Adversarial Imaging Pipelines
Supplemental Material

Buu Phan[1] Fahim Mannan[1] Felix Heide[1,2]
[1]Algolux [2]Princeton University

1. Experimental Image Processing Pipelines and Optics

Modern digital photography pipelines transform an incoming light field from a physical scene into an RGB image via three sub-modules: a compound optics module, a sensor and an image processing pipeline (ISP). The compound optics contains a set of lenses that focuses the light field onto a color filter array (CFA) in the sensor. Each photosensor on the CFA, depends on the type R(ed), G(reen) and B(lue), reads a specific color value by converting the photons into electrons, resulting a RAW image. In our experiments, we use a FLIR Blackfly S camera employing a Sony IMX249 sensor, which follows the Bayer CFA pattern. This sensor has a resolution of 1200×1920, and each pixel has a value of 12 bit ADC scaled to 16 bit. The image processing pipeline, through a series of operations, transform the RAW images into the RGB ones.

1.1. Image Signal Processor Pipelines

Image processing pipelines contain a sequence of operations to reconstruct the RGB images from RAW measurements. Typically, an ISP starts with the white-balance operation, which is responsible for adjusting the RAW measurements such that white objects will appear as white in the output image. The white-balanced RAW image is then going through the demosaicking operation, which constructs an initial RGB image. A set of image denoising operations are then applied to this RGB image, whose color is then adjusted by global operations in the color and tone correction stage. Finally, it is compressed into other formats such as JPEG and PNG.

In the following, we describe the ISPs that we used in our experiments. We used two black-box/non-differentiable hardware ISPs, the ARM Mali C71 and Movidius Myriad 2 ISPs. In addition to the two hardware ISPs, we also jointly evaluate two (differentiable) software ISPs. The former only performs bilinear demosaicing and will be referred to as the Demosaicing ISP. The latter performs bilinear demosaicing operation followed by bilateral filtering [7] and is referred to as Bilateral Filter ISP.

Demosaicing ISP. Given a RAW image $x \in \mathbb{R}^{m\times n}$, where x_{ij} is the value at row i and column j (zero-index), and white-balance coefficients $[r, g_1, g_2, b]$, the Demosaicing ISP first apply white-balancing function $w : \mathbb{R}^{m\times n} \rightarrow \mathbb{R}^{m\times n}$ on this image as follows for each pixels:

$$x_{wb}^{ij} = w^{ij}(x, [r, g_1, g_2, b]) = \begin{cases} r \cdot x^{ij}, & \text{where } i, j \bmod 2 = 0 \\ g_1 \cdot x^{ij}, & \text{where } i \bmod 2 = 0, j \bmod 2 = 1 \\ g_2 \cdot x^{ij}, & \text{where } i \bmod 2 = 1, j \bmod 2 = 0 \\ b \cdot x^{ij}, & \text{where } i, j \bmod 2 = 1 \end{cases} \tag{1}$$

where x_{wb} is the white-balanced RAW image. Then, we get the demosaiced image x_D by applying the bilinear demosaicing function $k : \mathbb{R}^{m\times n} \rightarrow \mathbb{R}^{m\times n\times 3}$ on x_{wb}:

$$x_D[R] = x_{wb,R} * \begin{bmatrix} 0.25 & 0.5 & 0.25 \\ 0.5 & 1.0 & 0.5 \\ 0.25 & 0.5 & 0.25 \end{bmatrix} ; x_D[G] = x_{wb,G} * \begin{bmatrix} 0.0 & 0.25 & 0.0 \\ 0.25 & 1.0 & 0.25 \\ 0.0 & 0.25 & 0.0 \end{bmatrix} ; x_D[B] = x_{wb,B} * \begin{bmatrix} 0.25 & 0.5 & 0.25 \\ 0.5 & 1.0 & 0.5 \\ 0.25 & 0.5 & 0.25 \end{bmatrix} \tag{2}$$

where $x_{wb,R}, x_{wb,G}, x_{wb,B}$ are RGB masked matrices from x_{wb}, following the Bayer pattern.

Bilateral Filter ISP. The Bilateral Filter ISP follows exactly the same steps as the Demosaicing ISP, with an additional bilateral filter operation [9]:

$$x_F^{ij} = \frac{1}{K} \sum_{u,v} G_{\sigma_s}(\|(u,v) - (i,j)\|) G_{\sigma_r}(\|x_D^{uv} - x_D^{ij}\|) x_D^{ij}, \tag{3}$$

where $K = \sum_{u,v} G_{\sigma_s}(\|(u,v) - (i,j)\|) G_{\sigma_r}(\|x_D^{uv} - x_D^{ij}\|)$, G_{σ_s} is a spatial Gaussian kernel with $\sigma_s = 1.0$ and G_{σ_r} is a range Gaussian kernel with $\sigma_r = 0.138$. Both kernels use a window size of 13×13.

Movidius Myriad 2 ISP. The Movidius Myriad 2 [1] is a high-performance and low-power vision processor from Intel that can achieve a throughput up to 600 megapixels per second. This platform allow the developer to construct a processing graph, which our is presented in Figure 1. Unlike the differentiable ISPs, this pipeline involves multiple complex processing steps [1], involving filter operations in both RAW and RGB domain. The white-balanced RAW image first goes through the sigma denoising, lens shading correction and RAW filter stages, before being demosaiced. Then the pipeline separately process the luma and chroma factors in the image. After that, they are merged together and the final global tonemapping operation is applied.

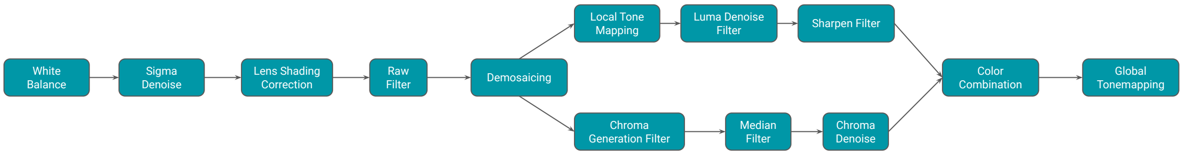


Figure 1: Movidius Myriad 2 ISP pipeline.

