Supplementary Material for "Towards Large yet Imperceptible Adversarial Image Perturbations with Perceptual Color Distance"

Zhengyu Zhao, Zhuoran Liu, Martha Larson
Radboud University, Nijmegen, Netherlands
{z.zhao, z.liu, m.larson}@cs.ru.nl

Approach	Budget	λ	
		Targeted	Untargeted
C&W [1]	3×100	1	0.1
	5×200	1	1
	9×1000	1	1
PerC-C&W (ours)	3×100	10	100
	5×200	10	100
	9×1000	10	10

Table 1: Selected initializations of λ via grid search.

Approach	Budget	Success Rate (%)	Perturbation Size		
			$\overline{L_2}$	$\overline{L_\infty}$	$\overline{C_2}$
I-FGSM [2]	-	100.0	1.94	1.02	255.92
C&W [1]	3×100	100.0	0.69	3.61	88.76
	5×200	100.0	0.45	3.79	59.88
	9×1000	100.0	0.41	3.74	54.17
PerC-C&W (ours)	3×100	100.0	1.47	6.78	78.25
	5×200	100.0	0.90	6.71	51.35
	9×1000	100.0	0.56	6.58	33.00
DDN [3]	100	100.0	0.35	4.03	49.43
	300	100.0	0.33	4.08	47.58
	1000	100.0	0.32	4.11	46.51
PerC-AL (ours)	100	100.0	0.53	5.58	30.39
	300	100.0	0.50	6.93	27.65
	1000	100.0	0.51	8.92	26.62

Table 2: Success rates and perturbation sizes on the 1000 images from the ImageNet-Compatible dataset, with varied budgets in the targeted setting. Perturbation size is quantified in terms of L_2 and L_∞ norms of the perturbations in RGB space ($\overline{L_2}$ and $\overline{L_\infty}$) and also in terms of image-level accumulated perceptual color difference ($\overline{C_2}$). For this relatively easy untargeted case, PerC-AL is initialized with $\alpha_c = 0.1$.

References

[1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy*, pages 39–57, 2017. 1

[2] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations*, 2017. 1

[3] Jérôme Rony, Luiz G. Hafemann, Luiz S. Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based L_2 adversarial attacks and defenses. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 4322–4330, 2019. 1

