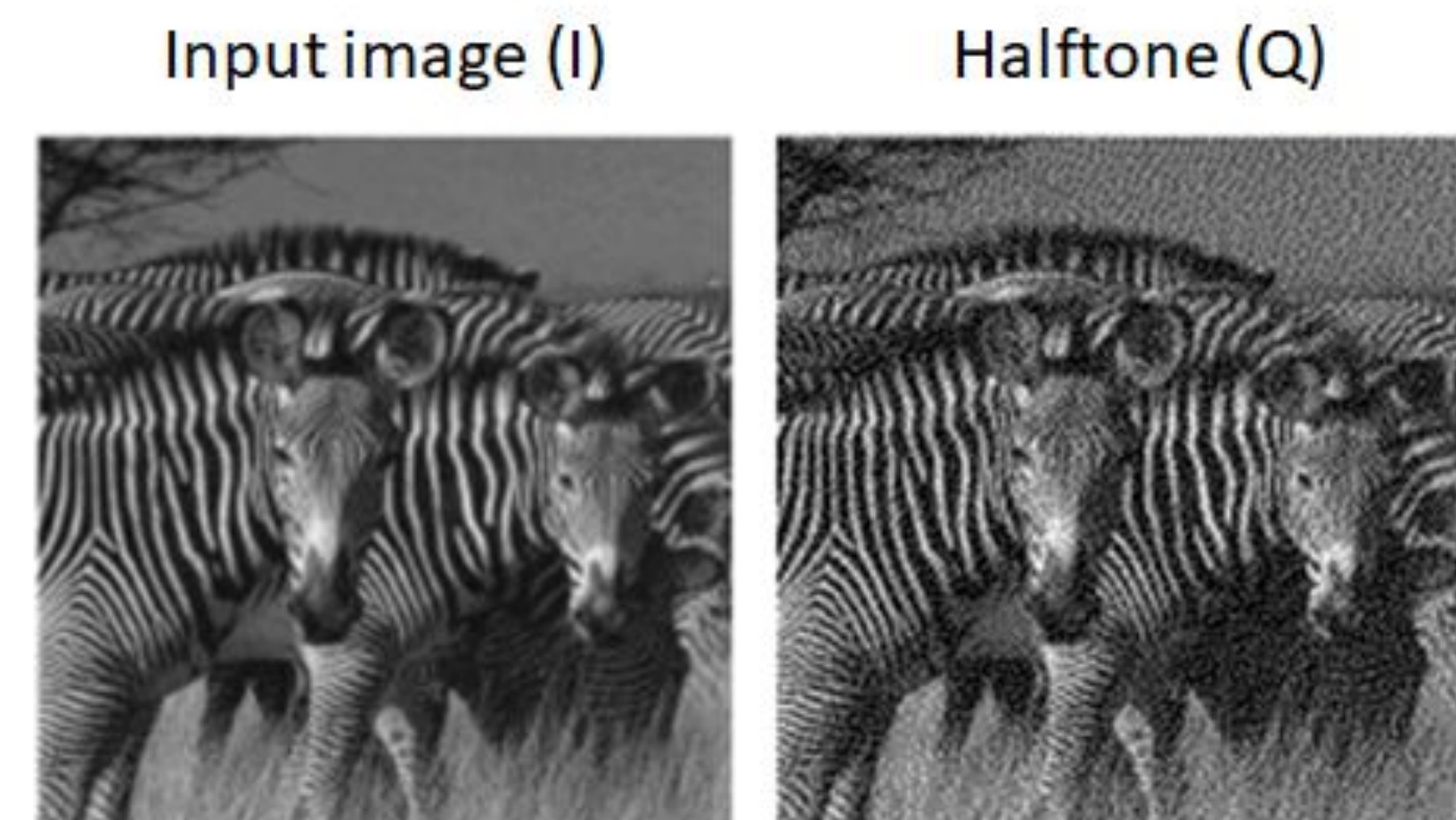


## Introduction

- Although image transformation-based defenses were widely considered at an earlier time, most of them have been defeated by adaptive attacks.
- We propose a new image transformation defense based on error diffusion halftoning, and combine it with adversarial training to defend against adversarial examples.
- Error diffusion halftoning projects an image into a 1-bit space and diffuses quantization error to neighboring pixels
- This process can remove adversarial perturbations from a given image while maintaining acceptable image quality in the meantime in favor of recognition.
- The proposed method can improve adversarial robustness even under advanced adaptive attacks, while most of the other image transformation-based defenses do not.