ARM C71 Mali. The ARM Mali-C71 ASIC ISP [5] is capable of processing 12 megapixel streams at up to 100 Hz with less than one Watt power consumption. The processing pipeline, as shown in Figure 2, contains 7 stages. Before being demosaiced, the RAW image first goes through the mutli-exposure fusion, RAW HDR processing, noise reduction, and vignetting and tonemapping stages. The noise reduction blocks contains a complex filtering algorithm that filters the noises while preserving the edges and texture. In this stage, chromatic abberration correction algorithm is also applied. The RAW image is then fed into the vignetting and tonemapping stage, which includes radial and mesh shading corrections. The demosaiced image finally goes through the HDR tonemapping and color correction stages.

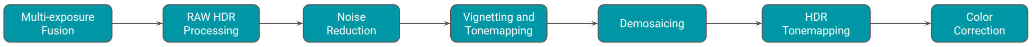


Figure 2: ARM C71 Mali ISP pipeline.

1.2. Optics

Next, we describe in detail the optics that we used in our experiments. Each lens-assembly is focused at infinity with the screen beyond the hyperfocal distance.

Fujinon CF12.5HA-1. We use the RAW image data captured using a Fujinon CF12.5HA-1 lens for our proxy ISP experiments. According to the manufacture, the lens is made of advanced glass composites and anti-reflective coatings, which enable it to produce high contrast, detailed images with resolution up to 1.5 megapixels. The field of view (FOV) is 54° on 1” CCD systems. The optical system contains a 1” 12.5mm focal length, which can be adjusted manually. The manual iris range is from F1.4 to F22.

Cooke Triplet. We use a Cooke triplet design [4] with a manually optimized starting point and optimize it further for spot size and Modulation Transfer Function (MTF) using Zemax’s Hammer optimization. The compound optics are fabricated using a positive polymethyl methacrylate (PMMA) as lens material. Additionally, antireflective material is also coating the lens. The lens has a 25° FOV, an effective focal length of 25mm and an aperture size of 5 mm. We note that PPMA has a refractive index of 1.493 at 550 nm wavelength. The design is optimized for the Sony IMX249 sensor. We will release the full Zemax design files for reproducibility.

2. Network Architectures

In this section, we explain the architecture and training details of the proxy network we used in our experiments. In both settings, $h(x)$ is a black-box function (ISP or optics and ISP) that maps an input x (RAW or RGB) to a post-processed RGB image.

Proxy ISPs. The proxy ISP function $\tilde{h}_\theta : \mathbb{R}^d \to \mathbb{R}^{d\times 3}$ depends on θ as learnable parameters (*i.e*. CNN weights) also maps a RAW image to a post-ISP image. We use bilinear demosaicing as a first layer in this proxy module. Then the demosaiced RGB image is fed into a U-Net [8]. Our U-Net version uses 32 channels in the first convolutional operation, with 3 downsampling steps where each step we double the convolutional channel, as in the original version. We use ReLU activation instead of sigmoid for the output layer, since we found this to improve the approximation performance. The value of the output image is then clipped to the range of $[0, 1]$.

Proxy Camera. The proxy function for a full camera maps a scene (displayed on a physical screen) to a post-ISP image, that is $\tilde{h}_\theta : \mathbb{R}^{d\times 3} \to \mathbb{R}^{d\times 3}$. Since the U-Net keeps the output dimension to be equal to the displayed image, which has the dimension of 448×448, we reduce the U-Net output by a factor of 2, using bilinear downsampling. This image is then clipped to the valid range of $[0, 1]$ and fed to the classifier.

Training Loss. Given a set of RAW captures (Proxy ISPs) or display images (Proxy Camera): $X = \{x_1, x_2, ..., x_N\}$, where each $x_i \in \mathbb{R}^d$, we train the proxy function $\tilde{h}_\theta$ by minimizing the ℓ_1 reconstruction loss.

$$\mathcal{L}_{proxy} = \frac{1}{N \times d} \sum_{i=0}^{N} ||h(x_i) - \tilde{h}(x_i)||_1. \tag{4}$$

We note that, unlike Tseng et al. [10], our loss function does not contain the perceptual loss term.

Training Details. For training our proposed local proxy function, we use the same loss on the new augmented dataset. For training the proxy ISP and proxy camera, we train by using a batch size of 16, a learning rate of 0.001, a decay rate of 10^{-1} every 4500 steps , for a total of 13500 steps. The training set size is 4200 images and the validation set size is 500, sampled randomly from the Imagenet dataset. To train the local proxy, we use a batch size of 8 and fix the learning rate to 10^{-5}.

3. Limitations of Query-Based Methods

We note that, theoretically, one can estimate the gradient $\nabla_\delta \mathcal{L}(f(x+\delta), t)$ by using query-based black-box attacks. This approach [2, 6, 3, 11], however, is impractical and not scalable in our scenario since it requires a *few thousands queries per image*, and the capturing and process time is 3 sec (physical ISP attack) and 5 sec (physical ISP and optics attack) per image. Leveraging the advantage of knowing the classifier g is also not trivial. For instance, instead of estimating $\nabla_\delta \mathcal{L}(f(x+\delta), t)$ directly, we can use the chain-rule and estimate $\nabla_\delta h(x+\delta)$, as

$$\nabla_\delta \mathcal{L}(f(x+\delta), t) = \frac{d\mathcal{L}(f(x+\delta), t)}{dh(x+\delta)} \cdot \frac{dh(x+\delta)}{d\delta}. \tag{5}$$

Estimating this is still expensive, if not more than the original problem since the dimension d in δ and $h(x+\delta)$ in the case of Imagenet classifiers is often around or more than 224×224, prohibiting us from effectively estimate the Jacobian gradient.

4. Proxy Approximation Quality

In this section, we first quantitatively validate that the the proposed local proxy outperforms the original version by Tseng *et al.* [10]. Then, we show that the proposed proxy accurately approximates the output of the hardware ISPs.

Effectiveness of Local Proxy. We compare the success rate of the untargeted camera attack between the our proposed local proxy $\hat{h}$ and the original version $\tilde{h}$ from Tseng *et al.* [10] for the Movidius Myriad 2 and the ARM Mali C71 ISP in Figure 3. We observe that our model $\hat{h}$ consistently outperforms the original version, with more than 12% success rate for every ϵ.

