Pulling Back the Curtain: Unsupervised Adversarial Detection via Contrastive Auxiliary Networks

Eylon Mizrahi, Raz Lapid, Moshe Sipper
Department of Computer Science, Ben-Gurion University of the Negev

BGU
Ben-Gurion University of the Negev

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Abstract

Deep learning models are widely employed in safety-critical applications yet remain highly vulnerable to adversarial attacks—imperceptible perturbations that can severely degrade performance. Existing defenses typically focus either on improving robustness or on detecting adversarial inputs separately.

We propose **U-CAN** (**U**nsupervised adversarial detection via **C**ontrastive **A**uxiliary **N**etworks), a framework that detects adversarial behavior within auxiliary feature representations, without requiring adversarial examples. U-CAN attaches lightweight networks to selected intermediate layers of a frozen backbone, using projection layers and ArcFace-based objectives to refine representations, thereby exposing adversarial shifts.

Experiments on CIFAR-10, Mammals, and ImageNet with ResNet-50, VGG-16, and ViT show that **U-CAN** consistently outperforms existing unsupervised detectors, achieving higher F1 scores across three powerful adversarial attacks, while remaining scalable and efficient.

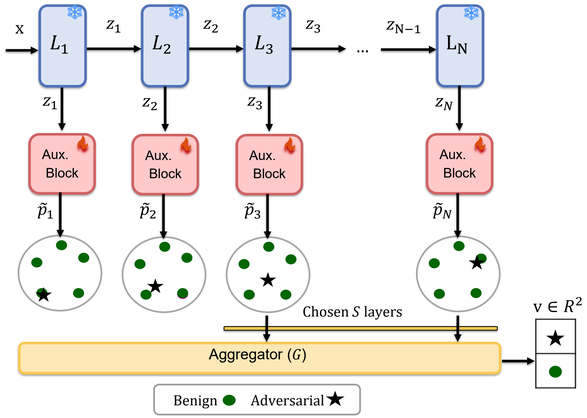
Introduction

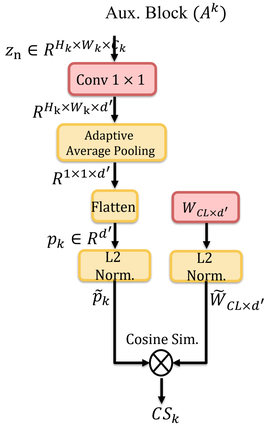
Defenses against adversarial perturbations follow two main paths:
1. **Robustness**: Making the model itself resilient.
2. **Detection**: Flagging adversarial inputs before misclassification.

We propose **U-CAN**, an unsupervised detection method that requires no adversarial examples. Small auxiliary networks are attached to selected layers of a frozen model, producing contrastive representations that expose adversarial shifts from benign inputs.

Key advantages of U-CAN:
- **Unsupervised:** Detects adversarial inputs without labeled data or prior attack knowledge, generalizable to unlimited attack types.
- **No modifications to the target model:** Leaves the target model weights and architecture unchanged, preserving original performance.
- **Layer-wise refinement:** Uses ArcFace-based refinement at multiple intermediate layers for clearer feature analysis, helping separate benign and adversarial inputs.
- **Efficient:** Lightweight auxiliary nets that detect adversarial inputs in a single forward pass with minimal overhead.
- **Compatible:** Integrates with other intermediate-layer adversarial detection methods to further boost performance.

Methodology





U-CAN Algorithm

Input: Frozen model $\mathcal{M}$ with layers $\{L_1, \ldots, L_N\}$,benign training data, and aggregator $\mathcal{G}$.

Output: Trained adversarial detector $\mathcal{D}$.

Training:

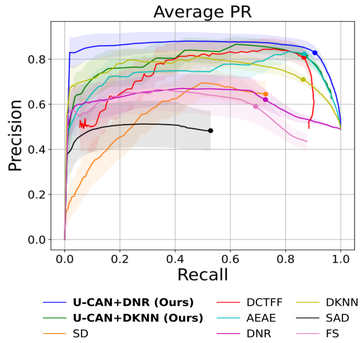
1. Initialize auxiliary blocks $\{A_k\}_{k=1}^{N}$.
2. Attach the auxiliary blocks to $\mathcal{M}$ and train them jointly using the L_{global}.
3. Compute validation $CS_k^{(avg)}$ per A_k ; select the best S auxiliary blocks $\{A_s \mid s \in S\}$.

