A. Dense Adversary Generation attack

In this section, we provide the algorithm of the Dense Adversary Generation (DAG) attack from [39]. The main difference with the original algorithm proposed in [39] is the stopping criterion based on the pixel success rate in steps 9 to 11. In the original method published in [39], the attack is supposed to stop once all the constraints are satisfied (*i.e.* all pixels in the mask $\boldsymbol{m}$ are adversarial). However, since this criterion is rarely satisfied, even after hundreds of iterations on a dataset like Cityscapes, the actual implementation made available in https://github.com/cihangxie/DAG uses the stopping criterion described in Algorithm 2: the attack stops once a threshold of pixel success rate set to 99% is reached. The threshold value used is thus identical to the one used in the experiments of this paper.

Algorithm 2 DAG attack

Require: Classifier f, original image $\boldsymbol{x} \in [0,1]^{C\times H\times W}$, true or target label $\boldsymbol{y} \in \mathbb{N}^{H\times W}$, binary mask $\boldsymbol{m} \in \{0,1\}^{H\times W}$
Require: Step size η, maximum number of iterations N, threshold of pixel success rate ν
1: Initialize $\boldsymbol{\delta}^{(0)} \leftarrow \boldsymbol{0}$
2: If targeted attack: $\mu \leftarrow -1$ else $\mu \leftarrow 1$
3: **for** $t \leftarrow 1, \dots, N$ **do**
4: $\quad \tilde{\boldsymbol{x}}^{(t)} \leftarrow \mathcal{P}_{[0,1]}(\boldsymbol{x} + \boldsymbol{\delta}^{(t-1)})$ // $\in [0,1]^{C\times H\times W}$
5: $\quad \boldsymbol{z} \leftarrow f(\tilde{\boldsymbol{x}}^{(t)})$ // $\in \mathbb{R}^{K\times H\times W}$
6: $\quad$ **for** $i \leftarrow 1, \dots, d$ **do**
7: $\quad\quad \boldsymbol{\Delta z}_i = \mu(\boldsymbol{z}_{\boldsymbol{y}_i,i} - \max_{j\neq \boldsymbol{y}_i} \boldsymbol{z}_{j,i})$ // Difference of logits
8: $\quad$ **end for**
9: $\quad r \leftarrow \frac{\boldsymbol{m}^\top [\boldsymbol{\Delta z} < 0]}{\|\boldsymbol{m}\|_1}$ // Pixel success rate
10: $\quad$ **if** $r \geq \nu$ **then**
11: $\quad\quad$ **return** $\tilde{\boldsymbol{x}}^{(t)}$ // Stop the attack
12: $\quad$ **end if**
13: $\quad \mathcal{L} \leftarrow \boldsymbol{m}^\top \max\{0, \boldsymbol{\Delta z}_i\}$
14: $\quad \boldsymbol{g} \leftarrow \nabla_{\boldsymbol{\delta}} \mathcal{L}$
15: $\quad \boldsymbol{\delta}^{(t)} \leftarrow \boldsymbol{\delta}^{(t-1)} - \frac{\eta}{\|\boldsymbol{g}\|_\infty} \boldsymbol{g}$ // Normalized gradient step
16: **end for**

B. Proof of Proposition 1

Proof. Problem (12) amounts to minimizing function

$$\Phi\colon (\boldsymbol{p},\beta) \mapsto \frac{1}{2}\|\boldsymbol{p}-\boldsymbol{\delta}\|_2^2 + \lambda\beta + \iota_{[0,+\infty[^{Cd}}(\beta \boldsymbol{1}_{Cd} - \boldsymbol{p}) + \iota_{[0,+\infty[^{Cd}}(\boldsymbol{p} + \beta \boldsymbol{1}_{Cd}) + \iota_\Lambda(\boldsymbol{p}), \quad (20)$$

defined on $\mathbb{R}^{Cd} \times \mathbb{R}$. This function is convex since it is a sum of elementary convex functions of $(\boldsymbol{p}, \beta)$. It follows that its marginal function

$$\underline{\Phi}\colon \beta \mapsto \inf_{\boldsymbol{p}\in\mathbb{R}^{Cd}} \Phi(\boldsymbol{p},\beta) \quad (21)$$

is convex. Since, for any given $\beta \in [0,1]$, $\boldsymbol{p} \mapsto \Phi(\boldsymbol{p},\beta)$ is strongly convex and proper, it admits a unique minimizer $\boldsymbol{p}_\beta$. More precisely, we have, for every $\beta \in \mathbb{R}$,

$$\underline{\Phi}(\beta) = \begin{cases} \frac{1}{2}\|\boldsymbol{p}_\beta - \boldsymbol{\delta}\|_2^2 + \lambda\beta & \text{if } \beta \in [0,1] \\ +\infty & \text{otherwise.} \end{cases} \quad (22)$$

The above function admits a unique minimizer $\beta^\star \in [0,1]$ since $\boldsymbol{p}^\star = \boldsymbol{p}_{\beta^\star}$ is uniquely defined (as it is the proximity point of a proper lower-semicontinuous convex function) and $\beta^\star = \|\boldsymbol{p}_{\beta^\star}\|_\infty$.

