Supplementary Materials

The Attack Generator: A Systematic Approach Towards Constructing Adversarial Attacks

In this document, we provide additional observations and results in support of the primary manuscript. The supplementary materials are divided into two parts: Appendix A is a comprehensive table which further illustrates the applicability of the attack generator concept. Here, we analyze existing adversarial attacks with respect to their building blocks, in particular their specificity, perturbation scope and perturbation imperceptibility measures. Appendix B shows further results of the two new semantic segmentation attacks. The first attack tries to enlarge the pedestrian class (confusion) while being imperceptible with respect to the L^∞-norm. The second attack erases the pedestrian class (dynamic target) by utilizing flow field perturbations (attention-based imperceptibility).

Due to the size of the table, Appendix A starts on the following page.

Attack Name	**Opt. Problem**	**Opt. Method**	**Specificity** $\mathcal{M}_{sp}(\delta)$	**Scope** $\mathcal{M}_{sc}(\delta)$	**Imperceptibility** $\mathcal{M}_{im}(\delta)$
FGSM [**?**]	$\min_{\delta} -l(F(x+\delta), y^x)$ $s.t.\ \|\delta\|_\infty \leq \epsilon,$ with y^x true label of image x and $l(\cdot,\cdot)$ loss function of neural network.	*First Order Method* Use scaled sign of gradient.	*Untargeted* Measure: $-l(F(x+\delta), y^x)$	*Individual Scope* Only considering single image. Measure: $\mathrm{id}(\mathcal{M}_{sp}(\delta))$	*L^p-based Imperceptibility* Measure: $\|\delta\|_\infty$
L-BFGS [**?**]	$\min_{\delta} l(F(x+\delta), y^x_{tar}) + \|\delta\|_2$ $s.t.\ x+\delta \in [0,1]^m$ with y^x_{tar} adversarial target.	*Second Order Method* Box-constrained L-BFGS is a quasi-Newton method.	*Targeted* y^x_{tar} is desired outcome. Measure: $l(F(x+\delta), y^x_{tar})$	*Individual Scope* Only considering single image. Measure: $\mathrm{id}(\mathcal{M}_{sp}(\delta))$	*L^p-based Imperceptibility* Measure: $\|\delta\|_2$
PGD [**?**]	$\min_{\delta} -l(F(x+\delta), y^x)$ $s.t.\ \|\delta\|_\infty \leq \epsilon,$ with y^x true label of x and $l(\cdot,\cdot)$ loss function of neural network.	*First Order Method* Multi-step gradient descent combined with projections.	*Untargeted* Measure: $-l(F(x+\delta), y^x)$	*Individual Scope* Only considering single image. Measure: $\mathrm{id}(\mathcal{M}_{sp}(\delta))$	*L^p-based Imperceptibility* Measure: $\|\delta\|_\infty$
Boundary Attack [**?**]	$\min_{\delta} \|\delta\|_2^2$ $s.t.\ c(\delta) = 1$ with $c(\cdot) \in \{0,1\}$ adversarial criterion for data point x.	*Evolution & Random Sampling* Method is initialized from an adversarial point and then random walk along decision boundary.	*Untargeted / Targeted* Depends on choice of adversarial criterion (and initialization point). Measure: $c(x+\delta)$	*Individual Scope* Only considering single image. Measure $\mathrm{id}(\mathcal{M}_{sp}(\delta))$	*L^p-based Imperceptibility* Measure: $\|\delta\|_2^2$
EOT [**?**]	$\min_{\delta} -\mathbb{E}_{t\sim T}[logP(y^x_{tar} \mid t(x+\delta))]$ $s.t.\ \mathbb{E}_{t\sim T}[d(t(x), t(x+\delta))] \leq \epsilon,$ $x \in [0,1]^m$ with distribution T of transformation functions, distance function $d(\cdot,\cdot)$ and y^x_{tar} target.	*First Order Method* Use projected gradient descent.	*Targeted* Measure also takes transformation t as input. Measure: $-logP(y^x_{tar} \mid t(x+\delta))$	*Contextual* Consider one image x and relevant transformations of x. Measure: $\mathbb{E}_{t\sim T}[\mathcal{M}_{sp}(t,\delta))]$	*L^p-based / Attention-based Imperceptibility*: Depends on choice of distance. Measure: $d(\cdot,\cdot)$

Table 1. Building block analysis of existing adversarial attacks.

