Randomized Attack

Analyze attack trends

Table 1: We run a logistic model regressing success against detection models, split by attack, in the randomized attack experiment. Both vanishing and mislabeling attacks obtain higher success on 1-stage (YOLOv3, SSD) than 2-stage (Faster R-CNN, Cascade R-CNN) detectors. However, the 1-stage RetinaNet is as resilient as 2-stage detectors. Table headers are explained in Appendix **??**.

Group	Regression							
Attack	term	sig	estimate	std.error	statistic	p.value	conf.low	conf.high
Vanishing	YOLOv3		0.000					
	SSD	*	-0.315	0.053	-5.956	0.000	-0.419	-0.211
	RetinaNet	*	-1.725	0.075	-22.889	0.000	-1.875	-1.579
	Faster R-CNN	*	-2.511	0.102	-24.732	0.000	-2.715	-2.317
	Cascade R-CNN	*	-1.953	0.082	-23.914	0.000	-2.116	-1.796
Mislabeling	YOLOv3		0.000					
	SSD		-0.051	0.068	-0.751	0.453	-0.185	0.083
	RetinaNet	*	-2.173	0.135	-16.124	0.000	-2.446	-1.917
	Faster R-CNN	*	-2.939	0.189	-15.521	0.000	-3.332	-2.587
	Cascade R-CNN	*	-1.959	0.123	-15.888	0.000	-2.207	-1.723
Untargeted	YOLOv3		0.000					
	SSD	*	0.587	0.079	7.460	0.000	0.433	0.742
	RetinaNet		0.038	0.087	0.433	0.665	-0.132	0.208
	Faster R-CNN	*	-0.319	0.094	-3.389	0.001	-0.504	-0.135
	Cascade R-CNN	*	-0.488	0.098	-4.954	0.000	-0.682	-0.296

Table 2: We run a logistic model regressing success against attacks, split by detection models in the randomized attack experiment. Targeted attacks obtain higher success than untargeted attacks on YOLOv3 only; within targeted attacks, vanishing attacks obtain higher success than mislabeling attacks on all models.. Table headers are explained in Appendix **??**.

Group	Regression							
Model	term	sig	estimate	std.error	statistic	p.value	conf.low	conf.high
YOLOv3	Vanishing		0.000					
	Mislabeling	*	-0.928	0.060	-15.542	0.000	-1.046	-0.812
	Untargeted	*	-1.561	0.071	-21.871	0.000	-1.703	-1.423

SSD	Vanishing		0.000					
	Mislabeling	*	-0.665	0.062	-10.658	0.000	-0.787	-0.543
	Untargeted	*	-0.660	0.062	-10.594	0.000	-0.783	-0.538
RetinaNet	Vanishing		0.000					
	Mislabeling	*	-1.376	0.142	-9.667	0.000	-1.663	-1.104
	Untargeted	*	0.201	0.090	2.237	0.025	0.025	0.378
Faster R-CNN	Vanishing		0.000					
	Mislabeling	*	-1.356	0.206	-6.571	0.000	-1.778	-0.966
	Untargeted	*	0.631	0.119	5.317	0.000	0.401	0.866
Cascade R-CNN	Vanishing		0.000					
	Mislabeling	*	-0.934	0.135	-6.901	0.000	-1.204	-0.673
	Untargeted		-0.096	0.106	-0.901	0.367	-0.304	0.112

Table 3: We run a logistic model regressing success against log(attack iterations) in the randomized attack experiment. Success rates increase with attack iterations for all models and attacks. Table headers are explained in Appendix **??**.

Group	Regression							
Attack	term	sig	estimate	std.error	statistic	p.value	conf.low	conf.high
YOLOv3								
Vanishing	log(iterations)	*	0.797	0.027	29.736	0	0.745	0.850
Mislabeling	log(iterations)	*	1.097	0.051	21.572	0	1.000	1.199
Untargeted	log(iterations)	*	0.347	0.036	9.615	0	0.277	0.419
SSD								
Vanishing	log(iterations)	*	0.852	0.032	26.573	0	0.790	0.915
Mislabeling	log(iterations)	*	0.922	0.044	20.885	0	0.837	1.010
Untargeted	log(iterations)	*	0.483	0.031	15.652	0	0.423	0.544
RetinaNet								
Vanishing	log(iterations)	*	0.880	0.062	14.229	0	0.762	1.005
Mislabeling	log(iterations)	*	0.903	0.115	7.855	0	0.688	1.139
Untargeted	log(iterations)	*	0.627	0.046	13.591	0	0.538	0.719
Faster R-CNN								
Vanishing	log(iterations)	*	0.707	0.082	8.664	0	0.552	0.872
Mislabeling	log(iterations)	*	0.975	0.191	5.111	0	0.627	1.378
Untargeted	log(iterations)	*	0.483	0.049	9.938	0	0.389	0.580
Cascade R-CNN								
Vanishing	log(iterations)	*	0.738	0.062	11.832	0	0.619	0.863
Mislabeling	log(iterations)	*	1.248	0.149	8.395	0	0.972	1.556
Untargeted	log(iterations)	*	0.450	0.050	9.040	0	0.354	0.549

Analyze individual cases

Table 4: We run a logistic model regressing success against target confidence in the randomized attack experiment. Lower target confidence significantly increases success rates for all models and attacks. Table headers are explained in Appendix **??**.