## Prior Works

- JPEG compression
- Bit-depth reduction
- Image denoising
  - o Gaussian blur
  - o Mean/median filter
  - o Non-local means
- ...etc [1]
- Most existing image transformation-based defenses are NOT robust against white-box attacks [2].

Defense	Dataset	Distance	Accuracy
Buckman et al. (2018)	CIFAR	0.031 ( $\ell_\infty$ )	0%*
Ma et al. (2018)	CIFAR	0.031 ( $\ell_\infty$ )	5%
Guo et al. (2018)	ImageNet	0.005 ( $\ell_2$ )	0%*
Dhillon et al. (2018)	CIFAR	0.031 ( $\ell_\infty$ )	0%
Xie et al. (2018)	ImageNet	0.031 ( $\ell_\infty$ )	0%*
Song et al. (2018)	CIFAR	0.031 ( $\ell_\infty$ )	9%*
Samangouei et al. (2018)	MNIST	0.005 ( $\ell_2$ )	55%**

## Proposed Method

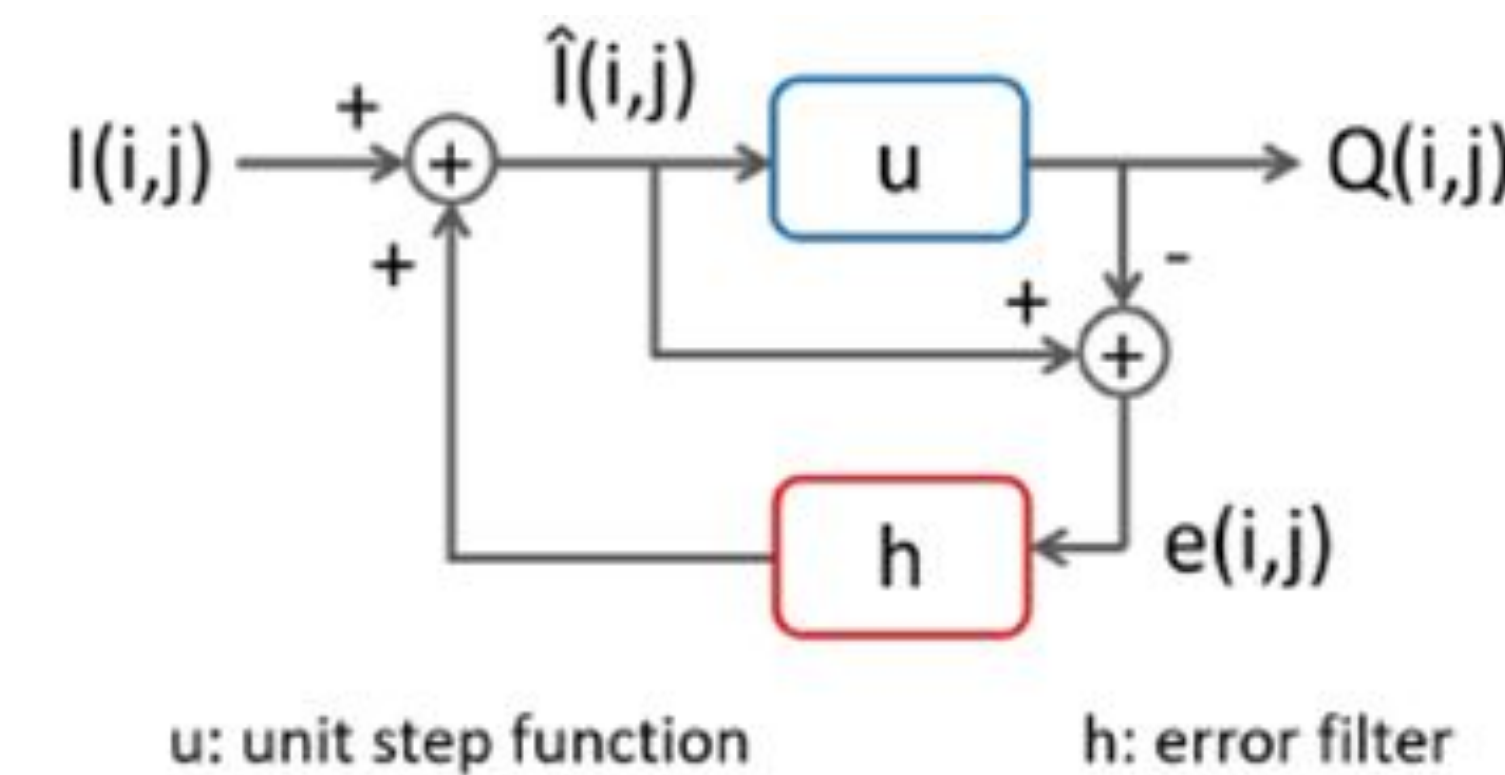
### Error diffusion halftoning: Floyd-Steinberg dithering

