

TsallisPGD: Adaptive Gradient Weighting for Semantic-Segmentation Attacks

Alexander Matyasko · Xin Lou · Indriyati Atmosukarto · Wei Zhang

Singapore Institute of Technology, Singapore

Main idea

TsallisPGD uses Tsallis cross-entropy to control gradient allocation across pixels, achieving the best average attack rank against PGD-style baselines across 21 settings spanning Cityscapes, PASCAL VOC, and ADE20K.

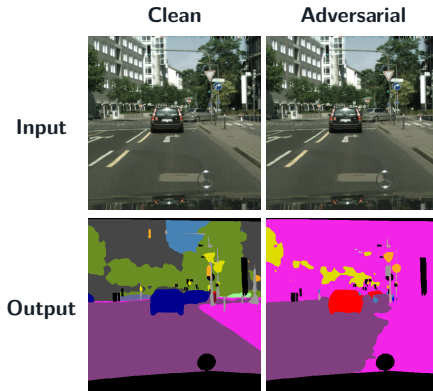


Project page

Background: Safety-Critical Segmentation

Safety-critical dense prediction

- ▶ **Autonomous systems:** depend heavily on pixel-level segmentation accuracy.
- ▶ **Coupled predictions:** one input perturbation affects thousands of pixel outputs simultaneously.
- ▶ **Catastrophic failures:** adversarial noise can hide pedestrians or misclassify roads, leading to safety hazards.



Dense attacks are a gradient-allocation problem

Dense attack story

Scene → Perturbation → Prediction
clean mask → $+\delta$ → **wrong labels**

What changes for segmentation

- ▶ One perturbation drives thousands of pixel losses.
- ▶ Pixel losses compete inside one aggregated objective.
- ▶ Attack strength depends on where gradients are spent.

Gradient waste

An attack can spend update budget on pixels that are already fooled while still-correct pixels remain under-attacked.

Talk thesis

TsallisPGD turns the loss into a control knob for gradient allocation across pixels.

Prior dense attacks also change pixel influence

Discrete weighting

- ▶ SegPGD anneals weights between correct and incorrect pixels (Gu et al., 2022).
- ▶ MaskedPGD removes already-misclassified pixels from CE (Croce et al., 2025).

Alternative dense losses

- ▶ CosPGD scales each pixel by prediction-target alignment (Agnihotri et al., 2024).
- ▶ JSPGD uses Jensen–Shannon divergence to reduce CE’s failure mode (Croce et al., 2025).

Gap

Existing attacks control pixel influence using masks, schedules, or alternative losses. TsallisPGD provides a continuous confidence-based knob while keeping the same optimizer and threat model.

Segmentation turns one loss into many competing losses

From one decision to many decisions

Classification

one label →

one loss →

flip label →

Segmentation

thousands of labels

many competing losses

degrade map

Toy segmentation map

Pixel group	Current state	Attack value
road	correct	useful
car edge	vulnerable	useful
sky	already fooled	low value
sign	correct	useful

Evaluation risk

If a loss keeps rewarding low-value pixels, the adversarial example can look weaker than the model really is. Robustness evaluation therefore depends on allocating gradient to pixels that still matter.

Croce et al. (2025) analyze this failure mode and motivate alternative dense-attack objectives.

Cross-Entropy overemphasizes already misclassified pixels

Cross-entropy gradient

Let p_y denote the softmax probability assigned to the ground-truth class.

$$L_{\text{CE}}(p, e_y) = -\log p_y, \quad \nabla_u L_{\text{CE}} = p - e_y.$$

The gradient's norm can be upper-bounded as:

$$\|\nabla_u L_{\text{CE}}(u, e_y)\|_2^2 \leq 1 - p_y + (1 - p_y)^2.$$

Lower p_y receives larger CE gradient.

Two-pixel toy example

Pixel	Status	p_y	CE upper	Share
A	still correct	0.80	0.24	11.5%
B	already fooled	0.05	1.85	88.5%

Takeaway

CE puts 88.5% of the weight on an already-fooled pixel.

