TsallisPGD: Adaptive Gradient Weighting for Adversarial Attacks on Semantic Segmentation

Alexander Matyasko, Xin Lou, Indriyati Atmosukarto, Wei Zhang
Singapore Institute of Technology, Singapore
{alexander.matyasko, lou.xin, indriyati, wei.zhang}@singaporetech.edu.sg

***Abstract*—Attacking semantic segmentation models is significantly harder than image classification models because an attacker must flip thousands of pixel predictions simultaneously. Standard pixel-wise cross-entropy (CE) is ill-suited to this setting: it tends to overemphasize already-misclassified pixels, which slows optimization and overstates model robustness. To address these issues, we introduce TsallisPGD, an adversarial attack built on the Tsallis cross-entropy, a generalization of CE parameterized by q, which adaptively reshapes the gradient landscape by controlling gradient concentration across pixels. By varying q, we steer the attack toward pixels at different confidence levels. We first show that no single fixed-q is universally optimal, as its effectiveness depends on the dataset, model architecture, and perturbation budget. Motivated by this, we propose a dynamic q-schedule that sweeps q during optimization. Extensive experiments on Cityscapes, Pascal VOC, and ADE20K show that TsallisPGD, using a single validation-selected schedule, achieves the best average attack rank across all evaluated settings and improves over CEPGD, SegPGD, CosPGD, JSPGD, and MaskedPGD in reducing accuracy and mIoU on both standard and robust models.**

***Index Terms*—adversarial attacks, adversarial robustness, semantic segmentation.**

I. Introduction

Deep neural networks (DNNs) have proven remarkably effective at tackling complex computer vision tasks, such as image classification [1, 2], object detection [3, 4], semantic segmentation [5, 6], depth estimation [7, 8], and optical flow estimation [9, 10]. These advances have significant applications in autonomous driving [11] and biomedical image analysis [12]. However, despite these impressive advancements, deep neural networks face critical challenges that limit their broader adoption in safety- and security-critical applications, such as self-driving cars and medical imaging. Specifically, DNNs are sensitive to distribution shifts such as changes in lighting and weather conditions [13, 14]. Most critically, state-of-the-art DNNs are vulnerable to adversarial attacks, which can craft small, imperceptible input perturbations that cause severe and unexpected model failures [15, 16].

Since the discovery of adversarial attacks against image classification models, significant progress has been made in evaluating and improving the robustness of image classification models [17, 18]. Recent research has broadened to consider a wider range of prediction tasks [19, 20] and input modalities [21, 22]. Semantic segmentation presents unique challenges for adversarial robustness compared to image classification. Although both types of attacks aim to optimize a loss function that measures attack success, adversarial attacks on semantic segmentation are significantly more challenging because they must alter predictions for all pixels simultaneously. Likewise, improving the robustness of semantic segmentation models is more difficult, as any adversarial defense must ensure that predictions for all pixels remain robust to all feasible perturbations.

Early approaches for attacking semantic segmentation models [19, 23] directly adapted the projected gradient descent (PGD) attack to the semantic segmentation setting. These attacks maximize the average per-pixel loss, e.g. cross-entropy, across all output pixels, mirroring the training objective. While maximizing average per-pixel loss is a straightforward extension of PGD, it ignores the multi-output nature of segmentation models, leading to suboptimal optimization dynamics. As a result, PGD-based attacks on semantic segmentation models often suffer from slow convergence, requiring hundreds of attack iterations and potentially leading to an overly optimistic assessment of model robustness.

To address these limitations, recent works have focused on adaptively reweighting per-pixel losses to improve optimization dynamics. SegPGD [24] assigns different weights to correctly and incorrectly classified pixels. SegPGD dynamically anneals pixel weights, initially focusing exclusively on correctly classified pixels and then gradually shifting attention to a mixture of correctly and incorrectly classified pixels. CosPGD [25] sets pixel weights proportional to the cosine similarity between the original and adversarial outputs, effectively prioritizing pixels where the model is more confident. Croce et al. [26] further observed that gradients of standard cross-entropy saturate on misclassified pixels, causing attacks to overemphasize pixels that are already successfully attacked. To mitigate this issue, they proposed alternative objectives with bounded gradients on misclassified pixels, such as the Jensen–Shannon divergence and masked cross-entropy, which yield more stable optimization dynamics.

