PRINCETON
COMPUTATIONAL IMAGING LAB

ECCV TEL AVIV 2022

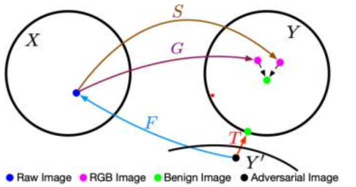
All You Need is RAW: Defending Against Adversarial Attacks with Camera Image Pipelines

Yuxuan Zhang, Bo Dong, Felix Heide

Motivation

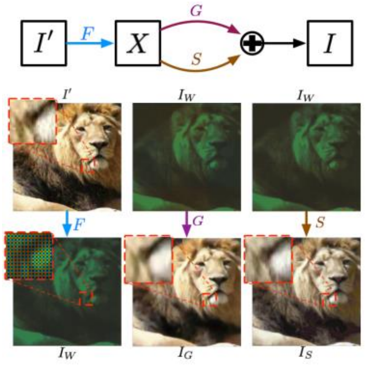
Existing neural networks for computer vision tasks are vulnerable to adversarial attacks: adding imperceptible perturbations to the input images can fool these methods to make a false prediction on an image that was correctly predicted without the perturbation. Various defense methods have proposed image-to-image mapping methods, either including these perturbations in the training process or removing them in a preprocessing denoising step. In doing so, existing methods often ignore that the natural RGB images in today's datasets are not captured but, in fact, recovered from RAW color filter array captures that are subject to various degradations in the capture. In this work, we exploit this RAW data distribution as an empirical prior for adversarial defense. Specifically, we proposed a model-agnostic adversarial defensive method, which maps the input RGB images to Bayer RAW space and back to output RGB using a learned camera image signal processing (ISP) pipeline to eliminate potential adversarial patterns. The proposed method acts as an off-the-shelf preprocessing module and, unlike model-specific adversarial training methods, does not require adversarial images to train. As a result, the method generalizes to unseen tasks without additional retraining. Experiments on large-scale datasets (e.g., ImageNet, COCO) for different vision tasks (e.g., classification, semantic segmentation, object detection) validate that the method significantly outperforms existing methods across task domains

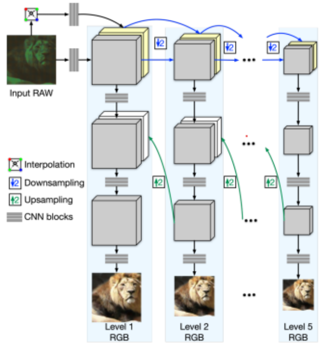
Method Overview



Existing defense approaches learn an RGB-to-RGB projection from an adversarial distribution (Y) to its natural RGB distribution (Y'): T : Y' → Y . In contrast, our approach learns a mapping via the intermediate natural RAW distribution (X), which is achieved by utilizing three specially designed operators : F : Y' → X, G : X → Y , and S : X → Y.

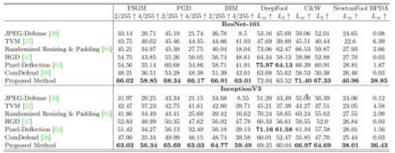
Model Details





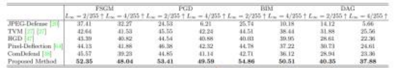
Departing from existing methods, we learn a mapping from Y' to Y via an intermediate RAW distribution, X, which incorporates these RAW statistics of natural images, such as sensor photon counts, multispectral color filter array distributions and optical aberrations. To this end, the approach leverages three specially designed operators: F : Y' → X, G : X → Y , and S : X → Y. Specifically, the F operator is a learned model, which maps an adversarial sample from its adversarial distribution to its corresponding RAW sample in the natural image distribution of RAW images. Operator G is another learned network that performs an ISP reconstruction task, i.e., it converts a RAW image to an RGB image. In theory, our goal can be achieved with these two operators by concatenating both G(F(·)) : Y' → X → Y . However, as these two operators are differentiable models, the potential adversary may still be able to attack the model if, under stronger attack assumptions, he has full access to the weight of preprocessing modules. To address this issue, we add the operator S, a conventional ISP, to our approach, which is implemented as a sequence of cascaded software-based sub-modules. In contrast to operator F, operator S is non-differentiable. For defending against an attack, the proposed approach first uses the F operator to map an input adversarial image, I', to its intermediate RAW measurements, IW . Then, IW is processed separately by the G and S operators to convert it to two images in the natural RGB distribution, denoted as as IG and IS, respectively. Finally, our method outputs a benign image, I, in the natural RGB distribution by combining I_G and I_S in a weighted-sum manner.