Approximation Quality. In Figure 4 we plot the cumulative distribution of the absolute confidence difference between the proxy pipeline and physical pipeline. This shows that the proposed local proxy function is able to approximate the hardware ISP well.

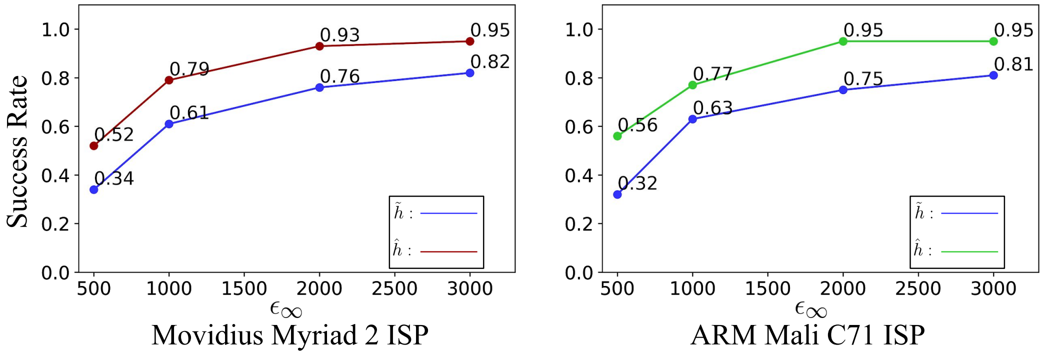


Figure 3: We show and compare the success rate of fooling a physical ISP by using δ predicted by $\tilde{h}$, $\hat{h}$.

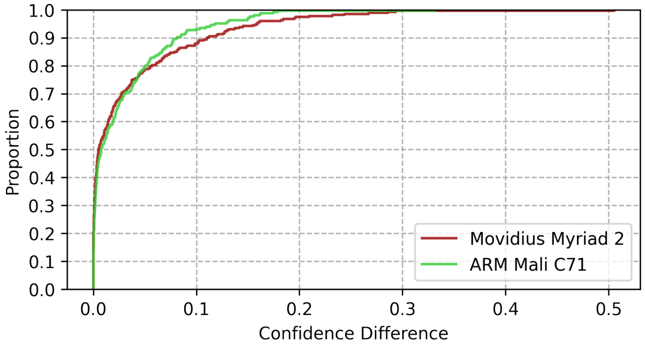


Figure 4: We show here that our local proxies accurately approximate the real physical ISPs. We calculate the absolute difference between the predicted confidence values from the physical and proxy pipeline for each image and plot their cumulative distribution. The plot shows that up to 90% of the images have lower than 0.1 absolute confidence difference.

5. Additional Experimental Results

5.1. Physical ISP Attack

We evaluate the targeted success rate for each individual ISP. The targeted success rate measures if an attack pattern changes the targeted pipeline's prediction to the target class while leaving other camera pipelines unaffected (class prediction does not change and the confidence difference between the original and adversarial RAW is below 0.15). It can be observed that for the hardware ISP, our attack reliably achieves more than 87% targeted success rate. For the differentiable ISPs, the targeted success rate is up to 91%.

Targeted ISPs	Movidius Myriad 2	ARM Mali C71	Bilateral Filter ISP	Demosaicing ISP
Targeted Success Rate	87.6%	88.9%	91.4%	92.1%

Table 1: Targeted success rate for each individual ISP.

5.2. Physical Optics Attack

Untargeted Camera Attack. We report the success and transfer rate for each individual optics in the case of untargeted optics attack in Table 2. We observe that the transfer rate for untargeted optics attacks is high. The attacks on Fujinon CF12.5HA-1, for example, have a transfer rate up to 41.4% when deployed to the Cooke Triplet.

Deployed Optics **Targeted Optics**	Fujinon CF12.5HA-1	Cooke Triplet
Fujinon CF12.5HA-1	**96.7%**	41.4%||53.7%
Cooke Triplet	34.3%||46.1%	**95.9%**

Table 2: Success and transfer rate for the untargeted physical optics attack.

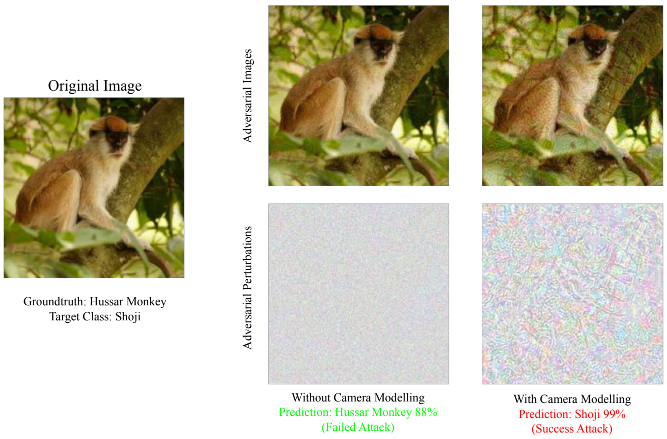


Figure 5: Visualization of attacks with and without camera modelling. The perturbation is more structured and visible when we take the camera transformation into account..

Effectiveness of Camera Modelling in Physical Adversarial Attack. As we stated in the paper, current approaches for adversarial attacks neglect the existence of the transformation happening in the camera, assuming that they preserve the adversarial pattern. We show here that this assumption is flawed by comparing the attack success rate when we include and do not include this transformation, using the Cooke Triplet optics and ARM C71 Mali ISP. When assuming that this transformation preserves the adversarial pattern, the success rate is 0.1% when $\epsilon = 0.08$. In contrast, when this transformation is considered, the attack achieves up to 95% for the same ϵ. We visualize and compare the perturbations in Figure 5, where it can be observed that although both use the same optimization algorithm and perturbation bound $\epsilon = 0.08$, the perturbation when we model the transformation in the camera is more visible and structured.

Targeted Optics Attack. We show the targeted success rate for the targeted optics attack in Table 3. Our attack is able to achieve more than 82% targeted success rate on both Fujinon CF12.5HA-1 and Cooke Triplet.

Targeted Optics	Fujinon CF12.5HA-1	Cooke Triplet
Targeted Success Rate	82.6%	83.9%

Table 3: Targeted success rate for each individual optics.

