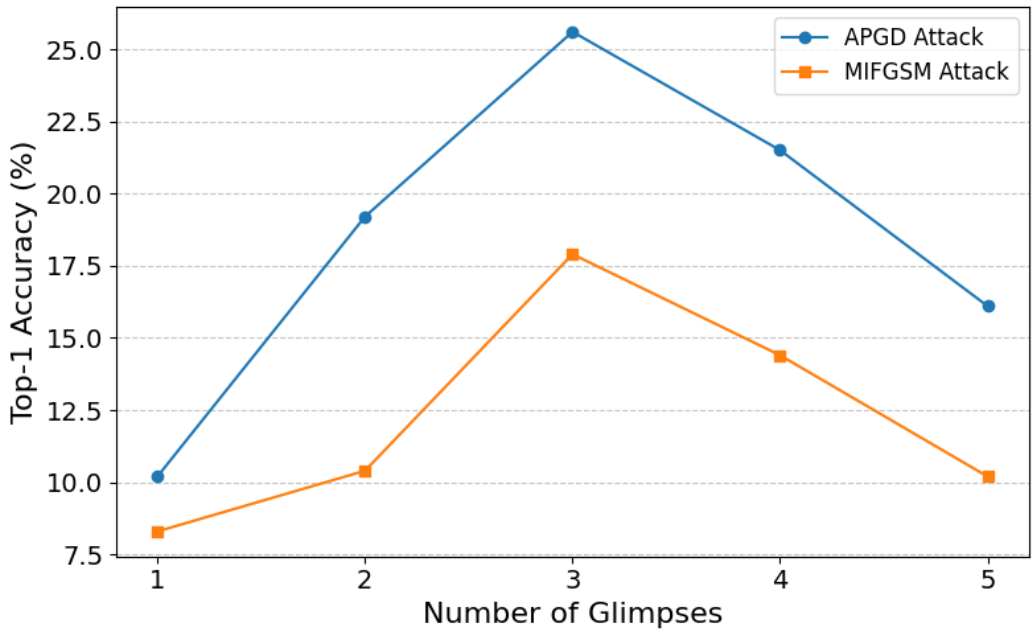

Figure 2. Accuracy comparison under different number of glimpse.