# SIFT-Graph: Benchmarking Multimodal Defense Against Image Adversarial Attacks With Robust Feature Graph

*Jingjie He, Weijie Liang, Zihan Shan, Matthew Caesar*
*University of Illinois Urbana-Champaign*

ICCV OCT 19-23, 2025 HONOLULU HAWAII

## Introduction

Adversarial attacks expose a fundamental vulnerability in modern deep vision models by exploiting their dependence on dense, pixel-level representations that are highly sensitive to imperceptible perturbations. Traditional defense strategies typically operate within this fragile pixel domain, lacking mechanisms to incorporate inherently robust visual features. In this work, we introduce SIFT-Graph, a multimodal defense framework that enhances the robustness of traditional vision models by aggregating structurally meaningful features extracted from raw images using both handcrafted and learned modalities. Specifically, we integrate Scale-Invariant Feature Transform keypoints with a Graph Attention Network to capture scale and rotation invariant local structures that are resilient to perturbations. These robust feature embeddings are then fused with traditional vision model, such as Vision Transformer and Convolutional Neural Network, to form a unified, structure-aware and perturbation defensive model. Preliminary results demonstrate that our method effectively improves the visual model robustness against gradient-based white box adversarial attacks, while incurring only a marginal drop in clean accuracy.

### Intuition behind our design

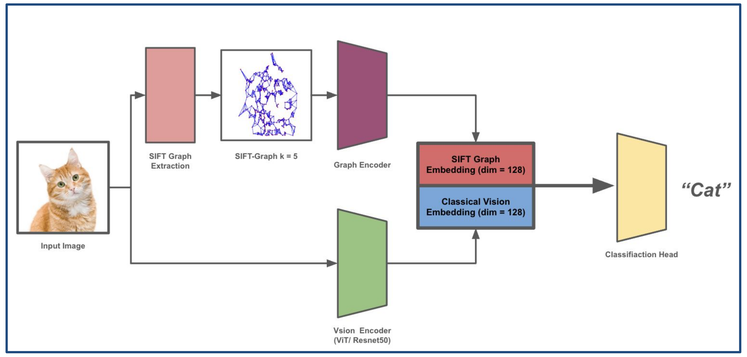
***Why adding SIFT-Graph?***

- SIFT features are scale and rotation invariant. By further aggregating these features into a topologically invariant structure, such as a graph, this invariance can be preserved for further .
- SIFT features naturally prevent gradient based attacks, as keypoint detection, argmax, and histogram binning do not provide tractable gradient flow
- SIFT feature extraction does not require additional training data, making it preferable augmentation under data limited scenarios.

## Methodology

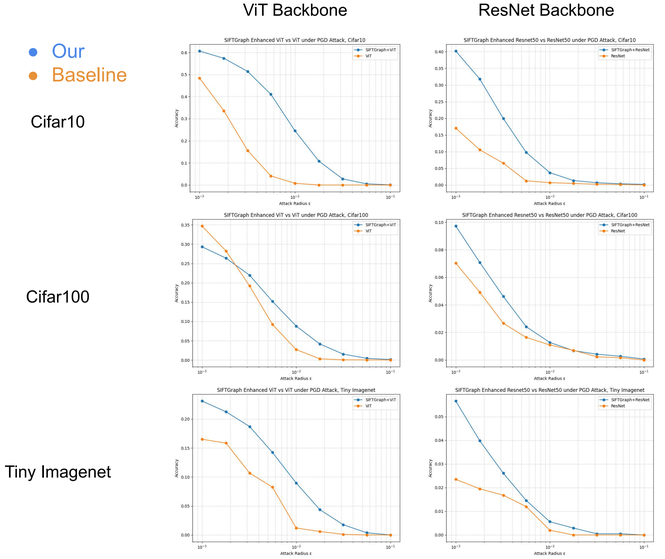
To retrieve the robust features from general image input, our method involve **Four** major component:

- **1.Robust Graph Visual Encoder**
  - **[Image->Vertices]** We leverage SIFT for graph vertices selection
  - **[Vertices->Graph]** We apply KNN for graph connection design
  - **[Graph->Embedding]** We use GAT to aggregate graph level information
- **2.Traditional Vision Encoder**
  - **[Image->Embedding]** We use traditional vision model, two models tested are **ViT** and **ResNet50**
- **3.Embedding Mixing:**
  - We apply simple concatenation with equal weight between graph and visual embedding to derive the final image representation.
- **4.Mixed-embedding Decoder**
  - We perform simple MLP object inference to perform the final classification of the input images.




## Experiment & Result

To evaluate practical adversarial robustness, we conduct PGD white-box attacks on both ViT and ResNet backbones across CIFAR-10, CIFAR-100, and Tiny ImageNet. Preliminary results show that our model consistently improved classifier robustness under various adversarial attack radius.




## Conclusion

We present SIFT-Graph, a multimodal image adversarial defense framework that require no extra data. Preliminary attack experiments on CIFAR-10, CIFAR-100, and Tiny-ImageNet demonstrate its stable enhancement in compared to original baselines.