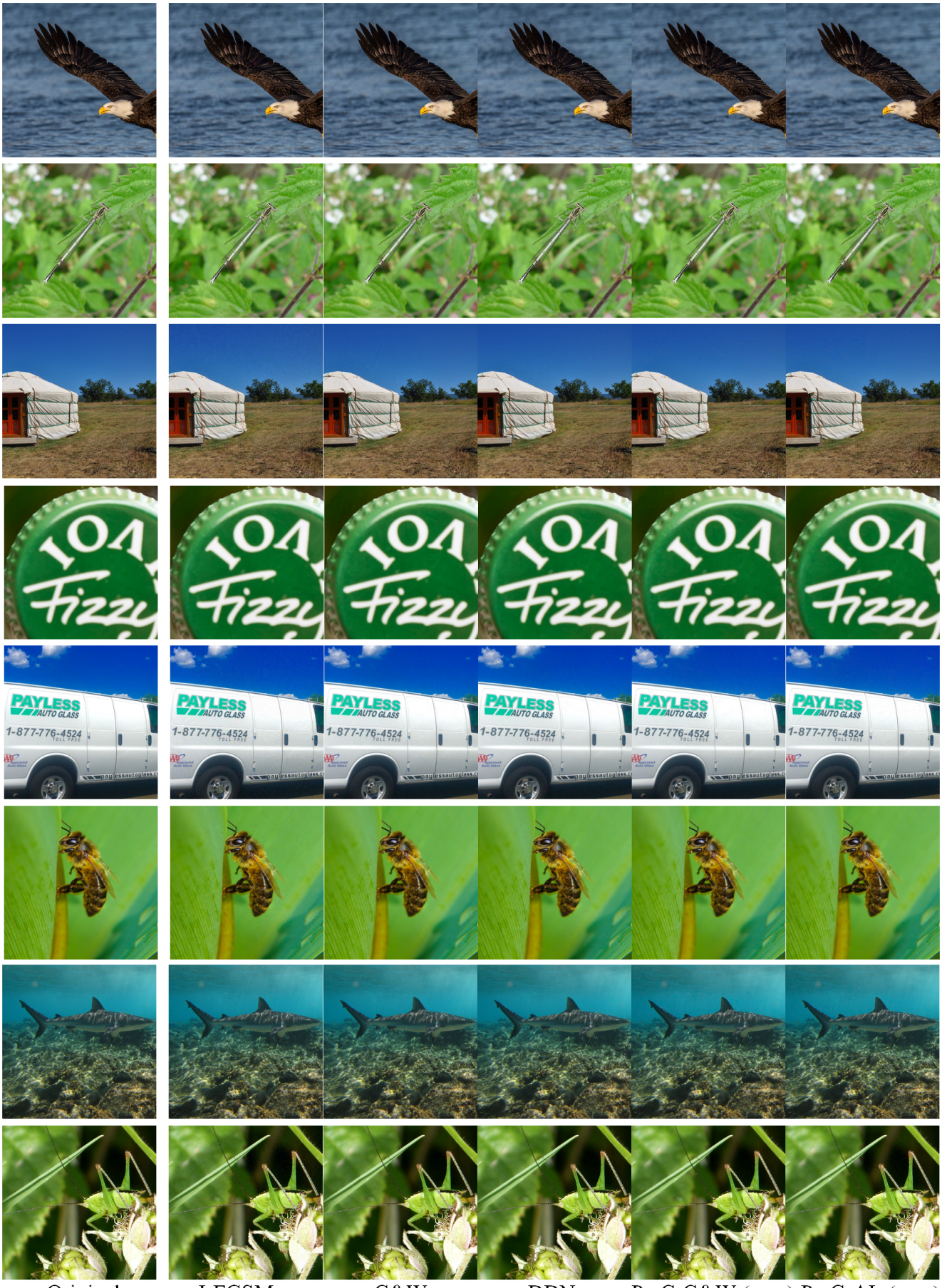

Figure 1: Examples of high-confidence adversarial images generated by five different approaches at $\kappa = 40$ (Set 1)

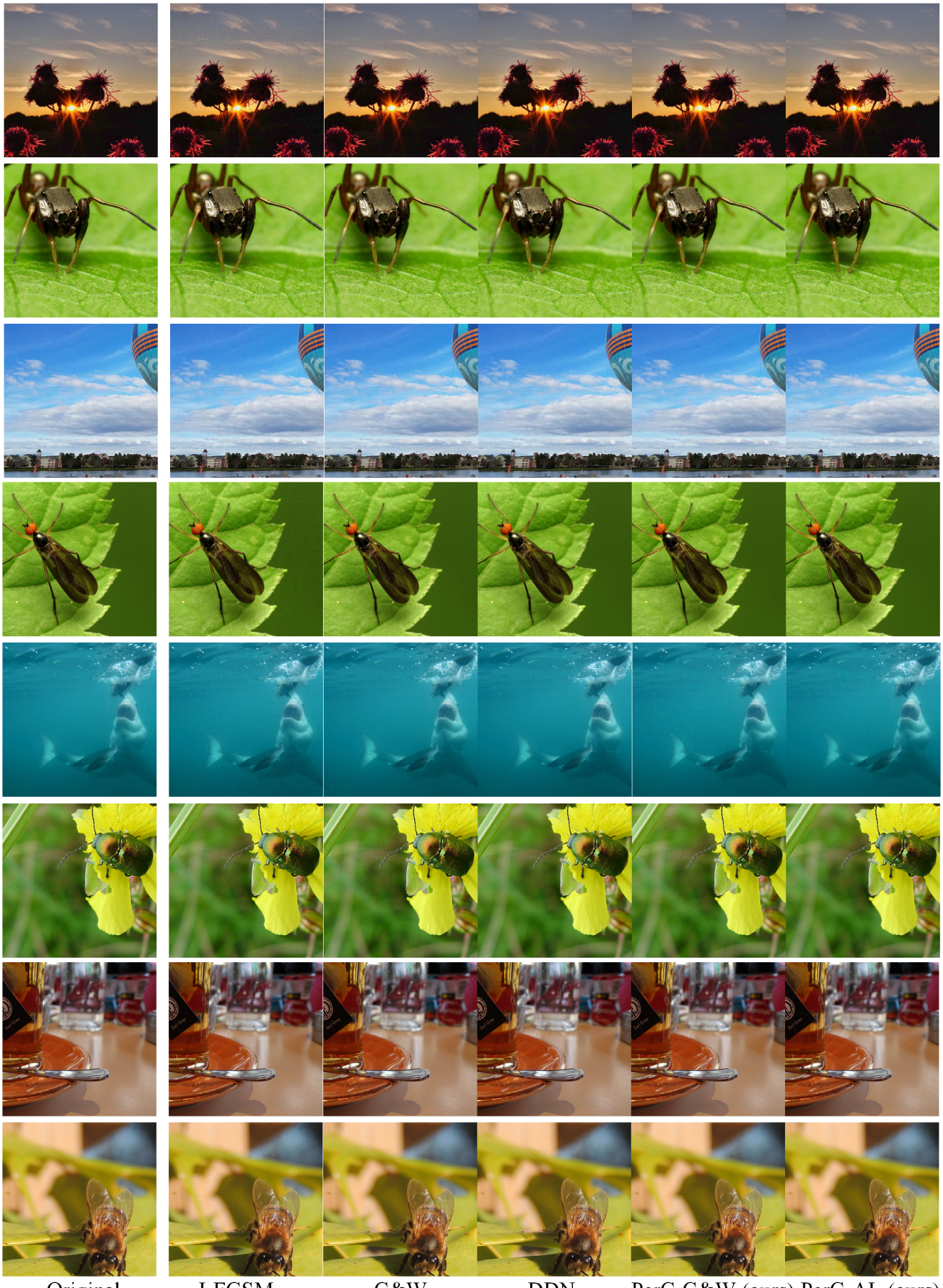

Figure 2: Examples of high-confidence adversarial images generated by five different approaches at $\kappa = 40$ (Set 2)