# Discerning the Chaos: Detecting Adversarial Perturbations while Disentangling Intentional from Unintentional Noises - Supplementary

## 1. Additional Dataset

### 1.1. Image Classification using CIFAR-10 Dataset

**Experimental Setup:** The dataset contains 45000 training images, 5000 validation images, and 10000 testing images. The entire training set is used to train for 10-class classification using a Vision Transformer. The classifier is trained for 5 epochs at a learning rate of $1e-4$ with a classification accuracy of $95.89\%$. The embedding dimension used for the CIAI detector is set as $d = 128$ with each image resized to $3 \times 224 \times 224$ and is trained for 3 epochs with a learning rate of $1e-4$. Similarly to the detector trained for the CelebA dataset the layers are frozen and only trained for the added linear layers for a 2-class and 3-class classification for another 3 epochs at a learning rate of $1e-4$.

**Seen and Unseen Noises:** For the CIFAR-10 dataset, the seen adversarial attacks are FGSM, PGD, PGDL2 [8], and CW, while the unseen attacks are Fast FGSM, RFGSM, MIFGSM [3], BIM, UPGD, and DeepFool. For seen unintended noises, Gaussian and salt & pepper noises are used while, for unseen unintended noises, the CIFAR-10-C dataset is used. The dataset gives noises with 5 different severity levels with level 5 as the most severe. Shot and Speckle noises are used as unseen unintended noises here. Except for PGDL2, CW, and DeepFool, which use $L_2$ as the distance measure, all the adversarial attacks used here use $L_\infty$ metric for attacking the image. For creating the FGSM, Fast FGSM, RFGSM, MIFGSM, BIM, PGD, and UPGD perturbations, we use the standard $\epsilon = 8/255$ and $steps = 10$ for all of them except FGSM. Further, for FFGSM, we use $\alpha = 10/255$; for RFGSM, BIM, and UPGD, we use $\alpha = 2/255$; and for PGD, we use $\alpha = 1/255$. Further, for PGDL2 the used parameters are $\epsilon = 1.0$, $\alpha = 0.2$, and $steps = 10$ while for CW the used parameters are $c = 1$ and $steps = 50$.

### 1.2. Detection Results on CIFAR-10 Dataset

We first train the CIAI network on a trained classifier with a classification accuracy of $95.89\%$. For experiments, two instances of the detectors are trained to showcase the efficiency of the proposed method. Further, we compare the results obtained with four high-performing detection methods: **LID [7]** or Local Intrinsic Dimensionality. **Hu [5]** works very well under white-box setting utilizing the high-density, occurring close to the boundary adversarial images for effective detection. **LNG [1]** created a Latent Neighborhood Graph for the attack estimation and detection in different settings. **NNIF [2]** is another nearest neighbor-based approach that pulls helpful/ harmful features from the original and adversarial images to train a detector. A major difference is that these are all detection results off of binary classification, while the results reported for CIAI are for 3-class classification.

$CIAI_1$**:** For, $L_{det}$, the values used for $\beta$, $\gamma$, and $\delta$ as $1/3$, $1/3$, and $1/3$ and the regularization value $\alpha = 0.3$. The trained classifier is used to create adversarially attacked images for FGSM and PGD attacks. Gaussian and Salt & Pepper noise is used for unintentional noises. After stage 1, the tSNE plot for all the five variations of images is plotted as seen in Figure 1(a). The detection results are further depicted in Table 1 as well as Table 2. For Table 2, the evaluation is made on the test set of Gaussian and Salt & Pepper noises as well as corruptions from the CIFAR-10-C dataset with different levels of severities. The detection network is trained for a 3-class classification, with 3 classes as original images, images modified intentionally, and images modified unintentionally. The detection results, especially for FGSM and PGD, are almost completely perfect and better than high-performing detection techniques. The other $L_\infty$-metric-based attacks are also detected with almost perfect accuracy. The $L_2$ attacks are not detected with as high accuracy.

