RL-SAF: REINFORCEMENT LEARNING-BASED SEMANTIC ADVERSARIAL FAIRNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarially trained models can achieve strong overall robustness while exhibiting substantial class-wise performance disparities under attack. Although this adversarial fairness problem has been studied in image classification, it remains largely unexplored in semantic segmentation. We show that such disparities occur across datasets, architectures, and adversarial training strategies, that adversarial perturbations amplify rather than uniformly degrade class-wise performance, and that class frequency alone does not explain which classes are most vulnerable. We further show that adapting a state-of-the-art adversarial fairness method from image classification to semantic segmentation reduces both worst-class and overall robustness, which we attribute to a mismatch between recall-based notions of class difficulty and the IoU objective of dense prediction. Motivated by these findings, we propose RL-SAF (Reinforcement Learning-Based Semantic Adversarial Fairness), which combines class-level loss aggregation with a policy-gradient controller that adaptively learns class-wise loss weights from validation feedback. On Cityscapes with DeepLabV3, RL-SAF improves the adversarial IoU of the most vulnerable classes across all evaluated attacks and preserves robust mIoU under the attack used for training, while trading off part of it under stronger attacks. These gains come at a substantial cost in clean accuracy, revealing a clear trade-off between adversarial fairness and standard segmentation accuracy. Overall, our results establish class-wise adversarial robustness disparities as an important challenge in semantic segmentation and provide initial evidence to motivate adaptive, fairness-aware adversarial training for dense prediction tasks.

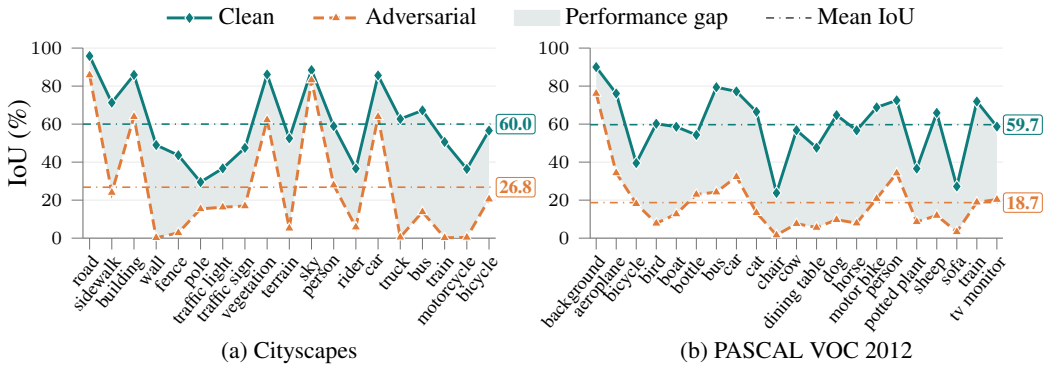
1 INTRODUCTION

Semantic segmentation models are increasingly evaluated under adversarial perturbations, with existing work primarily focusing on improving overall robustness measured by mean Intersection over Union (mIoU) (Halmosi et al., 2024; Gu et al., 2022; Chen et al., 2026; Arnab et al., 2020; Croce et al., 2025). However, strong overall robustness does not imply that robustness is distributed uniformly across semantic classes. In dense prediction tasks, large class-wise differences can be particularly important, since individual classes may be safety-relevant and their failures can be hidden by a single overall robustness score (Agnihotri et al., 2024; Chen et al., 2026; Halmosi et al., 2024).

Figure 1 illustrates the class-wise disparities for clean and adversarially perturbed images. Despite achieving high overall segmentation robustness, adversarially trained DeepLabV3 (Chen et al., 2017) models exhibit large variation in per-class IoU on both Cityscapes and PASCAL VOC 2012 (Cordts et al., 2016; Everingham et al., 2015). While some classes remain relatively robust under attack, others deteriorate to near-zero IoU. This suggests that overall robustness alone provides an incomplete view of adversarial performance in semantic segmentation.

In image classification, such disparities in class-wise adversarial robustness have been studied under the notions of robust fairness or adversarial fairness (Benz et al., 2021; Xu et al., 2021), and several methods have been proposed to mitigate them (Xu et al., 2021; Lee et al., 2024; Wei et al., 2023; Sun et al., 2023; Medi et al., 2025; 2026). In semantic segmentation, however, this problem remains largely unexplored. It is therefore unclear to what extent adversarial training induces or amplifies class-wise robustness disparities in dense prediction, and what factors contribute to these disparities. A natural explanation for disparities emerging in semantic segmentation, beyond of what is observed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Cityscapes (b) PASCAL VOC 2012

Figure 1: Per-class IoU (%) of DeepLabV3 trained with SegPGD-AT, evaluated on clean and adversarial inputs for (a) Cityscapes and (b) PASCAL VOC 2012. The shaded region indicates the performance gap between clean and adversarial inputs, and the dash-dotted lines mark the mean IoU. Robustness is evaluated using a 3-step SegPGD attack with $\epsilon = 8/255$.

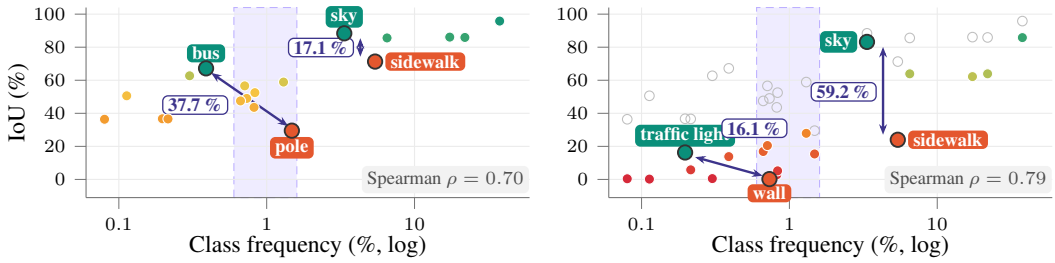
in image classification, is the strong class imbalance present in segmentation datasets (Arnab et al., 2020; Li et al., 2022; Zhang et al., 2023; Cordts et al., 2016; Everingham et al., 2015; Hoyer et al., 2022). However, our analysis shows that class frequency alone does not explain adversarial vulnerability as shown in Figure 2. By jointly examining class-wise IoU, frequency-normalized model prediction bias, and correct prediction mass, which are derived from metrics defined in a class-wise analysis of image classification in Cui et al. (2024), we find that vulnerable classes can simultaneously attract disproportionately many false-positives and receive too few true-positives, with both becoming more pronounced under adversarial perturbations.

In image classification, the robust fairness problem has been addressed through class-aware strategies such as class-wise loss reweighting, attack-margin scaling, optimization constraints, and class-dependent adversarial example generation (Wei et al., 2023; Lee et al., 2024; Xu et al., 2021; Sun et al., 2023). However, it remains unclear whether the assumptions underlying these methods transfer to semantic segmentation. We examine this question by adapting a recent adversarial fairness work, Distance-Aware Fair Adversarial Training (DAFA) (Lee et al., 2024), to semantic segmentation. DAFA characterizes class vulnerability through class-wise prediction similarities and confusion patterns, which appears compatible with the prediction bias observed in our analysis (section 3) (Lee et al., 2024). Nevertheless, our experiments show that its benefits do not transfer to dense prediction: all evaluated DAFA variants reduce robust worst class performance, overall robustness, and clean accuracy. This suggests that classification-oriented notions of class difficulty do not directly capture the class-wise behavior of semantic segmentation under adversarial perturbations.

These findings indicate that class difficulty in semantic segmentation cannot be reliably characterized using assumptions developed for image classification. Motivated by this limitation, we introduce RL-SAF, a policy-guided adversarial fairness framework for semantic segmentation. RL-SAF combines class-level loss aggregation with a policy-gradient controller that learns class-specific loss weights from the performance of the segmentation model on a held-out policy validation set. This allows the fairness objective to adapt to evolving class-wise robustness during training, rather than relying on a fixed definition of class difficulty.

Our main contributions are summarized as follows:

- ✓ We assess class-wise adversarial robustness disparities in semantic segmentation across datasets, attacks, and adversarial training strategies. We show that adversarial perturbations disproportionately degrade vulnerable classes and that class frequency alone does not explain this behavior.
- ✓ We examine the transferability of adversarial fairness methods from image classification to semantic segmentation. By adapting Distance-Aware Fair Adversarial Training (DAFA) to the dense prediction setting, we show that its benefits do not transfer effectively. We relate this



(a) Clean inputs (b) Adversarial inputs

Figure 2: Class frequency vs. class-wise IoU of DeepLabV3 trained with SegPGD-AT on Cityscapes. Points are coloured by IoU; grey rings in (b) mark clean IoU. Frequency correlates with IoU, more strongly under attack, but does not fully determine it. For example, sky (3.3%) and sidewalk (5.4%) have similar frequencies yet differ in IoU by 17.1% on clean inputs and 59.2% under attack, while the rare traffic light (0.2%) stays more robust than the more frequent wall (0.7%). Robustness is evaluated with a 3-step SegPGD attack ($\epsilon = 8/255$).

limitation to a mismatch between classification-based notions of class difficulty and IoU-based segmentation behavior with class imbalance.

✓ We propose RL-SAF, a reinforcement learning-based adversarial fairness framework that combines class-level loss aggregation with adaptively learned class-wise weighting using a policy. By guiding the fairness objective via model validation feedback, RL-SAF improves the adversarial robustness of the most vulnerable classes while largely preserving overall robustness.

2 Background

Adversarial training for semantic segmentation. Unlike image classification, semantic segmentation predicts a label for every pixel, such that the training objective aggregates losses over a dense set of spatial predictions (Minaee et al., 2022; Long et al., 2015). Adversarial training extends this objective by minimizing the worst-case segmentation loss within a constrained perturbation set, typically using projected-gradient methods such as PGD (Madry et al., 2018; Gu et al., 2022; Chen et al., 2026). In semantic segmentation, however, a single perturbation jointly affects many spatial locations and semantic classes, and robustness properties observed for image classification do not necessarily transfer directly to dense prediction (Arnab et al., 2020; Halmosi et al., 2024). This has motivated segmentation-specific attacks such as SegPGD and CosPGD (Gu et al., 2022; Agnihotri et al., 2024). SegPGD adapts projected-gradient optimization by dynamically balancing the contributions of correctly and incorrectly classified pixels to obtain stronger attacks within a limited number of iterations (Gu et al., 2022). More recent work has likewise shown that reliable robustness evaluation for pixel-wise prediction benefits from task-specific attack objectives and sufficiently strong optimization procedures (Agnihotri et al., 2024; Croce et al., 2025). In this work, we use both standard PGD and SegPGD to examine how adversarial training affects not only overall segmentation performance, but also the distribution of adversarial robustness across semantic classes.

