Role of Spatial Context in Adversarial Robustness for Object Detection: Supplementary Material

Aniruddha Saha* Akshayvarun Subramanya* Koninika Patil Hamed Pirsiavash
University of Maryland, Baltimore County
{anisaha1, akshayv1, koni1, hpirsiav}@umbc.edu

We show more results as supplementary material. Please refer to the captions of tables and figures for the description.

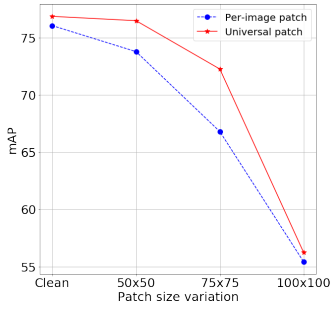

Figure A1: **Sensitivity to patch size -** We study the effect of the variation of patch size on our blindness attacks. We observe that as the patch size decreases, the attack success rate decreases. These results are also shown in Table A1.

	Mean	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	dtable	dog	horse	mbike	person	pplant	sheep	sofa	train	tv
Per-image																					
YOLOv2 (clean)	76.04	75.05	81.02	75.22	66.58	50.59	81.08	79.86	80.96	64.40	85.19	76.32	85.35	85.91	80.08	75.62	57.28	79.90	79.83	83.30	77.18
YOLOv2 (50x50 patch)	73.77	72.20	80.36	72.95	61.18	48.99	80.55	78.55	80.73	62.94	77.78	75.67	84.34	81.09	78.81	74.54	53.19	77.12	79.03	82.03	73.27
YOLOv2 (75x75 patch)	66.79	61.30	76.79	62.69	48.42	46.31	72.39	72.68	69.90	59.59	65.88	75.06	77.06	80.35	76.81	70.35	44.55	68.11	75.88	69.68	61.97
YOLOv2 (100x100 patch)	55.42	40.89	71.51	44.11	38.46	39.90	60.25	62.28	57.25	54.33	54.03	71.27	62.90	67.98	66.77	59.87	38.48	55.53	64.14	47.96	50.56
Universal																					
YOLOv2 (clean)	76.85	79.25	83.17	77.19	63.88	49.70	80.61	79.47	80.59	64.92	85.76	77.39	86.65	81.32	84.78	75.41	56.82	89.05	76.96	87.59	76.56
YOLOv2 (50x50 patch)	76.47	78.86	82.35	77.39	62.01	49.96	80.71	78.48	80.51	64.72	84.12	78.18	86.83	80.77	84.68	75.73	56.93	86.43	75.86	87.49	77.47
YOLOv2 (75x75 patch)	72.25	66.48	82.88	73.85	59.93	45.79	80.42	77.08	68.55	62.07	74.12	77.62	78.63	80.94	77.02	75.09	53.09	80.64	75.09	85.04	70.57
YOLOv2 (100x100 patch)	56.24	29.66	71.51	39.7	34.14	44.67	65.21	60.26	44.41	58.28	61.94	77.12	67.52	67.82	59.16	65.2	46.17	69.87	72.04	42.07	47.96

Table A1: **Sensitivity to patch size -** The first 4 rows are the per-image blindness attack and the last 4 rows are the universal blindness attack.

	Mean	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	dtable	dog	horse	mbike	person	pplant	sheep	sofa	train	tv
YOLOv2 (clean)	76.04	75.05	81.02	75.22	66.58	50.59	81.08	79.86	80.96	64.40	85.19	76.32	85.35	85.91	80.08	75.62	57.28	79.90	79.83	83.30	77.18
YOLOv2 (attacked)	58.49	38.60	73.86	49.39	34.27	41.12	60.50	61.50	71.91	53.38	56.74	73.24	71.94	75.01	70.69	58.30	38.48	58.99	70.61	58.23	52.98

Table A2: **Per-image objectness attack -** This attack is described in **Per-image objectness attack** paragraph of Section 4.3 in the main paper and the qualitative results are in Figure A4 of supplementary. We perform a different kind of adversarial patch attack by trying to fool YOLOv2 objectness confidence.