In this paper, we propose **TsallisPGD**, an adversarial attack for semantic segmentation built on Tsallis cross-entropy, a generalization of standard cross-entropy controlled by a parameter q. A key property of Tsallis cross-entropy is that it induces a confidence-dependent reweighting of pixels in gradient space: each pixel is effectively scaled by p_y^{1-q}, where p_y is the predicted probability of the ground-truth class. For $q < 1$, this reshapes optimization by emphasizing high-confidence, correctly classified pixels while down-weighting

high-confidence, incorrectly classified ones, making it well suited for segmentation attacks. We analyze the resulting gradient behavior and show that q controls both the magnitude and the location of the gradient peak. By adjusting q, the attack can thus target different subsets of pixels. We further show empirically that no single fixed-q is optimal across datasets, architectures, and perturbation budgets. This motivates a *dynamic* q-schedule that sweeps q during optimization enabling the attack to target pixels at different confidence levels. We select the best q-schedule using a held-out validation split and apply the same selected schedule across the entire test benchmark with no further tuning. Extensive experiments on CITYSCAPES, PASCAL-VOC, and ADE20K show that TsallisPGD achieves the best overall average rank across both standard and adversarially trained models. Our attack improves over CEPGD, SegPGD, CosPGD, JSPGD, and MaskedPGD in reducing both accuracy and mIoU on aggregate. Code will be available at https://github.com/aam-at/tsallis_pgd.

To summarize, our main contributions are as follows:

- We introduce TsallisPGD, a novel adversarial attack for semantic segmentation that leverages Tsallis cross-entropy to control the attack focus via parameter q.
- We provide a theoretical analysis showing that Tsallis cross-entropy induces a confidence-dependent gradient reweighting p_y^{1-q}, and demonstrate how q controls both the location and magnitude of a lower-bound proxy for the gradient peak.
- We propose a dynamic q-schedule that varies q over the course of optimization, enabling the attack to target pixels at different confidence levels.
- We demonstrate on CITYSCAPES, PASCAL-VOC, and ADE20K that TsallisPGD achieves the best overall average rank among strong baselines on both standard and adversarially trained models.

II. RELATED WORK

Adversarial attacks against classification models: Seminal work by Szegedy et al. [15] revealed that state-of-the-art image classifiers are highly vulnerable to adversarial perturbations—carefully crafted, small and often imperceptible input changes that can fool the model. Goodfellow et al. [16] introduced the Fast Gradient Sign Method (FGSM), a computationally efficient one-step attack that perturbs inputs in the direction of the gradient of the loss with respect to the input. Kurakin et al. [27] and Madry et al. [18] introduced stronger iterative attacks based on FGSM, such as BIM (I-FGSM) and PGD, respectively. Madry et al. [18] further formalized PGD as a universal first-order adversary, which has since become the de facto standard for adversarial robustness evaluation. AutoPGD [17] improved upon PGD by detecting optimization stalling and adaptively tuning the step size of projected gradient iterations. Despite these advances, reliably and accurately evaluating robustness remains challenging due to phenomena such as gradient masking and gradient obfuscation [28]. To address these issues, AutoAttack [17] combines multiple attacks into a unified, parameter-free evaluation suite and is now widely regarded as a standard benchmark for robustness evaluation in image classification.

Adversarial attacks against semantic segmentation models: Early approaches for attacking semantic segmentation models [19, 23] directly adapted classification attacks, such as projected gradient descent (PGD), by applying them in a pixel-wise manner and aggregating cross-entropy losses over all output pixels. Although this formulation is a natural extension of PGD to semantic segmentation, it ignored the structured, multi-output nature of segmentation models, resulting in slow convergence and suboptimal optimization dynamics. Consequently, these early works created the mistaken belief that segmentation models are inherently more robust than classification models [23]. However, later works identified that different pixels contribute unequally to the overall loss and showed that, when attacks are stronger and adaptive, segmentation models are as vulnerable as classifiers [24]. SegPGD [24] improves attack effectiveness by adaptively reweighting pixel-wise losses based on whether pixels are correctly or incorrectly classified. Using an adaptive annealing schedule, SegPGD achieves faster convergence and significantly outperforms standard PGD under the same iteration budget. CosPGD [25] weights pixel losses according to the cosine similarity between the original and adversarial outputs, effectively prioritizing pixels where the model is more confident. Further analyzing the shortcomings of standard objectives, Croce et al. [26] observed that the gradient of the cross-entropy loss saturates on misclassified pixels, causing attacks to overemphasize pixels that have already been successfully attacked. They introduced new segmentation attacks based on alternative loss functions, including JSPGD based on the Jensen–Shannon divergence (JS) and MaskedPGD based on masked cross-entropy (MaskedCE), whose gradients are bounded on already misclassified pixels and yield more stable optimization dynamics. Similar to AutoAttack for evaluating classification models, they introduced SEA [26], a robustness evaluation suite that combines multiple attacks to provide a more reliable evaluation of adversarial robustness for semantic segmentation models.

III. SEMANTIC SEGMENTATION ATTACKS

Let f_θ be a semantic segmentation model with parameters θ. For notational convenience, we represent an input image as a vector $\boldsymbol{x} \in [0,1]^D$, obtained by flattening an input image of spatial resolution $H \times W$ with C channels, so that $D = C \cdot H \cdot W$. The model maps $\boldsymbol{x}$ to per-pixel unnormalized class scores or logits $f_\theta(\boldsymbol{x}) \in \mathbb{R}^{K \times N}$, where $N = H \cdot W$ denotes the number of pixels (i.e., K logits per pixel). The predicted class at pixel i is $\hat{y}_i(\boldsymbol{x}) = \arg\max_k f_{k,i}(\boldsymbol{x})$. Given a ground-truth segmentation map $y \in \{1, \dots, K\}^N$, an adversarial attack seeks a perturbation $\delta \in \mathcal{U}$ that degrades segmentation performance on the perturbed image $\boldsymbol{x}' = \boldsymbol{x} + \delta$, as measured by standard metrics such as pixel accuracy or mean intersection-over-union (mIoU). In this work, we focus on ℓ_∞-bounded perturbations, i.e., $\mathcal{U} = \{\delta \in \mathbb{R}^D : \|\delta\|_\infty \leq \varepsilon\}$.