Adversarial fairness. Adversarial training improves overall robustness, but this objective can mask substantial class-wise disparities, where some classes remain considerably less robust than others. Importantly, such disparities have been observed even when class frequency is not the primary cause, meaning that a model with strong average robustness may still fail systematically on specific classes (Benz et al., 2021; Xu et al., 2021). This phenomenon is commonly referred to as the adversarial fairness problem. In image classification, prior work has sought to mitigate this issue through constrained optimization, source-target class modeling for adversarial attacks, class-wise loss reweighting, adaptive margins, and vulnerability-aware calibration (Lee et al., 2024; Wei et al., 2023; Sun et al., 2023; Xu et al., 2021). Despite these efforts, adversarial fairness has been studied far less extensively in dense prediction tasks. In this work, we provide a comprehensive analysis of adversarial fairness in semantic segmentation, examining how robustness disparities emerge at the

class level under adversarial perturbations. (Halmosi et al., 2024) re-evaluate well-known adversarially trained semantic segmentation models with an ensemble of new and existing strong attacks, showing that their robustness has been substantially overestimated in prior work; building on this benchmark, we use the models from their main results to analyze adversarial fairness across classes in segmentation.

Measuring class-wise prediction bias. Class-wise robustness scores show which classes degrade under adversarial perturbations, but not where the resulting prediction mass is redistributed. Following Cui et al. (2024), we therefore measure the relative incorrect probability mass, indicating the model prediction bias (MPB), and correct probability mass (CPM) associated with each class. To account for the strong class imbalance in semantic segmentation, we normalize both quantities by the relative class frequency (CF). The resulting scores,

$$\frac{\mathrm{MPB}_{\mathrm{rel}}(c)}{\mathrm{CF}_{\mathrm{rel}}(c)} \quad \text{and} \quad \frac{\mathrm{CPM}_{\mathrm{rel}}(c)}{\mathrm{CF}_{\mathrm{rel}}(c)}, \tag{1}$$

indicate whether class c attracts incorrect or correct probability mass disproportionately to its frequency. We use these measures in section 3 to analyze how adversarial perturbations redistribute predictions across classes. Full definitions are provided in section A.

Adaptive loss weighting and meta-learning. Reweighting strategies are commonly used in long-tailed learning to address unequal class or sample contributions directly in the optimization objective. Approaches such as class-balanced loss (Cui et al., 2019) or focal loss (Lin et al., 2017) derive fixed weights from class frequency or prediction confidence, while DAFA (Lee et al., 2024) similarly uses class-wise statistics collected during training. Meta-learning approaches make this weighting adaptive by incorporating feedback from the training process (Shu et al., 2019; Yoon et al., 2020; Huang et al., 2019). For example, Ren et al. (2018) and Shu et al. (2019) use validation performance to learn sample weights, while DVRL (Yoon et al., 2020) and Adaptive Loss Alignment (Huang et al., 2019) formulate adaptation as a sequential decision process. RL-SAF follows this principle at the class level by adapting loss weights over the course of training in response to evolving class-wise robustness disparities across semantic categories.

3 Adversarial Fairness in Semantic Segmentation

Following Cui et al. (2024), we compare class-wise behavior on clean and adversarially perturbed inputs for standard and adversarially trained DeepLabV3 and PSPNet models on Cityscapes and PASCAL VOC 2012 (Chen et al., 2017; Zhao et al., 2017; Cordts et al., 2016; Everingham et al., 2015), using the training configurations of Halmosi et al. (2024). We focus on Cityscapes here; VOC 2012 results are provided in Appendix A.

Training approaches. We evaluate standard training as well as adversarial training with PGD and the segmentation-specific SegPGD attack (Madry et al., 2018; Gu et al., 2022). Following Halmosi et al. (2024), we distinguish between adversarial training that combines clean and adversarial examples within each mini-batch and training performed exclusively on adversarial examples. We denote these settings as SegPGD-AT and SegPGD-AT-100, respectively, and use PGD-AT for the corresponding configuration based on standard PGD.

Class-wise scores. Semantic segmentation performance is typically summarized by mean Intersection over Union (mIoU), which can conceal substantial differences between individual classes. Figure 3 compares class-wise IoU on clean and adversarial inputs. As expected, adversarially trained models sacrifice some clean performance while substantially improving robustness under attack. More importantly, adversarial evaluation reveals a pronounced lower-performing tail and greater variation across classes. The change in class-wise performance cannot be explained by mIoU alone and motivates a closer analysis of how prediction errors are distributed across classes.

Model prediction bias. To understand how these disparities originate, we analyze the model prediction bias (section 2). In absolute terms, frequent classes attract most of the false-positive mass, consistent with findings in classification (Cui et al., 2024). Normalizing by pixel frequency changes this picture (Figure 4). Frequent classes such as *road* stay at or below the frequency-proportional level of 1, whereas infrequent classes such as *pole*, *traffic light*, and *motorcycle* receive roughly 5–7$\times$ more false-positive mass than their prevalence implies. Under attack, the normally trained model

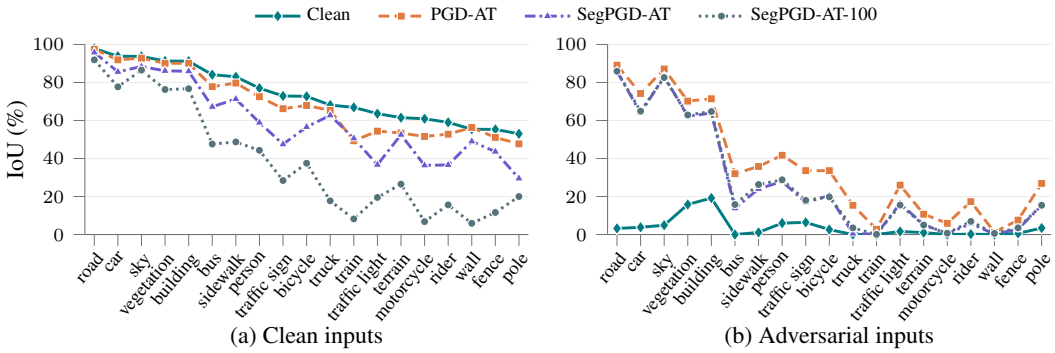


Figure 3: Per-class IoU on Cityscapes for a normally trained model (Clean) and three adversarially trained models, evaluated on (a) clean and (b) adversarial inputs. Classes are ordered by the clean-model IoU on clean inputs. Robust evaluation uses a 3-step SegPGD attack with $\epsilon = 8/255$.

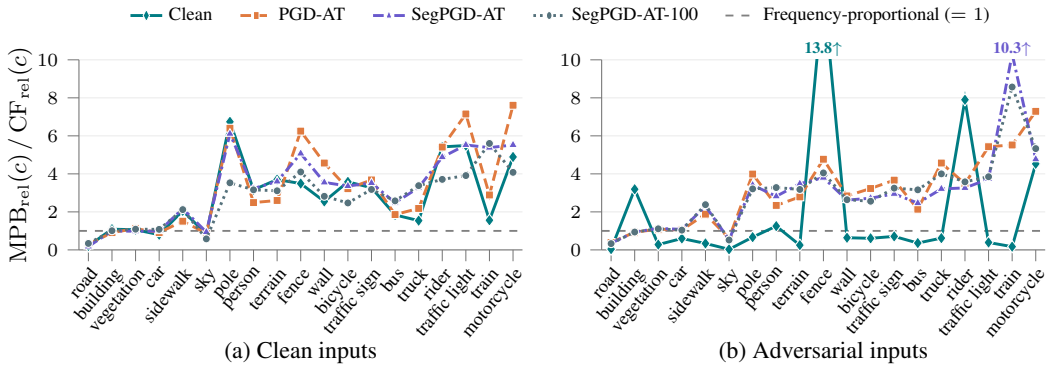


Figure 4: Frequency-normalized model prediction bias $\mathrm{MPB}_{\mathrm{rel}}(c)/\mathrm{CF}_{\mathrm{rel}}(c)$ for different training approaches on Cityscapes using DeepLabV3, on (a) clean and (b) adversarial inputs. Classes are sorted by decreasing pixel frequency (left to right). The dashed line at 1 marks the bias expected from class frequency alone; values above it indicate that a class attracts more false-positive probability mass than its pixel frequency implies. Robust evaluation uses 3-step SegPGD with $\epsilon = 8/255$.

distributes its errors into a few classes (e.g., $13.8\times$ for *fence*). Adversarially trained models keep a systematic skew toward rare classes, reaching $10.3\times$ for *train* with SegPGD-AT. Classes of similar frequency also differ widely, so frequency alone does not explain class-wise vulnerability. The frequency-normalized correct prediction mass shows the complementary trend: it decreases toward rare classes and drops further under attack (Figure 11). Vulnerable classes therefore absorb excess false positives while receiving too few correct predictions. Aggregate robustness metrics hide this disparity, and adversarial training alone does not remove it.

