SEGGUARD: DEFENDING SCENE SEGMENTATION AGAINST ADVERSARIAL PATCH ATTACK
SUPPLEMENTARY MATERIAL

Thomas Gittings[⋆] *Steve Schneider*[⋆] *John Collomosse*[⋆†]

[⋆] Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. UK.
[†] Adobe Research, San Jose. USA.

In this supplemental material we include the following additional experiments and visualisations:

1. Comparison of pixel optimised vs generator patch attacks at different sizes in the image
2. Comparison of overall performance and visualisation of the class performance of targeted deletion attacks
3. Analysis of defence performance against grey box adversarial attacks
4. Experiments comparing the performance of attacks and defences with different generator architectures and patch resolutions

1. PIXEL-OPTIMISED PATCH VS UNCONDITIONAL GENERATOR ATTACK

Table 4 compares the performance of untargeted patches produced by the baseline method A-Patch with those produced by our unconditional patch generator (A-UGen). Nesti *et al*.[25] train and test their patches constrained to a small region around the centre of the image, which is not realistic for practical scenarios therefore we additionally explore the patch located anywhere in the image (c.f. Brown *et al*.[4]). Nesti *et al*.[25] explore the effect of patch size on performance, and we extend this to test the performance of patches trained and tested at different sizes. We compare the sizes (S)mall/(M)edium/(L)arge, which correspond to adversarial patches of size $150 \times 300/200 \times 400/300 \times 600$, and covering 2.1%/3.8%/8.6% of the image respectively.

On DDRNet and BiSeNet, when testing patches at size L, the best results are achieved by A-UGen trained at size L. These attacks decrease the performance of the model by over 40 percentage points when compared to the best performing non-adversarial patch (A-Noise). For these networks the best performing A-Patch models, on the other hand, reduce the performance by less than 30 percentage points. The Centre/All positioning in testing/training makes little difference to the results for these networks. For ICNet the performance of A-Patch is significantly better than A-UGen, reducing the performance by over 40 percentage points. In this case, however, the attacks by A-Patch perform significantly worse in the Centre location than in the All location, although the location for training does not seem to make a difference.

For all three networks, there is little difference in performance between A-Patch and A-UGen when training and testing at smaller sizes. Most values are within 5 percentage points. The patches of both methods do have some ability to generalise to different sizes, decreasing the mIoU significantly more than A-Noise. The best generalisation can be seen for A-UGen on the smaller two sizes. For A-Patch there is little difference training with Centre vs All, but for A-UGen the impact is significant. In Section 4 we explore to what extent the difference in performance between A-Patch and A-Gen comes from the size of the patch (in terms of number of pixels).

2. DELETION ATTACK

In this section we compare the performance A-Patch, A-UGen and A-CGen when optimised to perform targeted deletion attacks. Table 5 compares the overall performance of these attacks. In this case the performance is very similar for all three attacks on each of the three networks. For BiSeNet and ICNet A-CGen performs the best by both metrics, whereas on DDRNet A-CGen is best on mmIoU but the other two methods are tied for best on mcIoU.

Figure 6 splits out the results by class in the manner explained in Section 4.1 of the main paper. In contrast to Figure 4 in the main paper, none of the three methods exhibit a dark diagonal in this figure, which implies that the effect of deletion is not always visible in the class ostensibly being deleted. Furthermore, when the attack is successful a common set of classes are affected by it, which can be seen by the fact that columns with dark squares are visually similar. Although A-CGen performed well in terms of the overall metrics in all cases, it does not respect the conditioning at all, note that every column of the matrix looks exactly the same. This is potentially caused by the fact that the loss function for this task contains nothing to discourage all the classes being attacked at once. Given these results we do not attempt to defend against targeted deletion attacks in the main paper, since the targeting has no impact, making them completely equivalent to untargeted attacks.

Table 4. Untargeted Attack Results: mean intersection over union (mIoU) evaluated over the Cityscapes evaluation dataset with different training and test settings.