Group	Regression							
Attack	term	sig	estimate	std.error	statistic	p.value	conf.low	conf.high
YOLOv3								
Vanishing	confidence	*	-1.017	0.162	-6.286	0	-1.334	-0.700
Mislabeling	confidence	*	-2.470	0.171	-14.445	0	-2.806	-2.136
Untargeted	confidence	*	-4.845	0.313	-15.476	0	-5.470	-4.241
SSD								
Vanishing	confidence	*	-1.505	0.163	-9.251	0	-1.825	-1.187
Mislabeling	confidence	*	-2.212	0.185	-11.970	0	-2.576	-1.852
Untargeted	confidence	*	-2.889	0.215	-13.462	0	-3.313	-2.471
RetinaNet								
Vanishing	confidence	*	-2.203	0.360	-6.124	0	-2.918	-1.507
Mislabeling	confidence	*	-4.778	0.682	-7.002	0	-6.173	-3.491
Untargeted	confidence	*	-5.816	0.439	-13.241	0	-6.701	-4.977
Faster R-CNN								
Vanishing	confidence	*	-3.442	0.390	-8.814	0	-4.213	-2.680
Mislabeling	confidence	*	-5.244	0.560	-9.361	0	-6.383	-4.178
Untargeted	confidence	*	-4.522	0.313	-14.433	0	-5.144	-3.915
Cascade R-CNN								
Vanishing	confidence	*	-1.647	0.303	-5.433	0	-2.237	-1.047
Mislabeling	confidence	*	-3.146	0.412	-7.635	0	-3.960	-2.341
Untargeted	confidence	*	-3.811	0.326	-11.692	0	-4.456	-3.177

Table 5: We run a logistic model regressing success against perturb-target distance (relative to image width/height) and perturb box size (relative to image width/height) in the randomized attack experiment. Larger perturb objects significantly increase success rates for all models and attacks, except for mislabeling attack on Faster R-CNN, after controlling for perturb-target distances; shorter perturb-target distances significantly increase success rates for all models and attacks, after controlling for perturb object sizes. Table headers are explained in Appendix **??**.

Group	Regression							
Attack	term	sig	estimate	std.error	statistic	p.value	conf.low	conf.high
YOLOv3								
Vanishing	distance	*	-8.536	0.694	-12.292	0.000	-9.929	-7.207
	size	*	26.831	1.719	15.610	0.000	23.555	30.294
	distance * size	*	-79.933	8.924	-8.957	0.000	-97.839	-62.847
Mislabeling	distance	*	-8.473	0.615	-13.778	0.000	-9.707	-7.297
	size	*	10.991	0.956	11.500	0.000	9.169	12.915
	distance * size	*	-24.117	5.917	-4.076	0.000	-35.972	-12.770

Untargeted	distance	*	-15.869	1.366	-11.614	0.000	-18.640	-13.284
	size		0.308	0.704	0.437	0.662	-1.087	1.678
	distance * size	*	39.532	6.522	6.061	0.000	26.743	52.347
SSD								
Vanishing	distance	*	-18.433	1.159	-15.903	0.000	-20.766	-16.222
	size	*	7.274	0.813	8.948	0.000	5.728	8.915
	distance * size		7.663	6.391	1.199	0.231	-5.139	19.931
Mislabeling	distance	*	-19.702	1.311	-15.023	0.000	-22.349	-17.208
	size	*	3.384	0.612	5.531	0.000	2.217	4.617
	distance * size	*	23.987	6.040	3.971	0.000	11.954	35.660
Untargeted	distance	*	-21.725	1.544	-14.069	0.000	-24.852	-18.799
	size	*	1.389	0.545	2.547	0.011	0.336	2.478
	distance * size	*	34.171	6.423	5.320	0.000	21.425	46.643
RetinaNet								
Vanishing	distance	*	-35.303	3.249	-10.864	0.000	-41.932	-29.191
	size	*	2.317	0.695	3.334	0.001	0.993	3.717
	distance * size	*	46.975	11.215	4.189	0.000	24.285	68.263
Mislabeling	distance	*	-49.847	6.486	-7.685	0.000	-63.277	-37.849
	size		1.056	1.187	0.889	0.374	-1.244	3.427
	distance * size		37.912	25.512	1.486	0.137	-15.784	84.709
Untargeted	distance	*	-13.895	1.412	-9.843	0.000	-16.788	-11.254
	size	*	2.989	0.539	5.544	0.000	1.938	4.054
	distance * size	*	28.072	5.111	5.493	0.000	18.127	38.241
Faster R-CNN								
Vanishing	distance	*	-21.030	3.204	-6.564	0.000	-27.739	-15.185
	size	*	6.096	1.228	4.962	0.000	3.747	8.571
	distance * size	*	-83.474	28.510	-2.928	0.003	-144.255	-31.915
Mislabeling	distance	*	-17.846	3.240	-5.507	0.000	-24.720	-12.034
	size		1.205	1.719	0.701	0.483	-2.408	4.397
	distance * size		-54.135	39.695	-1.364	0.173	-142.163	14.635
Untargeted	distance	*	-19.078	1.789	-10.665	0.000	-22.746	-15.729
	size		-0.274	0.719	-0.381	0.703	-1.711	1.113
	distance * size	*	61.468	6.966	8.824	0.000	48.369	75.700
Cascade R-CNN								
Vanishing	distance	*	-32.490	4.066	-7.991	0.000	-40.976	-25.029
	size	*	7.513	0.966	7.779	0.000	5.711	9.508
	distance * size	*	-106.218	31.092	-3.416	0.001	-172.083	-49.911
Mislabeling	distance	*	-27.708	4.732	-5.856	0.000	-37.836	-19.260
	size	*	4.898	0.797	6.146	0.000	3.354	6.485
	distance * size		-49.344	27.328	-1.806	0.071	-107.414	-0.192
Untargeted	distance	*	-22.497	2.467	-9.120	0.000	-27.587	-17.915