Adversarial perturbations amplify class-wise disparities. We analyze adversarial fairness using the training strategies benchmarked by Halmosi et al. (2024). We use their released checkpoints where available and replicate the remaining models with their settings. Table 1 shows that robust mIoU substantially underestimates class-level vulnerability: in 15 of 16 configurations, the amplification factor exceeds one ($A > 1$), i.e., the worst class degrades disproportionately relative to the average. The effect is strongest under adversarial training. For PGD-AT with DeepLabV3 on Cityscapes, the attack reduces mIoU by 48% but worst-class IoU by over 97% ($A \approx 2.0$), while the best class stays largely intact, widening the best–worst gap from 49.78 to 87.89%. This gap widens in all PGD-AT and SegPGD-AT configurations, and the precision gap widens in all 16, indicating that adversarial training mainly protects classes that are already well segmented. Stronger training does not help: SegPGD-AT-100 yields the highest amplification on Cityscapes ($A > 3.2$), and its

	mIoU			Worst-class IoU			Amplification factor A	Best–worst gap (pp)					
								IoU		Precision		Recall	
	Clean	Adv.	Drop (%)	Clean	Adv.	Drop (%)		Clean	Adv.	Clean	Adv.	Clean	Adv.
Normal Training													
PN+P	76.73	6.77	91.2	37.09	0.42	98.9	1.08	57.20	29.79	51.99	69.26↑	45.87	47.44↑
DL+P	77.21	7.32	90.5	37.91	0.20	99.5	1.10	56.46	29.01	51.07	70.69↑	43.06	36.00
PN+CS	73.25	3.76	94.9	53.60	0.00	100.0	1.05	44.46	17.78	24.06	75.40↑	39.38	64.29↑
DL+CS	73.76	3.67	95.0	53.00	0.00	100.0	1.05	44.90	19.19	24.87	73.38↑	37.09	68.28↑
PGD-AT													
PN+P	74.16	34.52	53.5	37.56	7.54	79.9	1.50	56.05	62.75↑	52.99	84.93↑	53.76	59.94↑
DL+P	70.59	42.08	40.4	30.23	9.80	67.6	1.67	62.27	72.11↑	59.03	79.43↑	44.08	61.97↑
PN+CS	66.94	34.52	48.4	29.76	0.49	98.4	**2.03**	66.95	87.21↑	32.03	84.51↑	68.86	97.13↑
DL+CS	68.85	35.92	47.8	47.69	1.21	97.5	**2.04**	49.78	87.89↑	27.07	84.99↑	45.56	95.10↑
SegPGD-AT													
PN+P	59.87	16.67	72.2	22.93	3.02	86.8	1.20	67.21	71.93↑	44.40	76.23↑	66.69	83.83↑
DL+P	59.67	18.75	68.6	23.80	1.85	92.2	1.34	66.27	74.32↑	39.93	78.28↑	66.59	86.29↑
PN+CS	58.49	25.14	57.0	29.17	0.03	99.9	1.75	66.56	84.39↑	48.80	88.31↑	61.26	94.91↑
DL+CS	60.04	26.78	55.4	29.46	0.18	99.4	1.79	66.32	85.60↑	47.84	89.54↑	55.82	94.60↑
SegPGD-AT-100													
PN+P	43.88	18.32	58.3	3.95	1.74	56.0	0.96	82.16	74.64	48.00	79.20↑	92.76	86.34
DL+P	42.07	19.05	54.7	4.88	1.74	64.3	1.18	80.84	75.35	56.74	80.02↑	84.31	87.07↑
PN+CS	36.71	25.63	30.2	1.31	0.00	100.0	**3.31**	89.07	84.79	49.43	89.33↑	96.35	95.23
DL+CS	39.34	27.42	30.3	5.90	0.07	98.8	**3.26**	85.96	85.87	44.01	90.69↑	91.15	93.96↑

Table 1: Adversarial fairness under a 3-step SegPGD attack with $\epsilon = 8/255$. Drop is the relative decrease from clean to adversarial performance. The amplification factor A is the ratio of the relative worst-class IoU drop to the relative mIoU drop, with $A > 1$ indicating disproportionate degradation of the most vulnerable class. Shading denotes A (light: $1 \leq A < 1.5$, medium: $1.5 \leq A < 2$, dark: $A \geq 2$). Best–worst gap is the difference between the best and worst class-wise IoU, precision, or recall; ↑ indicates a gap that widens under attack. DL = DeepLabV3, PN = PSPNet, P = VOC 2012, and CS = Cityscapes. More evaluations are reported in Appendix Tables 5 and 6.

lower performance on clean worst-class IoU reflect the need for fairer robustness. Overall metrics alone are therefore insufficient for assessing robust fairness.

4 Do Classification Adversarial Fairness Frameworks Transfer to Segmentation?

We adapt Distance-Aware Fair Adversarial Training (DAFA) (Lee et al., 2024), which attributes class-wise robustness disparities to predictive similarity due to confusion between similar classes and assigns larger loss weights and attack margins to difficult classes. For segmentation, we normalize its class-wise confusion statistics by per-class pixel counts and evaluate both the original TRADES-based objective (Zhang et al., 2019) (DAFA-TRADES) and a variant using class-weighted cross-entropy on SegPGD examples (Gu et al., 2022) (DAFA-CE); details are given in Appendix B.

DAFA does not reduce class-wise disparities. As shown in Table 2 (section 6), both variants of DAFA perform worse than SegPGD-AT on the worst, bottom-3, and bottom-5 classes, while robust mIoU drops by roughly half (26.78% → 13.75% and 13.05%) on SegPGD-3, and clean mIoU also decreases. Unlike in classification, where DAFA improves worst-class robustness at a small cost in average robustness (Lee et al., 2024), it benefits neither fairness nor overall robustness in segmentation. We attribute this to a metric mismatch. DAFA measures class difficulty as the mean predicted probability of the ground-truth class, a soft version of recall, which in classification coincides with the evaluation metric (Lee et al., 2024). Segmentation, however, is evaluated by

$$\mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \tag{2}$$

which penalizes false positives and false negatives equally (Long et al., 2015; Wang et al., 2023). Upweighting a difficult class reduces its false negatives but assigns more pixels of other classes to it, so the recall gain is offset by additional false positives. Since vulnerable classes already attract

excess false positives (section 3), DAFA addresses only one of their two failure modes and can worsen the other. Moreover, its weights are fixed after warm-up and cannot adapt as class-wise robustness evolves. Fairness-aware adversarial training for segmentation thus requires (i) objectives aligned with IoU rather than recall and (ii) weights that adapt to the model's current class-wise robustness, which motivates RL-SAF (section 5).

5 Reinforcement Learning for Adversarial Fairness

Our analysis identifies two major requirements for fairness-aware adversarial training in semantic segmentation. First, the training signal of vulnerable classes must not be dominated by frequent classes, since pixel-wise objectives let large regions govern the adversarial improvement. Second, class weights should adapt to the model's evolving class-wise robustness rather than rely on a fixed, classification-oriented notion of class difficulty. We address these requirements with RL-SAF, which combines class-level loss aggregation with a policy-gradient controller that learns class-wise loss weights from validation feedback. The overall procedure is summarized in Algorithm 1.

Class-Level Loss Aggregation. Segmentation models are typically trained with pixel-wise cross-entropy, where every pixel contributes equally to the gradient (Minaee et al., 2022). Under the strong class imbalance of segmentation datasets, even large weights on vulnerable classes are largely averaged out by the many pixels of frequent classes, which is one reason why reweighting alone, as in DAFA, fails to improve tail classes (Szabo et al., 2021; Kundu et al., 2026). We therefore first average the loss within each class and then across classes:

$$\mathcal{L}_{\text{class}} = \frac{1}{|\mathcal{C}_B|} \sum_{c \in \mathcal{C}_B} w_c \left(\frac{1}{n_c} \sum_{i:\, y_i = c} \ell_i \right), \tag{3}$$

where $\mathcal{C}_B$ is the set of classes present in mini-batch B, n_c is the number of pixels with ground-truth label c, ℓ_i is the cross-entropy loss of pixel i, and w_c is the weight of class c, which is estimated by the RL controller. This formulation implicitly normalizes by class frequency, so that each present class contributes equally before weighting, and the weights w_c act on class-level rather than pixel-level contributions. It is related to tilted cross-entropy Szabo et al. (2021), but aggregates losses at the class level without exponential tilting (Kundu et al., 2026). Following SegPGD-AT, each mini-batch contains clean and adversarial examples, and Equation 3 is applied to the adversarial part of the mini-batch to reduce the adversarial fairness problem.

Policy-Gradient Controller. Instead of prescribing class weights using a fixed heuristic, RL-SAF learns how to combine class-wise vulnerability signals through a REINFORCE-based controller (Williams, 1992). The controller intervenes at discrete decision steps t between training intervals and bases its decisions on a held-out policy validation set $\mathcal{D}_{\text{policy}}$. This set is sampled from the training split to approximately match its class distribution and is excluded from segmentation training and disjoint from the final evaluation set, which prevents data leakage. At decision step t, the controller observes the state

$$\mathbf{s}_t = \big[\, \mathbf{r}_t,\ \mathbf{q}_t,\ \mathbf{r}_t - \mathbf{r}_{t-1},\ \mathbf{q}_t - \mathbf{r}_t,\ \bar{r}_t,\ \bar{q}_t,\ r_t^{\min},\ \mathrm{std}(\mathbf{r}_t) \,\big], \tag{4}$$

where $\mathbf{r}_t, \mathbf{q}_t \in [0,1]^C$ denote the class-wise robust and clean IoU of the segmentation model on $\mathcal{D}_{\text{policy}}$ for C classes. The difference $\mathbf{r}_t - \mathbf{r}_{t-1}$ captures how the robustness of each class changed since the previous decision step, allowing the controller to assess the effect of its last action, while $\mathbf{q}_t - \mathbf{r}_t$ measures the clean–robust gap, i.e., how much of each class's clean performance is lost under attack. The remaining scalars summarize the class-wise distribution: $\bar{r}_t$ and $\bar{q}_t$ are the mean robust and clean IoU over classes, $r_t^{\min} = \min_c r_{t,c}$ is the worst-class robust IoU, and $\mathrm{std}(\mathbf{r}_t)$ is the standard deviation of class-wise robust IoU, which quantifies robustness disparity across classes.

Structured Action Space. Predicting one weight per class would require a C-dimensional action space, which is difficult to explore within the few decision steps available during training (Chandak et al., 2019). We therefore let the controller decide how to combine a small set of per-class vulnerability signals. The policy π_θ, a neural network with one hidden layer, maps $\mathbf{s}_t$ to the mean and

diagonal covariance of a Gaussian distribution, from which an action $\mathbf{a}_t \in \mathbb{R}^3$ is sampled. For each class c, we define three complementary signals:

$$u_{t,c}^{\text{weak}} = 1 - r_{t,c}, \qquad u_{t,c}^{\text{gap}} = q_{t,c} - r_{t,c}, \qquad u_c^{\text{rare}} = -\log f_c, \tag{5}$$

where $u_{t,c}^{\text{weak}}$ measures class-wise robust weakness, $u_{t,c}^{\text{gap}}$ the performance lost under attack, and u_c^{rare} measures class rarity based on the pixel frequency f_c of class c in the training set. The action is mapped to mixture coefficients $\boldsymbol{\alpha}_t = \mathrm{softmax}(\mathbf{a}_t)$, which combine the signals into a vulnerability score

$$v_{t,c} = \alpha_{t,1}\, u_{t,c}^{\text{weak}} + \alpha_{t,2}\, u_{t,c}^{\text{gap}} + \alpha_{t,3}\, u_c^{\text{rare}}. \tag{6}$$

The class weights $\mathbf{w}_t$ are then obtained by normalizing the exponentiated scores to unit mean and clipping them to $[w_{\min}, w_{\max}]$.

$$w_{t,c} = \mathrm{clip}\left(\frac{\exp(v_{t,c})}{\frac{1}{C}\sum_{c'=1}^{C} \exp(v_{t,c'})},\ w_{\min},\ w_{\max} \right). \tag{7}$$

Normalization preserves the overall loss scale, and clipping prevents individual classes from dominating the objective. Since the action dimension is independent of C, the controller remains compact, and $\boldsymbol{\alpha}_t$ directly reveals which notion of vulnerability is emphasized at each stage of training.