6. Additional Visualization Results

6.1. Targeted ISP Attacks

In Figure 6,7,8,9,10,11, we show visualizations for the targeted ISP attacks. Each figure shows attacks on different ISPs, using the same original RAW image. Interestingly, despite having the same adversarial RAW image as input, each ISP produces distinct RGB perturbations. Although the RGB perturbations seem to contain similar macro structures, only the one from the targeted ISP becomes adversarial to the classifier, while others pose no threat at all. Since the untargeted RGB perturbations do not change the prediction, this means that they are considered as noise by some hidden projections in the classifier. As such, the structured noise perturbations are tailored to a specific ISP.

6.2. Targeted Optics Attack

We provide additional examples of the targeted optics attack in Figure 12. We find that the perturbations show distinctive frequency-dependent patterns for each optical system. We interpret this attack as one that efficiently exploits the frequency bands specific to the optical transfer functions of the employed optics.

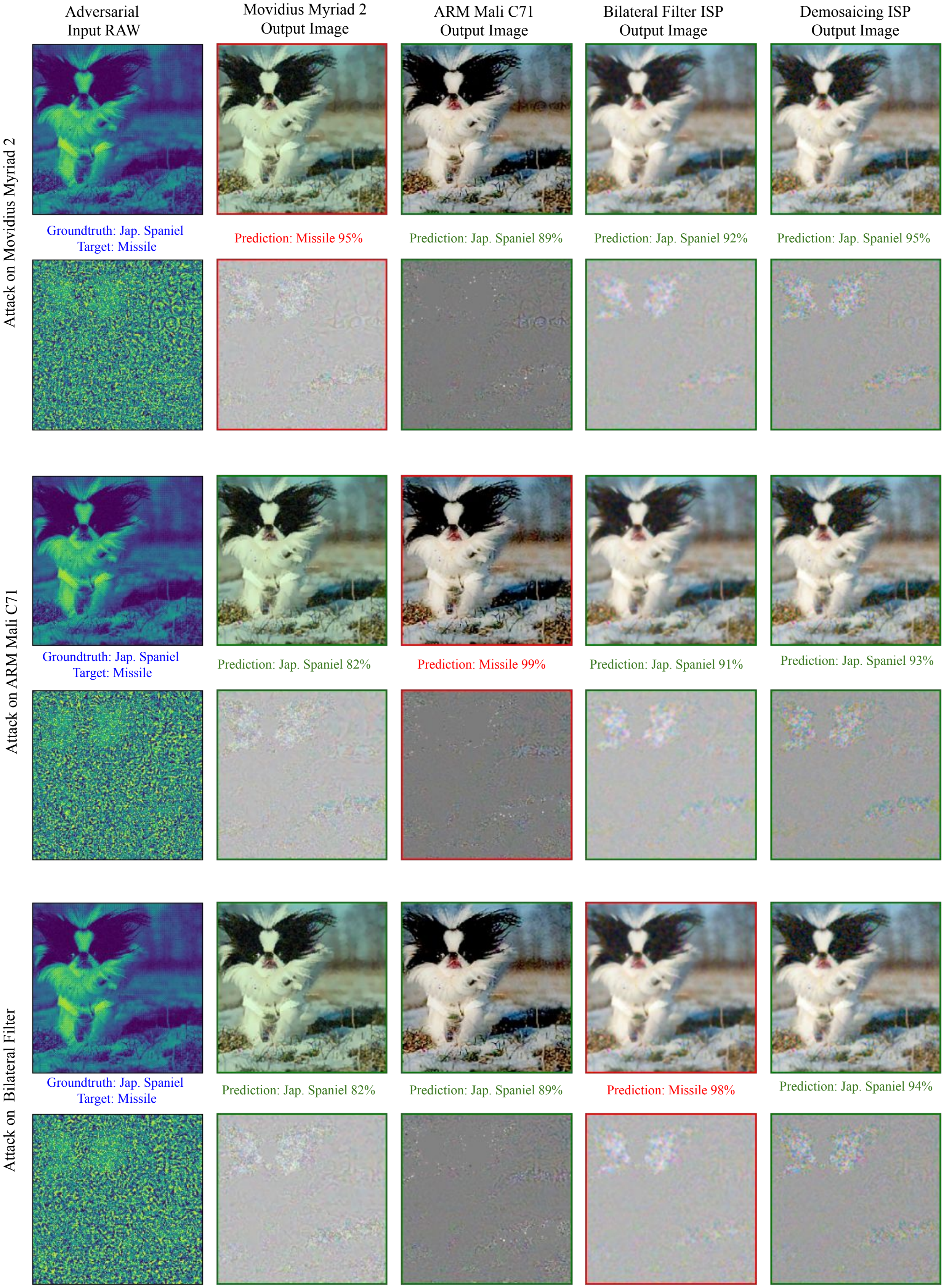

Adversarial Input RAW
Movidius Myriad 2 Output Image
ARM Mali C71 Output Image
Bilateral Filter ISP Output Image
Demosaicing ISP Output Image
Attack on Movidius Myriad 2
Groundtruth: Jap. Spaniel
Target: Missile
Prediction: Missile 95%
Prediction: Jap. Spaniel 89%
Prediction: Jap. Spaniel 92%
Prediction: Jap. Spaniel 95%
Attack on ARM Mali C71
Groundtruth: Jap. Spaniel
Target: Missile
Prediction: Jap. Spaniel 82%
Prediction: Missile 99%
Prediction: Jap. Spaniel 91%
Prediction: Jap. Spaniel 93%
Attack on Bilateral Filter
Groundtruth: Jap. Spaniel
Target: Missile
Prediction: Jap. Spaniel 82%
Prediction: Jap. Spaniel 89%
Prediction: Missile 98%
Prediction: Jap. Spaniel 94%

Figure 6: Visualization of the adversarial images and perturbations for the targeted ISP attack. Each pair of rows (top to bottom) shows the attack on the Movidius Myriad 2, ARM Mali C71 and Bilateral Filter ISP respectively. In each targeted ISP attack, we show in the first column the adversarial RAW (top) and perturbations (bottom). The next four columns show the associated RGB images and perturbations from the ISPs. The RGB perturbation is visualized by subtracting the ISP output of adversarial RAW from that of the unattacked output.