size	*	2.113	0.648	3.258	0.001	0.833	3.381
distance * size		5.873	11.482	0.512	0.609	-18.022	27.276

Table 6: We run a logistic model regressing success against mean COCO accuracy for the target class, with target confidence as covariate, in the randomized attack experiment. The results are mixed after controlling for target class confidence and the relatively large interaction terms make interpretation challenging. Table headers are explained in Appendix **??**.

Group	Regression							
Attack	term	sig	estimate	std.error	statistic	p.value	conf.low	conf.high
YOLOv3								
Vanishing	accuracy		0.842	0.747	1.127	0.260	-0.619	2.313
	confidence		0.368	0.671	0.548	0.584	-0.945	1.688
	accuracy * confidence	*	-2.046	1.007	-2.031	0.042	-4.026	-0.076
Mislabeling	accuracy		1.231	0.754	1.631	0.103	-0.247	2.712
	confidence		-0.139	0.700	-0.198	0.843	-1.514	1.234
	accuracy * confidence	*	-3.481	1.065	-3.270	0.001	-5.571	-1.396
Untargeted	accuracy		1.941	1.117	1.737	0.082	-0.240	4.143
	confidence		-1.715	1.230	-1.394	0.163	-4.155	0.671
	accuracy * confidence	*	-4.861	1.913	-2.541	0.011	-8.612	-1.112
SSD								
Vanishing	accuracy	*	3.774	0.582	6.485	0.000	2.640	4.923
	confidence	*	2.184	0.491	4.451	0.000	1.226	3.150
	accuracy * confidence	*	-6.655	0.854	-7.789	0.000	-8.340	-4.990
Mislabeling	accuracy	*	4.376	0.630	6.950	0.000	3.148	5.618
	confidence	*	2.449	0.538	4.550	0.000	1.395	3.506
	accuracy * confidence	*	-8.650	0.976	-8.864	0.000	-10.573	-6.746
Untargeted	accuracy	*	3.376	0.681	4.955	0.000	2.047	4.720
	confidence		0.423	0.626	0.677	0.499	-0.809	1.646
	accuracy * confidence	*	-6.063	1.106	-5.480	0.000	-8.239	-3.902
RetinaNet								
Vanishing	accuracy	*	3.267	1.389	2.353	0.019	0.576	6.018
	confidence		-0.776	2.077	-0.374	0.709	-4.879	3.260
	accuracy * confidence		-2.512	2.651	-0.948	0.343	-7.702	2.686
Mislabeling	accuracy	*	10.978	2.731	4.020	0.000	5.683	16.358
	confidence		3.473	4.602	0.755	0.450	-5.826	12.146
	accuracy * confidence	*	-11.692	5.707	-2.049	0.040	-22.608	-0.344
Untargeted	accuracy	*	3.553	1.292	2.751	0.006	1.029	6.093
	confidence		0.863	1.920	0.449	0.653	-2.964	4.566
	accuracy * confidence	*	-9.351	2.760	-3.388	0.001	-14.760	-3.935
Faster R-CNN								
Vanishing	accuracy		-1.752	1.802	-0.973	0.331	-5.202	1.874

	confidence	*	-6.201	2.110	-2.939	0.003	-10.372	-2.093
	accuracy * confidence		3.626	2.762	1.313	0.189	-1.797	9.030
Mislabeling	accuracy		2.740	2.469	1.110	0.267	-1.989	7.689
	confidence		-3.313	3.126	-1.060	0.289	-9.642	2.613
	accuracy * confidence		-2.724	4.126	-0.660	0.509	-10.668	5.473
Untargeted	accuracy		1.841	1.415	1.301	0.193	-0.897	4.655
	confidence		-2.543	1.607	-1.583	0.114	-5.733	0.572
	accuracy * confidence		-2.728	2.162	-1.262	0.207	-6.949	1.529
Cascade R-CNN								
Vanishing	accuracy	*	-4.247	1.491	-2.848	0.004	-7.156	-1.298
	confidence	*	-4.563	1.413	-3.229	0.001	-7.328	-1.779
	accuracy * confidence	*	4.330	1.956	2.214	0.027	0.483	8.158
Mislabeling	accuracy	*	-4.568	1.806	-2.530	0.011	-8.081	-0.985
	confidence	*	-6.823	1.939	-3.519	0.000	-10.663	-3.046
	accuracy * confidence	*	5.322	2.638	2.017	0.044	0.152	10.503
Untargeted	accuracy		-0.017	1.423	-0.012	0.990	-2.791	2.794
	confidence		-1.750	1.449	-1.207	0.227	-4.607	1.083
	accuracy * confidence		-2.732	2.037	-1.341	0.180	-6.726	1.265

Table 7: We run a logistic model regressing success against log(intended class probability) for the mislabeling attack, with predicted class's confidence as covariate, in the randomized attack experiment. Intended class probability does not predict success rates after controlling for target class confidence, except for RetinaNet. Table headers are explained in Appendix **??**.

