

## 348 A Additional Experimental Results

Table 5: White-box and Black-box Adversarial Robustness Evaluation (mean of the F1 score across 5 random seeds  $\pm$  *standard\_deviation*) - Adversarial attacks are ran against the model in each row and evaluated on models in each column.  $\dagger$  refers to detectors trained on the D<sup>3</sup> training set. Final row N - *model* refers to ensemble attacks targeting all models, excluding the evaluated against model. **bold** values corresponding to white-box evaluation in single detector attacks

	Ca24 $\dagger$	Ch24C	Ch24CN	Co23 $\dagger$	K24	O23 $\dagger$	W20 $\dagger$
Ca24 $\dagger$	<b>0.0</b> $\pm 0.0$	0.87 $\pm 0.0$	0.84 $\pm 0.01$	0.86 $\pm 0.0$	0.81 $\pm 0.02$	0.71 $\pm 0.02$	0.66 $\pm 0.02$
Ch24C	0.54 $\pm 0.01$	<b>0.0</b> $\pm 0.0$	0.24 $\pm 0.02$	0.75 $\pm 0.01$	0.08 $\pm 0.01$	0.23 $\pm 0.03$	0.56 $\pm 0.01$
Ch24CN	0.61 $\pm 0.01$	0.83 $\pm 0.01$	<b>0.26</b> $\pm 0.02$	0.73 $\pm 0.01$	0.81 $\pm 0.01$	0.73 $\pm 0.01$	0.56 $\pm 0.01$
Co23 $\dagger$	0.7 $\pm 0.0$	0.82 $\pm 0.0$	0.44 $\pm 0.01$	<b>0.0</b> $\pm 0.0$	0.75 $\pm 0.02$	0.63 $\pm 0.02$	0.02 $\pm 0.01$
K24	0.49 $\pm 0.02$	0.23 $\pm 0.03$	0.26 $\pm 0.02$	0.71 $\pm 0.01$	<b>0.07</b> $\pm 0.02$	0.35 $\pm 0.02$	0.6 $\pm 0.02$
O23 $\dagger$	0.51 $\pm 0.02$	0.38 $\pm 0.02$	0.35 $\pm 0.02$	0.73 $\pm 0.01$	0.29 $\pm 0.03$	<b>0.0</b> $\pm 0.0$	0.55 $\pm 0.01$
W20 $\dagger$	0.64 $\pm 0.01$	0.84 $\pm 0.01$	0.21 $\pm 0.02$	0.08 $\pm 0.01$	0.79 $\pm 0.01$	0.63 $\pm 0.01$	<b>0.0</b> $\pm 0.0$
N - <i>model</i>	0.61 $\pm 0.01$	0.3 $\pm 0.02$	0.19 $\pm 0.03$	0.47 $\pm 0.02$	0.09 $\pm 0.01$	0.21 $\pm 0.03$	0.3 $\pm 0.02$

349 We provide additional details regarding the variability of our experimental results due to different test  
350 splits. As such, we conducted the main experiments using five random seeds. The PGD attack is used  
351 with 10 steps, step size = 0.05, and  $\epsilon = 16/255$ . We report each evaluation metric’s mean and standard  
352 deviation in Tables 5 and 6. A similar trend can be observed where the ensemble attacks are a good  
353 estimate for the adversarial robustness of the AI-generated image detectors.

## 354 B Evaluation on Commercial Detectors

355 We use two commercial AI-generated image detectors provided by HIVE and Sightengine to evaluate  
356 RAID in real-world settings. We report the results set of 50 real images and 50 pairs of clean and  
357 adversarial images in Table 7. While these results are preliminary due to the dataset size, they  
358 still suggest that RAID retains effectiveness when assessing the adversarial robustness of detectors  
359 deployed in commercial detection APIs. We compute the Accuracy with a 0.5 threshold on the  
360 confidence score reported by the detectors. We also provide a few examples of the detection scores  
361 returned by commercial detectors on randomly selected images in Figures 3 and 4.