*Equal contribution

	Mean	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	dtable	dog	horse	mbike	person	pplant	sheep	sofa	train	tv
Clean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Targeted attack	18.61	13.09	13.55	22.30	15.48	11.99	16.36	20.06	20.47	19.50	21.05	19.86	21.65	20.96	15.23	28.64	15.28	19.40	23.54	16.03	17.73

Table A3: **Per-image targeted attack** on artificial ground-truth - This attack is described in **Per-image targeted attack** paragraph of Section 4.3 in the main paper and the qualitative results are in Figure A5 of supplementary. Our mAP before attack is approximately zero for all targets because we switch the ground truth labels during evaluation. We see an average increase in mAP of around 18 points. This means our adversarial patch successfully switches the detections of quite a few ground truth boxes to the target class. Note that this attack is more challenging than the blindness attack.

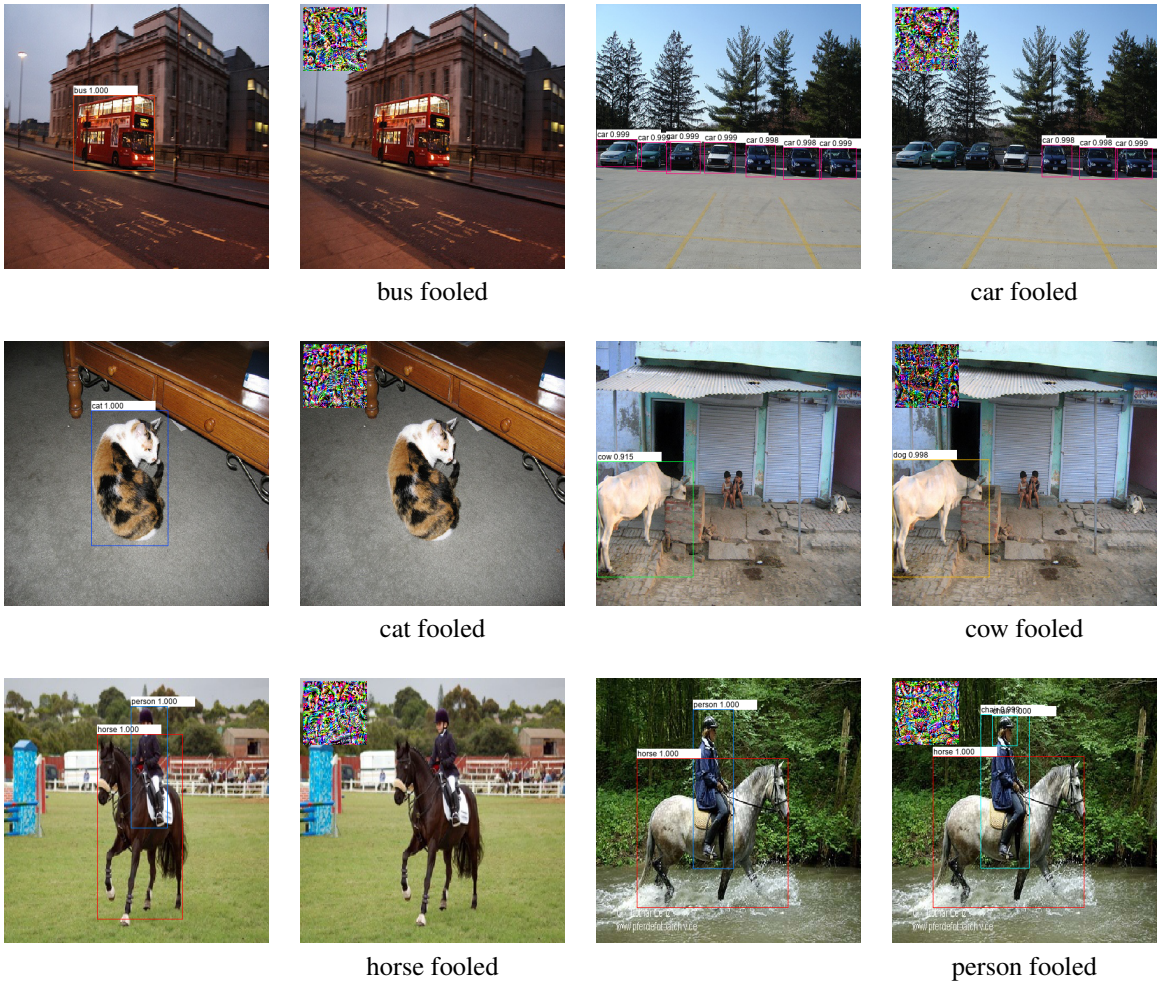

Figure A2: **Per-image blindness attack fooling results -** Additional results showcasing the fooling in **Per-image blindness** attack described in Section 3.1. These are similar to the fooling results in Figure 1 of the main paper. For every pair of columns, the left one is the original image and the right one is the attacked image. The attacked category is written below each example. A failure case is the right image of Row 1, where three out of seven instances of "cars" are detected correctly after attack.