Projected Gradient Descent (PGD) [18] is a widely used first-order method for generating adversarial examples. PGD approximately solves

$$\max_{\|\delta\|_\infty \le \varepsilon} \mathcal{L}(\boldsymbol{x}+\delta, y), \tag{1}$$

where $\mathcal{L}$ is a differentiable surrogate loss that measures attack success. Starting from an initialization $\boldsymbol{x}^{(0)}$, PGD iteratively takes a step in the direction of steepest ascent and projects the result onto the feasible set:

$$\boldsymbol{x}^{(t+1)} \;=\; \Pi_{\mathcal{B}_\infty(\boldsymbol{x},\varepsilon)}\Big(\boldsymbol{x}^{(t)} + \alpha\,\mathrm{sign}\big(\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^{(t)}, y)\big)\Big), \tag{2}$$

where α is the step size and $\Pi_{\mathcal{B}_\infty(\boldsymbol{x},\varepsilon)}$ denotes projection onto the ℓ_∞ ball around $\boldsymbol{x}$.

A natural choice for the surrogate loss $\mathcal{L}$ in semantic segmentation attacks is to use the sum or mean per-pixel cross-entropy, mirroring the standard training objective:

$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{x}, y) \;=\; \frac{1}{N}\sum_{i=1}^{N} \ell_{\mathrm{CE}}\big(p_{y_i,i}(\boldsymbol{x})\big), \qquad \ell_{\mathrm{CE}}(p) = -\log p.$$

However, standard cross-entropy has several limitations when used as the attack loss for semantic segmentation, often leading to suboptimal optimization dynamics:

a) Objective monotonically increasing as $p_y \to 0$: The per-pixel loss $\ell_{\mathrm{CE}}(p_{y_i,i}) = -\log p_{y_i,i}$ increases monotonically as $p_{y_i,i} \to 0$ and is unbounded. As a result, the average or mean loss can be dominated by a small subset of pixels with extremely small $p_{y_i,i}$, which are already misclassified, so improving the objective may largely reflect further decreasing p_y on these pixels rather than improving the overall attack success.

b) Gradient monotonically increasing as $p_y \to 0$: The per-pixel cross-entropy loss also leads to gradients that are highly concentrated on pixels with low predicted probability for the true class. For the cross-entropy loss, its gradient with respect to logits is bounded as follows (see Appendix A [26]):

$$\frac{K}{K-1}\,(1-p_y)^2 \;\le\; \big\|\nabla_u \mathcal{L}_{\mathrm{CE}}(u, e_y)\big\|_2^2 \;\le\; 1-p_y+(1-p_y)^2. \tag{3}$$

Thus, as $p_y \to 0$, both the lower and upper bounds on gradients are monotonically increasing, resulting in overemphasis on already misclassified pixels and away from high-confidence correctly classified regions.

Due to these effects, PGD with cross-entropy loss struggles to change high-confidence pixels and overemphasizes low-confidence pixels, leading to suboptimal attack performance. In the next section, we introduce a Tsallis cross-entropy objective, which is bounded and whose gradient peak can be adaptively controlled, enabling us to steer the attack and improve optimization dynamics.

IV. Gradient Rebalancing with Tsallis Cross-Entropy

This section introduces Tsallis cross-entropy, a one-parameter generalization of standard cross-entropy controlled by a parameter q, known as an entropic index. We show that Tsallis cross-entropy induces a confidence-dependent, per-pixel reweighting of the standard cross-entropy gradient, thereby rebalancing how optimization effort is distributed across pixels. We analyze how the Tsallis parameter q shapes the gradient. We derive a closed-form expression for the gradient peak location, which provides an intuitive rule for selecting q to target pixels within a desired confidence regime. This improved control leads to more effective PGD optimization and higher attack success rates for semantic segmentation.

Let $p_{y_i,i} \in (0,1]$ denote the predicted probability of the ground-truth class y_i at a pixel i. For $q \in \mathbb{R}$ with $q \neq 1$, we define the Tsallis cross-entropy loss as

$$\mathcal{L}_q(p_{y_i,i}) \;=\; \frac{1-p_{y_i,i}^{1-q}}{1-q}. \tag{4}$$

Tsallis cross-entropy is bounded between 0 and $\frac{1}{1-q}$ for all $q < 1$. It reduces to the standard cross-entropy in the limit $q \to 1$. For semantic segmentation attacks, we use the average of the per-pixel Tsallis losses.