Let $\boldsymbol{p}^\star = \operatorname{prox}_{\lambda\|\cdot\|_\infty + \iota_\Lambda}(\boldsymbol{\delta})$ be the minimizer of (12) and let $\boldsymbol{\delta}_\Lambda = \mathcal{P}_\Lambda(\boldsymbol{\delta})$. We have

$$\frac{1}{2}\|\boldsymbol{p}^\star - \boldsymbol{\delta}\|_2^2 + \lambda\|\boldsymbol{p}^\star\|_\infty \leq \frac{1}{2}\|\boldsymbol{\delta}_\Lambda - \boldsymbol{\delta}\|_2^2 + \lambda\|\boldsymbol{\delta}_\Lambda\|_\infty \quad (23)$$

Since $\boldsymbol{\delta}_\Lambda = \mathcal{P}_\Lambda(\boldsymbol{\delta}) = \operatorname*{arg\,min}_{\boldsymbol{y}\in\Lambda} \|\boldsymbol{y} - \boldsymbol{\delta}\|_2^2$, we also have

$$\|\boldsymbol{\delta}_\Lambda - \boldsymbol{\delta}\|_2^2 \leq \|\boldsymbol{p}^\star - \boldsymbol{\delta}\|_2^2 \quad (24)$$

Thus

$$\begin{aligned} \lambda\|\boldsymbol{p}^\star\|_\infty &\leq \underbrace{\frac{1}{2}(\|\boldsymbol{\delta}_\Lambda - \boldsymbol{\delta}\|_2^2 - \|\boldsymbol{p}^\star - \boldsymbol{\delta}\|_2^2)}_{\leq 0} + \lambda\|\boldsymbol{\delta}_\Lambda\|_\infty \\ &\leq \lambda\|\boldsymbol{\delta}_\Lambda\|_\infty \end{aligned} \quad (25)$$

We deduce that $\boxed{\beta^\star = \|\boldsymbol{p}^\star\|_\infty \leq \|\boldsymbol{\delta}_\Lambda\|_\infty}$. □

C. ALMA prox attack algorithm

Algorithm 3 ALMA prox attack (untargeted)

Require: Classifier f, original image $\boldsymbol{x} \in [0,1]^{C\times H\times W}$, true or target label $\boldsymbol{y} \in \mathbb{N}^{H\times W}$, binary mask $\boldsymbol{m} \in \{0,1\}^{H\times W}$
Require: Threshold of pixel success rate ν
Require: Penalty function P, initial multiplier $\boldsymbol{\mu}^{(0)} \in \mathbb{R}_{++}^{H\times W}$, initial penalty parameter $\boldsymbol{\rho}^{(0)} \in \mathbb{R}_{++}^{H\times W}$
Require: Minimum scale $w_{\min}$, scale adjustment rate $\gamma_w > 0$
Require: Number of iterations N, initial step size $\lambda^{(0)}$, penalty parameter increase rate $\gamma > 1$, constraint improvement rate τ, M number of steps between ρ increase, α smoothing parameter