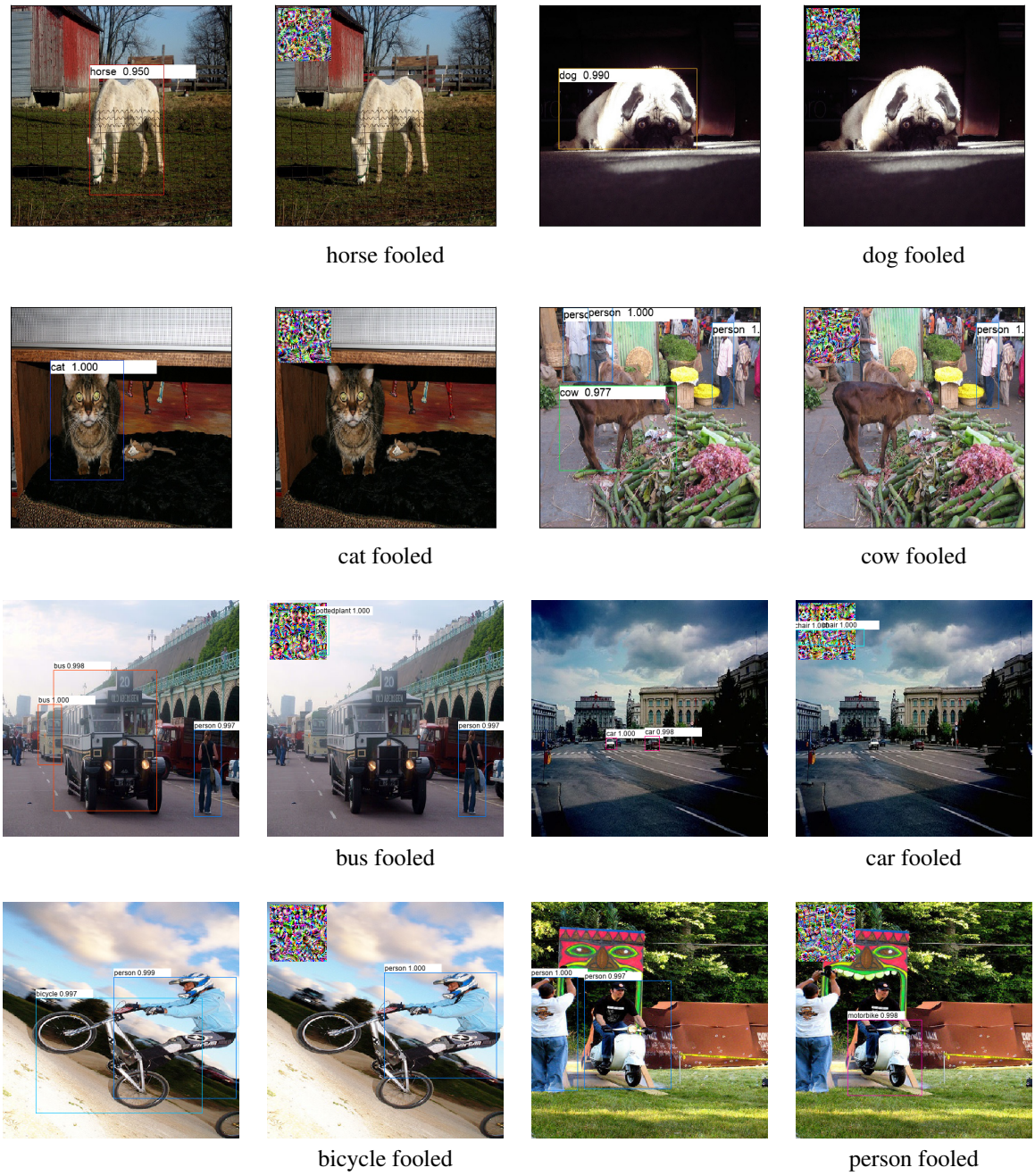

Figure A3: **Universal patch blindness attack -** Additional results showcasing **Universal patch blindness attack** described in Section 3.1 .These are similar to the fooling results in Figure 4 of main paper. For every pair of columns, the left one is the original image and the right one is the attacked image. The patch is always on the top-left corner. The attacked category is written below each example.