Proposition 1 (Tsallis cross-entropy is a confidence-weighted cross-entropy)**.** *The gradient of Tsallis cross-entropy with respect to p is a confidence-weighted version of the standard cross-entropy gradient:*

$$\frac{\partial \mathcal{L}_q(p)}{\partial p} \;=\; p^{1-q}\,\frac{\partial \mathcal{L}_{\mathrm{CE}}(p)}{\partial p}. \tag{5}$$

Proof. The gradient of the Tsallis loss $\mathcal{L}_q(p)$ from equation 4 with respect to p is computed as

$$\frac{\partial \mathcal{L}_q(p)}{\partial p} \;=\; -p^{-q}. \tag{6}$$

On the other hand, the gradient of standard cross-entropy is

$$\frac{\partial \mathcal{L}_{\mathrm{CE}}(p)}{\partial p} \;=\; -\frac{1}{p}. \tag{7}$$

Combining the two expressions gives

$$\frac{\partial \mathcal{L}_q}{\partial p} \;=\; -p^{-q} \;=\; p^{1-q}\left(-\frac{1}{p}\right) \;=\; p^{1-q}\,\frac{\partial \mathcal{L}_{\mathrm{CE}}(p)}{\partial p}, \tag{8}$$

which completes the proof of equation 5. □

When $q < 1$, this weight increases with p and therefore emphasizes high-confidence pixels; when $q > 1$, it emphasizes low-confidence pixels. For untargeted semantic segmentation attacks, we consider values of $q \in (-\infty, 1)$ to focus optimization on high-confidence pixels that are difficult to flip.

The next proposition characterizes how Tsallis cross-entropy reshapes the gradient magnitude with respect to logits.

Proposition 2 (Bounds for Tsallis Cross-Entropy Gradient)**.** *Let $u \in \mathbb{R}^K$ be the per-pixel logits, $p = \mathrm{softmax}(u)$, and let y be the ground-truth class with probability $p_y \in (0,1]$. The gradient of Tsallis cross-entropy (per pixel) satisfies*

$$\frac{K}{K-1}\,p_y^{2(1-q)}(1-p_y)^2 \le \big\|\nabla_u \mathcal{L}_q(u,y)\big\|_2^2 \le p_y^{2(1-q)}\big(1-p_y+(1-p_y)^2\big). \tag{9}$$

Equation 9 can be derived by combining equation 5 with the standard cross-entropy gradient bounds. For $q < 1$, the factor $p_y^{2(1-q)} \leq 1$ implies that the Tsallis gradient uniformly down-weights low-confidence pixels with $p_y \approx 0$, preventing gradients from concentrating on pixels that are already confidently misclassified.

Proposition 3 (Tsallis Cross-Entropy Gradient Peak)**.** *For Tsallis cross-entropy with $q < 1$, the lower-bound proxy of the squared ℓ_2-norm of the logit-gradient in equation 9 admits a unique maximum over $p_y \in (0, 1)$ at*

$$p_y^\star = \frac{1-q}{2-q}.$$

Proof. For simplicity, we analyze the lower bound in equation 9. Since constant factors do not affect the maximizer, it suffices to maximize $f(p) = p^{2(1-q)}(1-p)^2$ over $p \in (0, 1)$. Maximizing $\log f(p) = 2(1-q)\log p + 2\log(1-p)$ yields the following first-order condition:

$$\frac{d}{dp}\log f(p) = \frac{2(1-q)}{p} - \frac{2}{1-p} = 0 \implies p^\star = \frac{1-q}{2-q}.$$

□

The expression for $p_y^\star$ provides an intuitive mechanism for controlling gradient concentration. Smaller values of q shift the gradient peak toward high-confidence, correctly classified pixels, while increasing q moves the peak toward lower-confidence regions, recovering standard cross-entropy behavior as $q \to 1$. Figure 1 compares the per-pixel confidence-dependent gradients induced by Tsallis cross-entropy with those of standard cross-entropy, cross-entropy combined with cosine similarity [25], and Jensen–Shannon divergence [26]. Unlike prior approaches, Tsallis cross-entropy allows the location of the gradient peak to be explicitly controlled through the parameter q, thereby enabling precise control over which pixels the attack focuses on. How q should be chosen in practice is an empirical question. In Section V-A, we show that the best fixed-q varies across datasets, architectures, and threat budgets. In Section V-B, based on this observation, we propose a dynamic q-schedule and select the best schedule on a held-out validation split. Our schedule starts with $q < 0$ to emphasize high-confidence, correctly classified pixels in early iterations, and gradually increases $q \to 1$, recovering standard cross-entropy by the end of optimization.