Architecture	Train Settings			Test Settings					
				Centre			All		
				S	M	L	S	M	L
DDRNet	A-Noise		S	0.77	0.75	0.76	0.77	0.77	0.76
			M	0.77	0.75	0.75	0.77	0.76	0.76
			L	0.77	0.76	0.74	0.76	0.76	0.76
	A-Patch	Centre	S	**0.69**	**0.66**	0.72	0.69	**0.66**	0.73
			M	0.75	0.68	0.68	0.75	0.67	0.68
			L	0.77	0.75	0.54	0.77	0.75	0.52
		All	S	0.73	0.71	0.75	0.73	0.71	0.75
			M	0.75	0.68	0.68	0.75	0.67	0.68
			L	0.77	0.75	0.54	0.76	0.75	0.53
	A-UGen	Centre	S	**0.69**	0.68	0.71	**0.68**	0.68	0.72
			M	0.71	0.60	0.55	0.69	0.60	0.56
			L	0.75	0.70	0.32	0.75	0.71	0.28
		All	S	**0.69**	0.69	0.72	**0.68**	0.68	0.71
			M	0.73	0.68	0.67	0.73	0.68	0.67
			L	0.75	0.70	**0.30**	0.74	0.70	**0.26**
BiSeNet	A-Noise		S	0.67	0.66	0.64	0.68	0.67	0.64
			M	0.67	0.65	0.64	0.67	0.67	0.65
			L	0.68	0.67	0.64	0.67	0.66	0.65
	A-Patch	Centre	S	0.59	0.54	0.49	0.58	0.53	0.47
			M	0.61	0.54	0.45	0.60	0.53	0.43
			L	0.63	0.59	0.27	0.62	0.59	0.25
		All	S	0.58	0.54	0.45	0.58	0.52	0.45
			M	0.60	0.51	0.42	0.60	0.53	0.42
			L	0.65	0.62	0.43	0.65	0.62	0.42
	A-UGen	Centre	S	**0.51**	0.44	0.32	**0.51**	0.43	0.31
			M	0.55	**0.41**	0.21	0.55	**0.40**	0.21
			L	0.58	0.44	**0.17**	0.57	0.44	**0.16**
		All	S	**0.51**	0.44	0.32	**0.51**	0.44	0.31
			M	0.55	**0.41**	0.20	0.55	**0.40**	0.19
			L	0.58	0.44	**0.17**	0.58	0.44	**0.16**
ICNet	A-Noise		S	0.77	0.75	0.73	0.76	0.76	0.73
			M	0.77	0.75	0.74	0.77	0.75	0.74
			L	0.78	0.76	0.72	0.77	0.76	0.73
	A-Patch	Centre	S	0.74	0.71	0.69	0.72	0.70	0.69
			M	0.75	0.68	0.64	0.74	0.66	0.62
			L	0.77	0.73	**0.42**	0.75	0.71	**0.30**
		All	S	**0.65**	0.62	0.57	0.63	0.61	0.57
			M	0.69	**0.60**	0.51	0.66	0.76	0.51
			L	0.76	0.72	0.44	0.75	0.71	**0.30**
	A-UGen	Centre	S	0.67	0.63	0.57	**0.65**	0.63	0.57
			M	0.76	0.72	0.66	0.75	0.71	0.65
			L	0.77	0.74	0.59	0.75	0.73	0.53
		All	S	0.67	0.63	0.57	**0.65**	0.62	0.57
			M	0.68	0.62	0.57	0.66	**0.61**	0.55
			L	0.73	0.67	0.51	0.71	0.65	0.49

Table 5. Average performance of Deletion Attacks with three different kinds of optimisation: a pixel optimised-patch (A-Patch), unconditional generator (A-UGen) and conditional generator (A-CGen)

Architecture	Metric	A-Patch	A-UGen	A-CGen
DDRNet	mcIoU	**0.54**	**0.54**	0.60
	mmIoU	0.61	0.62	**0.59**
BiSeNet	mcIoU	0.29	0.28	**0.20**
	mmIoU	0.43	0.37	**0.20**
ICNet	mcIoU	0.52	0.57	**0.50**
	mmIoU	0.55	0.59	**0.51**

3. DEFENCES VS GREY BOX ATTACKS

In Table 6 we show results for our defended networks under attack from an attacker with reduced knowledge compared to the previous section. Here we consider a form of grey box attack, where the attacker knows the model architecture and has access to the published pre-trained model weights, but not the weights of the defended model, or the defence method in the case of baselines. In almost all cases the defence is completely successful in this setting, with mIoU reduced by only a few percentage points compared to the results on clean data. The baseline defences perform a lot better in this setting than in the white box setting, but they still do not meet the performance of our defences.

4. GENERATOR AND PATCH SIZE

We explore the impact of modifying the generator architecture via the addition of layers, the introduction of a rectangular filter in the first layer (see Section 3.2 of the main paper), and changing the value of the parameter d, which scales the number of channels in the generator. Table 7 compares the performance of the different generators as attacks, both conditional and unconditional. In general the more parameters the generator has, the better its performance, but in most cases the effect is not very significant.

Table 8 compares the effect of changing d in a SegGuard defence with a 7 layer rectangular generator generator, attacked with targeted A-Patch. In this case the performance difference is mostly minimal, with a gradual improvement being seen as the value of d is increased. In the rest of our experiments we chose to use $d = 32$ as a compromise between model size an performance.