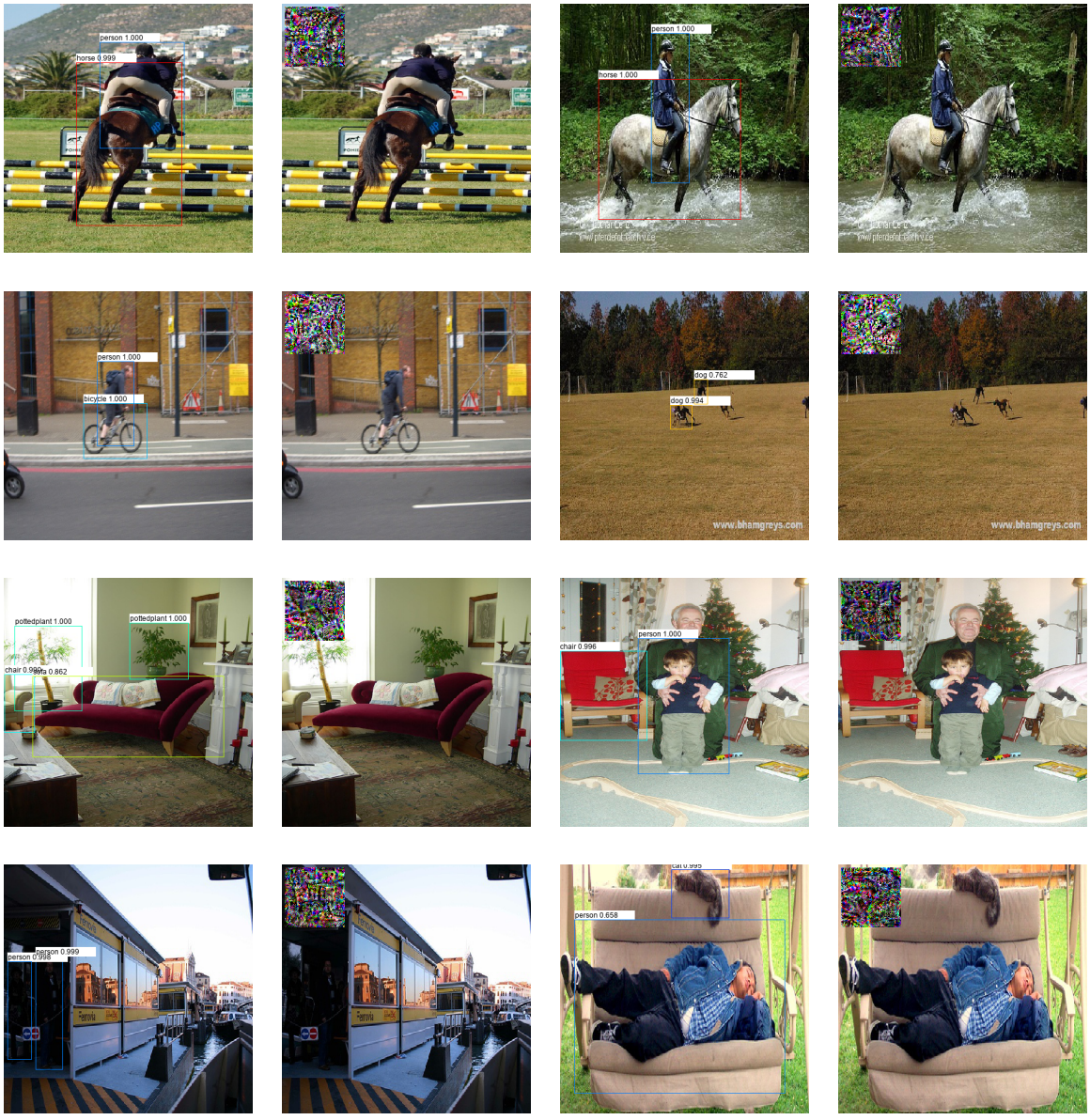

Figure A4: **Objectness attack -** This attack is described in **Per-image objectness attack** paragraph of Section 4.3 in the main paper and the quantitative results are in Table A2 of supplementary. For every pair of columns, the left one is the original image and the right one is the attacked image. Note that this attack is class agnostic.