- Quantize each pixel in the raster order (from left to right, top to bottom) one-by-one, and spread the quantization error to the neighboring pixels.
- Beginning with the top-left pixel, the pixel value is binarized by thresholding, then the quantization error is dispersed to neighboring pixels using pre-defined weights.
- Following the raster-scan indexing scheme, the procedure continues until the bottom-right pixel has been transformed.

$$\hat{I}(i, j) = I(i, j) + \sum_{m, n \in S} h(m, n)e(i - m, j - n)$$

$$Q(i, j) = u(\hat{I}(i, j) - \theta)$$

$$e(i, j) = \hat{I}(i, j) - Q(i, j)$$



### Algorithm 1: Floyd-Steinberg dithering

**Result:** Output halftone  $Q$   
Given an input image  $I$  with pixel values  $\in [0, 1]$ ,

```

for i from top to bottom do
  for j from left to right do
    oldValue = I[i][j]
    if oldValue > 0.5 then
      | newValue = 1
    else
      | newValue = 0
    end
    Q[i][j] = newValue
    error = oldValue - newValue
    I[i + 1][j] += error * 7/16
    I[i - 1][j + 1] += error * 3/16
    I[i][j + 1] += error * 5/16
    I[i + 1][j + 1] += error * 1/16
  end
end

```

- The **quantization operation** invalid the adversarial variations.
- **Updating the values of the neighboring pixels repeatedly** makes the adaptive attacks hard to identify the mapping between the original image and the corresponding halftone.
- **Spreading quantization errors** produces better halftoning quality and tends to enhance edges and object boundary in an image.
- Take **both** adversarial robustness and clean data performance.
- Complementary to adversarial training.