Initialize $\boldsymbol{\delta}^{(0)} \leftarrow \mathbf{0}$, $\boldsymbol{v}^{(0)} \leftarrow 0$, $w^{(0)} \leftarrow 1$
for $t \leftarrow 1, \dots, N$ **do**
  $\tilde{\boldsymbol{x}}^{(t)} \leftarrow \boldsymbol{x} + \boldsymbol{\delta}^{(t-1)}$ // $\in [0,1]^{C\times H\times W}$
  $\boldsymbol{d}^{(t)} \leftarrow \mathrm{DLR}^+(f(\tilde{\boldsymbol{x}}^{(t)}), \boldsymbol{y})$ // $\in \mathbb{R}^{H\times W}$
  if $\frac{\boldsymbol{m}^\top [\boldsymbol{d}^{(t)} \le 0]}{\|\boldsymbol{m}\|_1} < \tau$ **then** // Adjust constraint scale
    $\hat{w} \leftarrow \frac{w^{(t-1)}}{1-\gamma_w}$ // Increase scale
  else
    $\hat{w} \leftarrow \frac{w^{(t-1)}}{1+\gamma_w}$ // Decrease scale
  end if
  $w^{(t)} \leftarrow \mathcal{P}_{[w_{\min},1]}(\hat{w})$
  $\xi^{(t)} \leftarrow (1-(1-\nu)\frac{t-1}{N-1})$-percentile of $\boldsymbol{d}^{(t)}$
  $\tilde{\boldsymbol{m}}^{(t)} \leftarrow [\boldsymbol{d}^{(t)} \le \xi^{(t)}]$ // $\in \{0,1\}^{H\times W}$
  $\hat{\boldsymbol{\mu}} \leftarrow \nabla_{\boldsymbol{d}} \left((\tilde{\boldsymbol{m}}^{(t)})^\top P(w^{(t)}\boldsymbol{d}^{(t)}, \boldsymbol{\rho}^{(t-1)}, \boldsymbol{\mu}^{(t-1)}) \right)$
  $\boldsymbol{\mu}^{(t)} \leftarrow \mathcal{P}_{[\mu_{\min},\mu_{\max}]}(\alpha\boldsymbol{\mu}^{(t-1)} + (1-\alpha)\hat{\boldsymbol{\mu}})$ // $\in \mathbb{R}_{++}^{H\times W}$
  for $i \leftarrow 1, \dots, d$ **do** // ρ adjustment
    if $t \bmod M = 0$ **and** $\tilde{\boldsymbol{m}}_i^{(t)} = 1$ **and** $\left(\exists j \in \{0,\dots,M-1\} : \boldsymbol{d}_i^{(t-j)} \le 0 \text{ or } \boldsymbol{d}_i^{(t)} \le \tau \boldsymbol{d}_i^{(t-M)}\right)$ **then**
      $\boldsymbol{\rho}_i^{(t)} \leftarrow \boldsymbol{\rho}_i^{(t-1)}$ // Constraint improved or satisfied
    else
      $\boldsymbol{\rho}_i^{(t)} \leftarrow \gamma\boldsymbol{\rho}_i^{(t-1)}$
    end if
  end for
  $\mathcal{L} \leftarrow (\tilde{\boldsymbol{m}}^{(t)})^\top P(w^{(t)}\boldsymbol{d}^{(t)}, \boldsymbol{\rho}^{(t)}, \boldsymbol{\mu}^{(t)})$ // $\in \mathbb{R}$
  $\boldsymbol{g}^{(t)} \leftarrow \nabla_{\boldsymbol{\delta}}\mathcal{L}$ // $\in \mathbb{R}^{C\times H\times W}$
  $\boldsymbol{v}^{(t)} \leftarrow \alpha\boldsymbol{v}^{(t-1)} + (1-\alpha)(\boldsymbol{g}^{(t)})^2$ // $\in \mathbb{R}_+^{C\times H\times W}$
  $\mathsf{H} \leftarrow \mathrm{Diag}\left(\sqrt{\frac{\boldsymbol{v}^{(t)}}{1-\alpha^t}} + \varepsilon\right)$
  $\boldsymbol{\delta}^{(t)} \leftarrow \mathrm{prox}^{\mathsf{H}}_{\lambda^{(t)}\|\cdot\|_\infty + \iota_\Lambda}(\boldsymbol{\delta}^{(t-1)} - \lambda^{(t)}\mathsf{H}^{-1}\boldsymbol{g})$ // VMFB
end for
return $\tilde{\boldsymbol{x}}^{(t)}$ that is adversarial and has the smallest norm

D. Image size and performance of the models

For all models, except DeepLabV3 DDC-AT, the images of Pascal VOC 2012 are resized so that the smaller side is of length 512 while keeping the aspect ratio, and for Cityscapes, the images keep their original size of $2\,048\times1\,024$. For DeepLabV3 DDC-AT from [41], the images of Pascal VOC 2012 are resized so that the longer side is of length 512 while keeping the aspect ratio, and for Cityscapes, the images are resized to $1\,024 \times 512$.

Dataset	Model	mIoU (%)	Pixel Accuracy (%)
Pascal VOC 2012 (+Aug)	DeepLabV3+ ResNet-50 [11]	$77.4_{+0.8}$	$94.9_{+0.2}$
	DeepLabV3+ ResNet-101 [11]	$78.8_{+0.1}$	$95.3_{+0.1}$
	FCN HRNetV2 W48 [38]	$76.4_{+0.2}$	$94.7_{+0.1}$
	DeepLabV3 DDC-AT [41]	$75.2_{+0.0}$	94.4
Cityscapes	DeepLabV3+ ResNet-50 [11]	$80.1_{-0.2}$	$96.4_{-0.1}$
	FCN HRNetV2 W48 [38]	$80.5_{-0.2}$	$96.6_{-0.1}$
	SegFormer MiT-B0 [40]	$76.4_{-0.1}$	$95.9_{-0.0}$
	SegFormer MiT-B3 [40]	$81.8_{-0.0}$	$96.7_{-0.1}$
	DeepLabV3 DDC-AT [41]	$71.0_{-0.3}$	95.0

Table 3. Performance of the models used in the experiments on the validation sets. Numbers were obtained from our evaluation; subscripts correspond to the difference with the original evaluation protocol. For DeepLabV3 DDC-AT from [41], the pixel accuracy was not reported.

E. Cityscapes target label

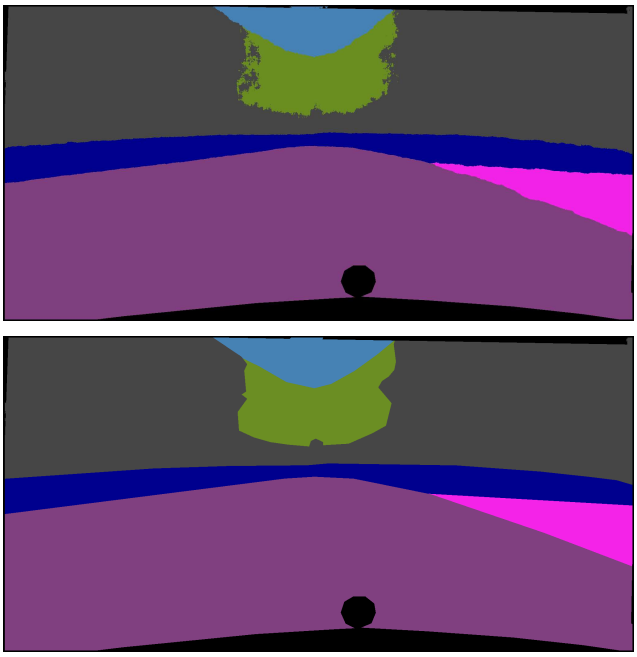

Figure 4. Cityscapes target segmentation used in our experiments. Top is the pixel majority label, bottom is the smoothed version. Classes in the target are: *road* (purple), *sidewalk* (pink), *car* (blue), *building* (dark gray), *vegetation* (green), *sky* (cyan) and *no label* (black).