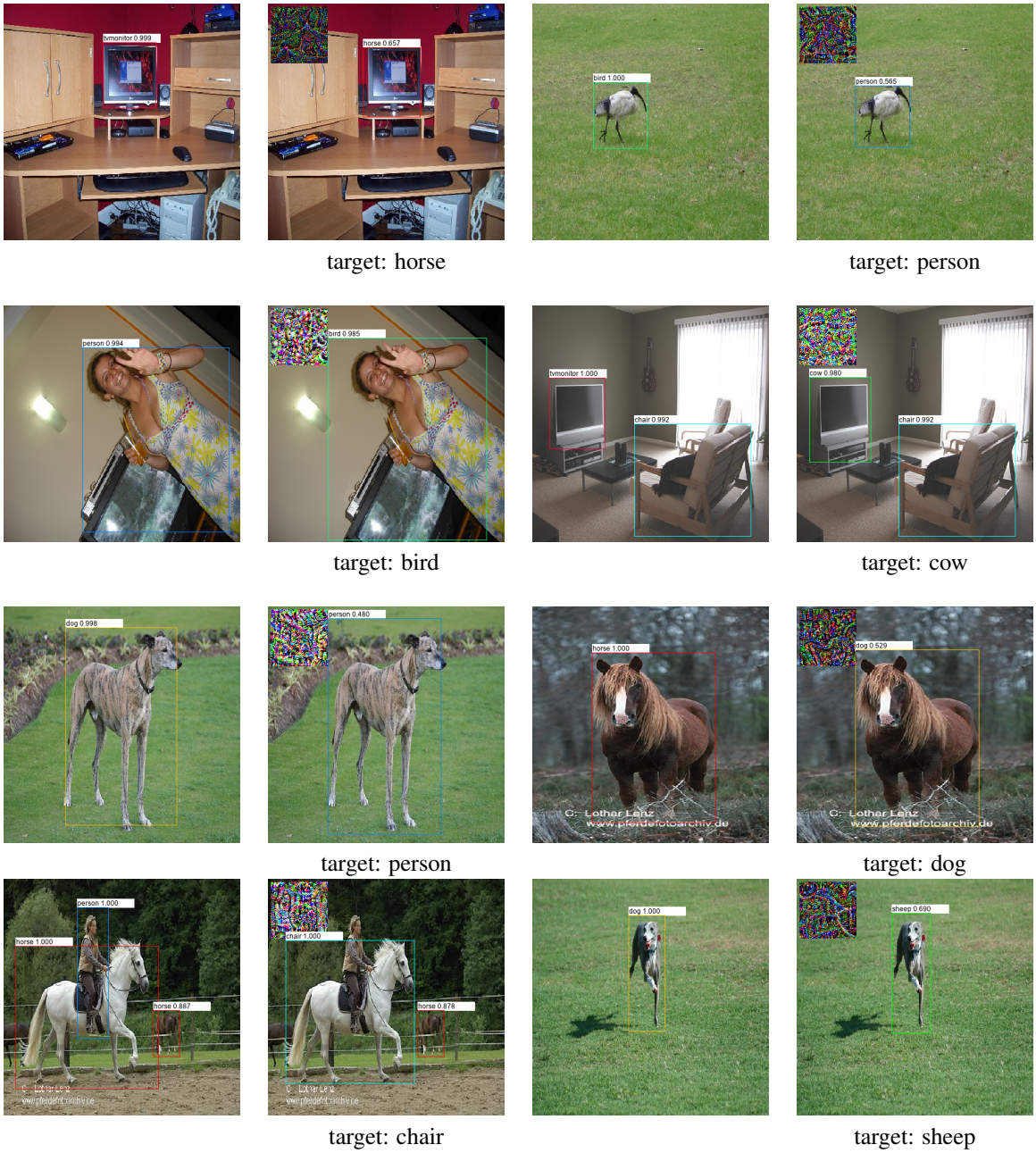

Figure A5: **Per-image targeted attack -** This attack is described in **Per-image targeted attack** paragraph of Section 4.3 in the main paper and the quantitative results are in Table A3 of supplementary. We attack the model to change the label of all objects to the target category. For every pair of columns, the left one is the original image and the right one is the attacked image. The target category is written below each example. As failure cases, a "horse" instance on the left image of Row 4 and a "chair" instance on the right image of Row 2 are still detected correctly.