Group	Regression							
Model	term	sig	estimate	std.error	statistic	p.value	conf.low	conf.high
Mislabeling								
YOLOv3	log(probability)	*	-0.183	0.042	-4.344	0.000	-0.266	-0.101
	confidence		0.119	0.522	0.227	0.820	-0.904	1.143
	log(probability) * confidence	*	0.317	0.062	5.140	0.000	0.196	0.438
SSD	log(probability)	*	0.196	0.055	3.574	0.000	0.089	0.304
	confidence	*	-1.546	0.503	-3.071	0.002	-2.532	-0.558
	log(probability) * confidence		0.011	0.078	0.146	0.884	-0.141	0.166
RetinaNet	log(probability)	*	1.117	0.373	2.993	0.003	0.374	1.837
	confidence	*	-8.002	1.997	-4.006	0.000	-11.970	-4.136
	log(probability) * confidence		-1.384	0.757	-1.828	0.067	-2.822	0.145
Faster R-CNN	log(probability)		0.158	0.120	1.314	0.189	-0.080	0.393
	confidence	*	-7.667	1.544	-4.964	0.000	-10.765	-4.692
	log(probability) * confidence		-0.330	0.196	-1.684	0.092	-0.709	0.061
Cascade R-CNN	log(probability)		0.096	0.111	0.864	0.388	-0.123	0.313
	confidence	*	-2.499	1.024	-2.440	0.015	-4.493	-0.470

log(probability) * confidence	0.020	0.153	0.133	0.894	-0.275	0.326

Table 8: We run a logistic model regressing success against target IOU for the untargeted attack in the randomized attack experiment. Target IOU for the untargeted attack increases success rates on all models. Table headers are explained in Appendix **??**.

Group	Regression							
Model	term	sig	estimate	std.error	statistic	p.value	conf.low	conf.high
Untargeted								
YOLOv3	bbox_iou_eval	*	-3.194	0.351	-9.098	0	-3.878	-2.501
SSD	bbox_iou_eval	*	-2.747	0.288	-9.539	0	-3.309	-2.180
RetinaNet	bbox_iou_eval	*	-3.085	0.328	-9.402	0	-3.725	-2.438
Faster R-CNN	bbox_iou_eval	*	-2.020	0.374	-5.403	0	-2.745	-1.278
Cascade R-CNN	bbox_iou_eval	*	-2.895	0.364	-7.953	0	-3.606	-2.177