F. Masking strategies for ALMA prox and PDPGD

In this section, we detail the modifications brought to PDPGD [30] and study the effect of the masking strategies on PDPGD and ALMA prox. Besides the addition of perturbation tracking logic (*i.e.* monitoring the best smallest perturbations during the optimization), we also incorporate the masking needed for the unlabeled regions.

Modifications for PDPGD For PDPGD, the dual variable $\lambda \in \mathbb{R}$ is replaced by a vector $\boldsymbol{\lambda} \in \mathbb{R}^d$. In [30], the gradient ascent on the dual variable is performed in the log domain, and is projected onto the 1-simplex (section IV of [30]). For the segmentation variant, this translates into adding 1 to the denominator of the $\mathrm{softmax}$ function used to project onto the $(d-1)$-simplex. This is equivalent to padding $\boldsymbol{\lambda}$ with a 0 and projecting it on the d-simplex $\Delta^d \subset [0,1]^{d+1}$. The weight of the norm $\|\boldsymbol{\delta}\|$ in Equation (10) of [30] becomes 1 minus the sum of the weights of the constraints. Adding the mask $\boldsymbol{m}$, this results in the following computations (introducing variables not described in [30]):

$$\begin{aligned} \boldsymbol{\lambda}_\Delta &= \frac{\boldsymbol{m} \odot \exp \boldsymbol{\lambda}}{1 + \boldsymbol{m}^\top \exp \boldsymbol{\lambda}} \quad \in [0,1]^d \\ \mathbb{L}_{\boldsymbol{\delta}}(\boldsymbol{\delta}, \boldsymbol{\lambda}) &= (1 - \boldsymbol{m}^\top \boldsymbol{\lambda}_\Delta) \|\boldsymbol{\delta}\| + \boldsymbol{\lambda}_\Delta^\top \mathcal{L}(\boldsymbol{x} + \boldsymbol{\delta}, \boldsymbol{y}). \end{aligned} \tag{26}$$

By replacing $\boldsymbol{m}$ by $\tilde{\boldsymbol{m}}^{(t)}$ in each iteration, we obtain the adaptive constraint masking described in Section 4.1.

The step-size is set to 0.01 for the primal variables and 0.1 for the dual variables with the same exponential and linear decays respectively, as in [30]. The dual variable is initialized such that the ratio of the weight of the norm term and the constraints terms is 1. From the above equations, we can derive the initial $\boldsymbol{\lambda}^{(0)} \in \mathbb{R}^d$ from the ratio $r \in \mathbb{R}_{++}$:

$$\begin{aligned} \frac{1 - \boldsymbol{m}^\top \boldsymbol{\lambda}_\Delta^{(0)}}{\boldsymbol{m}^\top \boldsymbol{\lambda}_\Delta^{(0)}} = r &\Leftrightarrow 1 - \boldsymbol{m}^\top \boldsymbol{\lambda}_\Delta^{(0)} = r \boldsymbol{m}^\top \boldsymbol{\lambda}_\Delta^{(0)} \\ &\Leftrightarrow (1+r) \boldsymbol{m}^\top \boldsymbol{\lambda}_\Delta^{(0)} = 1 \\ &\Leftrightarrow (1+r) \frac{\boldsymbol{m}^\top \exp \boldsymbol{\lambda}^{(0)}}{1 + \boldsymbol{m}^\top \exp \boldsymbol{\lambda}^{(0)}} = 1 \\ &\Leftrightarrow (1+r) \boldsymbol{m}^\top \exp \boldsymbol{\lambda}^{(0)} = \\ &\qquad 1 + \boldsymbol{m}^\top \exp \boldsymbol{\lambda}^{(0)} \\ &\Leftrightarrow \boldsymbol{m}^\top \exp \boldsymbol{\lambda}^{(0)} = \frac{1}{r} \end{aligned} \tag{27}$$

Assuming that $\boldsymbol{\lambda}^{(0)} = \omega \mathbf{1}_d$ with $\omega \in \mathbb{R}$, we get:

$$\begin{aligned} \boldsymbol{m}^\top \exp \boldsymbol{\lambda}^{(0)} = \frac{1}{r} &\Leftrightarrow \|\boldsymbol{m}\|_1 \exp \omega = \frac{1}{r} \\ &\Leftrightarrow \omega = -\log(r \|\boldsymbol{m}\|_1) \end{aligned} \tag{28}$$

Ablation To test the effect of these modifications on PDPGD and ALMA prox, we perform an ablation study on the masking strategy. We perform this experiment with DeepLabV3+ ResNet-50 on Cityscapes. We compare three different constraint masking strategies:

- masking only the unlabeled regions, denoted by $\boldsymbol{m}$;
- masking the unlabeled regions and $(100 - \nu)\%$ of the largest constraints, denoted by $\tilde{\boldsymbol{m}}$;
- masking the unlabeled regions and linearly decreasing the fraction of constraints to reach $\nu\%$ at the last iteration, corresponding to strategy described in Equation (6), denoted by $\tilde{\boldsymbol{m}}^{(t)}$.