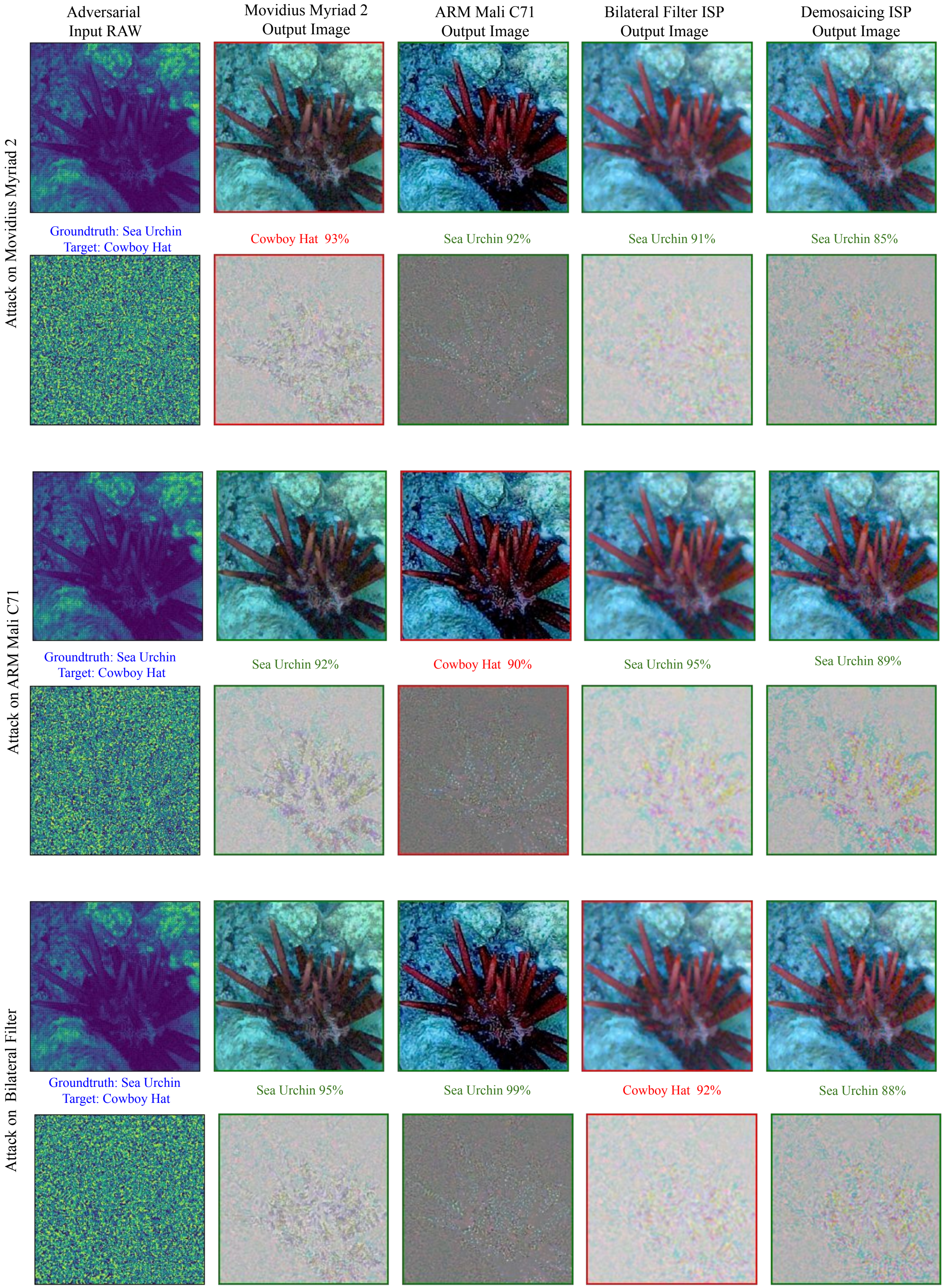


Figure 7: Addtional visualizations of the targeted ISP attack.