Reward. After the segmentation model has been trained with weights $w_{t,c}$ for K epochs, the controller receives the reward

$$R_t = \Delta\bar{r}_t + \Delta\bar{q}_t + \Delta\bar{r}_t^{\text{tail}} - \Delta r_t^{\max}, \tag{8}$$

where $\Delta\bar{r}_t = \bar{r}_{t+1} - \bar{r}_t$ and $\Delta\bar{q}_t = \bar{q}_{t+1} - \bar{q}_t$ reward improvements in mean robust and clean IoU, and

$$\Delta\bar{r}_t^{\text{tail}} = \frac{1}{|\mathcal{T}_t|} \sum_{c \in \mathcal{T}_t} \big(r_{t+1,c} - r_{t,c}\big), \qquad \Delta r_t^{\max} = \max_c \max\big(0,\ r_{t,c} - r_{t+1,c}\big). \tag{9}$$

Here, $\mathcal{T}_t$ denotes the lower half of classes ranked by robust IoU at step t. The tail term explicitly rewards gains on vulnerable classes, and the penalty term discourages improvements that come at the cost of a severe robustness drop for any single class. The mean terms prevent the controller from trading overall performance for tail improvements.

Policy Optimization. The controller maximizes the REINFORCE objective

$$J(\theta) = \mathbb{E}\Big[(R_t - b_t) \log \pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t) + \eta\, \mathcal{H}\big(\pi_\theta(\cdot \mid \mathbf{s}_t)\big) \Big], \tag{10}$$

where b_t is an exponential moving average baseline, updated as $b_{t+1} = \gamma b_t + (1-\gamma) R_t$ to reduce the variance of the gradient estimate, and the entropy term $\mathcal{H}$ with coefficient η encourages exploration of the action space (Williams, 1992; Greensmith et al., 2004; Andrychowicz et al., 2021).

Training Procedure. Fairness-oriented reweighting is most informative once the class-wise robustness of the model has stabilized. RL-SAF is therefore applied as a fine-tuning stage on a pretrained adversarially trained model, instead of starting from an early warm-up checkpoint as in DAFA. Starting from this model, the controller alternates between selecting class weights, training the segmentation model for K epochs with Equation 3, and updating the policy with the reward in Equation 8. Architectural details and hyperparameters of the RL-based controller are provided in Appendix C.

6 Evaluation of RL-SAF

We evaluate RL-SAF with DeepLabV3 on Cityscapes under 3-step and 10-step SegPGD and 10-step PGD attacks. RL-SAF is applied as a fine-tuning stage to a converged SegPGD-AT model checkpoint which has been pretrained for 200 epochs, allowing us to study whether adaptive class weighting can improve robustness of vulnerable classes.

Table 2: Robust IoU (%) of DeepLabV3 on Cityscapes under 3-step and 10-step SegPGD and 10-step PGD attacks. We use $\epsilon = 8/255$ for all attacks. B-k denotes the mean IoU of the k worst classes. Clean mIoU is reported for reference.

	SegPGD-3				SegPGD-10				PGD-10				
Method	Worst	B-3	B-5	mIoU	Worst	B-3	B-5	mIoU	Worst	B-3	B-5	mIoU	Clean
Standard	0.00	0.00	0.00	3.67	0.00	0.00	0.00	0.04	0.00	0.00	0.00	1.09	**73.76**
SegPGD-AT	0.18	0.29	0.83	26.78	0.00	0.01	0.21	**17.47**	0.00	0.00	0.16	**19.12**	60.04
DAFA-TRADES	0.08	0.17	0.18	13.75	0.00	0.01	0.06	5.01	0.00	0.00	0.05	6.27	46.30
DAFA-CE	0.02	0.11	0.52	13.05	0.00	0.00	0.01	1.97	0.00	0.00	0.00	3.94	52.96
RL-SAF	**0.58**	**2.00**	**3.40**	**27.05**	**0.04**	**0.32**	**0.64**	14.26	**0.71**	**0.84**	**1.06**	16.90	44.47

Robustness of vulnerable classes. As shown in Table 2, RL-SAF improves all worst class metrics under SegPGD-3. The worst-class IoU increases from 0.18% to 0.58%, while bottom-3 and bottom-5 IoU increase from 0.29% and 0.83% to 2.00% and 3.40%. Robust mIoU is preserved at 27.05%, compared with 26.78% for SegPGD-AT. These results suggest that adaptive reweighting can shift robustness toward vulnerable classes without reducing average robustness under the training attack. This comes at a substantial cost in clean performance.

Stronger attacks. Under stronger attacks, RL-SAF continues to improve worst class robustness but at cost of overall mIoU. For SegPGD-10, bottom-3 and bottom-5 IoU increase from 0.01% and 0.21% to 0.32% and 0.64%, while mIoU decreases from 17.47% to 14.26%. A similar pattern appears under PGD-10, where RL-SAF improves worst-class IoU from 0.00% to 0.71% and bottom-5 IoU from 0.16% to 1.06%, but reduces mIoU from 19.12% to 16.90%.

Overall, RL-SAF provides evidence that adaptive class weighting can reduce class-wise robustness disparities. The remaining trade-offs in clean and aggregate robust performance indicate that it should be viewed as a step toward fairness-aware adversarial training.

Uniform-weight ablation. To separate the effect of class-level loss aggregation from that of the learned policy, we replace the adaptive weights with uniform class weights ($w_c = 1$). As shown in Table 3, uniform weighting improves the bottom-3 and bottom-5 robust IoU over SegPGD-AT, showing that class-level aggregation itself redistributes robustness toward vulnerable classes. RL-SAF further improves worst-class and tail robustness while recovering overall robust mIoU, indicating that the learned weighting provides benefits beyond uniform class balancing.

Table 3: Uniform-weight ablation on Cityscapes under SegPGD-3. All values are IoU (%).

Method	Worst	B-3	B-5	mIoU	Clean
SegPGD-AT	0.18	0.29	0.83	26.78	60.04
Uniform	0.02	1.78	3.10	24.81	40.72
RL-SAF	**0.58**	**2.00**	**3.40**	**27.05**	**44.47**

7 Conclusions

In this work, we investigate class-wise adversarial robustness disparities in semantic segmentation and show that adversarial perturbations amplify existing inter-class differences rather than uniformly degrading performance. Vulnerable classes can both attract disproportionate false-positive mass and receive too few correct predictions, effects not fully explained by class frequency. Directly adapting DAFA from image classification proves ineffective for segmentation, motivating RL-SAF, which combines class-level loss aggregation with adaptive, validation-guided weighting. Across evaluated attacks, RL-SAF consistently improves robustness for vulnerable classes and preserves overall robust mIoU under the training attack, while stronger attacks and clean evaluation expose trade-offs with aggregate robustness and standard accuracy. These results establish class-wise adversarial robustness as an important consideration in semantic segmentation and highlight adaptive, class-aware optimization as a promising direction for fairer robust models.

We anticipate that our observations encourage future work to evaluate new segmentation models across datasets, and in cases where robust fairness is required, study the stability and generalization of the learned policy, investigate further objectives and reward formulations to optimize task-specific trade-offs between clean and aggregate robust performance while improving vulnerable classes.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on methodology and experiments. This was limited to recommendations regarding what to include in our experimental discussion. We did not use generative AI tools for developing theoretical models or conceptual frameworks, proposing or refining hypotheses, implementing methods, cleaning and reformating datasets, or supporting qualitative and thematic data analysis. Other tasks that require disclosure, i.e., generating synthetic datasets, formulating and proving mathematical claims, and assisting with translation, are not applicable to this work. Additionally, we used generative AI tools to create and modify figures comprising our experimental results, reviewing and debugging code, and drafting and editing parts of the paper. We have reviewed all AI-assisted work. All authors critically reviewed text passages and figures that have been created with the assistance of AI tools. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shashank Agnihotri, Steffen Jung, and Margret Keuper. CosPGD: an efficient white-box adversarial attack for pixel-wise prediction tasks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 416–451. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/agnihotri24b.html.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Leonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier Bachem. What matters for on-policy deep actor-critic methods? A large-scale study. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=nIAxjsniDzg.

Anurag Arnab, Ondrej Miksik, and Philip H. S. Torr. On the robustness of semantic segmentation models to adversarial attacks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 42 (12):3040–3053, December 2020. ISSN 1939-3539. doi: 10.1109/TPAMI.2019.2919707. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2019.2919707.

Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. In Luca Bertinetto, João F. Henriques, Samuel Albanie, Michela Paganini, and Gül Varol (eds.), *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, volume 148 of *Proceedings of Machine Learning Research*, pp. 325–342. PMLR, 11 Dec 2021. URL https://proceedings.mlr.press/v148/benz21a.html.

Yash Chandak, Georgios Theocharous, James Kostas, Scott Jordan, and Philip Thomas. Learning action representations for reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 941–950. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/chandak19a.html.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *ArXiv*, abs/1706.05587, 2017. URL https://api.semanticscholar.org/CorpusID:22655199.

Zhaoyu Chen, Zhengyang Shan, Jingwen Chang, Kaixun Jiang, Dingkang Yang, Yiting Cheng, Jinglun Li, and Wenqiang Zhang. Delving into the adversarial robustness of semantic segmentation with decision-based black-box attacks. *Neural Networks*, 202:109020, 2026. ISSN 0893-6080. doi: 10.1016/j.neunet.2026.109020. URL https://www.sciencedirect.com/science/article/pii/S0893608026004806.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016. doi: 10.1109/CVPR.2016.350.

Francesco Croce, Naman D. Singh, and Matthias Hein. Towards reliable evaluation and fast training of robust semantic segmentation models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 180–197, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72986-7.

