Rethinking Adversary in Semantic Segmentation: An Out-of-Distribution Perspective

Yuxuan Zhang[1], Shuchang Wang[1,3,†], Xiaoman Liu[1], Zhenbo Shi[1], Zhidong Yu[1], Wei Song[4], and Wei Yang[1,2,3,†]

[1] School of Computer Science and Technology, University of Science and Technology of China
[2] Suzhou Institute for Advanced Research, University of Science and Technology of China
[3] Hefei National Laboratory, University of Science and Technology of China
[4] School of Artificial Intelligence and Computer Science, Jiangnan University

Abstract. Semantic segmentation models remain vulnerable to adversarial perturbations, which can degrade pixel-wise predictions despite being imperceptible. Existing adversarial training methods primarily rely on iterative prediction-driven attacks that push features across decision boundaries between in-distribution (ID) classes while remaining close to the ID manifold. This limits their ability to expose more severe representation failures. In this paper, we propose SegOOD, a novel adversarial attack framework that explicitly encourages out-of-distribution (OOD)-like deviations in latent feature space. SegOOD introduces a Feature Wasserstein Separation objective to increase distributional discrepancy between adversarial features and ID semantic prototypes, along with a Weighted K-Nearest-Neighbor Separation objective to enforce local feature-level deviation. These objectives are efficiently implemented using prototype aggregation and feature-aware superpixel clustering. Furthermore, we integrate SegOOD with conventional prediction-driven attacks during adversarial training, improving robustness against both existing ID-style and our OOD-like perturbations. Extensive experiments demonstrate state-of-the-art robustness while maintaining competitive clean performance and no additional inference cost.

Keywords: Semantic segmentation · Adversary · Out of distribution

1 Introduction

Semantic segmentation has achieved remarkable progress over the past decade. Despite these advances, modern segmentation models remain highly vulnerable to adversarial perturbations [27], where quasi-imperceptible input changes can lead to severe degradation in pixel-wise predictions. Given the safety-critical nature of applications such as autonomous driving, enhancing robustness against adversarial attacks has become a crucial research problem.

[†] Corresponding authors: Shuchang Wang and Wei Yang (E-mail: qubit@ustc.edu.cn)

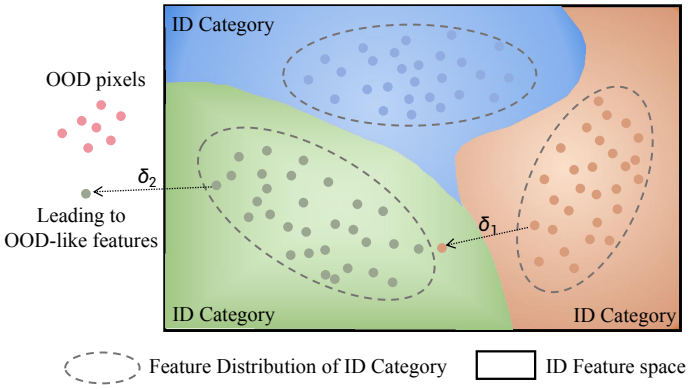

Fig. 1: Pixel-wise feature space of semantic segmentation model. Previous adversarial perturbation δ_1 misleads the feature to other ID categories. We explore whether the tiny perturbation δ_2 can lead to OOD-like properties.

Adversarial training (AT) has emerged as one of the most effective strategies for improving robustness in semantic segmentation. DDC-AT [28] first introduced AT to this domain, followed by SegPGD [12], CosPGD [2], SEA [13], and RPPGD [30], which focus on strengthening the underlying attack mechanisms used during training. These methods demonstrate improved robustness under strong white-box and black-box evaluations.

However, most underlying attacks are prediction-driven. Specifically, they optimize objectives that aim to misclassify as many pixels as possible [12,16,30,32]. Under mild regularity assumptions on the data manifold and feature mapping, such objectives typically generate perturbations that move along tangent directions of the in-distribution (ID) feature manifold. Consequently, adversarial features tend to cross decision boundaries toward other ID categories while remaining close to the original high-density ID manifold in representation space [23] (see Figure 1).

Such behavior effectively confines AT to a limited region of the ID feature manifold, where perturbations primarily induce decision-boundary crossings among ID classes rather than genuine manifold departure. As a result, models trained under this paradigm may fail to encounter adversarial examples that exhibit atypical, low-density representations in feature space. This observation prompts a fundamental question: *Can norm-bounded adversarial perturbations induce feature representations that deviate significantly from the ID manifold, exhibiting OOD-like characteristics*? If such perturbations exist, they represent a qualitatively distinct failure mode compared to conventional prediction-driven attacks. Exploring this regime is essential for broadening the spectrum of adversarial representations encountered during training and for identifying complementary weaknesses in current robust segmentation models.

Motivated by this observation, we propose to explicitly encourage OOD-like representation deviations during adversarial optimization. To this end, we intro-

duce SegOOD, a novel adversarial attack framework for semantic segmentation. SegOOD is built upon two complementary objectives that operate in the latent feature space.

First, we propose the Feature Wasserstein Separation (FWS) objective, which encourages adversarial representations to deviate from clean ID semantic prototypes in the latent space. While the Wasserstein distance provides a principled metric for measuring distributional discrepancy, directly computing pixel-level distances is computationally prohibitive for high-resolution segmentation. To address this challenge, we introduce a dual-sided compression strategy: clean features are aggregated into ID semantic prototypes, while adversarial features are grouped via a feature-aware superpixel clustering method, termed F-SLIC. This design significantly reduces computational overhead while preserving essential structural information. Second, to prevent trivial global drift and ensure meaningful local deviation, we introduce a Weighted K-Nearest Neighbor Separation (WKS) objective that enforces each adversarial superpixel to stay separated from its closest ID prototypes. This constraint provides a local guarantee that complements the global Wasserstein separation.