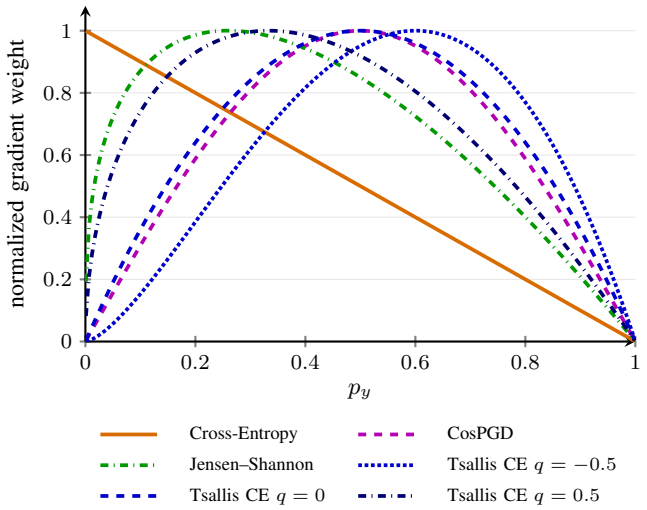

Fig. 1: **Comparison of per-pixel gradient weightings.** Best viewed in color. We compare per-pixel gradient weightings for different attacks as a function of the predicted probability p_y of the ground-truth class. Varying q in Tsallis cross-entropy shifts the location of the gradient peak, enabling targeted optimization over pixels at different confidence levels.

V. Experiments

We evaluate and compare TsallisPGD against state-of-the-art segmentation attacks. Our evaluation proceeds in three stages to motivate and validate the dynamic q-schedule. First, in Section V-A, we study *fixed-q* and show that no single value is universally optimal: the best q depends on the dataset, architecture, and threat budget. This motivates a dynamic schedule that sweeps q within a single attack. In Section V-B, we define a linear q-schedule parameterized by $(q_{\text{start}}, q_{\text{end}})$ and select these parameters using a held-out validation split, disjoint from the test set used in the final comparison. Finally, in Section V-C, we fix this validation-selected schedule and evaluate TsallisPGD against state-of-the-art segmentation attacks on the test set.

Datasets: We conduct experiments on PASCAL-VOC [30], ADE20K [31], and CITYSCAPES [32], following standard protocols and evaluating on the central crop [26].

Models: We evaluate diverse segmentation architectures, both standard and adversarially trained: PSPNet [33] (ResNet-50 backbone) with clean and DDCAT-robust weights, UPerNet [34] with a ConvNeXt-T backbone [35], and Segmenter [36] with a ViT-S backbone [37].

Threat model and metrics: We consider the ℓ_∞ threat model with perturbation radius ϵ_∞ (reported in the tables). We report (i) average pixel accuracy (ACC) and (ii) mean intersection-over-union (mIoU), where lower values indicate stronger attacks.

Baselines and protocol: We compare TsallisPGD against strong PGD-based objectives for segmentation: CEPGD (pixel-wise cross-entropy), SegPGD [24], CosPGD [25], JSPGD [26] (Jensen–Shannon divergence PGD), and MaskedPGD [26] (masked cross-entropy PGD). All attacks, including TsallisPGD, use APGD [17] for step-size selection, a fixed budget of T=300 iterations, a single random start, and the multi-ε trick from [26]. This ensures that performance differences are attributable to the attack objective.

A. Analysis of Fixed-q TsallisPGD

The Tsallis objective (Section IV) introduces a continuous parameter q that controls the concentration of per-pixel gradients. A natural question is whether a single fixed value of q is sufficient, or whether stronger attacks require varying q during optimization. To study this, we evaluate

TABLE I: **Effect of q on attack strength.** We evaluate TsallisPGD with fixed-q on the strongest robust model for each dataset: PSPNet (DDCAT) on Cityscapes, UPerNet (ConvNeXt-T) on VOC, and Segmenter (ViT-S) on ADE20K. *Best-of-q* reports the lowest value obtained by aggregating multiple fixed-q variants using SEA [26], representing the strongest combination of fixed-q attacks. All values are rounded to one decimal place; attacks tied at the rounded best value are all marked best. Lower values indicate stronger attacks. For each row and metric, the best fixed-q result is shown in **bold**.