The results of this experiment are provided in Figure 5. While the effect is small for this particular model and dataset, the $v\tilde{m}$ and $v\tilde{m}^{(t)}$ strategies do result in smaller perturbations. However, PDPGD does not seem to benefit from the more advanced masking strategies. It obtains the best results when masking only the unlabeled regions. This issue comes from the projection of the dual variables onto the d-simplex (26): as different constraints get discarded in subsequent iterations, their relative weights vary drastically, leading to oscillations of the dual variables. This phenomenon does not occur with an Augmented Lagrangian based attacks, as the penalty multipliers are not projected together. This does not create a dependency between the multipliers, resulting in a more stable optimization. Therefore, for the experiments, ALMA prox is used with the adaptive masking strategy with a linear decay, whereas PDPGD is used with the unlabeled region masking only.

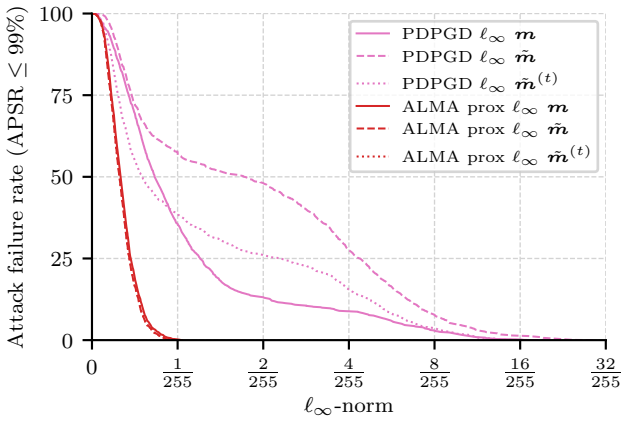

Figure 5. Influence of the constraint masking strategies on PDPGD and ALMA prox for untargeted attacks on Cityscapes with DeepLabV3+ ResNet-50. $\boldsymbol{m}$ corresponds to masking the pixels with no labels, $\tilde{\boldsymbol{m}}$ corresponds to additionally masking the top $(100 - \nu)\%$ constraints, and $\tilde{\boldsymbol{m}}^{(t)}$ corresponds to the strategy in Equation (6). The curves for PDGD and ALMA prox $\tilde{\boldsymbol{m}}$ and $\tilde{\boldsymbol{m}}^{(t)}$ overlap.

G. Attack complexities and run-times

Tables 4 and 5 report the average complexities (in terms of forward and backward) and run-time per sample for the attacks. The run-times for the targeted attacks are similar to their targeted variant, as the only difference lies in the loss, which is far from being the most computationally expensive part of the attacks. These results show that ALMA prox has slightly higher run-times compared to the other attacks with similar budgets, while being more effective. Note that some variance come from the fact that the experiments were run on a shared compute cluster.

Attack	Forwards	Backwards
I-FGSM [27] 13×20 and MI-FGSM [22]	260	260
PGD 13×40 and $13 \times 4 \times 10$ [29]	520	520
DAG $\eta = 0.003$ [39]	63	62
DAG $\eta = 0.001$ [39]	155	154
FMN ℓ_∞ [34]	500	500
PDPGD ℓ_∞ [30]	500	500
ALMA prox	500	500

Table 4. Average complexity of the attacks in terms of number of forward and backward propagations. All the attacks, except I-FGSM, MI-FGSM and PGD have a 500 iteration budget. DAG is the only attack that uses an early stopping criterion.

H. Complete attack results

Table 6 reports the median and average ℓ_∞-norm (multiplied by 255 for readability) of the perturbations produced by the attacks for all regular models. As in Section 6, the perturbation norm is considered to be 1 for unsuccessful attacks. This means that attack with less than 50% success have a 1 (*i.e.* 255) median ℓ_∞-norm in the table. Additionally, Figures 6 and 7 show the percentage of unsuccessful attacks on Pascal VOC 2012 and Cityscapes for all models considered.