By integrating these objectives, SegOOD adopts an iterative PGD-based optimization scheme [20] to drive adversarial features away from the ID manifold while increasing their free energy scores. Unlike conventional prediction-driven attacks, SegOOD explicitly encourages representation-level deviations, enabling it to expose new robustness deficiencies in existing AT-trained models.

Furthermore, we employ both SegOOD and the classic prediction-driven attack SegPGD [12] as underlying attacks during adversarial training. This hybrid strategy significantly improves robustness against both ID-style and OOD-like adversarial perturbations, while maintaining competitive clean performance and introducing no additional inference overhead.

- We rethink the adversarial attack pattern in semantic segmentation, and propose SegOOD, a novel underlying attack strategy that promotes OOD-like feature deviations to deceive current robust segmentation models.
- We design a Feature Wasserstein Separation objective and a Weighted K-Nearest Neighbor Separation objective, which together encourage adversarial features to deviate from ID semantic prototypes in latent space.
- Extensive experiments demonstrate that AT with SegOOD achieves state-of-the-art robustness against various attacks across multiple segmentation robust models and datasets, while maintaining efficient inference and minimal performance degradation on clean data.

2 Related Work

2.1 Adversarial Attack in Semantic Segmentation

Several attack strategies originally developed for image classification can be adapted to semantic segmentation models [3, 4, 20]. To further strengthen attack capability, existing methods primarily aim to mislead as many pixels as

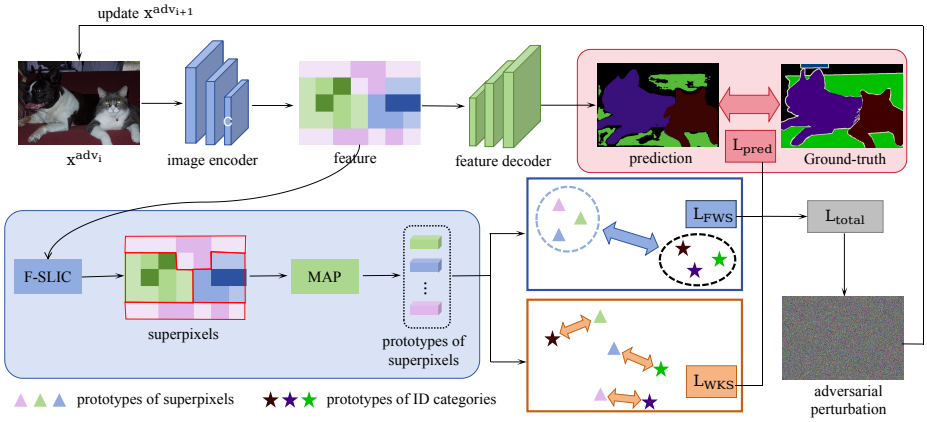

Fig. 2: Overall pipeline of SegOOD. In each attack iteration, SegOOD jointly calculates L_{FWS}, L_{WKS}, and L_{pred} to perform a comprehensive perturbation, resulting in OOD-like representations to enhance attack capability.

possible. For example, SegPGD [12] enhances attack effectiveness via adaptive regional weighting, CosPGD [2] introduces an alignment score to modulate optimization smoothly, and RPPGD [30] refines attacks through fine-grained region division. Beyond white-box settings, transfer-based black-box attacks leverage surrogate models, with several approaches improving transferability in feature space [17, 23]. In addition, adversarial patches [22, 24] provide a practical and real-world attack paradigm.

Despite their effectiveness, these approaches remain confined to ID-style adversarial perturbations, as they primarily induce feature transitions among ID categories rather than encouraging departure from the learned feature manifold. In contrast, we rethink adversarial attack design from a representation-level perspective and propose SegOOD, which explicitly drives OOD-like deviations in latent space to explore a qualitatively different adversarial regime.

2.2 Adversarial Defense in Semantic Segmentation

To defend against adversarial attacks in semantic segmentation, various strategies have been proposed. Perturbation detection [7,26] identifies adversarial patterns to reject unreliable predictions, while model pruning [29, 35] removes vulnerable neurons to improve robustness. Among these approaches, AT has proven the most effective, initially introduced by DDC-AT [28]. Later works, such as SegPGD [12], CosPGD [2], SEA [13], and RPPGD [30], focus on enhancing the underlying attack for AT, as weak attacks may give a false sense of robustness. However, existing AT strategies are insufficient to defend against the proposed OOD-like adversarial samples. In this work, we investigate AT schemes that can simultaneously resist both ID-style and OOD-like attacks.

3 Methodology

3.1 Task Preview

Given an input image $x \in \mathbb{R}^{3\times N} \subset \mathcal{D}_{\text{train}}$ ($N = H \times W$), a semantic segmentation model $F(x;\theta)$ aims to predict each pixel in one of C predefined categories to approximate the ground-truth $y \in \{1, 2, ..., C\}^N$, where θ is the model parameter of F. The segmentation model can be split as $F = \text{Softmax}\,(h \circ g)$, where $g : \mathbb{R}^{3\times N} \to \mathbb{R}^{d\times N}$ is the segmentation encoder that extracts pixel-wise features with d dimensions, and $h : \mathbb{R}^{d\times N} \to \mathbb{R}^{C\times N}$ is the segmentation decoder that produces pixel-wise logits for prediction. $\circ$ represents function composition.

However, segmentation models are vulnerable to imperceptible adversarial perturbation δ, where $\arg\max_c F(x+\delta;\theta) \neq \arg\max_c F(x;\theta)$. To defend against such attacks, AT enforces prediction consistency by training on adversarial sample x^{adv} ($x^{\text{adv}} := x + \delta$) as a min-max problem:

$$\min_{\theta} \max_{||\delta||_\infty \leq \epsilon} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}} L_{\text{task}}(x+\delta, y; \theta) \tag{1}$$

The inner maximization problem aims to generate adversarial samples that maximize the objective L, while the outer minimization seeks optimal model parameters θ to defend against such attacks. To ensure imperceptibility, perturbations are constrained within an ϵ-ball with l_∞-norm.