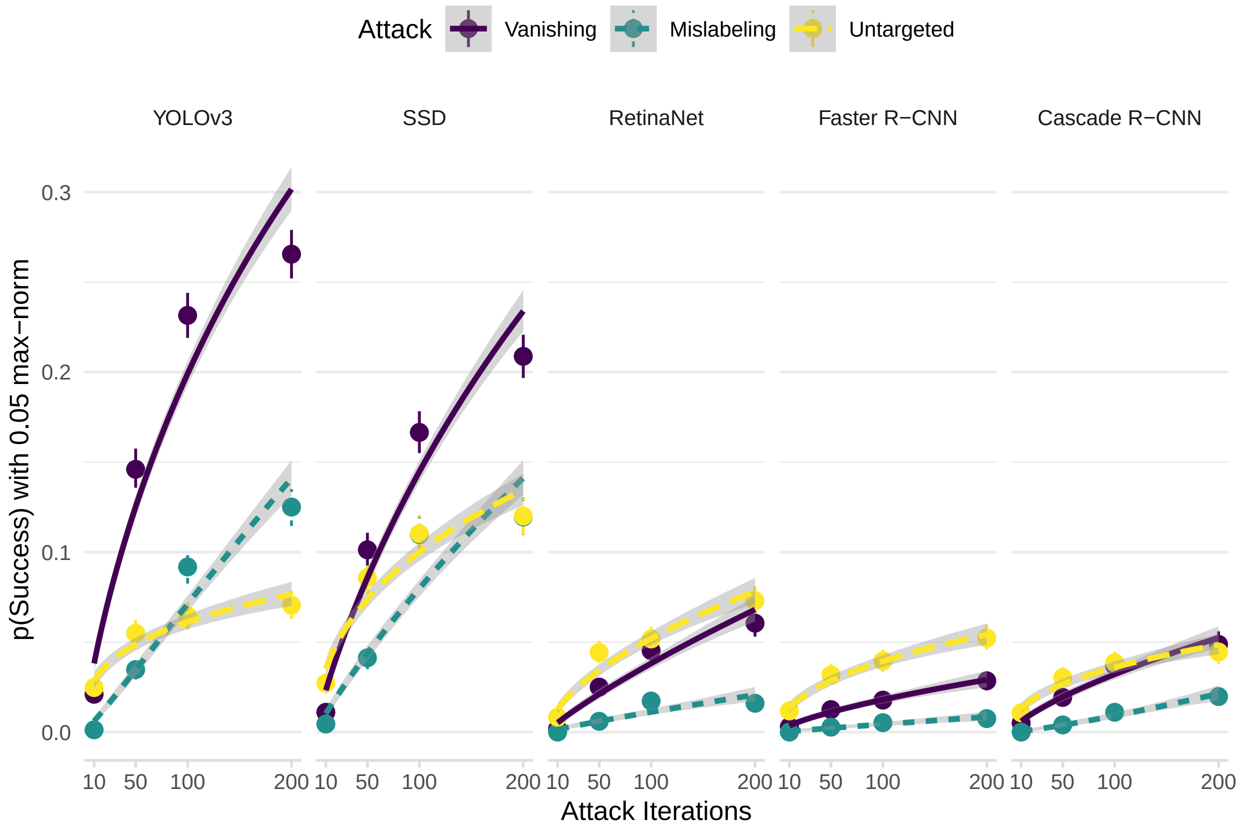

Figure 1: **Intent obfuscating attack is feasible for all models and attacks:** We conduct a randomized experiment by resampling COCO images, and within those images randomly sampling correctly predicted target and perturb objects. Then we distort the perturb objects to disrupt the target objects varying the attack iterations. The binned summaries and regression trendlines graph success proportion against attack iterations in the randomized attack experiment. Errors are 95% confidence intervals. and every point aggregates success over 4,000 images. Targeted vanishing and mislabeling attacks obtain significantly greater success on the 1-stage YOLOv3 and SSD than the 2-stage Faster R-CNN and Cascade R-CNN detectors. However, the 1-stage RetinaNet is as resilient as the 2-stage detectors. Additionally, targeted attacks are significantly more successful than untargeted attacks on YOLOv3 and SSD, but the pattern does not exist for RetinaNet, Faster R-CNN, and Cascade R-CNN. Within targeted attacks, vanishing achieves significantly greater success than mislabeling attack on all models except YOLOv3. Moreover, success rates significantly increase with larger attack iterations. Significance is determined at $\alpha < 0.05$ using a Wald z-test on the logistic estimates. Full details are given in Section **??**.