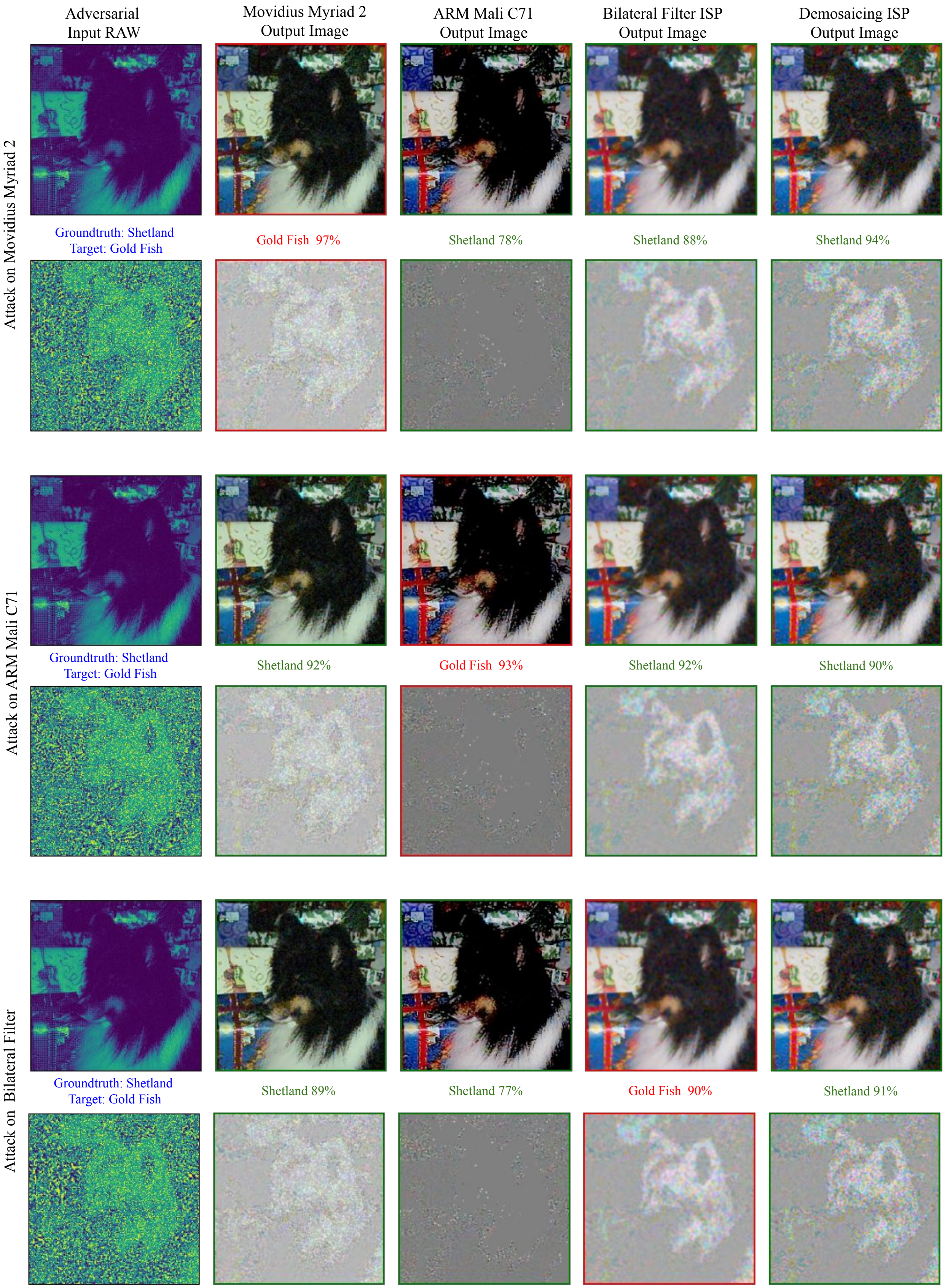


Figure 8: Addtional visualizations of the targeted ISP attack.

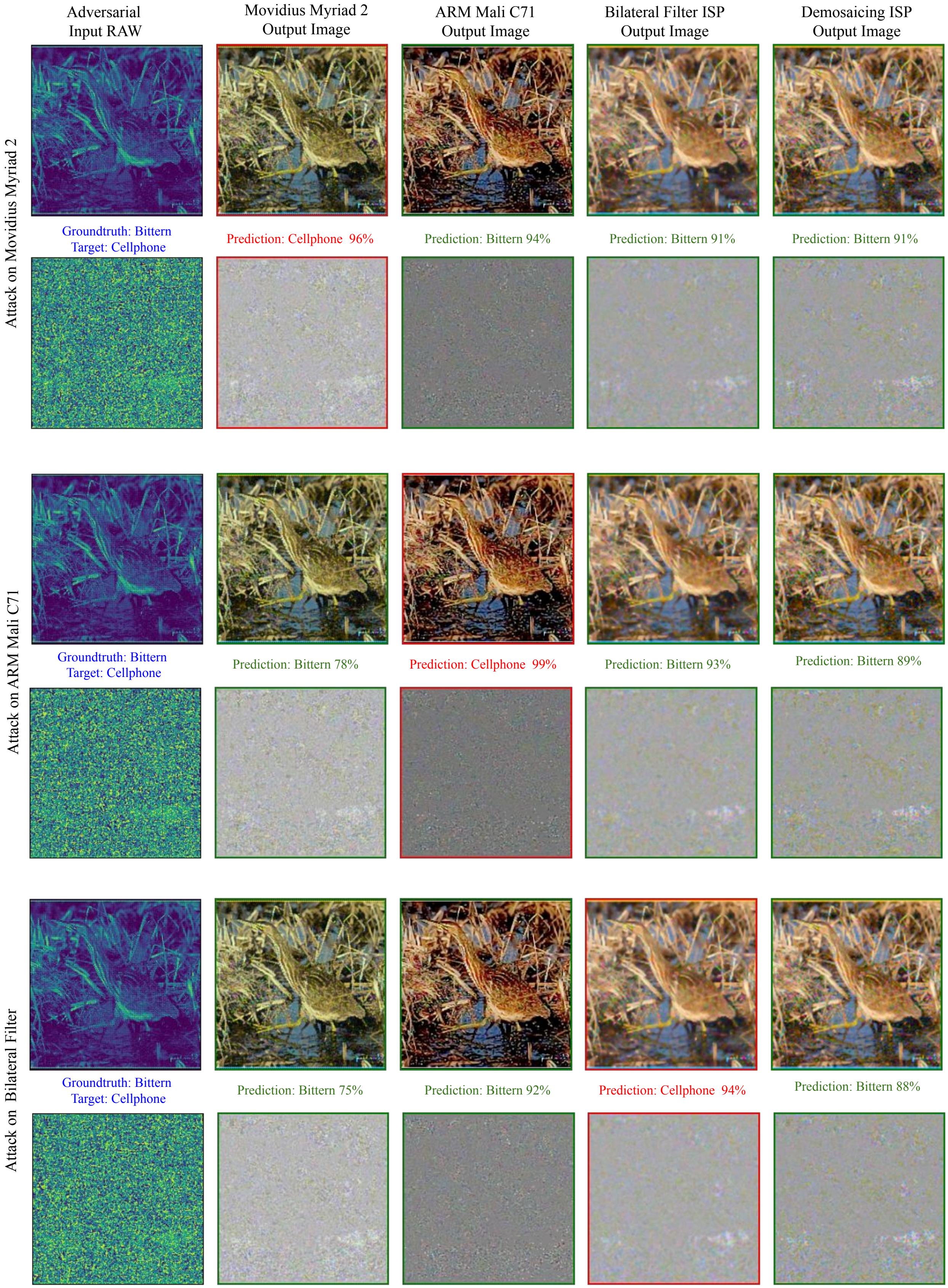


Figure 9: Addtional visualization for targeted ISP attack.