$CIAI_2$**:** For, $L_{det}$, the values used for $\beta$, $\gamma$, and $\delta$ as $0.8$, $0.15$, and $0.05$ and the regularization value $\alpha = 0.3$. For this variation, we use two different sets of adversarial attacks The tSNE plot for all the five variations of images is plotted as seen in Figure 1(b). The detection results are further depicted in Table 1 where the detector is trained for a 3-class classification. While the $L_\infty$ attacks are detected with almost perfect accuracy, the $L_2$ attacks are also detected with comparable accuracy. it can be seen in the tSNE

| | Org | FGSM | PGD | FFGSM | RFGSM | MIFGSM | BIM | UPGD | CW | PGDL2 | DeepFool |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cl acc | 95.89 | 15.59 | 2.35 | 16.06 | 3.72 | 3.76 | 1.40 | 1.28 | 6.67 | 14.84 | 0.62 |
| LID [7] | - | 73.56 | 67.95 | - | - | - | - | - | 55.60 | - | - |
| Hu [5] | - | 84.44 | 58.55 | - | - | - | - | - | 90.99 | - | - |
| LNG [1] | - | 99.88 | 91.39 | - | - | - | - | - | 89.74 | - | - |
| NNIF [2] | - | 87.75 | 98.31 | - | - | - | - | - | **98.98** | - | **97.98** |
| $CIAI_1$ (Ours) | 99.64 | **99.97** | **99.06** | 99.98 | 99.89 | 99.99 | 99.90 | 99.92 | 6.68 | 6.25 | 44.49 |
| $CIAI_2$ (Ours) | 90.90 | **100.0** | **99.86** | 99.99 | 100.0 | 100.0 | 100.0 | 100.0 | 85.90 | 88.56 | 81.52 |

Table 1. Detection results for proposed network, CIAI in two different instances for CIFAR-10 dataset. $CIAI_1$ is trained using FGSM and PGD as adversarial attacks and Gaussian and Salt & Pepper noise as unintentional noises. While $CIAI_2$ is trained using FGSM and PGD as one adversarial group and CW and PGDL2 as another group. (Cl acc is the classification accuracy for the task of image classification)

| | Gaussian | SP | Gaussian Noise | | | | | Impulse Noise (SP) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Severity | | | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Cl acc | 93.05 | 91.69 | 86.93 | 76.89 | 65.74 | 59.51 | 54.52 | 89.23 | 83.94 | 79.35 | 68.10 | 57.61 |
| $CIAI_1$ (Ours) | 94.32 | 99.97 | 96.79 | 99.03 | 97.45 | 96.12 | 93.96 | 99.99 | 100.0 | 100.0 | 99.94 | 99.59 |
| | | | Shot Noise | | | | | Speckle Noise | | | | |
| Severity | | | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Cl acc | - | - | 90.98 | 85.89 | 72.90 | 67.64 | 59.30 | 91.35 | 82.74 | 78.59 | 69.56 | 61.18 |
| $CIAI_1$ (Ours) | - | - | 39.79 | 92.54 | 98.45 | 97.74 | 96.68 | 32.41 | 90.19 | 95.41 | 97.30 | 98.45 |

Table 2. Detection Results for $CIAI_1$ which is trained using FGSM and PGD as adversarial attacks and Gaussian and Salt & Pepper noise as unintentional noises. The results are demonstrated on noises from the CIFAR-10-C dataset. (Cl acc is the classification accuracy for the task of image classification)

| | Org | FGSM | PGD | Gaussian | SP |
|---|---|---|---|---|---|
| Cl acc | 95.89 | 15.59 | 2.35 | 93.05 | 91.69 |
| CIAI | 99.64 | 100.0 | 100.0 | 98.62 | 100.0 |
| | | FFGSM | RFGSM | BIM | UPGD |
| Cl acc | | 16.06 | 3.72 | 1.40 | 1.28 |
| CIAI | | 100.0 | 100.0 | 100.0 | 100.0 |

Table 3. Classification accuracy and detection results on the CIFAR dataset for a 2-class classification setting.

plot as well.