	Attack	Pascal VOC 2012				Cityscapes				
		DeepLabV3+ ResNet-50	DeepLabV3+ ResNet-101	FCN HRNetV2 W48	DeepLabV3 DDC-AT	DeepLabV3+ ResNet-50	FCN HRNetV2 W48	SegFormer MiT-B0	SegFormer MiT-B3	DeepLabV3 DDC-AT
Untargeted	I-FGSM 13 × 20 [27]	13.3	19.2	28.8	–	87.0	43.2	31.5	83.4	–
	MI-FGSM 13 × 20 [22]	13.3	19.2	29.1	–	87.6	41.5	31.1	83.2	–
	PGD CE 13 × 40 [29]	26.4	38.2	57.0	–	171.9	81.9	61.7	164.6	–
	PGD CE 13 × 4 × 10 [29]	26.6	38.3	57.2	–	174.1	81.9	61.6	164.5	–
	PGD DLR 13 × 40 [29]	27.7	39.3	60.9	–	179.5	88.1	67.6	170.7	–
	PGD DLR 13 × 4 × 10 [29]	27.9	39.4	57.6	–	179.3	88.2	67.6	170.6	–
	DAG $\eta = 0.003$ [39]	2.1	5.2	6.0	7.0	19.5	14.4	9.2	49.2	22.4
	DAG $\eta = 0.001$ [39]	4.8	12.9	14.8	12.7	45.5	37.3	23.4	107.3	44.7
	FMN ℓ_∞ [34]	24.4	32.0	41.3	24	155.2	79.1	59.1	158.2	64.5
	PDPGD ℓ_∞ [30]	28.1	35.5	41.0	25.5	161.0	84.5	64.1	162.4	64.7
	ALMA prox	28.1	36.1	43.4	28	161.9	96.8	77.6	176.3	69.8
Targeted	I-FGSM 13 × 20 [27]	13.3	19.2	29.8	–	87.5	41.4	31.1	83.0	–
	MI-FGSM 13 × 20 [22]	13.3	19.4	28.8	–	87.8	41.4	31.1	83.1	–
	PGD CE 13 × 40 [29]	27.0	38.1	56.8	–	171.9	81.9	61.4	165.0	–
	PGD CE 13 × 4 × 10 [29]	26.7	38.4	56.5	–	171.9	82.0	61.5	164.3	–
	PGD DLR 13 × 40 [29]	27.7	39.4	58.9	–	179.6	88.2	68.4	172.0	–
	PGD DLR 13 × 4 × 10 [29]	27.8	39.8	58.7	–	179.4	88.2	69.7	171.0	–
	DAG $\eta = 0.003$ [39]	0.8	1.5	1.9	2.2	47.6	18.6	21.9	54.1	44.5
	DAG $\eta = 0.001$ [39]	2.1	4.2	4.6	4.8	111.7	44.9	52.0	132.9	61.0
	FMN ℓ_∞ [34]	24.4	32.1	41.1	25.4	155.1	79.2	58.7	157.9	60.4
	PDPGD ℓ_∞ [30]	27.6	36.9	39.0	26.8	158.8	81.4	64.1	159.1	63.2
	ALMA prox	28.4	36.3	43.0	28.5	161.9	97.2	78.4	179.5	67.6

Table 5. Average run-times per image for the attacks, in seconds.

	Attack	Pascal VOC 2012						Cityscapes							
		DeepLabV3+ ResNet-50		DeepLabV3+ ResNet-101		FCN HRNetV2 W48		DeepLabV3+ ResNet-50		FCN HRNetV2 W48		SegFormer MiT-B0		SegFormer MiT-B3	
Untargeted	I-FGSM 13 × 20 [27]	146.32	136.55	124.14	131.34	227.11	145.69	255.00	242.15	255.00	194.98	106.34	121.12	255.00	208.97
	MI-FGSM 13 × 20 [22]	195.35	145.96	188.37	150.13	255.00	157.63	255.00	244.43	255.00	218.91	202.15	159.52	255.00	228.43
	PGD CE 13 × 40 [29]	80.10	120.93	100.42	123.17	152.08	136.05	255.00	236.77	255.00	176.94	9.31	48.10	255.00	188.15
	PGD CE 13 × 4 × 10 [29]	24.90	74.15	20.05	30.28	66.93	118.76	255.00	244.92	255.00	216.38	34.77	42.49	255.00	231.51
	PGD DLR 13 × 40 [29]	7.22	26.42	9.00	21.29	11.11	23.85	17.75	24.23	13.29	16.15	23.53	32.34	84.22	102.80
	PGD DLR 13 × 4 × 10 [29]	4.42	24.99	6.41	11.02	10.46	29.75	255.00	227.89	12.08	45.17	22.37	35.49	110.82	134.41
	DAG $\eta = 0.003$ [39]	5.69	6.63	6.65	8.74	8.80	10.61	6.30	7.51	8.42	9.91	6.41	6.48	8.83	9.02
	DAG $\eta = 0.001$ [39]	5.23	8.22	6.17	21.14	8.49	14.61	5.95	9.39	8.20	20.29	6.08	13.57	8.59	53.65
	FMN ℓ_∞ [34]	0.46	38.34	0.56	51.60	0.91	46.39	1.97	96.57	0.97	35.12	0.58	1.16	1.08	6.42
	PDPGD ℓ_∞ [30]	0.73	1.77	1.17	3.49	1.52	2.43	1.06	2.82	1.69	2.75	0.87	1.17	1.39	3.05
	ALMA prox	**0.32**	**0.34**	**0.37**	**0.41**	**0.51**	**0.56**	**0.24**	**0.26**	**0.40**	**0.41**	**0.26**	**0.26**	**0.33**	**0.33**
Targeted	I-FGSM 13 × 20 [27]	0.47	0.50	0.65	0.67	0.59	0.64	255.00	255.00	115.76	128.65	5.91	31.96	51.73	88.24
	MI-FGSM 13 × 20 [22]	0.59	0.66	0.77	0.82	0.77	0.85	4.26	55.35	4.23	4.47	2.48	2.53	2.54	2.60
	PGD CE 13 × 40 [29]	0.37	0.43	0.50	0.55	0.50	0.54	3.49	3.71	3.49	3.62	2.30	2.34	1.87	1.92
	PGD CE 13 × 4 × 10 [29]	0.50	0.51	0.62	0.70	0.62	0.65	255.00	255.00	255.00	255.00	255.00	255.00	255.00	255.00
	PGD DLR 13 × 40 [29]	0.62	0.68	0.81	0.92	0.81	0.94	8.37	67.86	7.41	7.52	4.79	4.96	3.98	4.09
	PGD DLR 13 × 4 × 10 [29]	0.87	1.07	1.18	2.03	1.12	1.28	255.00	255.00	255.00	255.00	255.00	255.00	255.00	255.00
	DAG $\eta = 0.003$ [39]	4.21	4.50	4.84	5.09	5.32	5.66	11.34	12.96	9.87	10.49	8.05	8.44	9.82	10.06
	DAG $\eta = 0.001$ [39]	3.92	4.21	4.55	5.80	5.07	5.36	10.96	40.28	9.43	14.42	255.00	145.61	11.53	119.47
	FMN ℓ_∞ [34]	0.42	0.45	0.46	0.56	0.47	0.49	255.00	254.36	2.25	2.34	255.00	255.00	255.00	255.00
	PDPGD ℓ_∞ [30]	0.28	0.36	0.31	0.40	0.35	0.46	14.51	14.41	16.71	16.75	17.93	17.87	19.26	19.20
	ALMA prox	**0.25**	**0.26**	**0.29**	**0.30**	**0.32**	**0.34**	**1.15**	**1.17**	**1.11**	**1.12**	**0.70**	**0.70**	**0.65**	**0.66**