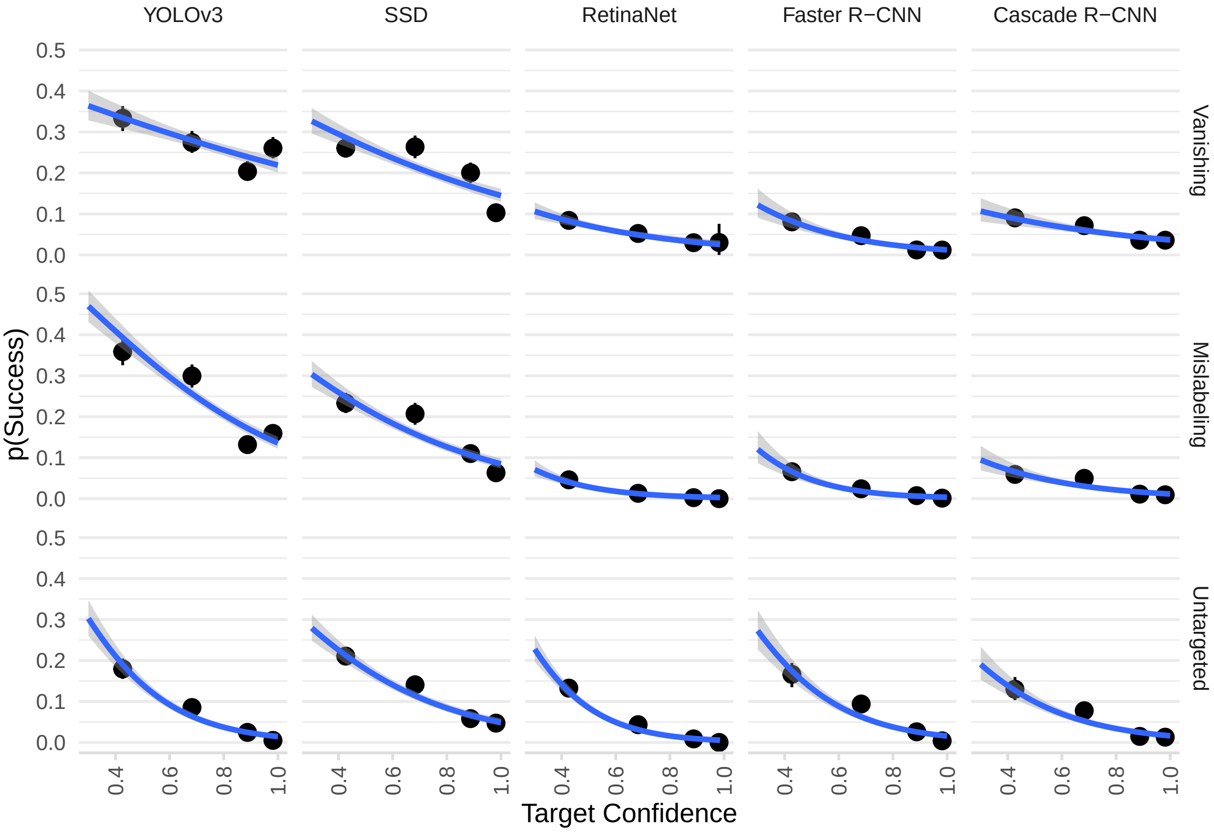

Figure 2: **Lower target confidence significantly increases success rates for all models and attacks:** The binned summaries and regression trendlines graph success proportion against target confidence in the randomized attack experiment. Bins are split into quantiles. Errors are 95% confidence intervals.

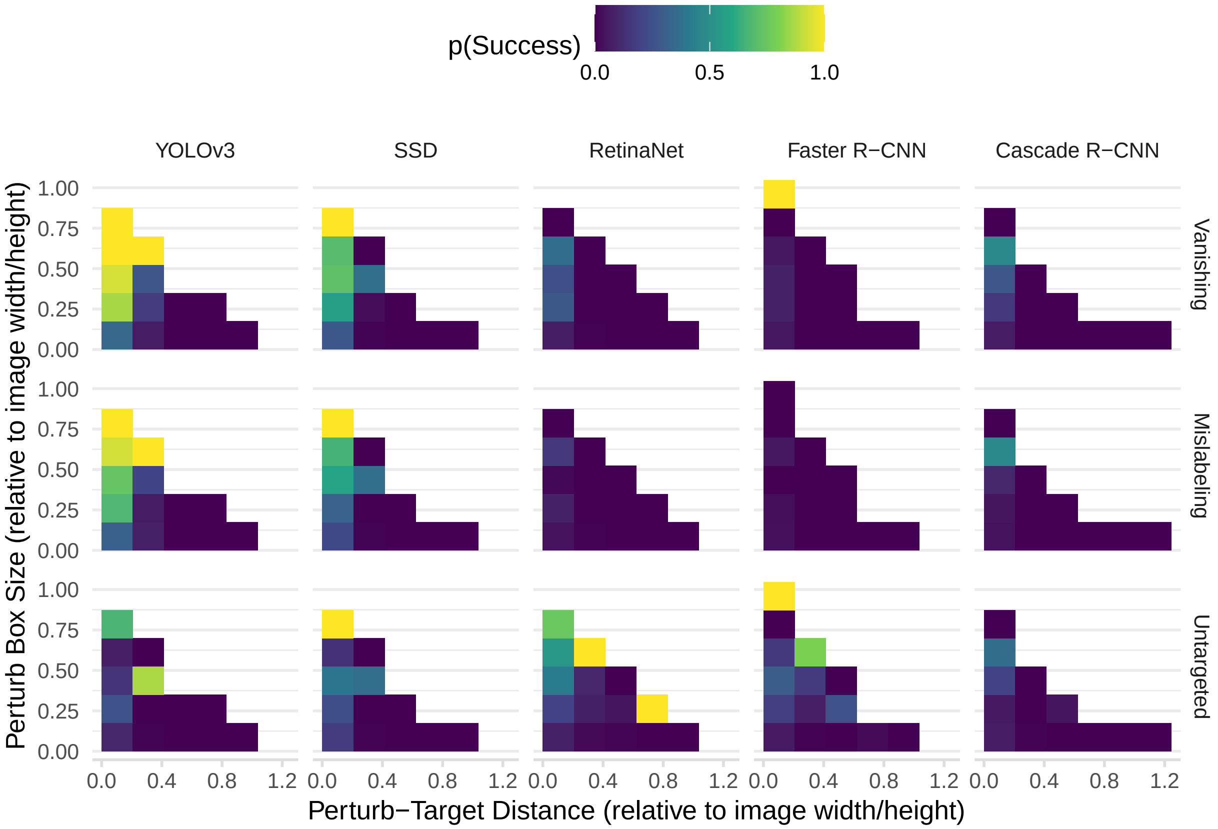

Figure 3: **Larger perturb objects significantly increase success rates for all models and attacks, except for mislabeling attack on Faster R-CNN, after controlling for perturb-target distances; Shorter perturb-target distances significantly increase success rates for all models and attacks, after controlling for perturb object sizes:** The binned summaries graph success proportion against perturb-target distance (relative to image width/height) and perturb box size (relative to image width/height) in the randomized attack experiment.