Table 6: White-box and Black-box Adversarial Robustness Evaluation (mean of the AUROC score across 5 random seeds  $\pm$  *standard\_deviation*) - Adversarial attacks are ran against the model in each row and evaluated on models in each column.  $\dagger$  refers to detectors trained on the D<sup>3</sup> training set. Final row N - *model* refers to ensemble attacks targeting all models, excluding the evaluated against model. **bold** values corresponding to white-box evaluation in single detector attacks

	Ca24 <sup>†</sup>	Ch24C	Ch24CN	Co23 <sup>†</sup>	K24	O23 <sup>†</sup>	W20 <sup>†</sup>
Ca24 <sup>†</sup>	<b>0.50</b> $\pm$ <b>0.00</b>	0.62 $\pm$ 0.02	0.81 $\pm$ 0.01	0.85 $\pm$ 0.0	0.69 $\pm$ 0.02	0.75 $\pm$ 0.01	0.75 $\pm$ 0.01
Ch24C	0.53 $\pm$ 0.01	<b>0.5</b> $\pm$ <b>0.0</b>	0.57 $\pm$ 0.01	0.79 $\pm$ 0.01	0.52 $\pm$ 0.0	0.56 $\pm$ 0.01	0.7 $\pm$ 0.0
Ch24CN	0.51 $\pm$ 0.02	0.62 $\pm$ 0.02	<b>0.49</b> $\pm$ <b>0.01</b>	0.78 $\pm$ 0.0	0.61 $\pm$ 0.02	0.74 $\pm$ 0.01	0.7 $\pm$ 0.01
Co23 <sup>†</sup>	0.48 $\pm$ 0.02	0.62 $\pm$ 0.02	0.61 $\pm$ 0.01	<b>0.5</b> $\pm$ <b>0.0</b>	0.63 $\pm$ 0.01	0.7 $\pm$ 0.01	0.51 $\pm$ 0.0
K24	0.53 $\pm$ 0.01	0.51 $\pm$ 0.01	0.57 $\pm$ 0.0	0.77 $\pm$ 0.0	<b>0.51</b> $\pm$ <b>0.01</b>	0.59 $\pm$ 0.01	0.72 $\pm$ 0.01
O23 <sup>†</sup>	0.53 $\pm$ 0.01	0.53 $\pm$ 0.01	0.6 $\pm$ 0.01	0.78 $\pm$ 0.0	0.55 $\pm$ 0.02	<b>0.5</b> $\pm$ <b>0.0</b>	0.69 $\pm$ 0.01
W20 <sup>†</sup>	0.51 $\pm$ 0.01	0.65 $\pm$ 0.01	0.53 $\pm$ 0.01	0.52 $\pm$ 0.0	0.66 $\pm$ 0.01	0.71 $\pm$ 0.01	<b>0.5</b> $\pm$ <b>0.0</b>
N - <i>model</i>	0.42 $\pm$ 0.02	0.53 $\pm$ 0.01	0.55 $\pm$ 0.01	0.65 $\pm$ 0.01	0.52 $\pm$ 0.01	0.55 $\pm$ 0.01	0.59 $\pm$ 0.01

Table 7: Performance on Commercial Deepfake Detectors. Mean score refers to the mean of the scores given by the detector on the AI-generated images. Adversarial variant of each metric refers to the evaluation on the adversarial RAID images. *metric\_real* refers to the metrics reported for real images

Detector	Score	Score_real	Adversarial Score	Acc	Acc_real	Adversarial Acc
Sightengine <sup>1</sup>	0.65	0.99	0.01	0.66	1.0	0.0
HIVE <sup>2</sup>	0.59	0.99	0.45	0.6	1.0	0.44