Table 6. Median and average $\|\boldsymbol{\delta}\|_\infty \times 255$ for each adversarial attack on Pascal VOC 2012 and Cityscapes for the regular models.

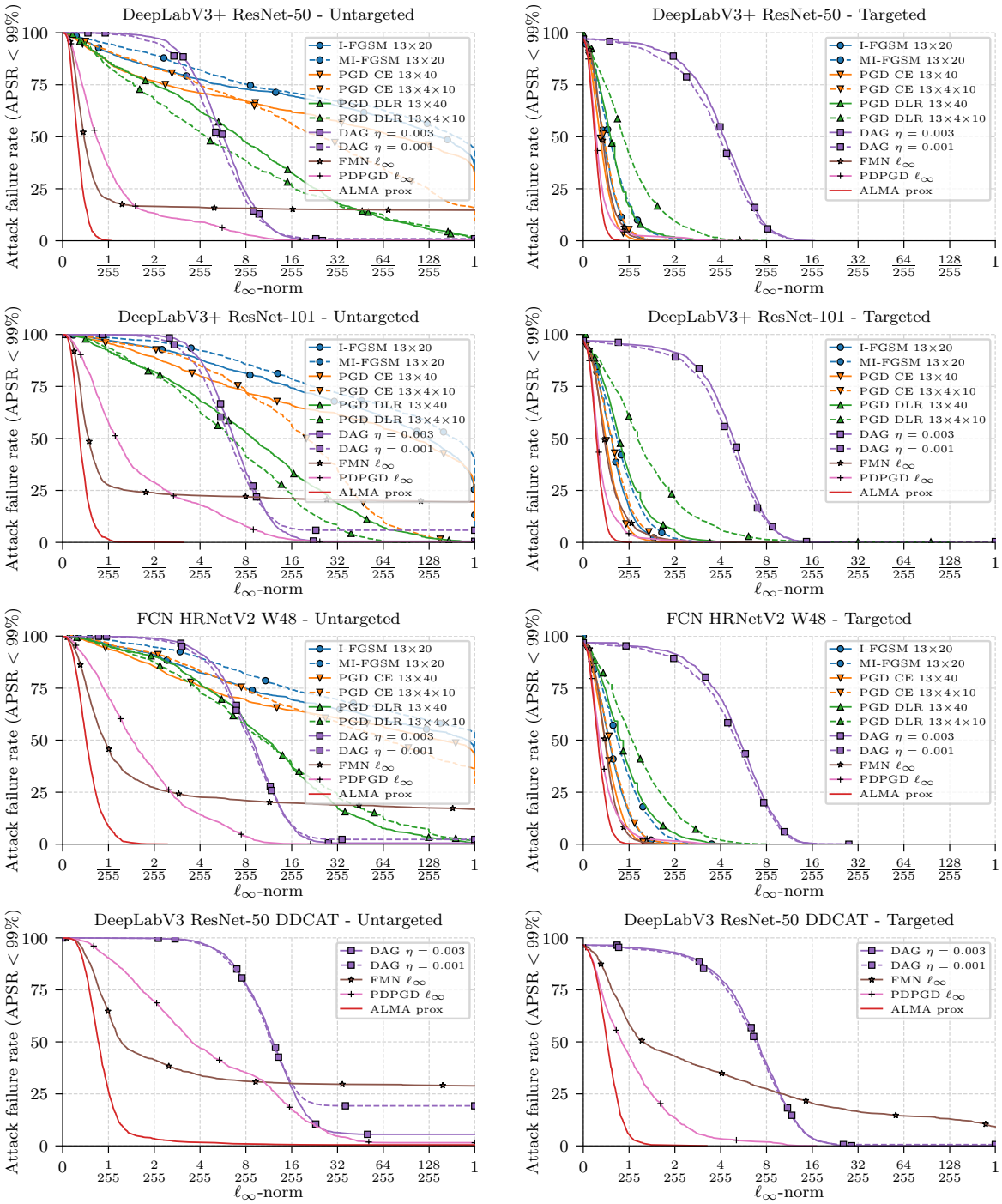

Figure 6. Percentage of unsuccessful ℓ_∞ attacks on **Pascal VOC 2012** (*i.e.* with APSR $\leq 99\%$). A stronger attack has a lower curve; a more robust model has a higher