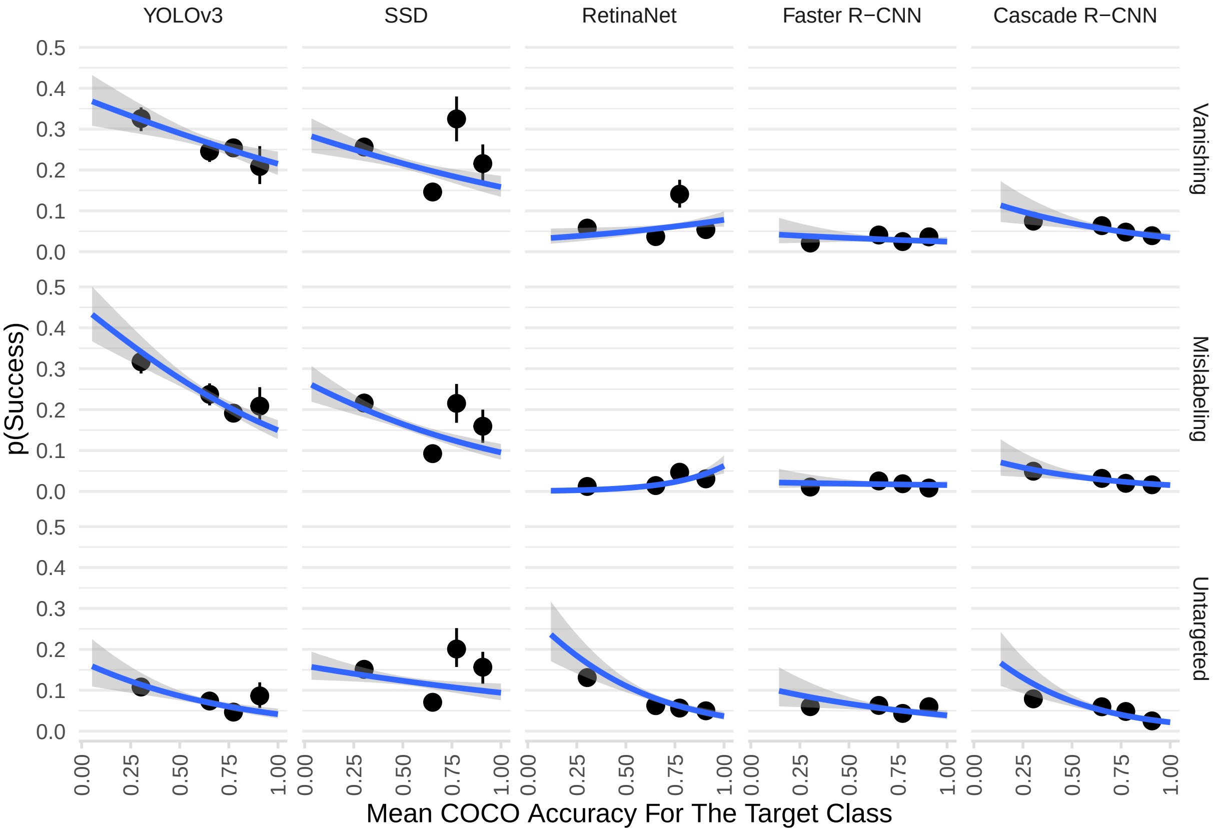

Figure 4: **Although higher mean COCO accuracy for the target class seem to decrease success rates, the results are mixed after controlling for target class confidence (Table 6):** The binned summaries and regression trendlines graph success proportion against mean COCO accuracy for the target class in the randomized attack experiment. Bins are split into quantiles. Errors are 95% confidence intervals.