Jiequan Cui, Beier Zhu, Xin Wen, Xiaojuan Qi, Bei Yu, and Hanwang Zhang. Classes are not equal: An empirical study on image recognition fairness. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23283–23292, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.02197. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.02197.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 9268–9277. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00949. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Cui_Class-Balanced_Loss_Based_on_Effective_Number_of_Samples_CVPR_2019_paper.html.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111:98 – 136, 2015. URL https://api.semanticscholar.org/CorpusID:207252270.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, November 2004. ISSN 1532-4435.

Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip H. S. Torr. SegPGD: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIX*, volume 13689 of *Lecture Notes in Computer Science*, pp. 308–325. Springer, 2022. doi: 10.1007/978-3-031-19818-2_18. URL https://doi.org/10.1007/978-3-031-19818-2_18.

Levente Halmosi, Bálint Mohos, and Márk Jelasity. Evaluating the adversarial robustness of semantic segmentation: Trying harder pays off. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXIX*, pp. 1–18, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72985-0. doi: 10.1007/978-3-031-72986-7_1. URL https://doi.org/10.1007/978-3-031-72986-7_1.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9924–9935, June 2022.

Chen Huang, Shuangfei Zhai, Walter Talbott, Miguel Bautista Martin, Shih-Yu Sun, Carlos Guestrin, and Josh Susskind. Addressing the loss-metric mismatch with adaptive loss alignment. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2891–2900. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/huang19f.html.

Soumya Snigdha Kundu, Florian Kofler, Marina Ivory, Hendrik Möller, Jonathan Shapey, and Tom Vercauteren. Instance awareness of multi-class semantic segmentation loss functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 6680–6688, June 2026.

Hyungyu Lee, Saehyung Lee, Hyemi Jang, Junsung Park, Ho Bae, and Sungroh Yoon. DAFA: Distance-aware fair adversarial training. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 28780–28809, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/7bd4a7d0e6773072c2e3c77b11d93065-Paper-Conference.pdf.

Ruihuang Li, Shuai Li, Chenhang He, Yabin Zhang, Xu Jia, and Lei Zhang. Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–16, June 2022. ISSN 1063-6919. doi: 10.1109/CVPR52688.2022.01130.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 2999–3007. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.324. URL https://doi.org/10.1109/ICCV.2017.324.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, Los Alamitos, CA, USA, June 2015. IEEE Computer Society. doi: 10.1109/CVPR.2015.7298965. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298965.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Tejaswini Medi, Steffen Jung, and Margret Keuper. FAIR-TAT: Improving model fairness using targeted adversarial training. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2025)*, pp. 7827–7836, Tucson, AZ, USA, 2025. IEEE. ISBN 979-8-3315-1083-1. doi: 10.1109/WACV61041.2025.00760.

Tejaswini Medi, Levan Mikeladze, and Margret Keuper. RL-FAT: Reinforcement learning for fair adversarial training, 2026. URL https://arxiv.org/abs/2608.29247.

Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022. doi: 10.1109/TPAMI.2021.3059968.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ren18a.html.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-Weight-Net: Learning an explicit mapping for sample weighting. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/e58cc5ca94270acaceed13bc82dfedf7-Paper.pdf.

Chunyu Sun, Chenye Xu, Chengyuan Yao, Siyuan Liang, Yichao Wu, Ding Liang, Xianglong Liu, and Aishan Liu. Improving robust fairness via balance adversarial training. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i12.26769. URL https://doi.org/10.1609/aaai.v37i12.26769.

Attila Szabo, Hadi Jamali-Rad, and Siva-Datta Mannava. Tilted Cross-Entropy (TCE): Promoting fairness in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2305–2310, June 2021.

Zifu Wang, Maxim Berman, Amal Rannen-Triki, Philip H.S. Torr, Devis Tuia, Tinne Tuytelaars, Luc Van Gool, Jiaqian Yu, and Matthew B. Blaschko. Revisiting evaluation metrics for semantic segmentation: optimization and evaluation of fine-grained intersection over union. In *Proceedings*

of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Zeming Wei, Yifei Wang, Yiwen Guo, and Yisen Wang. CFA: Class-wise calibrated fair adversarial training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8193–8201, 2023. doi: 10.1109/CVPR52729.2023.00792.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11492–11501. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/xu21b.html.

Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data valuation using reinforcement learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10842–10851. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/yoon20a.html.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/zhang19p.html.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023. doi: 10.1109/TPAMI.2023.3268118.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239, Los Alamitos, CA, USA, July 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.660. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.660.

Appendix

A Additions to the Adversarial Fairness Analysis

In this section, we provide additional results and details of the analysis of adversarial fairness presented in section 3.

Datasets. We use Cityscapes (Cordts et al., 2016) and PASCAL VOC 2012 (Everingham et al., 2015) as datasets for our analysis because they are both widely adopted in related work and central to the research conducted by Halmosi et al. (2024), which we adapt for our experimental studies. Cityscapes has been designed for urban scene understanding and it contains 2975 training images that have been recorded from the perspective of a driving car across several European cities. Each pixel belongs to one of 19 semantic classes representing the most commonly occurring objects that are relevant to autonomous driving such as traffic signs, buildings, and cars. The Cityscapes test set is not publicly available, which is why we use the validation set that contains 500 additional images for reporting final performance. The PASCAL Visual Object Classes (VOC) 2012 dataset has originally been designed as a competition dataset. We use the segmentation part of this dataset, which comprises 1464 training and 1449 validation images. Segmentation masks are annotated with 21 different semantic classes, which represent general objects such as bicycles, chairs, and dining tables in real world settings. Dataset statistics are summarized in Table 4.

Dataset	Classes	Most frequent	Least frequent	90% of pixels
Cityscapes	19	30.23%	0.19%	7 (36.84%)
VOC 2012	21	64.38%	0.43%	12 (57.14%)

Table 4: Analysis of class frequencies in Cityscapes and VOC 2012. Most frequent refers to the class label that occurs most frequently, whereas least frequent refers to the class that is observed most infrequently. The last column indicates the top-k classes occupying at least 90% of space across all pixels.

Experimental setup. The models we investigate follow those established in Halmosi et al. (2024). The authors provide checkpoints for clean training and PGD-AT, while we replicated SegPGD-AT and SegPGD-AT-100 using the settings stated in their work. Cityscapes training images are augmented using a 449×449 random crop, horizontal flipping with $p = 0.5$, and random resizing with scale range $[0.75, 1.75]$. Validation images are cropped to size 512×1024 without any further augmentations. VOC 2012 training images are cropped to size 473×473 with similar additional augmentations as Cityscapes, while validation images are resized such that their larger dimension has a length of 513 pixels. All configurations use a standard SGD optimizer with momentum of 0.9, weight decay factor of $4e-5$ and a polynomially decaying learning rate scheduler with an initial value of 0.01 and power 0.9. Experiments on Cityscapes are trained for 400 epochs while we use a shorter 50 epoch run for models on VOC 2012, as described by Halmosi et al. (2024). This general setup is also used for training DAFA and RL-SAF, with task-specific changes being described in the corresponding sections in the appendix.

Comprehensive evaluation results. In addition to the results presented in Table 1, we provide the full set of evaluation results comprising IoU, recall and precision. We report mean scores as well as performance of the worst and best class to visualize exisiting disparities. Table 5 contains results of our analyzed training techniques for PSPNet and DeepLabV3 on Cityscapes and VOC 2012 on clean data. The Normal Training and PGD-AT models are evaluated based on the checkpoints provided by Halmosi et al. (2024), whereas SegPGD-AT and SegPGD-AT-100 represent our experimental results. The entries in the table show moderate disparity in clean data, which get increasingly larger for models that have been trained with stronger adversarial attacks. SegPGD-AT-100, which has been exclusively trained on adversarial examples, exhibits the largest degree of class-wise disparity. Results on adversarially perturbed evaluation data are shown in Table 6. Comparing these results,

		mIoU	Precision	Recall
Normal Training	PN+P	76.73 (94.29 / 37.09)	86.63 (**97.25** / 45.26)	85.62 (97.34 / 51.47)
	DL+P	**77.21** (**94.37** / **37.91**)	**86.93** (96.90 / 45.83)	**86.16** (97.47 / 54.41)
	PN+CS	73.25 (**98.06** / **53.60**)	**87.68** (98.96 / **74.90**)	80.65 (99.07 / 59.69)
	DL+CS	**73.76** (97.90 / 53.00)	87.38 (**99.00** / 74.13)	**81.51** (98.87 / **61.78**)
PGD-AT	PN+P	74.16 (93.61 / 37.56)	84.07 (96.79 / 43.80)	85.04 (**97.56** / **56.10**)
	DL+P	70.59 (92.50 / 30.23)	80.06 (96.88 / 37.85)	83.95 (95.37 / 51.29)
	PN+CS	66.94 (96.71 / 29.76)	84.52 (97.28 / 65.25)	75.57 (**99.40** / 30.54)
	DL+CS	68.85 (97.47 / 47.69)	84.90 (98.29 / 71.22)	77.04 (99.15 / 53.59)
SegPGD-AT	PN+P	59.87 (90.14 / 22.93)	79.74 (93.15 / 48.75)	69.12 (96.54 / 29.85)
	DL+P	59.67 (90.07 / 23.80)	79.11 (93.87 / **53.94**)	69.26 (96.47 / 29.88)
	PN+CS	58.49 (95.73 / 29.17)	78.16 (98.58 / 49.78)	67.47 (97.07 / 35.81)
	DL+CS	60.04 (95.78 / 29.46)	78.04 (98.73 / 50.89)	69.65 (96.97 / 41.15)
SegPGD-AT-100	PN+P	43.88 (86.11 / 3.95)	74.34 (88.52 / 40.52)	51.24 (96.95 / 4.19)
	DL+P	42.07 (85.72 / 4.88)	74.37 (97.10 / 32.98)	49.12 (89.72 / 5.41)
	PN+CS	36.71 (90.38 / 1.31)	74.62 (97.87 / 48.44)	42.41 (97.67 / 1.32)
	DL+CS	39.34 (91.86 / 5.90)	72.63 (94.38 / 50.37)	45.77 (97.18 / 6.03)

Table 5: Clean validation performance of different training schemes on VOC 2012 and Cityscapes. Each entry reports the average metric followed by the best / worst class-wise score in parentheses. Blocks correspond to training schemes and rows to model–dataset combinations, where DL = DeepLabV3, PN = PSPNet, P = VOC 2012, and CS = Cityscapes. All scores are reported in %. For each dataset, bold marks the highest value in each column across all training schemes and models, determined separately for the average, best-class, and worst-class scores (blue: VOC 2012; green: Cityscapes).