ϵ_∞	Fixed-q												Best-of-q	
	$q=-3$		$q=-2$		$q=-1$		$q=0$		$q=0.5$		$q=1$			
	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU
CITYSCAPES: PSPNet+ResNet-50 **(DDCAT [29])** (Acc 94.9, mIoU 64.7)														
0.25/255	63.6	34.9	**63.2**	33.8	63.3	33.0	64.8	**32.5**	66.6	33.3	72.3	35.3	**61.9**	**30.4**
0.5/255	28.6	16.1	28.1	15.7	**27.8**	**14.7**	28.9	14.9	31.2	15.3	43.0	21.0	**24.6**	**12.4**
1/255	5.4	3.4	4.5	2.8	**4.1**	2.4	4.2	**2.0**	5.8	2.6	17.8	7.2	**2.0**	**1.1**
PASCAL-VOC: UPerNet+ConvNeXt-T **(PIR-AT [26])** (Acc 92.7, mIoU 75.2)														
4/255	90.0	67.7	89.7	67.0	89.3	66.1	88.9	65.2	**88.8**	**64.9**	89.0	65.5	**88.6**	**64.4**
8/255	76.4	43.5	75.1	41.9	73.9	40.5	**73.1**	**40.0**	73.9	41.4	77.4	46.1	**71.2**	**36.3**
12/255	32.3	10.6	31.6	**10.2**	**31.5**	**10.2**	34.2	12.1	38.5	15.0	52.4	22.9	**27.2**	**7.3**
ADE20K: Segmenter+ViT-S **(PIR-AT [26])** (Acc 69.1, mIoU 28.6)														
4/255	59.4	20.3	58.5	19.6	57.4	18.5	56.1	17.3	**55.8**	**16.8**	57.1	17.3	**55.2**	**15.9**
8/255	40.3	10.9	38.5	10.0	36.2	9.0	**34.5**	8.0	34.9	**7.8**	39.9	9.3	**32.9**	**6.5**
12/255	11.7	2.8	10.8	2.4	**9.8**	2.2	10.2	**2.1**	11.6	2.3	20.3	4.0	**8.3**	**1.4**

TsallisPGD with constant $q \in \{-3, -2, -1, 0, 0.5, 1\}$, which correspond to gradient peak locations ranging from 0.8 to 0.0, on the strongest robust model for each dataset: PSPNet (DDCAT [29]) on Cityscapes, UPerNet (ConvNeXt-T, PIR-AT [26]) on Pascal VOC, and Segmenter (ViT-S, PIR-AT [26]) on ADE20K. Here, $q = 1$ recovers the standard cross-entropy objective. We also report the *Best-of-q* result obtained by aggregating several fixed-q variants using SEA [26], which represents the best combination of multiple fixed-q attacks.

Table I shows that no single fixed value of q is optimal across datasets, models, perturbation budgets, or even evaluation metrics. For example, on Cityscapes, the lowest accuracy is achieved by $q = -2$ at $\epsilon_\infty = 0.25/255$, but by $q = -1$ at $0.5/255$ and $1/255$. On the other hand, the best mIoU is obtained with $q = 0$ at $0.25/255$ and $1/255$, and with $q = -1$ at $0.5/255$. The table also shows that standard CEPGD ($q = 1$) is consistently suboptimal. On VOC at $12/255$, accuracy drops from 52.4 with $q = 1$ to 31.5 with the best fixed choice of q. These results confirm that replacing standard cross-entropy with Tsallis cross-entropy, and tuning q, can substantially strengthen the attack.

Finally, the *Best-of-q* baseline is stronger than the best single fixed-q attack in every row, and often yields substantially stronger results. This suggests that different values of q expose different vulnerable subsets of pixels, so no single fixed choice can capture all of them. This directly motivates a *linear* dynamic q-schedule: by sweeping q during optimization, a single attack run can traverse multiple gradient-concentration regimes and target pixels at different confidence levels. In the next subsection, we describe how we choose the endpoints of this linear schedule, $(q_{\text{start}}, q_{\text{end}})$.

B. Selection of the Dynamic q-Schedule

The observation that the best fixed value of q varies across settings motivates sweeping q during optimization, allowing the attack to focus on pixels at different confidence levels throughout adversarial generation. This removes the need to tune q separately for each dataset, model, and perturbation budget and, when chosen appropriately, can approach the performance of the best combination of fixed-q attacks.

We consider a linear schedule parameterized by the range $[q_{\text{start}}, q_{\text{end}}]$. Specifically, we evaluate schedules with $q_{\text{start}} \in \{-3, -2, -1\}$ and $q_{\text{end}} \in \{0.5, 1.0\}$, using linear interpolation, for a total of six candidate schedules. Each candidate is evaluated on the *validation* split of the corresponding dataset, following the protocol in Section V-A. For each schedule, we compute its rank on every validation row and then average these ranks across all rows. The candidate with the best average validation rank is the linear sweep of q over $[-2, 1]$. We therefore select this schedule as the TsallisPGD schedule and use it unchanged for the final benchmark in Section V-C across all datasets, architectures, and perturbation budgets.

C. Comparison with State-of-the-Art Attacks

With the q-schedule fixed by the selection procedure in Section V-B, we compare TsallisPGD against state-of-the-art segmentation attacks on the full test benchmark. The same schedule is used unchanged across all datasets, architectures, and perturbation budgets, with no per-row tuning.

TABLE II: **Comparison of attacks.** We evaluate all attacks for 300 iterations on clean and robust models trained on PASCAL-VOC, ADE20K, and CITYSCAPES. TsallisPGD uses the dynamic q-schedule selected in Section V-B, applied identically across all rows without per-row tuning. We report average pixel accuracy (Acc) and mean IoU (mIoU), where lower values indicate stronger attacks. Clean performance is shown next to each model. Attacks are ranked independently for each row and metric, with lower ranks indicating stronger attacks; **Avg. Rank** reports the mean rank across all rows, datasets, models, and ϵ_∞ budgets. For each metric, the best result is shown in **bold** and the second best is underlined.