## References

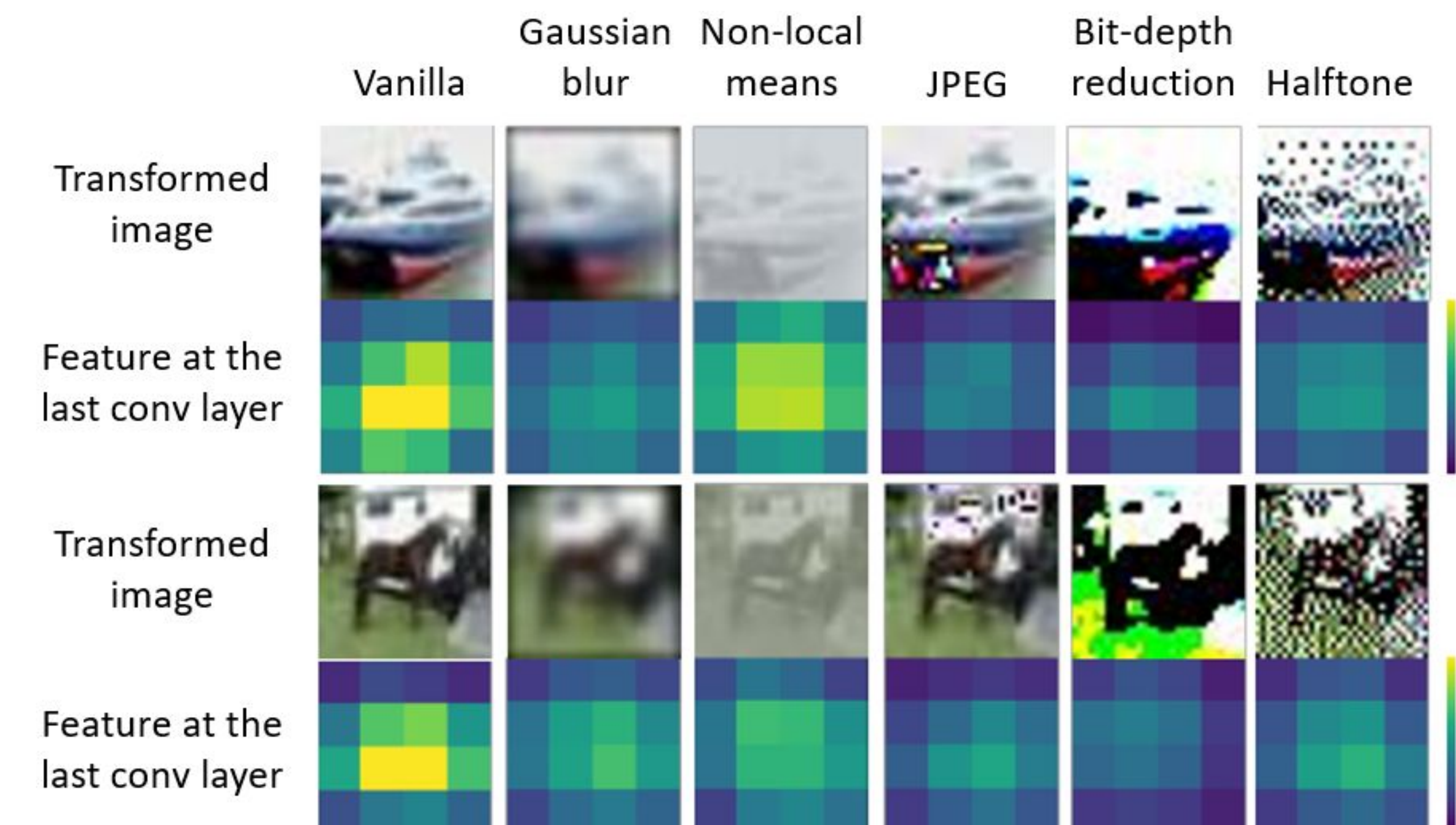
- [1] E. Raff, J. Sylvester, S. Forsyth, and M. McLean, "Barrage of random transforms for adversarially robust defense," in IEEE Conference on Computer Vision and Pattern Recognition, 2019
- [2] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in International Conference on Machine Learning, 2018.