3.2 Motivation

Existing adversaries mainly focus on misleading the predictions, without fully exploring the properties in the latent space. We experimentally observe that these attacks push features across decision boundaries toward other ID classes. That is, the latent representations remain near the ID space (refer to the energy and softmax probabilities in Figures 5 and 6).

To give a more theoretical explanation of this phenomenon, we first make the following assumptions:

Assumption 1 *(Manifold and feature map regularity) The data support locally lies on a compact m-dimensional embedded submanifold $M \subset \mathbb{R}^N$. The feature map g is twice continuously differentiable in a neighbourhood of M. Let $S = g(M)$, For every $x \in M$ the Jacobian $J_g(x) \in \mathbb{R}^{d\times N}$ has constant rank m and*

$$range(J_g(x)) = T_z S, \quad z = g(x). \tag{2}$$

So the image of the input tangent space under J_g equals the tangent space of S at z.

Assumption 2 *(Loss gradient lies in tangent directions) There exists a neighborhood U of S in Z such that for every $z \in U$ the feature-space gradient $\nabla_z L_{task}(x)|_z$ belongs to the tangent space $T_z S$. Equivalently, the directional derivative of the loss in any normal direction to S is zero at points of S.*

Assumption 3 *(Uncertainty on low-likelihood features) There exist constants* $\eta \geq 0$ *and* $\zeta > 0$ *such that for any* $z \in Z$ *with* $p_{id}(z) < \zeta$ *the model's max-class probability is bounded:*

$$\max_c softmax(h_c(z)) \leq \frac{1}{C} + \eta, \tag{3}$$

with η *small. This formalizes the hypothesis OOD features cause near-uniform predictions.*

Assumption 4 *(Small-*ϵ *linearization control) There exists* C_1, $C_2 > 0$ *such that for all* $||\delta||_p \leq \epsilon$ *small enough,*

$$g(x+\delta) = g(x) + J_g(x)\delta + r_g(\delta), \tag{4}$$

with $||r_g(\delta)|| \leq C_1||\delta||^2$, and

$$L_{\text{task}}(x+\delta) = L_{\text{task}}(x) + \nabla_z L_{\text{task}}(x)^\top J_g(x)\delta + r_L(\delta), \tag{5}$$

with $||r_L(\delta)|| \leq C_2||\delta||^2$ based on Taylor's theorem with remainder.

Theorem 1. *For small* ϵ*, any adversarial perturbation* δ^* *that locally maximizes* L_{task} *produces a feature* $z^{adv} = g(x+\delta^*)$ *that remains within an* $O(\epsilon^2)$ *neighborhood of the ID manifold* S*. Consequently, if* p_{id} *is continuous on* S*,* $p_{id}(z^{adv})$ *remains close to* $p_{id}(z)$ *and in particular is not below the OOD threshold* ζ*.*

The detailed proof is provided in the supplementary material. For semantic segmentation models, previous iterative adversarial attacks that maximize the pixel-wise cross-entropy loss primarily move along semantic manifold directions. Because off-manifold perturbations only change the loss at second order and produce uncertain predictions, and the adversarial feature z^{adv} stays near the ID manifold. These observations motivate the introduction of an enhanced objective in later sections that explicitly promotes distributional separation.

To explicitly introduce OOD-like properties in adversarial samples, we propose SegOOD, a novel adversarial attack for semantic segmentation. As shown in Figure 2, SegOOD introduces two attack objectives to emulate OOD-like behavior, namely the FWS and WKS objectives, detailed in the following subsections.

3.3 Feature Wasserstein Separation

To design an attack objective for inducing OOD-like features, an intuitive idea is to separate the pixel-wise representations between x and x^{adv}. Specifically, the pixel-wise representation set of clean image x and adversarial image x^{adv} can be defined as $P := \{g_i(x) \in \mathbb{R}^D | i = 1, ..., N\}$ and $Q := \{g_i(x^{\text{adv}}) \in \mathbb{R}^D | i = 1, ..., N\}$, respectively, where i represents the pixel index and $N = HW$ is the number of pixels in the image. Accordingly, Wasserstein distance can be served

as an attack objective to increase the difference between these two feature distributions [25], formulated by:

$$L_{FWS} = \min_{\pi \in \Pi(P,Q)} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \pi_{ij} \cdot ||P_i - Q_j||^2 \right)^{\frac{1}{2}}, \tag{6}$$

where π denotes a transport plan and Π is the solution space. Wasserstein distance calculates the minimum cost required to transport all points from P to Q. The existence of a minimizer is guaranteed because the solution space is finite. However, the computational complexity is extremely high, with $O(N^3)$ based on the classic linear program [9]. Therefore, it is essential to reduce the points of both P and Q to alleviate the computational burden.

Motivated by [10, 34], features corresponding to a single class exhibit high intra-class similarity, allowing for the use of a global descriptor to represent the average feature of each category, a.k.a the prototype. The prototype of category c is calculated by masked average pooling:

$$p_c = \frac{\sum_{x \in \mathcal{D}_{\text{train}}} \sum_{i=1}^{N} g_i(x) \cdot \mathbb{1} \left[F_i^c(x;\theta) > \tau \right]}{\sum_{x \in \mathcal{D}_{\text{train}}} \sum_{i=1}^{N} \mathbb{1} \left[F_i^c(x;\theta) > \tau \right]}, \tag{7}$$

where $\mathbb{1}[\cdot]$ is an indicator function, $F_i^c(\cdot)$ denotes the softmax probability of category c at the i-th pixel, and τ is a threshold. The prototype of each category is derived based on the average of the training set. Accordingly, clean feature set P can be simplified to $P := \{p_i | i = 1, ..., C\}$.