we observe larger gaps between best and worst class performance in the adversarial results, with frequent near-zero worst-class IoU.

		mIoU	Precision	Recall
Normal Training	PN+P	6.77 (30.21 / 0.42)	11.62 (69.77 / 0.51)	17.28 (49.70 / 2.26)
	DL+P	7.32 (29.21 / 0.20)	12.33 (70.94 / 0.25)	18.63 (36.93 / 0.93)
	PN+CS	3.76 (17.78 / 0.00)	14.89 (75.40 / 0.00)	8.08 (64.29 / 0.00)
	DL+CS	3.67 (19.19 / 0.00)	14.18 (73.38 / 0.00)	7.60 (68.28 / 0.00)
PGD-AT	PN+P	34.52 (70.29 / 7.54)	44.30 (**94.35** / 9.42)	**63.22** (85.45 / **25.51**)
	DL+P	**42.08** (**81.91** / **9.80**)	**53.73** (93.29 / **13.86**)	63.08 (87.03 / 25.06)
	PN+CS	34.52 (87.70 / 0.49)	53.04 (**92.37** / **7.86**)	42.06 (**97.65** / 0.52)
	DL+CS	**35.92** (**89.10** / **1.21**)	**53.24** (92.10 / 7.11)	**43.83** (96.48 / **1.38**)
SegPGD-AT	PN+P	16.67 (74.95 / 3.02)	28.73 (83.59 / 7.36)	24.79 (87.88 / 4.05)
	DL+P	18.75 (76.17 / 1.85)	32.50 (84.15 / 5.87)	27.28 (88.93 / 2.64)
	PN+CS	25.14 (84.42 / 0.03)	43.77 (88.38 / 0.07)	31.16 (94.95 / 0.04)
	DL+CS	26.78 (85.78 / 0.18)	42.72 (89.96 / 0.42)	33.13 (94.86 / 0.26)
SegPGD-AT-100	PN+P	18.32 (76.38 / 1.74)	32.25 (84.45 / 5.25)	26.72 (88.88 / 2.54)
	DL+P	19.05 (77.09 / 1.74)	33.76 (84.41 / 4.39)	26.88 (**89.88** / 2.81)
	PN+CS	25.63 (84.79 / 0.00)	42.37 (89.33 / 0.00)	31.69 (95.23 / 0.00)
	DL+CS	27.42 (85.94 / 0.07)	42.85 (90.87 / 0.18)	34.32 (94.07 / 0.11)

Table 6: Robust validation performance of different training schemes on VOC 2012 and Cityscapes under a 3-step SegPGD attack with $\epsilon = 8/255$. Each entry reports the average metric followed by the best / worst class-wise score in parentheses. Blocks correspond to training schemes and rows to model–dataset combinations, where DL = DeepLabV3, PN = PSPNet, P = VOC 2012, and CS = Cityscapes. All scores are reported in %. For each dataset, bold marks the highest value in each column across all training schemes and models, determined separately for the average, best-class, and worst-class scores (blue: VOC 2012; green: Cityscapes).

Class-wise prediction bias. Let $\mathbf{p}_i = \sigma(f_\theta(\mathbf{x}_i))$ denote the predicted class probabilities for pixel i and $\mathbf{y}_i$ its one-hot ground-truth label. Following Cui et al. (2024), we accumulate the incorrect and correct probability mass assigned to each class as

$$\mathbf{G}_n = \sum_i \mathbf{p}_i \odot (\mathbf{1} - \mathbf{y}_i), \qquad \mathbf{G}_p = \sum_i \mathbf{p}_i \odot \mathbf{y}_i. \tag{11}$$

Here, $\mathbf{G}_n(c)$ measures the false-positive probability mass assigned to class c, whereas $\mathbf{G}_p(c)$ measures the probability mass correctly assigned to class c. We normalize these quantities across classes:

$$\mathrm{MPB}_{\mathrm{rel}}(c) = \frac{\mathbf{G}_n(c)}{\sum_{c'} \mathbf{G}_n(c')}, \qquad \mathrm{CPM}_{\mathrm{rel}}(c) = \frac{\mathbf{G}_p(c)}{\sum_{c'} \mathbf{G}_p(c')}. \tag{12}$$

Because semantic segmentation datasets are typically strongly imbalanced, these quantities are further normalized by the relative pixel frequency of each class,

$$\mathrm{CF}_{\mathrm{rel}}(c) = \frac{\mathbf{CF}(c)}{\sum_{c'} \mathbf{CF}(c')}. \tag{13}$$

We therefore report

$$\frac{\mathrm{MPB}_{\mathrm{rel}}(c)}{\mathrm{CF}_{\mathrm{rel}}(c)} \qquad \text{and} \qquad \frac{\mathrm{CPM}_{\mathrm{rel}}(c)}{\mathrm{CF}_{\mathrm{rel}}(c)}. \tag{14}$$

Values above one indicate that a class receives more incorrect or correct probability mass, respectively, than expected from its relative frequency, while values below one indicate less mass than expected. These normalized quantities allow us to distinguish effects caused by class prevalence from systematic shifts in the model's predictions under adversarial perturbations.

Class-wise scores, prediction bias, and correct predictions. The plots we included in section 3 show results for the Cityscapes dataset. Here, we additionally provide such plots for these training methods on the VOC 2012 dataset. We include plots for class-wise IoU, model prediction bias, and correct predictions for DeepLabV3 and PSPNet. VOC 2012 is less unbalanced than Cityscapes because of its dominant background class. If this class is not taken into account, the remaining classes are more balanced than those of Cityscapes. This is visible in the class-wise IoU plots in Figure 5 and Figure 6. On clean data, there is a smaller observable decrease than in Cityscapes, while adversarial validation performance is distributed more uniformly over all classes.

This also affects the model prediction bias plots in Figure 7 and Figure 8. The overall patterns favoring infrequent classes are not as pronounced as they are for Cityscapes. Although variance between classes and mean prediction bias values are still larger than in the clean validation case, the more balanced nature of VOC 2012 make this pattern less obvious.

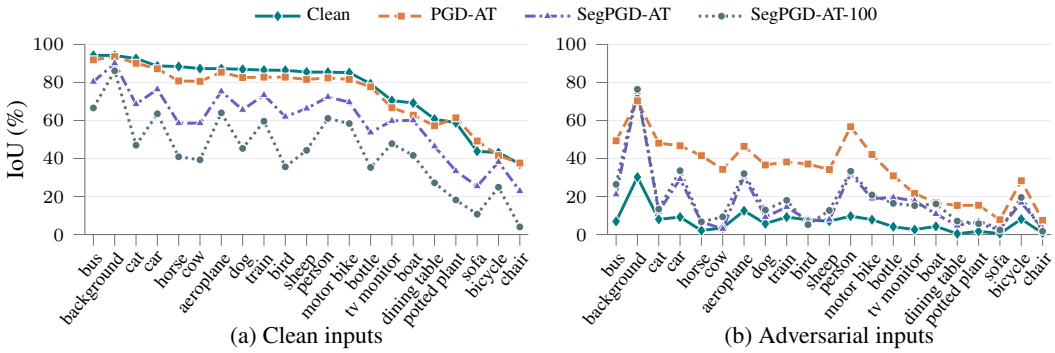
Frequency-normalized correct prediction mass reveals the effect of adversarial examples in more detail. Results are shown in Figure 9 and Figure 10. On clean examples, VOC 2012 is relatively balanced, with some outliers. Most classes are at or near a score one, indicating that they are correctly predicted as confidently as their frequency suggests. The more frequent background class receives slightly more correct prediction mass, as it presents the main source of imbalance. On the other classes, all training configurations exhibit behavior that is expected on frequency-balanced data. Adversarial evaluation breaks this pattern, which illustrates the effect of adversarial examples we discussed in our adversarial fairness analysis. Most classes are now recognized disproportionately unreliably, with a large degree of variance between classes.

Cityscapes-Correct predictions. We additionally examine the frequency-normalized correct prediction mass to measure how reliably each class is recognized. As shown in Figure 11, correct prediction mass generally decreases toward less frequent and more difficult classes, with adversarial perturbations producing larger variation and further suppressing already weak classes. Together with the prediction-bias analysis, these results show that vulnerable classes can simultaneously receive insufficient correct prediction mass and attract disproportionate false-positive mass.

Taken together, our analysis establishes a clear class-wise adversarial fairness problem in semantic segmentation. Adversarial perturbations amplify existing disparities rather than degrading all classes uniformly, and class frequency alone is insufficient to explain which classes are most affected. These observations motivate fairness-aware training strategies that account for class-specific behavior beyond pixel frequency.

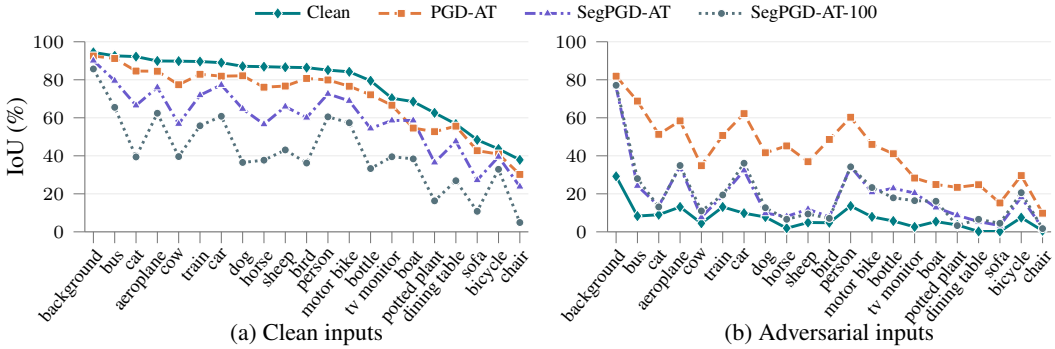
B DAFA Implementation Details

Experimental setup. We follow the experimental setup established in Appendix A for our DAFA experiments. Class-wise statistics are computed from adversarial training results after 70 warmup



(a) Clean inputs (b) Adversarial inputs

Figure 5: Per-class IoU on PASCAL VOC for a normally trained (Clean) and three adversarially trained PSPNet models, evaluated on (a) clean and (b) adversarial inputs. Classes are sorted by clean-model IoU on clean inputs. We use a 3-step SegPGD attack with $\epsilon = 8/255$ for robust evaluation.