B. Sample Results

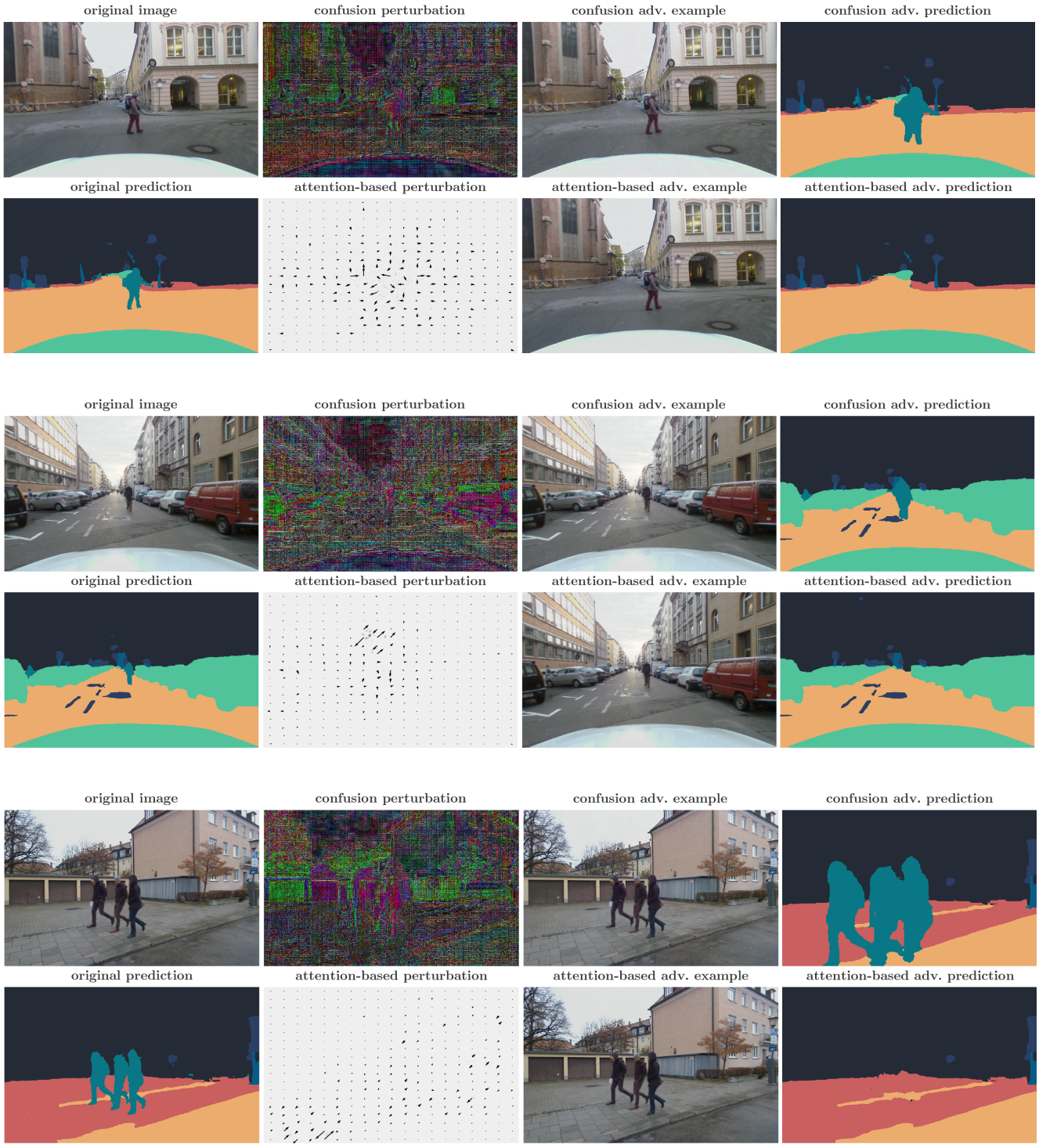

Figure 1. Sample results of adapted semantic segmentation attacks with confusion target or attention-based imperceptibility. Attack input is a self-trained ICNet and a single image (N=1, ϵ = 15).