On the other hand, adversarial inputs may corrupt the semantic feature space, making category-wise prototypes unreliable as cluster centers. To tackle this, we propose a feature-aware superpixel method, termed F-SLIC, to cluster adversarial features and reduce the dimensionality of Q. Unlike the classic SLIC algorithm [1] that groups pixels based on RGB values and spatial proximity, F-SLIC incorporates feature similarity and pixel coordinates to form clusters. This is inspired by the fact that similar RGB values may correspond to distinct semantic representations under adversarial perturbations, making traditional SLIC unsuitable for feature-based clustering. Accordingly, the distance between two pixels i and j is defined as:

$$D_{i,j} = \sqrt{(d_f)^2 + (d_c/m)^2}, \tag{8}$$

where $d_f = (1 - \frac{g_i \cdot g_j}{||g_i|| \cdot ||g_j||})/2$ denotes the normalized negative cosine similarity between two pixel-wise representations, and $d_f \in [0, 1]$. A larger d_f indicates a lower similarity between two pixel-wise representations. Similarly, d_c represents the Euclidean distance between two pixels in the coordinate space, also normalized to $[0, 1]$ for consistency. The parameter m controls the trade-off between feature similarity d_f and spatial proximity d_c, balancing their contributions to the overall distance metric for clustering.

Based on the distance metric in Eq. (8), we leverage K-means to cluster the pixels in an adversarial image into C groups $\{s_i | i = 1, ..., C\}$, and call them

feature-aware superpixels. To reduce the data points in Q, we calculate the prototype q_i of each feature-aware superpixel via masked average pooling (analogous to Eq. (7)), resulting in a set of superpixel-wise prototypes $Q := \{q_i | i = 1, ..., C\}$. Based on the dimension reduction on both sides, we revisit the formulation of L_{FWS} in Eq. (6). The computational complexity is thereby reduced from $O(N^3)$ to $O(C^3)$, achieving efficient computation while maintaining effective results. As to the detailed algorithm of F-SLIC, please see the supplementary material.

3.4 Weighted K-Nearest-Neighbor Separation

While the Feature Wasserstein Separation objective enlarges the global distributional discrepancy between clean and adversarial representations, it does not prevent individual adversarial superpixel prototypes from remaining close to certain ID prototypes. To further enforce local deviation from the ID feature manifold, we introduce a Weighted K-Nearest-Neighbor Separation constraint.

KNN distance. Let $P = \{p_1, \ldots, p_C\}$ denote the set of ID category prototypes defined in Eq. (7), and let $Q = \{q_1, \ldots, q_C\}$ denote the superpixel-wise adversarial prototypes obtained via F-SLIC. For each $q_i \in Q$, we define its K-nearest-neighbor distance to P as

$$d_K(q_i, P) = \frac{1}{K} \sum_{p \in \mathcal{N}_K(q_i)} \cos(q_i, p), \tag{9}$$

where $\mathcal{N}_K(q_i) \subset P$ denotes the set of K nearest prototypes to q_i in Euclidean distance, and cos represents cosine similarity. Based on this distance, a naive KNN separation objective can be formulated in a margin-based separation constraint:

$$L_{KNN} = \frac{1}{C} \sum_{i=1}^{C} \max\big(0,\ -\gamma + d_K(q_i, P)\big), \tag{10}$$

where $\gamma > 0$ is a predefined margin, and we set $\gamma = 0.3$ empirically. This objective penalizes adversarial prototypes whose local neighborhood similarity to ID prototypes exceeds the margin γ.

However, we observe a uniform superpixel averaging bias here:

Theorem 2 (Uniform Superpixel Averaging Bias). *Let $d(q_i) \coloneqq d_K(q_i, P)$, assume that:*

***1. One dominant superpixel**: There exists q_l corresponding to a large superpixel with size $|s_l| \gg |s_i|$ for all $i \neq l$.*

***2. Near-ID dominant region**: $d(q_l) = \alpha$, with $\alpha \approx 1$.*

***3. Small superpixels strongly separated**: $d(q_i) = \beta \quad \forall i \neq l$, with $\beta \approx 0$.*

In this case, we have $\beta < \gamma < \alpha$. Considering a large C thus:

$$L_{KNN} = \frac{-\gamma + \alpha}{C} \to 0. \tag{11}$$

Based on Theorem 2, even when the dominant superpixel remains close to the ID manifold, the uniformly averaged objective can become arbitrarily small due to the dilution effect induced by numerous small superpixels [31]. This weak optimization signal prevents effective deviation of large semantic regions toward OOD-like representations. To address this issue, we assign each superpixel a weight proportional to its spatial area, yielding the size-aware formulation:

$$L_{WKS} = \frac{1}{H \times W} \sum_{i=1}^{C} |s_i| \cdot \max\big(0,\ -\gamma + d_K(q_i, P)\big). \tag{12}$$

In this manner, dominant superpixels receive stronger optimization signals, encouraging larger spatial regions to deviate toward OOD-like representations in the latent space.

3.5 Overall SegOOD Attack for AT

Based on the above objectives, we formulate the overall attack objective as:

$$L_{total} = \beta_1 L_{FWS} + \beta_2 L_{WKS} + \beta_3 L_{pred}, \tag{13}$$

where $(\beta_1, \beta_2, \beta_3) = (1, 1, 1)$ in this paper unless explicitly stated, and L_{pred} denotes a standard prediction-based attack objective designed to mislead pixel-wise predictions of the segmentation model. In this work, we employ the formulation used in SegPGD [12]. For detailed formulations, see the supplementary material.

As SegOOD adopts a PGD-like iterative paradigm [20] to organize the attack, the adversarial sample of the i-th attack iteration can be formulated by:

$$x^{\mathrm{adv}_{i+1}} = \Phi_\epsilon \left(x^{\mathrm{adv}_i} + \alpha \cdot \mathrm{sign}\left(\bigtriangledown_{x^{\mathrm{adv}_i}} L_{total} \right) \right), \tag{14}$$

where x^{adv_i} denotes the adversarial sample at the i-th iteration, α is the step size, and Φ_ϵ projects the adversarial sample into the ϵ-ball around the clean image under the l_∞-norm constraint. A detailed description of the SegOOD attack procedure is given in the supplementary material for a clearer understanding.