TsallisCE reweights CE by confidence

One-parameter replacement for per-pixel CE

TsallisCE changes the loss geometry without changing the attack optimizer:

$$\mathcal{L}_q(p_y) = \frac{1 - p_y^{1-q}}{1 - q}, \quad q \neq 1.$$

The derivative gives an exact confidence-dependent rescaling:

$$\nabla_u \mathcal{L}_q(u, e_y) = p_y^{1-q} \nabla_u \mathcal{L}_{CE}(u, e_y).$$

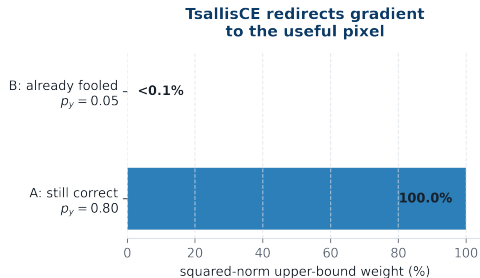
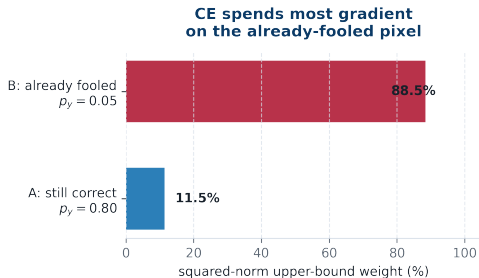
- ▶ $q \rightarrow 1$ recovers standard cross-entropy.
- ▶ $q < 1$ softly downweights low- p_y pixels.
- ▶ q steers which confidence band receives gradient.

TsallisPGD

- ▶ Same APGD update machinery.
- ▶ Same perturbation budgets.
- ▶ Same random-start protocol.
- ▶ Only the objective changes.

Toy example: TsallisCE redirects gradient

Same two pixels, different gradient allocation

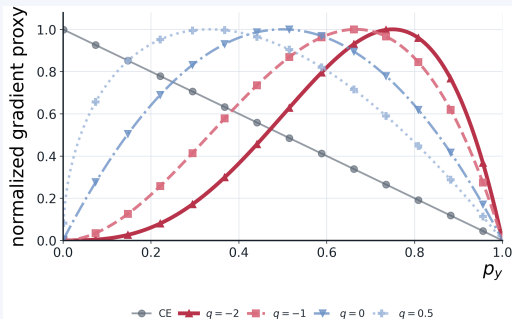


Mechanism

With $q = -1$, TsallisCE multiplies CE gradients by p_y^2 , so squared-norm weights get a p_y^4 factor.

q selects which confidence band to attack

Gradient emphasis across p_y



q as a gradient-allocation knob

$$\|\nabla_u \mathcal{L}_q(u, e_y)\|_2^2 = p_y^{2(1-q)} \|\nabla_u L_{\text{CE}}(u, e_y)\|_2^2.$$

The proxy gradient is maximized at

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

Peak is for the squared-norm upper-bound proxy.