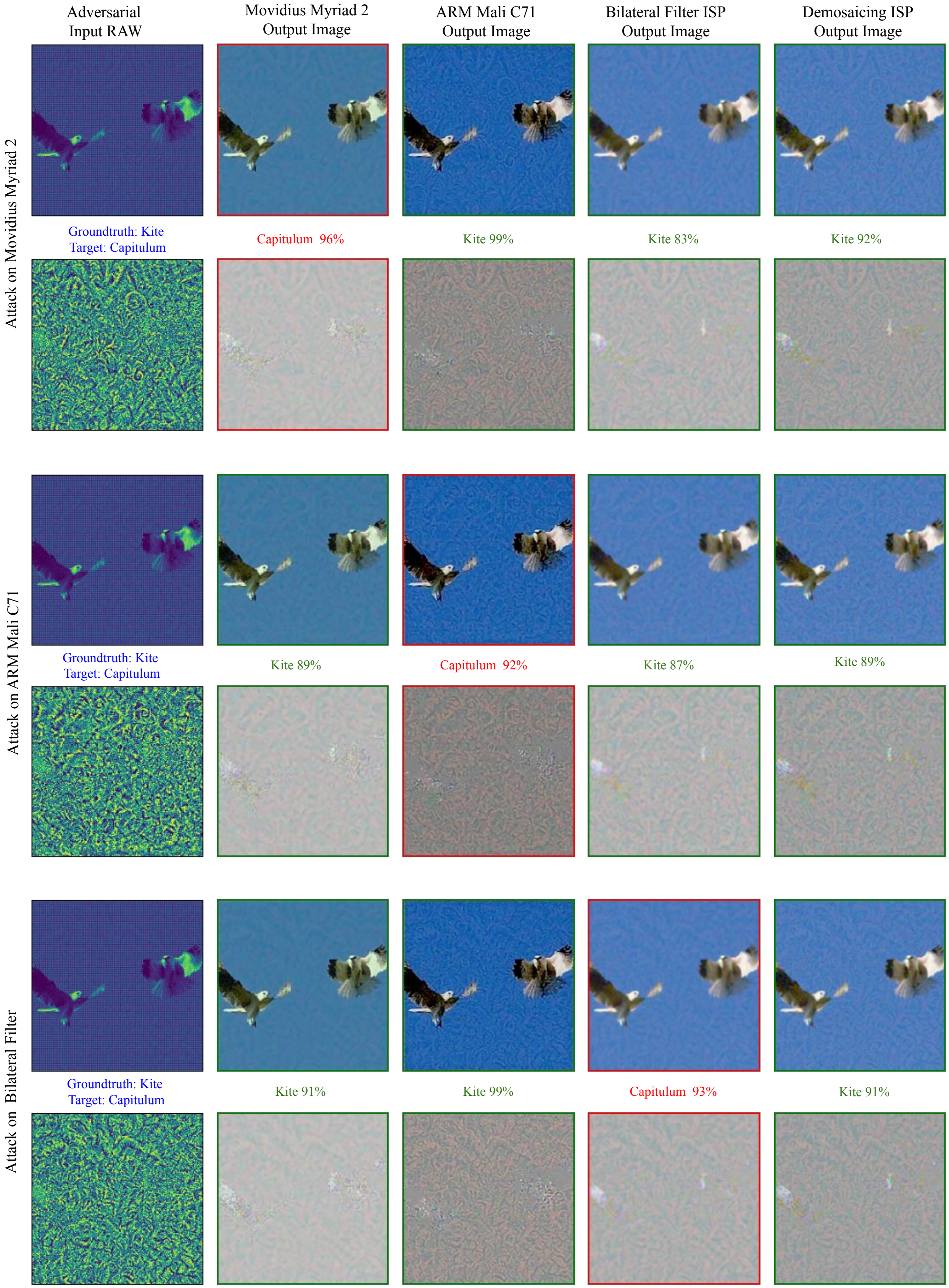


Figure 10: Addtional visualization for targeted ISP attack.