After generating adversarial samples using SegOOD, we inject them into the clean training set for AT. By introducing OOD-like properties that present greater challenges for segmentation models, SegOOD enables AT to more effectively enhance model robustness against a broad range of adversarial attacks in semantic segmentation.

4 Experiment

4.1 Setup

Dataset: Following previous benchmarks, we adopt two commonly used semantic segmentation datasets for evaluation. Pascal VOC [11] contains 1,464, 1,499, and 1,456 images for training, validation, and test, respectively, with 20 predefined categories. Following SBD [14], the training set is extended to 10,582

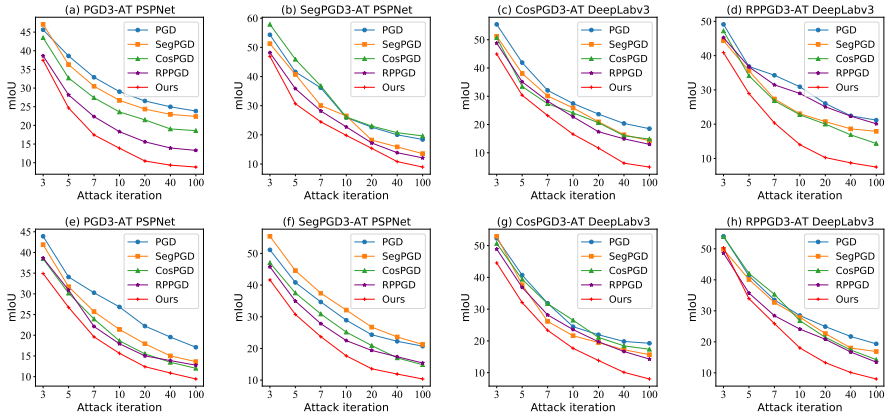

Fig. 3: Attack efficacy of SegOOD compared with several top-performing adversaries. Figures (a)-(d) illustrate the mIoU results on Pascal VOC, and the results on Cityscapes are presented in figures (e)-(h). Compared with other adversaries. SegOOD performs a stronger attack capability on various existing AT strategies.

images. Cityscapes [8] focuses on urban scene understanding and includes 2,975 training, 500 validation, and 1,525 test images annotated with 19 categories.

Metric: We use mean Intersection over Union (mIoU) to assess the semantic segmentation performance, where a larger mIoU drop indicates stronger attack effectiveness. Moreover, we employ energy scores [18] and maximum softmax probability distributions to evaluate OOD properties.

Implementation Details: To keep consistency with previous works, we select PSPNet [33] and DeepLabv3 [6] as semantic segmentation models to assess the attack efficacy and model robustness with AT. Besides, we also adopt extra models and datasets for comprehensive evaluation on OOD-ness. These segmentation models are based on the ResNet50 backbone [15]. Moreover, we use the top-performing iterative attacks PGD [20], SegPGD [12], CosPGD [2], and RPPGD [30] as our baseline attacks. For better demonstration, we employ PGDn to represent PGD attack with n iterations, and PGDn-AT denotes conducting AT on adversarial samples generated by PGD with n attack iterations. For example, SegPGD3-AT PSPNet represents PSPNet adversarially trained with the adversarial samples generated by SegPGD with 3 iterations. Our attack follows the commonly used l_∞-norm with the step size $\alpha = 3/255$ and perturbation budget $\epsilon = 8/255$. Note that these settings strictly follow the DDC-AT protocol. For the ID prototypes, we update them using an exponential moving average based on clean samples from each mini-batch, with a momentum factor of 0.99.

4.2 Attack Effectiveness

Attack Capability: We evaluate the attack capability of SegOOD by reporting mIoU at iterations 3, 5, 7, 10, 20, 40, and 100, and compare it with other

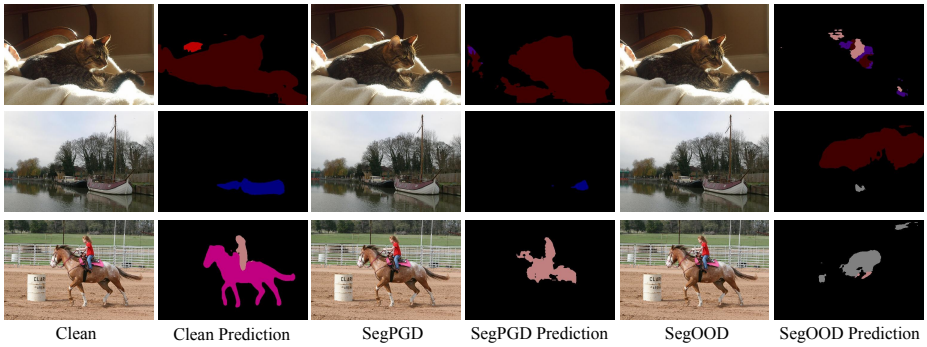

Fig. 4: Qualitative results of segmentation robustness on PGD3-AT PSPNet under adversarial attacks. SegOOD demonstrates superior attack effectiveness on robust segmentation model compared to SegPGD.

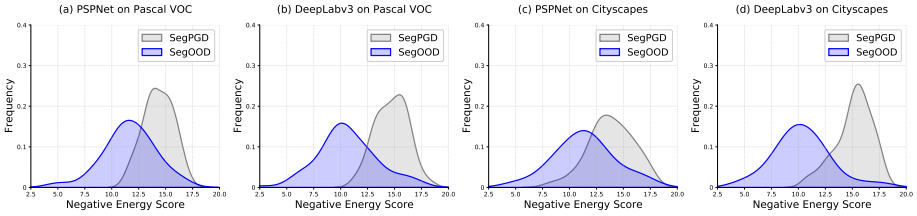

Fig. 5: Negative energy score. SegOOD produces more OOD-like energy distributions.

adversaries. As shown in Figure 3, SegOOD shows stronger attack performance across various robust models with multiple AT strategies on Pascal VOC and Cityscapes. This superiority is mainly attributed to two factors: 1) SegOOD simulates OOD-like properties in adversarial samples, making segmentation current robust models AT with ID-style underlying attack difficult to defend against; 2) By explicitly guiding features away from the ID space, SegOOD prevents mispredicted pixels from reverting to their benign states, which troubles prior attacks like SegPGD (please refer to the supplementary material for more details). We also illustrate the attack efficacy of SegPGD and our SegOOD in Figure 4, SegOOD exhibits stronger attack capability on the robust segmentation model PGD3-AT PSPNet.