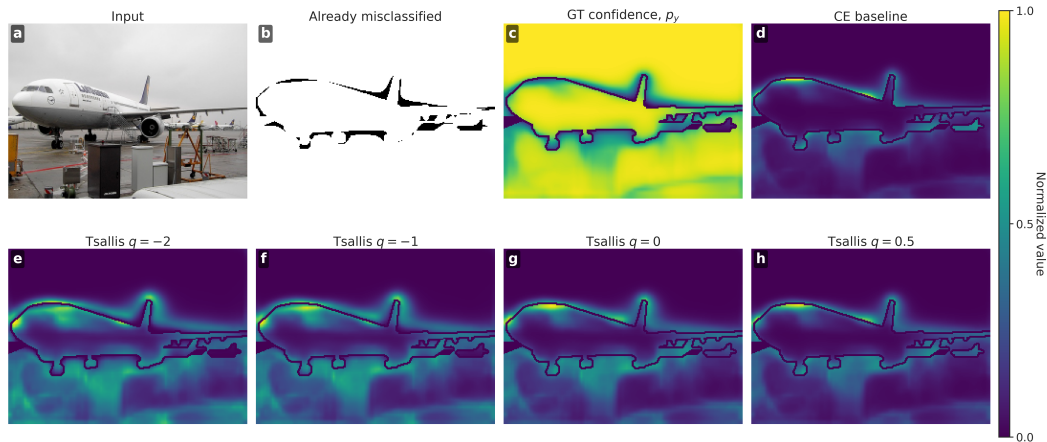
Confidence peak for plotted q values

q	-2	-1	0	0.5
p_y^*	0.75	0.67	0.50	0.33

higher confidence $\rightarrow$ lower confidence

Smaller q : higher-confidence targets.
Larger q : CE-like refinement.

Gradient maps show the allocation shift



Lower (q) suppresses misclassified/low-confidence regions and shifts gradient mass toward high-confidence ground truth.

A fixed q -schedule sweeps across confidence bands

Attack strategy

No single q is universally best:

- ▶ Early iterations attack confident correct pixels, which CE under-emphasizes.
- ▶ Middle iterations sweep through confidence bands.
- ▶ Late iterations return to CE-like geometry to refine the adversarial example.

Validation-selected schedule

$$q_t = q_{\text{start}} + (q_{\text{end}} - q_{\text{start}}) \frac{t}{T-1}, \quad t = 0, \dots, T-1.$$

Selected once on validation data, then frozen across all test datasets, architectures, and perturbation budgets.

Early

$$q = -2$$

high-confidence correct pixels

Middle

q sweep

multiple confidence regimes

Late

$$q = 1$$

CE-like optimization geometry
CE limit, special-cased

Benchmark isolates the objective change

Only the attack objective changes.

All methods use the same optimizer, iteration budget, random-start protocol, threat model, and evaluation metrics.

Evaluation scope

Datasets Cityscapes, PASCAL VOC, ADE20K
Models PSPNet, UPerNet, Segmenter
Metrics pixel accuracy, mIoU

Attack protocol

Baselines CE-PGD, SegPGD, CosPGD, JSPGD, MaskedPGD
Optimizer APGD step-size selection
Budget $T = 300$, one random start

Evaluation criterion

Under the same protocol, lower pixel accuracy and lower mIoU indicate a stronger attack.

Source coverage

Datasets, models, and baselines follow standard semantic-segmentation robustness protocols.

Results: one fixed schedule is strongest on average

Main outcome

Across 21 dataset/model/ ϵ_∞ settings, one attack with q -schedule gives the best average rank for both accuracy and mIoU.

Average rank across all settings

Metric	TsallisPGD	Best baseline	Wins/ties
Acc	1.38	MaskedPGD 2.43	16/21
mIoU	1.57	CosPGD 2.57	13/21

Robust models

- ▶ PASCAL VOC UPerNet: best across budgets.
- ▶ ADE20K: strongest robust-model gains.
- ▶ Known exception: MaskedPGD edges TsallisPGD on Cityscapes DDCAT at 1/255.

Clean models

Competitive on Cityscapes/PASCAL VOC;
MaskedPGD can match it at larger budgets.

Interpretation

The largest gains are on robust models, where
gradient allocation matters most.

Conclusion and Future Work

Conclusion

TsallisPGD shows that loss geometry can control gradient allocation across pixels, improving robustness evaluation for semantic segmentation.

Adaptive attacks

- ▶ Select q_t from confidence histograms.

Beyond segmentation

- ▶ Extend to VLMs and multimodal models.

Robust training

- ▶ Use Tsallis-style weighting for robust training.

Takeaway: TsallisPGD turns the loss function into a controllable gradient-allocation mechanism for dense adversarial evaluation.

Thank you



Project page

github.com/aam-at/tsallis_pgd