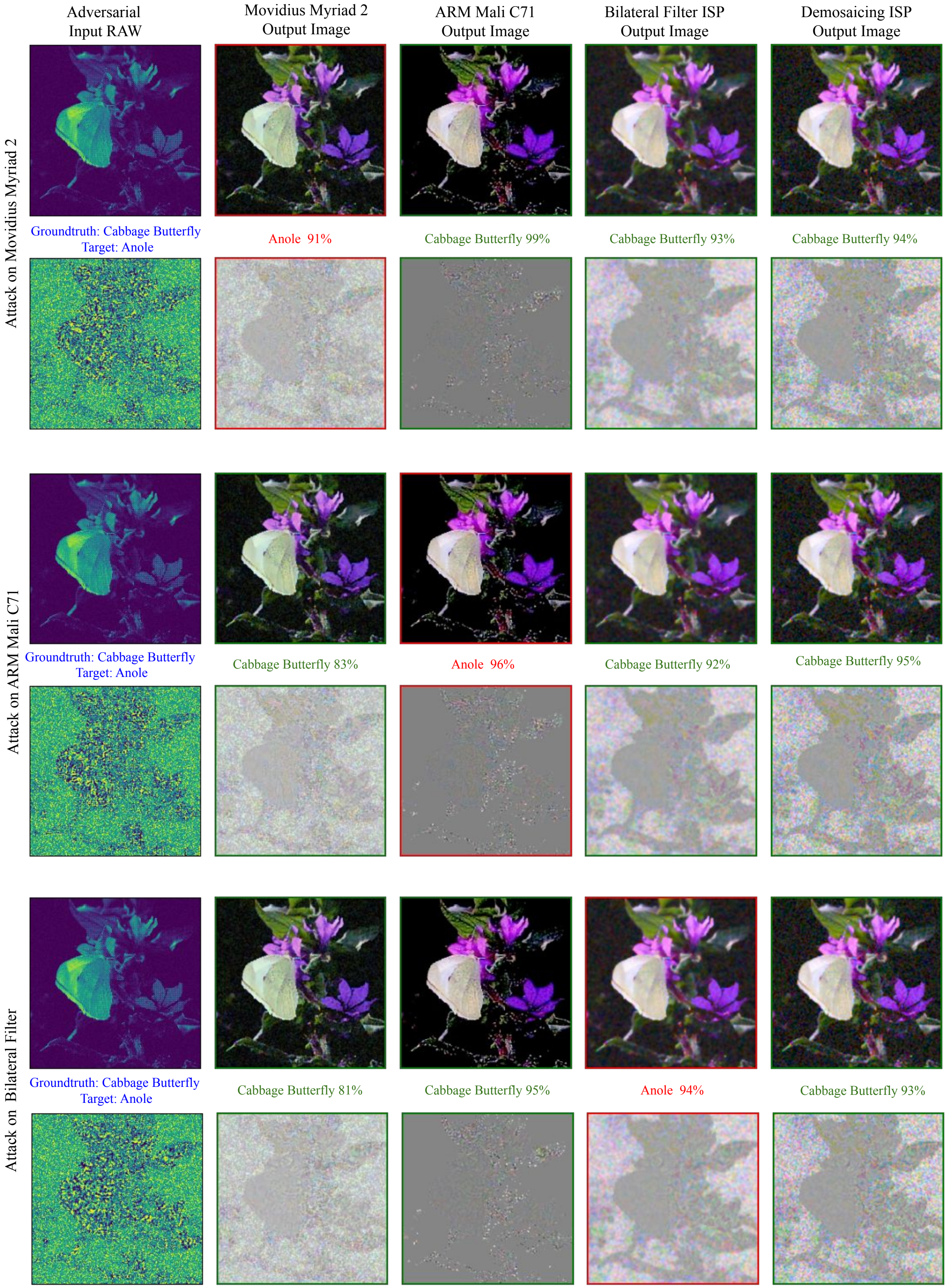
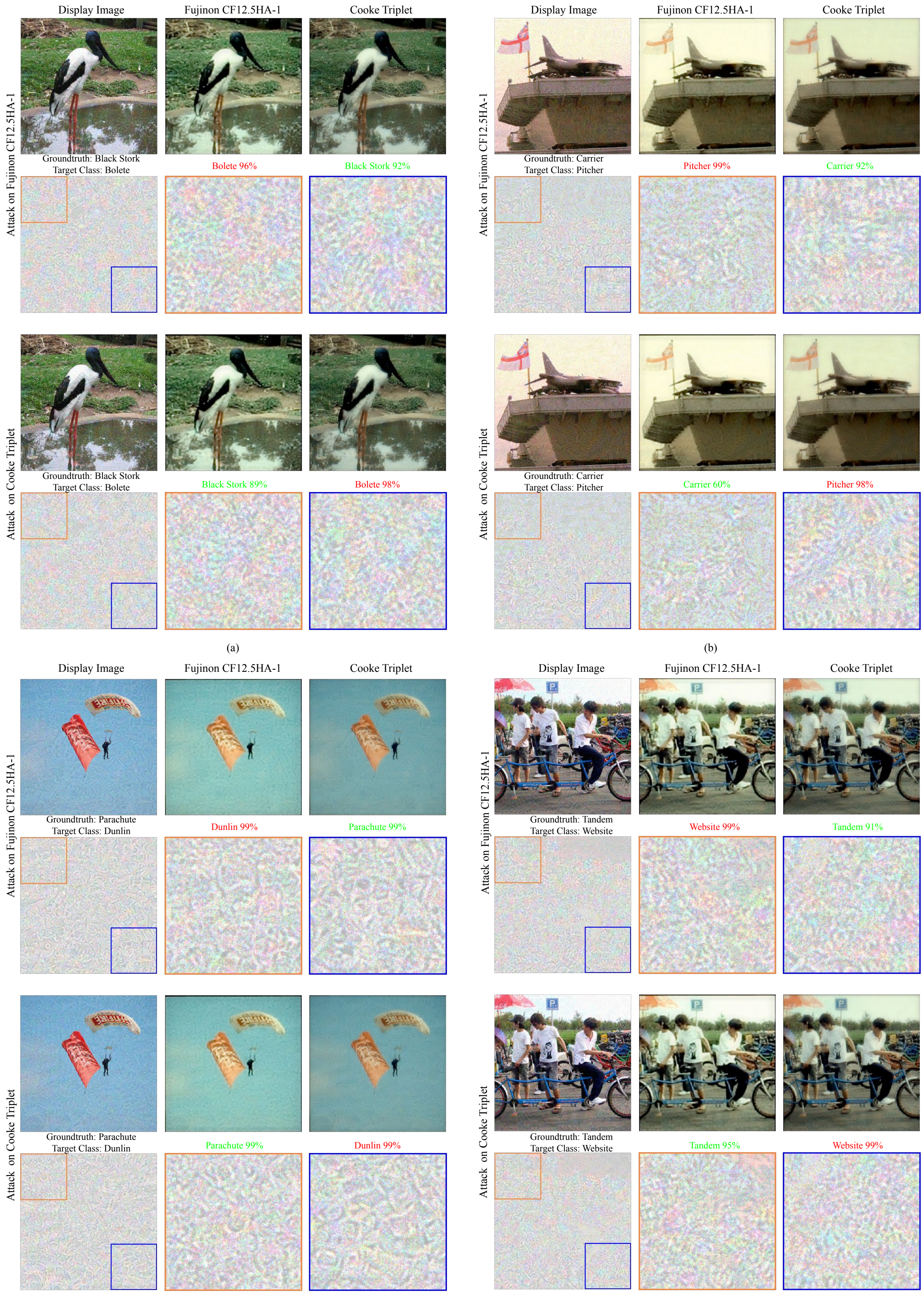


Figure 11: Addtional visualization for targeted ISP attack.



(a) (b) (c) (d)

Figure 12: Visualization (a),(b),(c),(d) of the targeted optics attack on the Fujinon CF12.5HA-1 and Cooke Triplet optics. For each attack, we show the displayed adversarial and post-processed images (top row). In the bottom row, we visualize (from left to right) the additive perturbations on the display image and its zoomed in 150 × 150 top-left and bottom right region.

References

[1] Movidius Myriad 2 Documentation. https://usermanual.wiki/Document/MDKMA2x5xSIPPUserManual.995467580. 2

[2] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017. 3

[3] Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. *arXiv preprint arXiv:1911.07140*, 2019. 3

[4] Michael J Kidger. Fundamental optical design. SPIE Bellingham, 2001. 2

[5] Arm Ltd. Mali-c71. https://www.arm.com/products/silicon-ip-multimedia/image-signal-processor/mali-c71. Accessed: 2019-08-30. 2

[6] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–169, 2018. 3

[7] Sylvain Paris, Pierre Kornprobst, Jack Tumblin, and Frédo Durand. *Bilateral filtering: Theory and applications*. Now Publishers Inc, 2009. 1

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[9] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pages 839–846. IEEE, 1998. 2

[10] Ethan Tseng, Felix Yu, Yuting Yang, Fahim Mannan, Karl ST Arnaud, Derek Nowrouzezahrai, Jean-François Lalonde, and Felix Heide. Hyperparameter optimization in black-box image processing using differentiable proxies. *ACM Trans. Graph.*, 38(4):27–1, 2019. 3

[11] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 742–749, 2019. 3