curve. Horizontal axis is linear on $[0, 2/255]$ and logarithmic on $[2/255, 1]$.

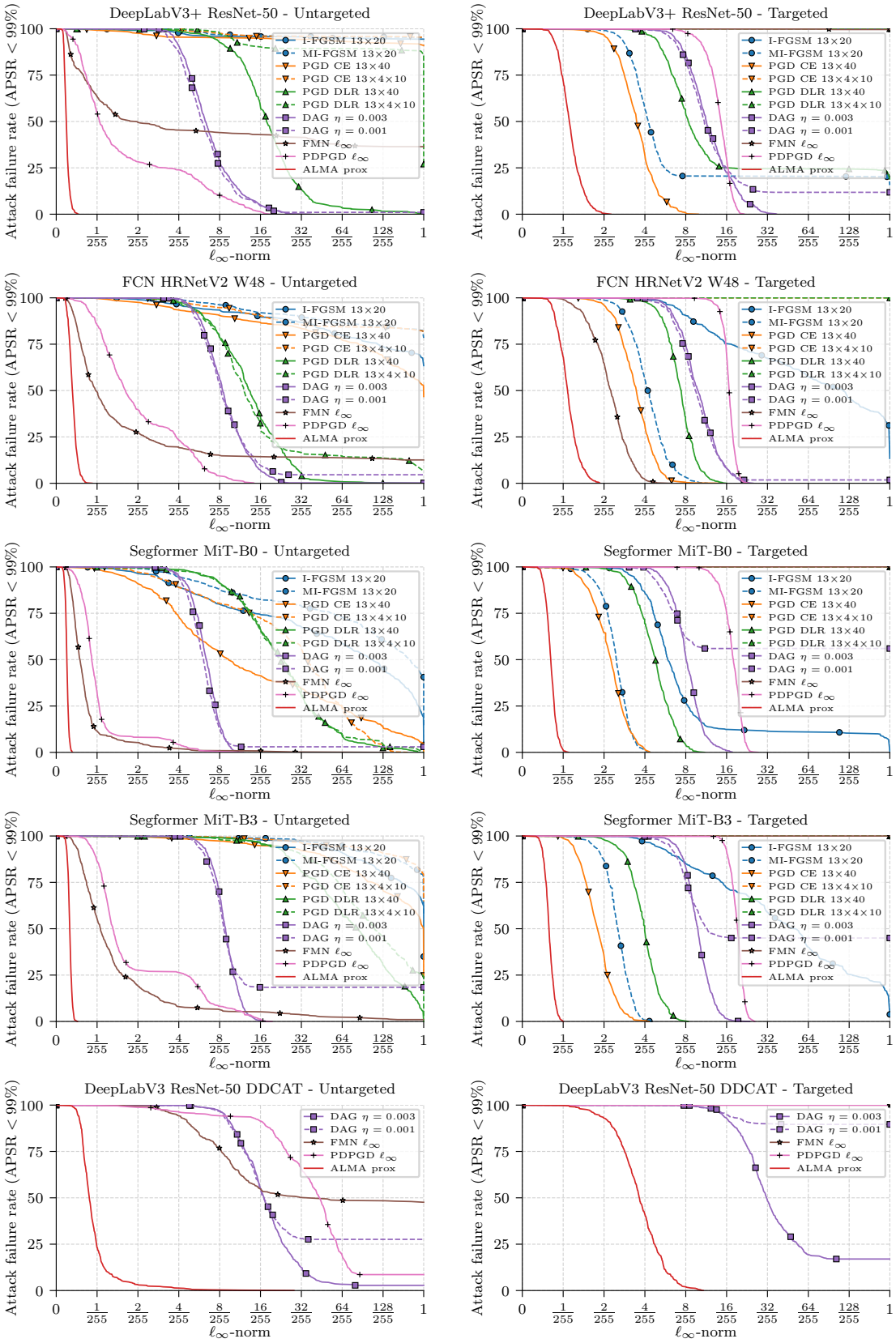

Figure 7. Percentage of unsuccessful ℓ_∞ attacks on **Cityscapes** (*i.e.* with APSR $\leq 99\%$). A stronger attack has a lower curve; a more robust model has a higher curve. Horizontal axis is linear on $[0, 2/255]$ and logarithmic on $[2/255, 1]$.