In Table 9 we explore the impact of patch size/resolution on attack performance, both targeted and untargeted. For targeted patches it is clear that the larger the patch the better the performance. For untargeted patches an intermediate size seems to provide the best performance. These results imply that the size of the patch is not the only factor explaining the different behaviour of A-Patch vs A-Gen.

Table 6. Mean intersection over union (mIoU) of defended networks against large grey box adversarial patches applied in a random position. We report figures for our SegGuard defence trained with untargeted attacks (D-UGen) and targeted insertion attacks (D-CGen) with both large and small size generators.

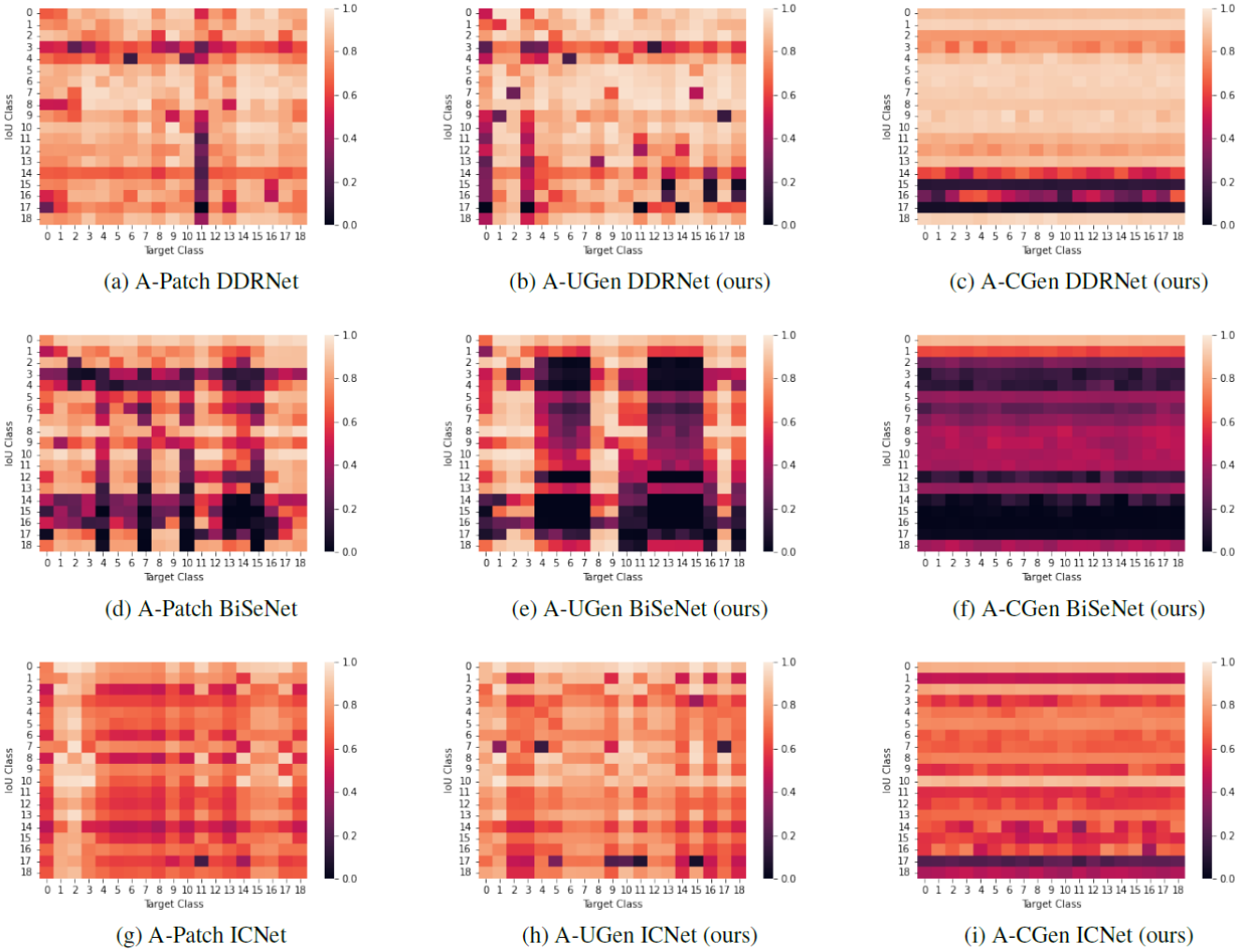
Architecture	Defence	A-Patch	A-UGen	A-CGen		A-Patch	
				mcIoU	mmIoU	mcIoU	mmIoU
DDRNet	D-UGen (S)	0.73	0.75	0.76	**0.76**	0.49	0.72
	D-UGen (L)	0.75	0.75	0.75	0.76	**0.75**	0.75
	D-CGen (S)	**0.76**	**0.76**	**0.76**	**0.77**	0.69	**0.76**
	D-CGen (L)	0.75	0.75	0.75	0.76	**0.75**	0.75
	D-Patch	0.69	0.63	0.61	0.68	0.61	0.69
	D-LGS	0.67	0.63	0.60	0.69	0.57	0.69
BiSeNet	D-UGen (S)	0.64	0.59	0.63	0.64	0.58	0.63
	D-UGen (L)	0.61	0.52	0.61	0.62	0.61	0.61
	D-CGen (S)	**0.66**	0.64	**0.67**	**0.67**	**0.66**	**0.66**
	D-CGen (L)	**0.66**	**0.65**	0.66	**0.67**	**0.66**	**0.66**
	D-Patch	0.54	0.42	0.47	0.51	0.48	0.52
	D-LGS	0.54	0.30	0.45	0.51	0.48	0.56
ICNet	D-UGen (S)	0.74	0.76	0.70	0.77	0.61	0.76
	D-UGen (L)	**0.77**	0.77	0.77	0.77	0.77	0.77
	D-CGen (S)	**0.77**	**0.78**	**0.78**	**0.79**	**0.78**	**0.78**
	D-CGen (L)	**0.77**	0.77	**0.78**	0.78	**0.78**	**0.78**
	D-Patch	0.68	0.64	0.63	0.71	0.62	0.68
	D-LGS	0.64	0.56	0.61	0.69	0.51	0.67