OOD-like Properties: To verify that SegOOD generates adversarial samples with OOD-like properties, we compare it with the baseline attack SegPGD in terms of energy scores [18], maximum softmax probability distributions, and pixel-wise representations. Here, we employ standard models to exhibit the intrinsic distributional deviation of the adversarial samples. As shown in Figure 5, SegOOD produces lower negative energy scores and exhibits a clear gap compared to SegPGD, indicating stronger OOD-like behavior. Moreover, we visualize the pixel-wise maximum softmax probability distributions in Figure 6. Compared to SegPGD, our attack exhibits lower maximum softmax probabilities, suggest-

Table 1: Ablation results of attack effectiveness on Pascal VOC.

(a) Objective combinations.

L_{pred}	L_{FWS}	L_{WKS}	SegPGD3-AT PSPNet	CosPGD3-AT DeepLabv3
✓			26.48	25.88
✓	✓		22.47	20.68
✓		✓	24.70	23.55
	✓	✓	25.12	22.80
✓	✓	✓	**19.82**	**16.57**

(b) Different clustering methods for adversarial features.

Cluster	SegPGD3-AT PSPNet	CosPGD3-AT DeepLabv3
Prototype	34.85	30.17
SLIC	24.91	25.54
F-SLIC	**19.82**	**16.57**

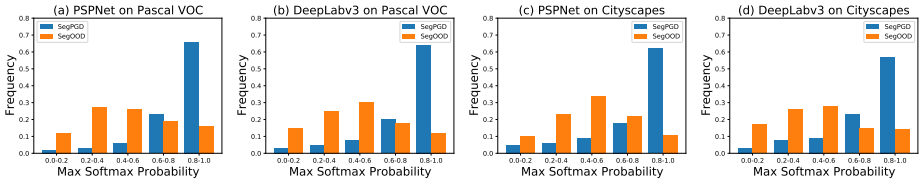

Fig. 6: Max softmax distribution. SegOOD produces more OOD-like distributions.

ing reduced prediction confidence and more OOD-like properties. In addition, we visualize pixel-wise features using t-SNE [19] in the supplementary material. The results show that SegPGD features stay close to ID prototypes, while SegOOD generates representations that deviate significantly, showcasing stronger OOD-like characteristics in the adversarial samples.

4.3 Ablation Study

To further assess SegOOD, we conduct extensive ablation experiments on Pascal VOC with two robust segmentation models, i.e., SegPGD3-AT PSPNet and CosPGD3-AT DeepLabv3, to explore the optimal choices for several key components, using 10 attack iterations for evaluation.

Attack Combination: We investigate the effectiveness of different combinations of attack objectives. As shown in Table 2a, the combination of L_{FWS}, L_{WKS}, and L_{pred} yields the strongest attack performance. Specifically, both L_{FWS} and L_{WKS} achieve strong attack effectiveness, inducing OOD-like characteristics that make the adversarial examples difficult for existing robust models to defend against, thereby significantly improving the overall attack performance. When combined with L_{pred}, the attack becomes even more effective, as it introduces an additional optimization direction that further strengthens the attack.

F-SLIC vs. Other Clustering Methods: We compare F-SLIC with two alternatives: category-wise prototypes and SLIC [1]. The former uses semantic prototypes for clustering, while the latter relies on RGB color. As shown in Table 2b, F-SLIC outperforms both. Semantic prototypes are unreliable under adversarial attacks, and RGB-based clustering overlooks contextual information, which is critical for semantic segmentation. Therefore, F-SLIC serves as an effective approach for reducing the dimension of pixel-wise features.

For more ablation studies, please refer to the supplementary material.

Table 3: mIoU of PSPNet and DeepLabv3 under different AT strategies against several white-box attacks on Pascal VOC.

AT Strategy	PSPNet								DeepLabv3							
	Clean	CW	DF	PGD	SegPGD	CosPGD	RPPGD	SegOOD	Clean	CW	DF	PGD	SegPGD	CosPGD	RPPGD	SegOOD
N/A	76.64	4.72	14.20	5.21	2.03	2.94	1.76	1.47	77.36	5.24	13.57	4.36	1.80	2.36	1.39	1.07
DDC-AT	76.44	55.37	57.16	22.94	23.63	22.81	20.54	15.20	75.81	56.46	60.72	26.18	21.90	20.74	23.30	17.44
SegPGD3-AT	75.38	56.52	59.47	26.60	27.09	26.85	22.78	19.82	75.01	59.55	62.12	26.29	28.03	24.62	20.35	16.87
CosPGD3-AT	74.04	53.37	57.19	25.33	23.87	30.51	24.47	14.87	74.60	58.01	62.32	28.26	25.87	24.11	22.85	16.57
SEA3-AT	73.17	59.85	60.73	28.61	26.14	27.25	24.11	16.73	72.91	60.48	63.08	29.87	27.10	26.08	25.40	18.29
RPPGD3-AT	75.17	59.34	**62.37**	30.25	28.19	30.08	**27.14**	13.98	75.20	61.21	**64.40**	31.22	23.04	22.79	**28.61**	14.03
Ours-Co	73.85	58.64	61.30	31.14	29.57	30.85	26.41	25.83	75.02	59.71	63.89	30.83	30.17	25.94	27.68	27.03
Ours-Sep	74.39	**61.26**	62.28	**32.54**	**30.37**	**31.40**	27.03	**30.15**	74.88	**63.18**	64.23	**34.07**	**32.29**	**27.51**	28.40	**32.35**
SegPGD7-AT	74.45	48.79	51.44	25.73	28.33	23.98	22.04	14.83	74.46	51.42	51.47	30.95	29.55	27.99	24.61	15.48
CosPGD7-AT	75.12	47.43	49.86	26.10	25.84	27.17	21.90	17.39	72.88	50.66	52.37	28.70	28.06	29.13	23.93	17.15
SEA7-AT	73.30	52.17	58.84	30.27	29.30	**29.84**	27.03	15.11	73.09	52.80	57.28	32.87	31.27	29.79	28.14	20.29
RPPGD7-AT	74.93	55.65	60.32	29.15	31.24	27.85	32.70	14.93	74.01	57.58	59.60	33.52	31.47	26.56	**33.21**	15.34
Ours-Co	74.40	58.11	59.73	34.28	31.75	28.61	30.46	25.82	73.53	57.42	55.18	33.39	32.12	29.51	31.62	29.63
Ours-Sep	73.90	**60.42**	**61.16**	**35.38**	**33.64**	29.80	**33.17**	**31.10**	73.88	**59.57**	**61.74**	**35.27**	**34.86**	**30.17**	32.97	**35.11**