### 1.3. Observations and Attention Maps

For intentional and unintentional noises, we find that detecting the latter is easier even when they are closer to the distribution for the original images. Moreover, it is trickier to translate the detection to the unseen unintended noises over unseen intended noises from a similar group of attacks.

**2-class Classification:** The results reported in Tables 1, 2, and 3 of the main paper are for a 3-class classification setting. The classification is between original images, intended noises, and unintended noises. Table 3 reports results for a 2-class classification, differentiating between original and modified images, on the CIFAR dataset. Training the CIAI detector in stage 1 is done similarly to the 3-class classification setting. For stage 2, the detector is trained between 2 classes, original and modified. The table report the classification accuracy for the classifier trained on the recognition task as well as the detection accuracy between the two classes. We report the accuracy separately for all different noises. Standard cross-entropy loss is used for the training with 4 modified images for every original/ unmodified image. The entire training set is used for the training, and the results here are reported on the testing set of the respective datasets.

**Use of Vision Transformers and Generalizability;** In the literature, Transformers have been known to generalize better than CNNs. We performed experiments on ConvNeXt [6], a CNN-based architecture. The detector was not able to detect the images as accurately as the Transformer architecture, often overfitting on the training dataset. The detection accuracy is less than 90% for each set - original as well as modified images. We, therefore, only use Vision Transformer for the detector network.

## 2. Formulation of Loss

**Center Loss:** Center loss helps with separating classes by maximizing the intra-class distance and minimizing inter-class distance. It has shown impressive performance, especially in face recognition models [9]. Based on this, center

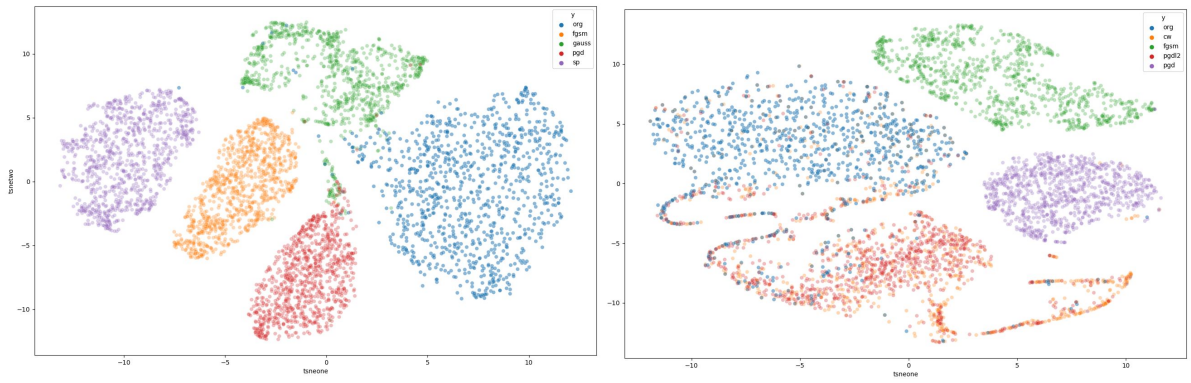


Figure 1. tSNE Plot for CIAI Detector trained on CIFAR dataset, (a) Trained with FGSM and PGD as adversarial attacks and Gaussian Noise and Salt & Pepper Noise as unintentional noises (b) Trained with FGSM and PGD as one family of adversarial attacks and CW and PGDL2 as another family of adversarial attacks.