(a) Clean inputs (b) Adversarial inputs

Figure 6: Per-class IoU on PASCAL VOC for a normally trained model (Clean) and three adversarially trained models using DeepLabV3, evaluated on (a) clean and (b) adversarial inputs. Classes are sorted by clean-model IoU on clean inputs. We use a 3-step SegPGD attack with $\epsilon = 8/255$ for robust evaluation.

epochs. We set $\lambda = 1.5$, following Lee et al. (2024), who state that a larger value is beneficial for strong disparities. DAFA-CE follows the regular SegPGD-AT training setup, with class weights applied to the adversarial loss term after warmup. For DAFA-TRADES, we set $\beta = 2.0$ and use identical attack settings as the original version of DAFA.

Class weights. DAFA (Lee et al., 2024) aggregates class-wise prediction probabilities in a matrix $\mathbf{P} \in \mathbb{R}^{C \times C}$, where $p_i(j) = \mathbf{P}_{ij}$ is the average probability assigned to class j for samples of class i. The diagonal entries $p_i(i)$ measure class difficulty, and the off-diagonal entries measure class-wise confusion. The weight of class i is

$$\mathcal{W}_i = 1 + \lambda \sum_{j \neq i} \Big(\mathbb{I}[p_i(i) < p_j(j)]\, p_i(j)\, p_j(j) \;-\; \mathbb{I}[p_i(i) > p_j(j)]\, p_j(i)\, p_i(i) \Big), \tag{15}$$

where $\mathbb{I}[\cdot]$ is the indicator function. For each pair of classes, the more difficult class gains weight and the easier class loses weight, so difficult classes generally receive weights larger than one. The weights scale both the class-wise loss and the adversarial attack margin, which pushes decision boundaries away from the centers of difficult classes.

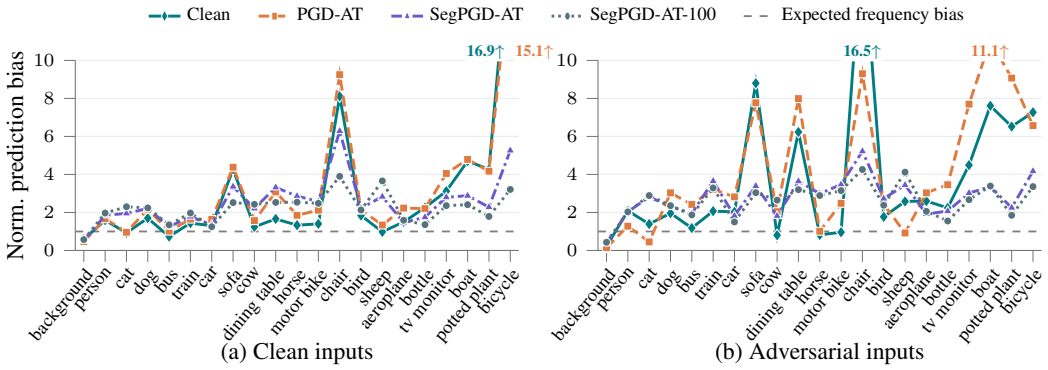


Figure 7: Frequency-normalized prediction bias for different training approaches on PASCAL VOC using PSPNet, on (a) clean and (b) adversarial inputs. Classes are sorted by decreasing frequency (left to right); the dashed line marks the bias expected from class frequency alone. We use a 3-step SegPGD attack with $\epsilon = 8/255$ for robust evaluation.

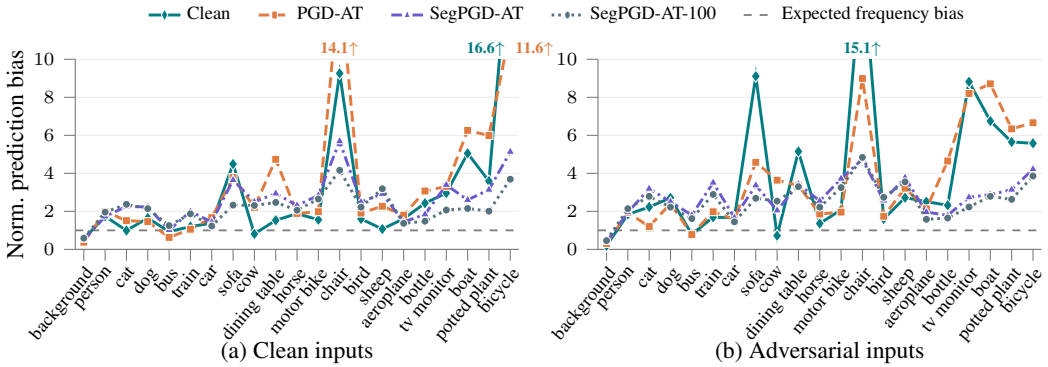


Figure 8: Frequency-normalized prediction bias for different training approaches on PASCAL VOC using DeepLabV3, on (a) clean and (b) adversarial inputs. Classes are sorted by decreasing frequency (left to right); the dashed line marks the bias expected from class frequency alone. We use a 3-step SegPGD attack with $\epsilon = 8/255$ for robust evaluation.

Adaptation to segmentation. Since each image contains multiple classes with different pixel counts, we compute $\mathbf{P}$ at the pixel level and normalize by the number of pixels per class:

$$p_i(j) = \frac{1}{|\Omega_i|} \sum_{u \in \Omega_i} \mathrm{softmax}\big(f_\theta(x)_u\big)_j, \tag{16}$$

where Ω_i is the set of pixels with ground-truth label i across the dataset. $\mathbf{P}$ is aggregated during the final warm-up epoch, and the resulting weights are kept fixed for the rest of training.

Training objectives. DAFA-TRADES follows the original formulation and uses the TRADES objective (Zhang et al., 2019), averaged over pixels:

$$\mathcal{L}(x, y) = \mathcal{W}_y \, \mathcal{L}_{\mathrm{CE}}\big(f_\theta(x), y\big) + \beta \max_{\|\delta\| \le \mathcal{W}_y \epsilon} \mathrm{KL}\big(f_\theta(x) \,\|\, f_\theta(x+\delta)\big). \tag{17}$$

DAFA-CE replaces TRADES with class-weighted cross-entropy on adversarial examples generated by SegPGD (Gu et al., 2022), providing a direct comparison to the SegPGD-AT baseline. Both variants use a 3-step attack during training and are evaluated with a 3-step SegPGD attack.

Additional results for PSPNet. We provide evaluation results of our DAFA experiments on Cityscapes using a PSPNet model. These can be found in Table 7. We observe similar patterns

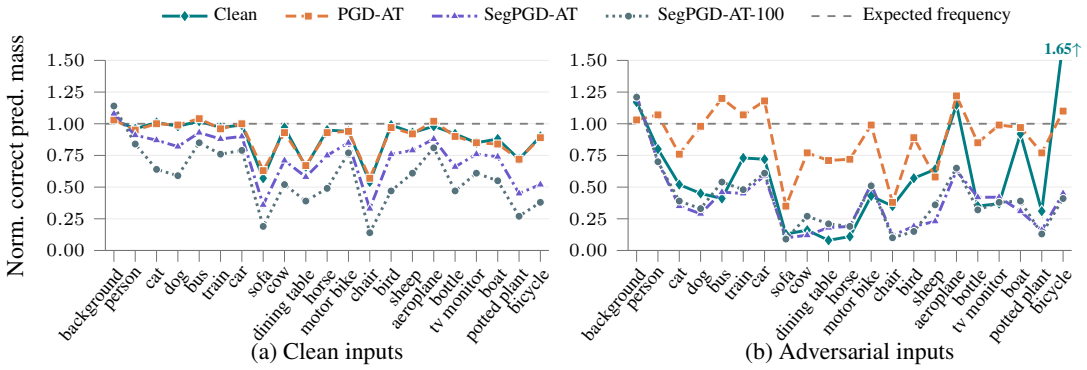


Figure 9: Frequency-normalized correct prediction mass for different training approaches on PASCAL VOC using PSPNet, on (a) clean and (b) adversarial inputs. Classes are sorted by decreasing frequency (left to right); the dashed line marks the value expected from class frequency alone. We use a 3-step SegPGD attack with $\epsilon = 8/255$ for robust evaluation.

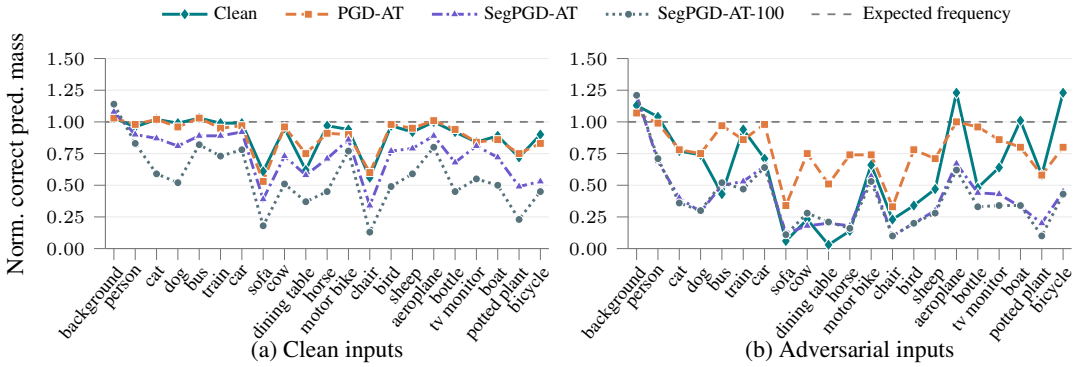
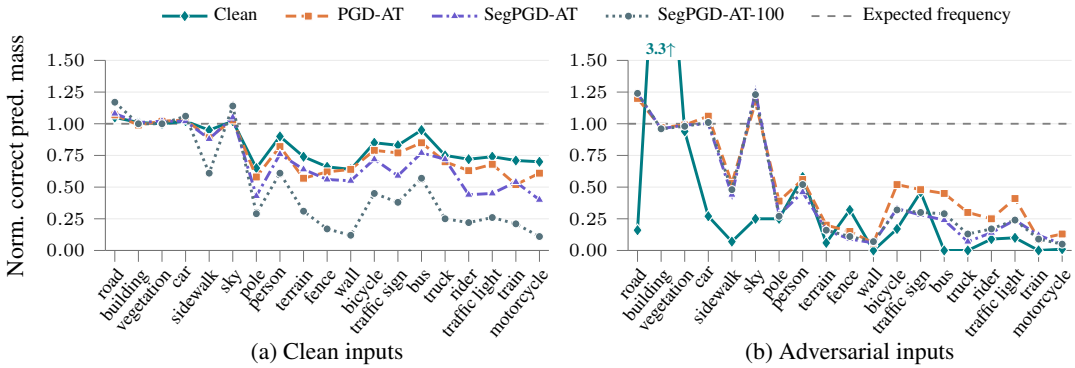


Figure 10: Frequency-normalized correct prediction mass for different training approaches on PASCAL VOC using DeepLabV3, on (a) clean and (b) adversarial inputs. Classes are sorted by decreasing frequency (left to right); the dashed line marks the value expected from class frequency alone. We use a 3-step SegPGD attack with $\epsilon = 8/255$ for robust evaluation.