Table 4: mIoU of PSPNet and DeepLabv3 under different AT strategies against several white-box attacks on Cityscapes.

AT Strategy	PSPNet								DeepLabv3							
	Clean	CW	DF	PGD	SegPGD	CosPGD	RPPGD	SegOOD	Clean	CW	DF	PGD	SegPGD	CosPGD	RPPGD	SegOOD
N/A	73.98	5.94	12.68	0.96	0.60	0.83	0.61	0.51	73.82	8.24	14.26	1.07	0.89	1.54	0.72	0.61
DDC-AT	73.31	34.59	36.82	31.17	25.85	23.90	20.20	17.46	71.59	34.85	38.13	30.77	19.67	22.74	18.41	15.18
SegPGD3-AT	71.01	36.30	38.27	33.52	32.11	25.16	22.48	17.56	71.04	37.93	37.63	32.11	27.93	25.56	20.44	20.17
CosPGD3-AT	72.51	34.70	34.18	30.76	27.79	26.41	23.36	19.03	70.74	36.83	39.07	31.81	21.63	26.52	23.57	17.67
SEA3-AT	70.86	37.74	38.10	34.57	31.42	26.27	23.27	16.71	70.44	35.71	39.01	34.94	25.84	24.97	23.61	15.59
RPPGD3-AT	71.58	38.64	39.70	35.19	33.57	28.69	26.40	14.72	71.25	**38.67**	38.39	36.26	27.94	26.84	24.11	18.04
Ours-Co	71.12	37.22	41.01	34.02	33.41	26.88	25.32	25.06	71.84	37.05	36.92	34.78	28.91	24.61	25.08	27.57
Ours-Sep	72.14	**40.11**	**42.36**	**35.41**	**36.52**	**29.96**	**27.71**	**32.02**	71.21	38.64	**41.27**	**37.92**	**33.48**	**29.49**	**27.02**	**30.71**
SegPGD7-AT	70.21	29.59	30.68	27.13	33.84	27.39	24.64	14.79	69.93	29.73	31.30	30.43	32.17	27.73	24.50	13.96
CosPGD7-AT	69.18	29.32	30.87	26.69	25.83	28.53	23.17	16.21	69.26	28.38	30.70	31.08	26.90	29.22	22.82	15.77
SEA7-AT	67.80	31.79	32.68	31.55	27.27	24.92	21.65	13.30	68.80	31.27	34.19	34.75	30.04	26.65	25.30	20.05
RPPGD7-AT	70.08	31.47	33.82	30.74	32.05	27.26	**29.51**	13.74	69.67	30.33	34.07	36.16	30.79	29.23	**27.00**	18.16
Ours-Co	68.55	30.77	33.18	30.90	31.04	27.43	26.12	28.64	70.15	29.98	32.84	35.73	31.77	29.05	24.67	27.80
Ours-Sep	69.92	**35.88**	**35.04**	**31.86**	**34.18**	**28.81**	29.32	**35.38**	69.55	**33.44**	**35.21**	**38.58**	**35.84**	**31.41**	26.78	**33.29**

4.4 Adversarial Training with SegOOD

We further probe into whether AT with SegOOD has a positive effect on enhancing the robustness of segmentation models. To ensure a fair comparison, we adopt the same setting of AT in previous works [12, 28, 30]. We employ the underlying attacks SegPGD, CosPGD, RPPGD, and SegOOD with 3 and 7 attack iterations to generate adversarial samples for AT, respectively. In addition, we also adopt an ensemble attack SEA [13] for AT. We evaluate two AT strategies: 1) **Ours-Co**: We use the unified objective in Eq. (13) to jointly optimize prediction and OOD objectives when generating adversarial samples for AT; 2) **Ours-Sep**: We separately use L_{pred} $(\beta_1, \beta_2, \beta_3 = (0, 0, 1))$ and L_{ood} $(\beta_1, \beta_2, \beta_3 = (1, 1, 0))$ to generate adversarial samples, where each objective is used for half of the training samples. The details of the latter AT process (Ours-Sep) are provided in the supplementary material.

Against White-box Attacks: We first evaluate the model robustness under various white-box attacks, including CW [5], DeepFool (DF) [21], PGD [20], SegPGD [12], CosPGD [2], RPPGD [30], and our proposed SegOOD attacks with 10 iterations. The results on Pascal and Cityscapes are summarized in Tables 3 and 4, respectively. As shown in the tables, segmentation models without AT are highly vulnerable to white-box attacks. While existing AT strategies

Table 5: mIoU results of PSPNet and DeepLabv3 under different AT strategies against black-box attacks. A→B means adversarial samples are crafted on A and used to attack B with different AT strategies.