## Results

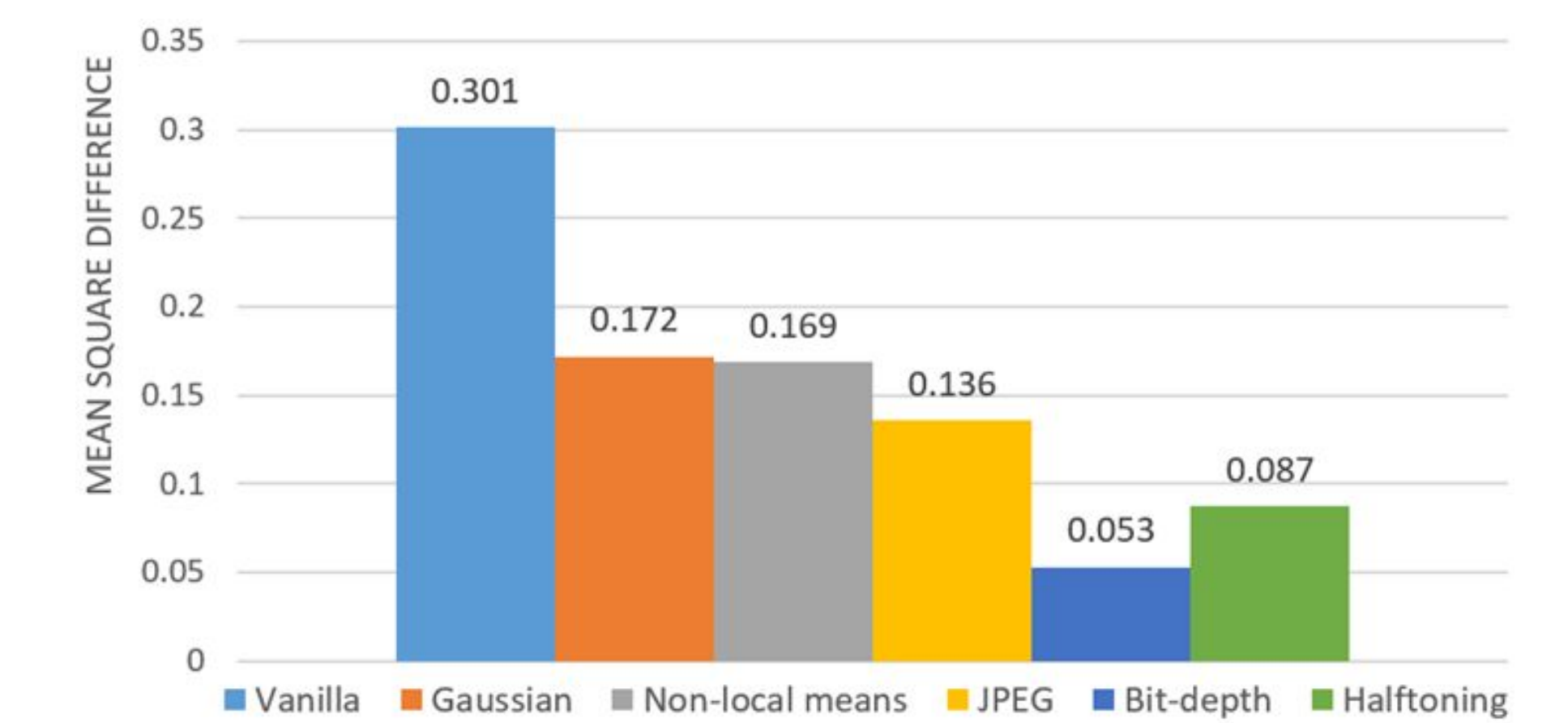
### Quantitative Results

Method	Training	Clean	PGD- $\ell_\infty$	PGD- $\ell_2$	Mult- $\ell_\infty$	Mult- $\ell_2$	Avg <sub>adv</sub>	Avg <sub>all</sub>
Vanilla		94.03	0.01	0.20	0.05	0.01	0.07	18.86
Gaussian blur	Standard training	90.17	0.20	1.34	0.17	0.05	0.44	18.39
Non-local means		88.66	0.02	0.49	0.03	0.00	0.14	17.84
JPEG compression		90.06	2.97	4.82	1.81	0.22	2.46	19.98
Bit-depth reduction		78.87	15.26	10.84	10.79	4.52	10.35	24.06
Halftoning (ours)		88.57	9.53	11.98	5.54	1.07	7.03	23.34
Vanilla	Adversarial training	83.31	51.15	50.68	54.10	40.29	49.06	55.91
Gaussian blur		75.96	44.59	47.12	45.07	32.48	42.32	49.04
Non-local means		75.47	44.67	45.29	16.59	14.53	30.27	39.31
JPEG compression		24.97	38.99	43.72	59.15	44.72	46.65	42.31
Bit-depth reduction		71.66	47.34	42.40	48.50	41.63	44.97	50.31
Halftoning (ours)	84.37	60.01	56.56	67.37	88.44	68.10	71.35	

### Feature Visualization



### Feature Analysis



## Acknowledgement

This work was supported by the DARPA GARD Program HR001119S0026-GARD-FP-052

## Code

htt [https:// github.com/shaoyuanlo/Halftoning-Defense](https://github.com/shaoyuanlo/Halftoning-Defense)