Table 7. Comparing the performance of unconditional generators with different numbers of layers and channels in terms of their ability to be used in an attack.

Task	d	Square			Rectangle		
		5	6	7	5	6	7
Untargeted ↓	8	0.41	0.31	0.20	0.59	0.34	0.19
Targeted mcIoU ↓		0.51	0.34	0.26	0.43	0.31	0.23
Targeted mmIoU ↓		0.70	0.66	0.61	0.70	0.64	0.60
Untargeted ↓	16	0.32	0.25	0.18	0.33	0.20	0.18
Targeted mcIoU ↓		0.36	0.28	0.20	0.32	0.24	0.21
Targeted mmIoU ↓		0.66	0.61	0.54	0.64	0.59	0.53
Untargeted ↓	32	0.30	0.21	0.20	0.34	0.19	0.51
Targeted mcIoU ↓		0.31	0.27	0.18	0.28	0.26	0.18
Targeted mmIoU ↓		0.61	0.55	0.48	0.58	0.54	0.49
Untargeted ↓	64	0.30	0.23	**0.15**	0.27	0.19	0.51
Targeted mcIoU ↓		0.26	0.24	0.18	0.27	0.19	**0.17**
Targeted mmIoU ↓		0.61	0.53	0.48	0.57	0.52	0.50
Untargeted ↓	128	0.31	0.22	**0.15**	0.31	0.18	0.16
Targeted mcIoU ↓		0.26	0.23	0.19	0.24	0.21	**0.17**
Targeted mmIoU ↓		0.59	0.53	**0.45**	0.57	0.51	0.49

Table 8. Comparing the performance of unconditional generators with different numbers of channels in terms of their ability to be used as part of a SegGuard defence against targeted pixel-optimised patches.

Task	d				
	8	16	32	64	128
mcIoU ↑	0.68	0.72	0.73	0.73	**0.75**
mmIoU ↑	0.74	0.75	0.75	0.75	**0.76**



(a) A-Patch DDRNet (b) A-UGen DDRNet (ours) (c) A-CGen DDRNet (ours)

(d) A-Patch BiSeNet (e) A-UGen BiSeNet (ours) (f) A-CGen BiSeNet (ours)

(g) A-Patch ICNet (h) A-UGen ICNet (ours) (i) A-CGen ICNet (ours)

Fig. 6. Comparison of targeted deletion attacks, broken down by target class. Each column corresponds to a target class, and then each row is the IoU for a class when the network is attacked with that patch. A value of 1 (white) means that the patch did not change the performance of the network on that class, and a value of 0 (black) means the IoU was 0 for that class.

Table 9. Comparing the performance of pixel-optimised patch attacks with different resolutions (*i.e.*the size of the patch before it is applied to the image).

Task	Square			Rectangle		
	64 × 64	128 × 128	256 × 265	64 × 128	128 × 256	512 × 512
Untargeted ↓	0.34	0.25	0.54	0.32	**0.21**	0.54
Targeted mcIoU ↓	0.24	0.21	0.17	0.22	0.19	**0.16**
Targeted mmIoU ↓	0.58	0.52	**0.47**	0.55	0.50	**0.47**