ϵ_∞	Previous SOTA attacks										Our attack	
	CEPGD		SegPGD		CosPGD		JSPGD		MaskedPGD		TsallisPGD	
	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU
CITYSCAPES: PSPNet+ResNet-50 **(clean training)** (Acc 95.3, mIoU 68.6)												
0.25/255	22.1	12.9	14.1	9.2	7.9	5.8	9.0	6.5	**6.5**	**5.0**	7.2	5.6
0.5/255	5.3	1.5	1.0	0.3	0.5	0.2	0.7	0.2	**0.2**	**0.1**	**0.2**	**0.1**
1/255	3.9	1.0	0.5	0.3	0.2	0.1	0.1	0.1	**0.0**	**0.0**	**0.0**	**0.0**
CITYSCAPES: PSPNet+ResNet-50 **(DDCAT [29])** (Acc 94.9, mIoU 64.7)												
0.25/255	72.0	34.9	66.9	34.1	64.7	**32.1**	66.0	32.4	63.7	32.8	**62.9**	32.8
0.5/255	42.2	20.0	36.2	18.3	27.1	14.0	29.2	14.3	**26.2**	**13.9**	27.3	14.1
1/255	16.2	6.4	12.4	5.8	3.0	**1.5**	3.8	1.8	**2.7**	1.6	4.1	2.1
PASCAL-VOC: PSPNet+ResNet-50 **(clean training)** (Acc 91.9, mIoU 85.9)												
0.25/255	48.1	22.1	36.5	14.4	33.7	13.9	36.5	16.1	32.5	13.1	**32.4**	**12.9**
0.5/255	23.1	11.2	8.3	3.3	5.6	2.7	7.7	4.2	**4.2**	**1.5**	4.3	1.7
1/255	13.2	7.3	1.1	0.9	0.2	0.3	0.7	0.8	**0.1**	**0.1**	**0.1**	**0.1**
PASCAL-VOC: PSPNet+ResNet-50 **(DDCAT [29])** (Acc 94.1, mIoU 76.9)												
0.25/255	72.3	36.8	68.3	33.8	67.3	**33.1**	68.2	33.6	**67.0**	33.4	**67.0**	33.2
0.5/255	46.5	18.0	33.8	12.5	30.3	11.6	33.1	12.9	**27.7**	10.6	28.1	**10.3**
1/255	22.7	8.3	10.1	3.8	3.1	1.5	5.0	2.8	2.4	**1.1**	**2.3**	**1.1**
PASCAL-VOC: UPerNet+ConvNeXt-T **(PIR-AT [26])** (Acc 92.7, mIoU 75.2)												
4/255	89.0	65.5	**88.7**	**64.7**	88.8	65.0	**88.7**	64.8	89.2	65.8	**88.7**	**64.7**
8/255	77.2	45.8	73.0	39.1	73.1	40.1	74.0	41.5	73.8	40.4	**72.8**	**39.0**
12/255	53.1	23.2	35.9	11.6	34.6	12.6	38.4	15.2	31.5	10.1	**30.7**	**9.7**
ADE20K: UPerNet+ConvNeXt-T **(PIR-AT [26])** (Acc 70.5, mIoU 31.8)												
4/255	57.5	19.8	55.9	**18.9**	55.9	19.0	55.9	19.0	57.0	20.0	**55.8**	**18.9**
8/255	35.2	8.9	28.4	7.2	27.6	**7.0**	28.4	7.3	28.5	7.5	**27.5**	7.1
12/255	13.2	2.5	5.3	1.2	3.9	1.0	4.8	1.2	**3.4**	**0.8**	**3.4**	**0.8**
ADE20K: Segmenter+ViT-S **(PIR-AT [26])** (Acc 69.1, mIoU 28.6)												
4/255	57.0	17.3	55.5	**16.5**	55.7	16.8	55.6	16.6	56.8	17.8	**55.4**	**16.5**
8/255	39.8	9.4	34.3	**7.5**	**34.0**	7.6	34.4	7.7	36.0	8.5	**34.0**	7.6
12/255	19.8	3.9	10.4	**2.0**	9.6	**2.0**	10.7	2.1	9.8	2.1	**9.1**	**2.0**
Avg. Rank	5.95	5.86	4.05	3.52	2.81	2.57	3.71	3.86	2.43	2.76	**1.38**	**1.57**

Results: Table II shows that TsallisPGD is the strongest attack on average, achieving the best average rank across all 21 evaluation settings: 1.38 in ACC and 1.57 in mIoU. For ACC, this improves over the strongest baseline, MaskedPGD (2.43), while for mIoU it outperforms CosPGD (2.57), establishing consistent gains across both metrics. At the per-setting level, TsallisPGD attains the best performance in $16/21$ settings for ACC and $13/21$ settings for mIoU (counting ties as best), indicating that its advantage is not limited to a small subset of configurations but holds broadly.