Table 7: Robust IoU (%) of PSPNet on Cityscapes under 3-step and 10-step SegPGD and 10-step PGD attacks. B-k denotes the mean IoU of the k worst classes. Clean mIoU is reported for reference.

	SegPGD-3				SegPGD-10				PGD-10				
Method	Worst	B-3	B-5	mIoU	Worst	B-3	B-5	mIoU	Worst	B-3	B-5	mIoU	Clean
Standard	0.00	0.00	0.00	3.76	0.00	0.00	0.00	0.76	0.00	0.00	0.00	1.13	**73.25**
SegPGD-AT	0.03	0.13	0.40	25.14	0.00	0.02	0.09	16.07	0.00	0.01	0.11	17.54	58.49
DAFA-TRADES	0.00	0.09	0.36	13.82	0.00	0.01	0.05	5.91	0.00	0.02	0.10	7.04	41.64
DAFA-CE	0.00	0.07	0.33	23.24	0.00	0.01	0.05	13.99	0.00	0.00	0.00	16.44	38.37

as we did for DeepLabV3, with DAFA not managing to improve adversarial fairness while exhibiting reduced overall robustness and clean performance.



(a) Clean inputs (b) Adversarial inputs

Figure 11: Frequency-normalized correct prediction mass for different training approaches on Cityscapes using DeepLabV3, on (a) clean and (b) adversarial inputs. Classes are sorted by decreasing frequency (left to right); the dashed line marks the value expected from class frequency alone. We use a 3-step SegPGD attack with $\epsilon = 8/255$ for robust evaluation.

Algorithm 1 RL-SAF

Require: Segmentation model f_ϕ, policy π_θ, policy validation set $\mathcal{D}_{\text{policy}}$, training set $\mathcal{D}_{\text{train}}$, action interval K

Initialize reward baseline $b_0 \leftarrow 0$
Initialize class frequencies $\mathbf{f}$
Evaluate f_ϕ on $\mathcal{D}_{\text{policy}}$ to get initial metrics $\mathbf{r}_0$ and $\mathbf{q}_0$
for decision step $t = 0, \ldots$ **do**
  Construct state $\mathbf{s}_t \leftarrow \left[\mathbf{r}_t,\ \mathbf{q}_t,\ \mathbf{r}_t - \mathbf{r}_{t-1},\ \mathbf{q}_t - \mathbf{r}_t,\ \bar{r}_t,\ \bar{q}_t,\ r_t^{\min},\ \mathrm{std}(\mathbf{r}_t)\right]$
  Sample action $\mathbf{a}_t \sim \pi_\theta(\cdot \mid \mathbf{s}_t)$
  Compute class weights $\mathbf{w}_t \leftarrow \texttt{action_to_weights}(\mathbf{a}_t, \mathbf{q}_t, \mathbf{r}_t, \mathbf{f})$
  Train f_ϕ on $\mathcal{D}_{\text{train}}$ for K using loss weights $\mathbf{w}_t$
  Evaluate the updated model on $\mathcal{D}_{\text{policy}}$ to get $\mathbf{q}_{t+1}$ and $\mathbf{r}_{t+1}$
  Compute reward $R_t \leftarrow \texttt{calculate_policy_reward}(\mathbf{q}_t, \mathbf{r}_t, \mathbf{q}_{t+1}, \mathbf{r}_{t+1})$
  Compute policy loss $\mathcal{L}_\pi = -(R_t - b_t)\log \pi(\mathbf{a}_t \mid \mathbf{s}_t) - \eta \mathcal{H}[\pi(\cdot \mid \mathbf{s}_t)]$
  Update θ using a gradient-based update to minimize $\mathcal{L}_\pi$
  Update the baseline $b_{t+1} \leftarrow \gamma b_t + (1-\gamma) R_t$
end for
return f_ϕ and π_θ

C RL-SAF Overview and Experimental Setup

Overview Algorithm 1 provides a pseudocode overview of the proposed RL-SAF. Computation of class weights from vulnerability scores and the reward computation are formulated as abstract functions for brevity.

Training Configuration and Experimental Setup We provide an overview of the experimental setup used to train RL-SAF. The hyperparameter values that have been chosen for our experiments can be found in Table 8.

Table 8: RL-SAF configuration

Hyperparameter	Value
Total epochs (SegPGD-AT pretraining + RL-SAF)	400
SegPGD-AT pretraining epochs	200
Action interval K	1
Policy hidden size	32
Policy optimizer	Adam
Policy learning rate	0.01
Entropy regularization factor η	0.001
Weight range $[w_{\min}, w_{\max}]$	$[0.5, 2.0]$
Baseline factor γ	0.9
Attack iterations k	3
Attack margin ϵ	$\frac{8}{255}$
Attack step size α	$\frac{2}{255}$

Table 9: Class-wise robust IoU results for clean training, the SegPGD baseline, DAFA, RL-SAF and uniform class weighting. Bottom-3 and Bottom-5 indicate the mean of the worst 3 and 5 class scores respectively. Clean mIoU is additionally included.

Method	Worst-class IoU	Bottom-3 IoU	Bottom-5 IoU	mIoU	Clean mIoU
Standard	0.00	0.00	0.00	3.67	**73.76**
SegPGD-AT	0.18	0.29	0.83	26.78	60.04
DAFA-TRADES	0.08	0.17	0.18	13.75	46.30
DAFA-CE	0.02	0.11	0.52	13.05	52.96
RL-SAF	**0.58**	**2.00**	**3.40**	**27.05**	44.47
Uniform Weights	0.02	1.78	3.10	24.81	40.72

Different weighting schemes We compare RL-SAF to a method which uses the same loss function with uniform class weights to isolate the effects of our learned policy and to determine whether the observed increases of robustness on vulnerable classes are actually caused by reweighting. Results of this experiments are shown in Table 9.

The uniform weighting experiment shows that the class-balanced loss objective can improve performance on vulnerable classes, with the method achieving similar bottom-3 and bottom-5 mean scores. In the observed setting, uniform weighting negatively impacts class scores of the worst class and overall robust and clean performance. We conclude that our proposed weighting scheme improves the natural redistribution effect introduced by class-balanced loss while retaining overall performance. Class weights are guided by a reward function that explicitly rewards maintained clean performance and worst-class scores, whereas uniform weighting uses an uniformed reweighting strategy that treats each class equally.

We additionally performed experiments to determine the effect of the structured action space and the scale of the class-balanced loss term. For the resulting RL-SAF formulations, we introduce a scaling factor s for $\mathcal{L}_{\text{class}}$, which we set to 0.1 to test how the strength of the fairness-oriented intervention influences results. Class-balanced loss is not scaled in the original RL-SAF formulation. These methods additionally use an unstructured action space and produce one weight per class directly. Results are presented in Table 10.

RL-Adv applies class weights to an adversarial class-based loss term, whereas RL-Separate introduces an additional balancing loss term for clean samples. We also include a uniform weighting baseline to determine the efficacy of these methods. While both configurations achieve improvements on vulnerable classes compared to SegPGD-AT and the uniform baseline, these increases are small compared to those achieved by RL-SAF. Noticeably, these methods retain increased clean performance, which suggests that the strength of fairness-oriented class balancing involves a trade-off with overall performance. Applying class weights to the clean loss improves clean performance, but it also severely impacts overall robustness and tail performance. Since these reinforcement learning-based methods also outperform the uniformly weighted baseline, we get additional evidence for the advantages of adaptive reweighting.

Table 10: Class-wise robust IoU results for clean training, the SegPGD baseline, DAFA, RL-SAF, and experiments with different class loss scaling and action space. Bottom-3 and Bottom-5 indicate the mean of the worst 3 and 5 class scores respectively. Clean mIoU is additionally included.

Method	Worst-class IoU	Bottom-3 IoU	Bottom-5 IoU	mIoU	Clean mIoU
Standard	0.00	0.00	0.00	3.67	**73.76**
SegPGD-AT	0.18	0.29	0.83	26.78	60.04
DAFA-TRADES	0.08	0.17	0.18	13.75	46.30
DAFA-CE	0.02	0.11	0.52	13.05	52.96
RL-SAF	**0.58**	**2.00**	**3.40**	**27.05**	44.47
RL-Adv	0.29	0.81	1.66	25.86	47.57
RL-Separate	0.26	0.61	0.98	13.71	50.04
Uniform Weights (Loss scaling)	0.23	0.29	0.66	25.18	44.68

D Ethics Statement

This work studies adversarial robustness and class-wise performance disparities in semantic segmentation. Improving robustness across semantic classes may be beneficial in applications where failures on particular object categories have disproportionate consequences. At the same time, techniques for generating and evaluating adversarial perturbations are dual-use: they can support robustness assessment and model development, but knowledge of model vulnerabilities may also be used to degrade the performance of deployed vision systems. Our experiments are conducted on established semantic-segmentation benchmarks and concern model-level class disparities rather than demographic or human-subject fairness. No human-subject study is conducted in this work. We use the term adversarial fairness specifically to refer to disparities in adversarial robustness across semantic classes, rather than fairness with respect to protected demographic groups. We encourage the use of these methods for robustness evaluation.

E Reproducibility Statement

We provide the methodological and experimental details needed to reproduce our results in the main paper and appendix, including training settings, attack configurations, datasets, architectures, evaluation metrics, RL-SAF formulation, implementation details, and hyperparameters. We also provide details of the DAFA adaptation and supplementary experiments. Where available, we use released checkpoints; otherwise, models are reproduced following the corresponding published settings. Code and evaluation scripts will be released upon acceptance.