Inference:

for test sample x **do:**
1. Feed x through $\mathcal{M}$.
2. Extract embeddings from $\{A_s\}_{s\in S}$.
3. Flag x as adversarial or benign.

Experiments and Results

We evaluate **U-CAN** on CIFAR-10, Mammals, and an ImageNet subset, with ResNet-50, VGG-16, and ViT-B-16 backbones, under PGD, C&W, and AutoAttack standard attacks and ADA-DKNN, ADA-AEAE, ADA-DCTFF and ADA-SD adaptive attacks; F1 is taken at the PR-curve–maximizing threshold. The following adversarial detection methods are compared: FS, SAD, DKNN, DNR, AEAE, DCTFF, SD, and ours–**U-CAN+DKNN** and **U-CAN+DNR** (combined with DKNN or DNR $\mathcal{G}$). We also evaluated the overhead of our method against others, including the inference latency.



Score \ Method	FS	SAD	DKNN	DNR	AEAE	DCTFF	SD	U-CAN +DKNN	U-CAN+ DNR
Latency	3.129	**0.701**	1.459	0.804	1.643	1.436	70.841	1.075	0.836
F1	74.63	48.35	82.09	78.28	88.91	86.40	65.27	87.27 ↑5.2	**90.13** ↑11.8
Adaptive F1	-	-	**71.20**	67.50	66.75	54.83	59.91	-	70.00

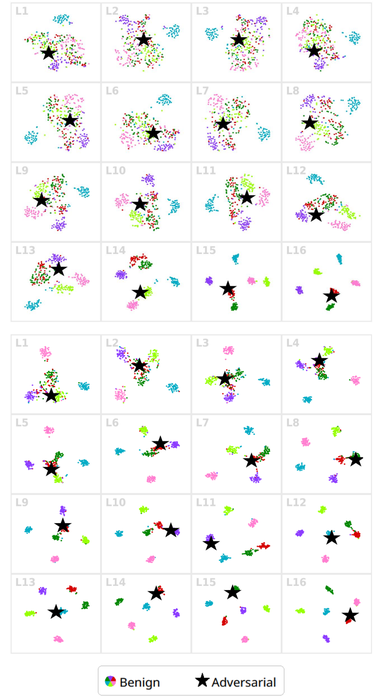
Average results across all dataset/model/attack/ε: top row shows average latency, middle row average F1 on standard attacks, and bottom row average F1 on adaptive attacks (only top-performing novel methods). **Boldface** result indicates the best score, underline the second-best, and **green** arrow/text the gain from stacking detectors on U-CAN.

Summary. Our method **U-CAN+DNR outperforms** all compared methods on average and shows no real difference in the adaptive scenario, with low overhead. Notably, **U-CAN boosts** both DKNN and DNR layer-wise detectors' performance.

Conclusions & Future Work

We introduced **U-CAN**, an unsupervised adversarial detection framework that attaches lightweight ArcFace-based auxiliaries to intermediate layers of a frozen model. Trained only on benign data, **U-CAN** boosts adversarial detection performance by exposing small adversarial shifts, while requiring only negligible overhead. When combined with detectors like DNR or DKNN, **U-CAN** consistently improves F1-scores and stability across all mentioned datasets, architectures and attacks. These results suggest that U-CAN can significantly enhance adversarial detection in safety critical applications.

Future work will focus on exploring alternative aggregation strategies such as utilizing the temporal relations between the layer-wise features, and extending **U-CAN** to additional tasks, modalities and architectures (LLMs, etc.). Overall, **U-CAN** provides a solid foundation for adversarial detection but can be further strengthened through complementary techniques and deeper feature analysis.



Visualizations. Layer-wise t-SNE reduced features of ResNet-50 on the ImageNet validation set. **Top:** raw features $\{z_n\}_1^{16}$. **Bottom:** **U-CAN**'s refined features $\{\tilde{p}_n\}_1^{16}$. From L_0 (top-left) to L_{16} (bottom-right), the plots show the benign clusters (colors) and a single adversarial sample (black star). Without **U-CAN**, adversarial points blend in; with **U-CAN**, refined features sharpen the boundaries, exposing adversarial shifts.