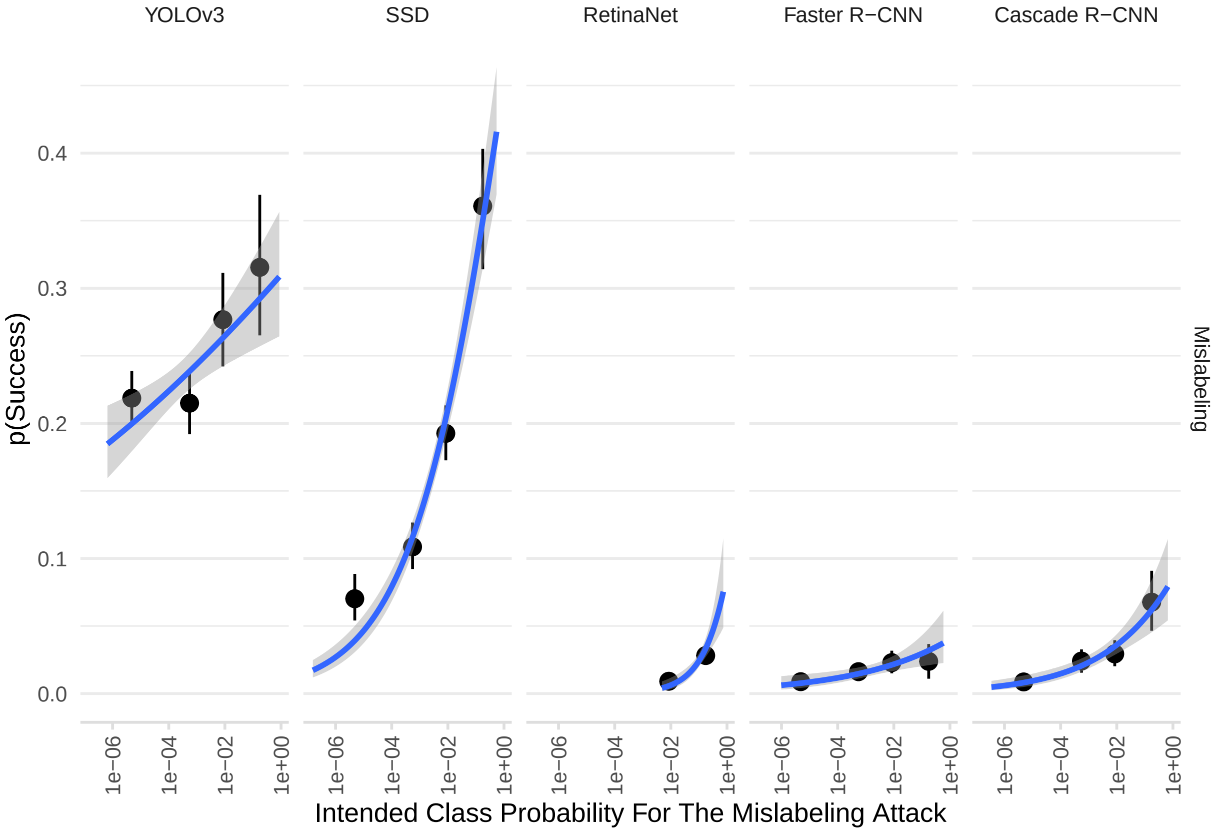

Figure 5: **Although intended class probability seem to increase success rates for the mislabeling attack, it does not predict success rates after controlling for target class confidence, except for RetinaNet (Table 7):** The binned summaries and regression trendlines graph success proportion against intended class probability in the randomized attack experiment. Bins are split into quantiles. Errors are 95% confidence intervals.

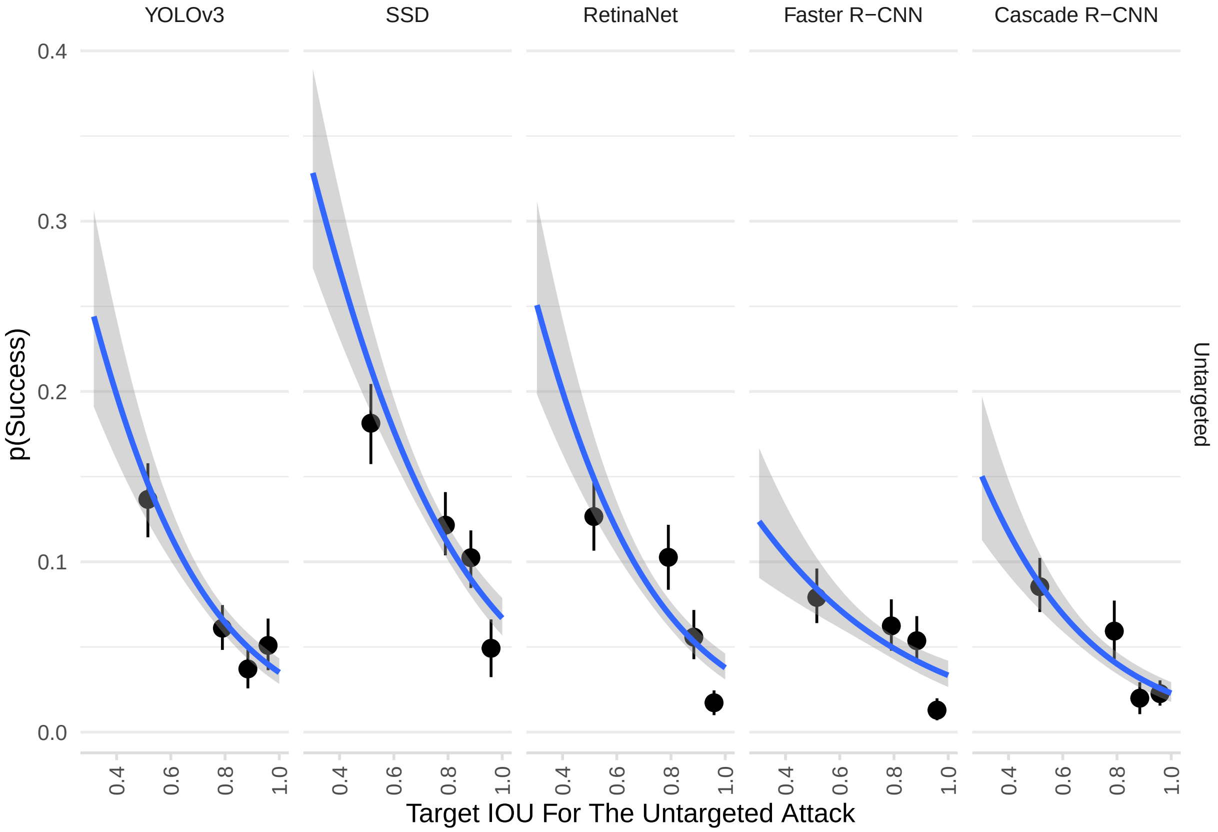

Figure 6: **Target IOU for the untargeted attack increases success rates on all models:** The binned summaries and regression trendlines graph success proportion against target IOU for the untargeted attack in the randomized attack experiment. Bins are split into quantiles. Errors are 95% confidence intervals.