Quantitative Result



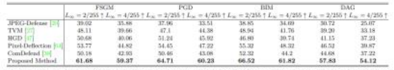
	FSGM		PGD		BIM		DeepFool		C&W		NewtonFool	BPDA
	2/255 ↑	4/255 ↑	2/255 ↑	4/255 ↑	2/255 ↑	4/255 ↑	L_∞ ↑	L_2 ↑	L_∞ ↑	L_2 ↑	L_∞ ↑	L_∞ ↑
ResNet-101												
JPEG-Defense [20]	33.14	20.71	45.19	21.74	36.78	8.5	53.16	45.69	59.06	52.01	24.65	0.08
TVM [27]	43.75	40.02	45.46	44.35	44.86	41.93	47.69	39.89	45.51	40.44	22.6	6.39
Randomized Resizing & Padding [81]	45.21	34.97	45.38	27.75	40.04	18.04	73.06	62.47	66.53	59.87	27.93	2.66
HGD [47]	54.75	43.85	55.26	50.05	56.74	48.61	64.34	58.13	59.98	52.88	27.70	0.03
Pixel-Deflection [64]	54.56	35.14	60.68	34.86	58.71	41.91	**75.97**	**64.13**	66.29	60.91	28.81	1.87
ComDefend [38]	48.21	36.51	53.28	48.38	51.39	42.01	63.68	55.62	58.53	50.38	26.46	0.03
Proposed Method	**66.02**	**58.85**	**68.34**	**66.17**	**66.91**	**63.01**	72.04	63.52	**71.40**	**67.33**	**40.96**	**38.85**
InceptionV3												
JPEG-Defense [20]	31.97	20.25	43.34	21.15	34.68	8.55	51.20	43.49	55.00	50.39	24.06	0.12
TVM [27]	42.47	37.23	42.75	41.61	42.80	39.71	45.21	37.39	43.27	37.51	23.05	4.58
Randomized Resizing & Padding [81]	41.86	34.49	43.41	25.60	39.42	16.62	70.24	58.65	63.24	55.62	27.55	2.09
HGD [47]	52.83	40.99	50.35	47.62	56.02	47.78	60.33	56.61	59.55	52.0	26.84	0.03
Pixel-Deflection [64]	51.42	34.27	56.13	32.49	56.18	39.13	**71.16**	**61.58**	61.94	57.58	28.01	1.56
ComDefend [38]	47.00	35.34	49.99	46.15	48.74	39.58	60.01	52.47	55.85	47.70	25.44	0.03
Proposed Method	**63.03**	**56.34**	**65.69**	**63.03**	**64.77**	**59.49**	69.25	60.04	**66.97**	**64.69**	**38.01**	**36.43**

Table 1: **Quantitative Comparisons on ImageNet** We evaluate Top-1 Accuracy on ImageNet and compare the proposed method to existing input-transformation methods. The best Top-1 accuracies are marked in bold. Our defense method offers the best performance in all settings, except for the DeepFool attack.



	FSGM		PGD		BIM		DAG	
	$L_\infty = 2/255$ ↑	$L_\infty = 4/255$ ↑	$L_\infty = 2/255$ ↑	$L_\infty = 4/255$ ↑	$L_\infty = 2/255$ ↑	$L_\infty = 4/255$ ↑	$L_\infty = 2/255$ ↑	$L_\infty = 4/255$ ↑
JPEG-Defense [20]	37.41	32.27	24.53	6.21	25.74	10.18	14.12	5.66
TVM [27] [27]	42.64	41.53	45.55	42.24	44.51	38.44	31.88	25.56
HGD [47]	43.39	40.82	44.54	40.88	40.03	39.95	28.61	22.36
Pixel-Deflection [64]	44.13	41.88	46.38	42.32	44.78	37.22	30.73	24.61
ComDefend [38]	45.57	39.23	44.85	41.14	42.71	36.12	28.94	23.36
Proposed Method	**52.35**	**48.04**	**53.41**	**49.59**	**54.86**	**50.51**	**40.35**	**37.88**

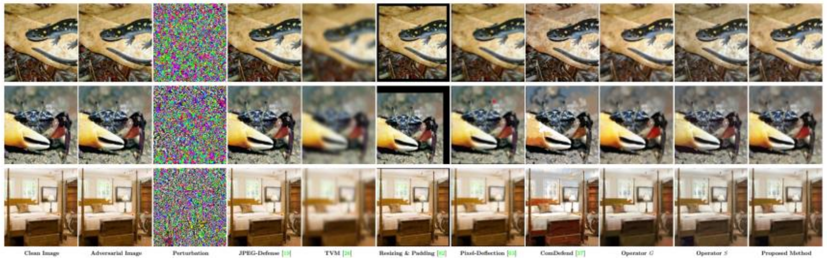
Table 2: **Quantitative Comparison to SOTA Input-Transformation Defense Methods on the COCO dataset.** We evaluate all methods on mean IoU (mIoU) and mark the best mIoU in bold. Our defense method offers the best performance in all settings.



	FSGM		PGD		BIM		DAG	
	$L_\infty = 2/255$ ↑	$L_\infty = 4/255$ ↑	$L_\infty = 2/255$ ↑	$L_\infty = 4/255$ ↑	$L_\infty = 2/255$ ↑	$L_\infty = 4/255$ ↑	$L_\infty = 2/255$ ↑	$L_\infty = 4/255$ ↑
JPEG-Defense [20]	39.02	35.88	37.96	33.51	38.85	34.69	30.72	25.07
TVM [27]	48.11	39.66	47.1	44.38	48.94	41.76	39.20	33.18
HGD [47]	50.68	40.06	51.24	45.92	46.80	39.74	41.15	37.23
Pixel-Deflection [64]	53.77	44.82	54.45	47.22	55.32	48.32	46.52	39.87
ComDefend [38]	50.18	42.93	50.46	43.08	52.32	44.2	44.68	37.22
Proposed Method	**61.68**	**59.37**	**64.71**	**60.23**	**66.52**	**61.82**	**57.83**	**54.12**

Table 3: **Quantitative Comparison to SOTA Input-Transformation Defenses on the Pascal VOC dataset**. We evaluate all compared methods for mean average precision (mAP) on Pascal VOC dataset. The best mAP are marked in bold. Our defense method offers the best performance in all settings.

Qualitative Result



Qualitative Result: Comparison of outputs of the proposed method along with both G and S operators and state-of-the-art defense methods on the ImageNet dataset.

Summary

- We propose, to the best of our knowledge, the first adversarial defense method that exploits the natural distribution of RAW domain images.
- The proposed method avoids the tedious generation of adversarial training images and can be used as an off-the-shelf preprocessing module for diverse tasks.
- We validate that the method achieves state-of-the-art defense accuracy for input transformation defenses, outperforming existing approaches