The gains are most pronounced on adversarially trained models, where optimization is typically more challenging. On VOC with UPerNet (ConvNeXt-T, PIR-AT), TsallisPGD achieves the best ACC at all three perturbation budgets and is best (or tied) in mIoU across all budgets. On ADE20K with UPerNet (ConvNeXt-T, PIR-AT), it is again best in ACC at all budgets and best (or tied) in mIoU at $4/255$ and $12/255$, while at $8/255$ it is narrowly second to CosPGD (7.1 vs. 7.0). Similarly, on ADE20K with Segmenter (ViT-S, PIR-AT), TsallisPGD is strongest at $4/255$ and $12/255$ on both metrics; at $8/255$ it achieves the best ACC but is marginally second in mIoU, where SegPGD performs best. On clean models, the results are more mixed but remain favorable: TsallisPGD is competitive across Cityscapes and VOC and achieves several clear wins, although MaskedPGD occasionally matches or slightly exceeds its performance at larger perturbation budgets. This behavior suggests that Tsallis-based reweighting is especially effective in more challenging robustness-evaluation regimes against adversarially trained models, where controlling the gradient focus across pixels becomes central to attack optimization. Overall, the results indicate that a single validation-selected dynamic Tsallis schedule generalizes across datasets, architectures, and perturbation budgets, outperforming prior state-of-the-art segmentation attacks. A natural direction for future work is to make the choice of q adaptive, for example by selecting q automatically at each attack iteration rather than relying on a fixed, predetermined schedule.

VI. Conclusion

We analyze the limitations of standard cross-entropy as an attack objective for semantic segmentation and introduce **TsallisPGD**, which replaces pixel-wise cross-entropy with Tsallis cross-entropy. We show that the parameter q controls which confidence levels the attack emphasizes, allowing it to target different subsets of pixels during optimization. Because the best choice of q depends on the dataset, model, and perturbation budget ϵ_∞, no single fixed value is optimal across settings. This motivates a *dynamic* q-schedule that varies q throughout the attack. Using a held-out validation set, we select the best linear schedule, which sweeps q from -2 to 1. Across CITYSCAPES, PASCAL-VOC, and ADE20K, TsallisPGD with this schedule achieves the best overall average rank among CEPGD, SegPGD, CosPGD, JSPGD, and MaskedPGD in both global accuracy and mIoU. These results show that controlling gradient concentration through Tsallis cross-entropy provides a simple and general way to construct stronger adversarial attacks for semantic segmentation. In future work, we plan to make the choice of q more adaptive and incorporate TsallisPGD into adversarial training to improve robustness.

Acknowledgment

This research is supported by the National Research Foundation, Singapore, under its AI Singapore Programme (AISG Award No: AISG4-GC-2023-006-1B), the Ministry of Education, Singapore, under its Academic Research Tier-1 Grant (Award No: 1091/R-MA124-R205-0002), A*STAR under its MTC Individual Research Grant (Award No: M23M6c0113), and MTC Programmatic Grant (Award No: M23L9b0052).

References

[1] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, 2018.

[6] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: simple and efficient design for semantic segmentation with transformers," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021.

[7] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019.

[8] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021.

[9] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[10] Z. Teed and J. Deng, "RAFT: recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[11] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, 2020.

[12] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep

learning in medical image analysis," *Med. Image Anal.*, vol. 42, 2017.

[13] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, "Measuring robustness to natural distribution shifts in image classification," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.

[14] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do ImageNet classifiers generalize to ImageNet?" in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019.

[15] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014.

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[17] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[19] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017.

[20] S. Thys, W. Van Ranst, and T. Goedemé, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019.

[21] Z. Jin, X. Ji, Y. Cheng, B. Yang, C. Yan, and W. Xu, "PLA-LiDAR: physical laser attacks against LiDAR-based 3D object detection in autonomous vehicle," in *Proc. IEEE Symp. Secur. Privacy (S&P)*, 2023.

[22] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, "Deep text classification can be fooled," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018.

[23] A. Arnab, O. Miksik, and P. H. S. Torr, "On the robustness of semantic segmentation models to adversarial attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.

[24] J. Gu, H. Zhao, V. Tresp, and P. H. S. Torr, "SegPGD: an effective and efficient adversarial attack for evaluating and boosting segmentation robustness," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022.

[25] S. Agnihotri, S. Jung, and M. Keuper, "CosPGD: an efficient white-box adversarial attack for pixel-wise prediction tasks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024.

[26] F. Croce, N. D. Singh, and M. Hein, "Towards reliable evaluation and fast training of robust semantic segmentation models," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.

[27] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. Int. Conf. Learn. Represent. Workshop (ICLR Workshop)*, 2017.

[28] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018.

[29] X. Xu, H. Zhao, and J. Jia, "Dynamic divide-and-conquer adversarial training for robust semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021.

[30] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[31] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," *Int. J. Comput. Vis.*, vol. 127, no. 3, 2019.

[32] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[33] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[34] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.

[35] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.

[36] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.