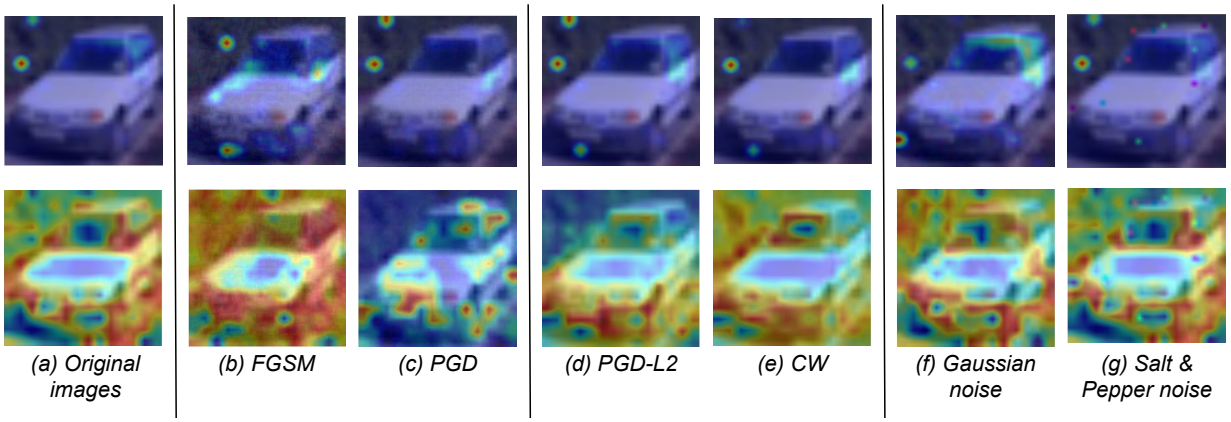


Figure 2. Attention Maps for the CIFAR-10 dataset. The top row indicates the attention maps for the 10-class classification task, and the second row indicates the maps for the detection task. For FGSM, PGD, Gaussian noise, and salt & pepper noise, $CIAI_1$ is used, while for CW and PGLD2, $CIAI_2$ is used for creating attention maps for detection.

loss was picked to not just differentiate between natural and attacked images but also between intentionally attacked and unintentionally attacked images. It helps with formulating as many classes as required during detection, which helps with our 2, 3, and 5-class setting during the detection phase.

**MMD Loss:** With a deep kernel, MMD loss has been shown to be aware of adversarial attacks [4]. Empirically also, MMD performed better over other metrics for measuring distance between distributions like KL divergence.

**Limitations and Applications:** The proposed method is untested for generative attacks and it is also architecture-dependent. Attacks formulated from a transformer-based classifier cannot be detected with as high accuracy on a CIAI detector trained on samples from a convolution-based classifier and vice-versa. By differentiating between benign and malicious samples, we can better develop defense strategies and better understand system behavior. It can also help in recognizing tampered samples, helping to establish the authenticity of samples, a persistent issue in the Internet world today.

## References

[1] A. Abusnaina, Y. Wu, S. S. Arora, Y. Wang, F. Wang, H. Yang, and D. Mohaisen. Adversarial example detection using latent neighborhood graph. In *ICCV*, pages 7667–7676, 2021.

[2] G. Cohen, G. Sapiro, and R. Giryes. Detecting adversarial samples using influence functions and nearest neighbors. In *IEEE/CVF CVPR*, pages 14441–14450, 2020.

[3] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *IEEE/CVF CVPR*, pages 9185–9193, 2018.

[4] R. Gao, F. Liu, J. Zhang, B. Han, T. Liu, G. Niu, and M. Sugiyama. Maximum mean discrepancy test is aware of

adversarial attacks. In *ICML*, pages 3564–3575, 2021.

[5] S. Hu, T. Yu, C. Guo, W. Chao, and K. Q. Weinberger. A new defense against adversarial images: Turning a weakness into a strength. In *NeurIPS*, pages 1633–1644, 2019.

[6] Liu et al. A convnet for the 2020s. In *IEEE CVPR*, 2022.

[7] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. N. R. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *ICLR*, 2018.

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[9] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *ECCV*, volume 9911 of *Lecture Notes in Computer Science*, pages 499–515. Springer, 2016.