<sup>1</sup><https://dashboard.sightengine.com/ai-image-detection>

<sup>2</sup><https://hivemoderation.com/ai-generated-content-detection>

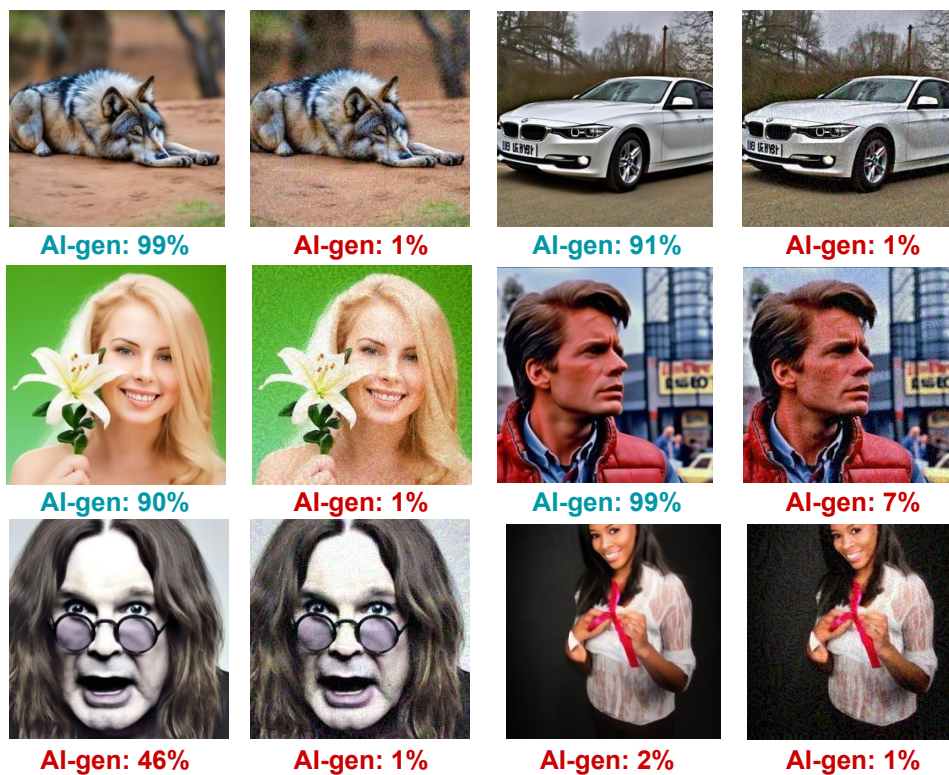


Figure 3: Detection scores returned by Sightengine\* detector on a subset of clean and adversarial AI-generated images. Higher scores indicate higher confidence that the image is AI-generated.

\*<https://dashboard.sightengine.com/ai-image-detection>

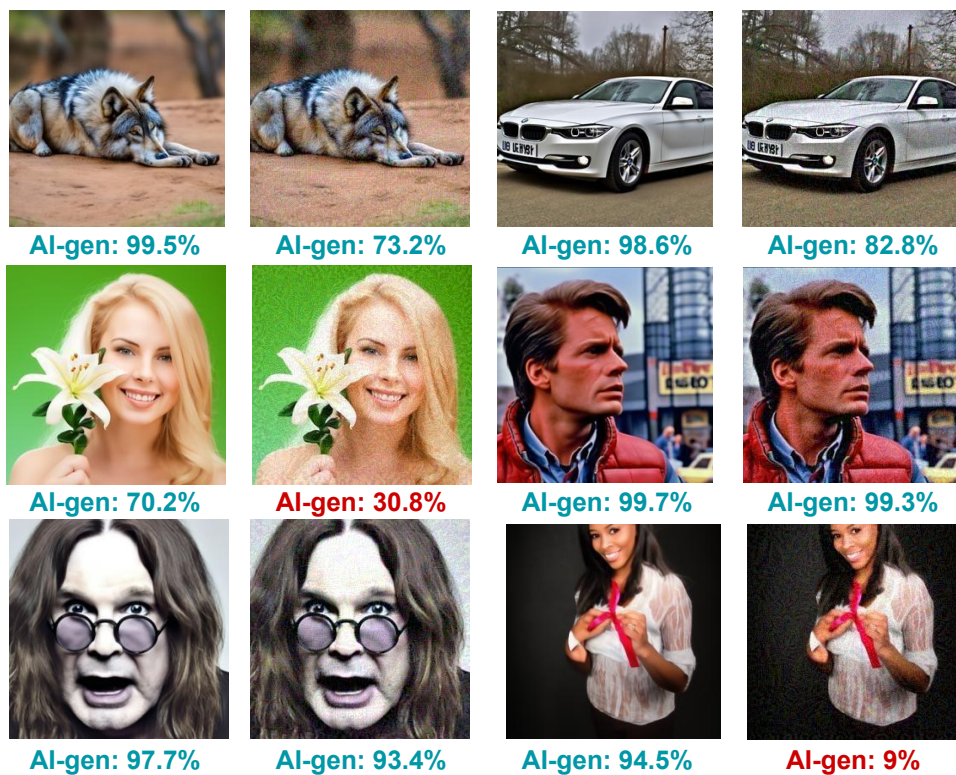


Figure 4: Detection scores returned by Hive Moderation\* detector on a subset of clean and adversarial AI-generated images. Higher scores indicate higher confidence that the image is AI-generated.

\*<https://hivemoderation.com/ai-generated-content-detection>