Dataset	AT Strategy	PSPNet→DeepLabv3							DeepLabv3→PSPNet						
		SegPGD	RPPGD	TranSegPGD	FSPGD	RWAE	AdvPatch	SegOOD	SegPGD	RPPGD	TranSegPGD	FSPGD	RWAE	AdvPatch	SegOOD
Pascal VOC	N/A	8.78	5.91	7.04	4.70	26.75	24.93	3.31	8.46	4.72	4.51	3.60	24.83	27.15	2.77
	DDC-AT	11.94	8.92	11.45	8.07	31.14	26.14	6.58	9.73	6.86	8.40	5.91	28.73	28.61	6.19
	SegPGD3-AT	13.34	9.17	12.69	10.77	35.39	27.19	8.05	10.51	8.96	9.37	8.75	30.46	25.62	6.94
	CosPGD3-AT	12.81	9.35	10.93	9.60	33.07	26.84	6.12	11.27	8.70	10.62	9.33	35.26	28.85	5.27
	SEA3-AT	13.86	11.54	11.93	10.78	34.06	30.09	7.59	14.68	11.86	11.24	11.03	34.00	37.81	8.80
	RPPGD3-AT	17.79	15.10	12.20	11.89	37.49	31.96	10.41	15.31	15.40	10.14	11.35	32.74	35.17	11.74
	Ours-Sep	**22.96**	**19.55**	**16.99**	**16.44**	**42.05**	**37.28**	**19.30**	**18.50**	**18.04**	**16.11**	**15.96**	**41.07**	**44.18**	**23.85**
Cityscapes	N/A	10.33	8.29	5.91	5.02	26.08	23.17	4.61	9.56	8.12	6.64	5.31	23.14	22.90	3.69
	DDC-AT	13.32	11.35	9.57	7.19	23.52	25.15	6.45	11.75	7.62	8.23	7.50	26.13	24.41	5.81
	SegPGD3-AT	20.11	12.55	13.84	9.74	30.46	23.82	8.18	12.82	9.03	12.60	11.46	32.05	26.14	7.07
	CosPGD3-AT	17.62	11.88	10.64	13.43	32.83	27.61	7.52	12.37	8.86	10.44	13.31	29.17	30.79	8.16
	SEA3-AT	22.61	13.39	12.42	15.18	29.84	33.64	9.72	13.34	16.11	13.95	14.45	30.57	36.60	11.28
	RPPGD3-AT	23.56	16.73	15.36	12.25	33.76	35.26	12.27	16.21	14.43	11.71	14.64	31.93	40.09	10.86
	Ours-Sep	**25.90**	**20.70**	**23.62**	**19.17**	**38.47**	**35.77**	**21.73**	**23.69**	**21.40**	**18.57**	**18.22**	**35.50**	**44.09**	**24.15**

improve robustness against previous ID-style iterative attacks, they still struggle to defend against our proposed SegOOD attack. In contrast, AT with SegOOD (Ours-Sep) significantly enhances segmentation robustness against a wide range of adversarial attacks, especially our OOD-style attack SegOOD.

Moreover, we observe that AT with SegOOD solely leads to a slight mIoU drop on clean images (refer to column "Clean"), while preserving promising segmentation accuracy. We also show in the supplementary material that models AT with SegOOD achieve competitive performance on clean images compared to standard models. Additionally, even with only 3 attack iterations, AT with SegOOD outperforms other approaches that use 7 iterations in most cases, demonstrating its efficiency in training a robust model. Note that, AT does not incur any additional inference cost, ensuring consistent efficiency to standard models.

Against Black-box Attacks: To evaluate robustness under more realistic settings, we test several transfer-based black-box attacks with 100 iterations, including SegPGD [12], RPPGD [30], TranSegPGD [17], and FSPGD [23]. These attacks generate adversarial samples from a surrogate model (e.g., PSPNet) to attack a target model (e.g., DeepLabv3), and *vice versa*. We also include two adversarial patch attacks, RWAE [22] and AdvPatch [24], with patch size 150×300. Moreover, we adopt SegOOD to perform a transfer-based black-box attack and evaluate the model robustness against such a potent attack. As shown in Table 5, AT with SegOOD (Ours-Sep) consistently achieves superior robustness across both models and datasets, especially against transfer-based SegOOD. This is because SegOOD exposes the model to feature-level distributional deviations, encouraging the encoder to learn representations that are robust to distribution shifts and thus improving transfer robustness. Moreover, separating SegPGD and OOD-based attacks during AT provides complementary supervision: SegPGD enhances robustness against prediction-level perturbations, while OOD-based attacks improve robustness at the feature level. Their combination leads to more comprehensive robustness against both types of adversarial perturbations.

5 Conclusion

In this paper, we rethink adversarial attacks in semantic segmentation and propose an OOD-like attack, termed SegOOD, to challenge existing robust models. By introducing a Feature Wasserstein Separation objective and a Weighted K-Nearest-Neighbor Separation objective, SegOOD generates adversarial samples with clear OOD-like characteristics, making them more difficult for robust segmentation models to defend against. Furthermore, incorporating these adversarial samples into adversarial training significantly improves model robustness, and combining them with other attacks further enhances robustness against diverse threats. Extensive experiments demonstrate that adversarial training with SegOOD surpasses state-of-the-art methods by a large margin, while maintaining strong clean performance and efficient inference.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No. 62172385, National Natural Science Foundation of China (T2541014), CCF-Baidu Open Fund, Fundamental Research Funds for the Central Universities (WK2150250041), the Innovation Program for Quantum Science and Technology under Grant No. 2021ZD0302900, and the Anhui Provincial Department of Science and Technology under Grant No. 202103a05020009.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence **34**(11), 2274–2282 (2012)
2. Agnihotri, S., Keuper, M.: Cospgd: a unified white-box adversarial attack for pixel-wise prediction tasks. In: ICML (2024)
3. Bai, T., Cao, Y., Xu, Y., Wen, B.: Stealthy adversarial examples for semantic segmentation in remote sensing. IEEE Transactions on Geoscience and Remote Sensing (2024)
4. Bar, A., Klingner, M., Varghese, S., Huger, F., Schlicht, P., Fingscheidt, T.: Robust semantic segmentation by redundant networks with a layer-specific loss contribution and majority vote. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 332–333 (2020)
5. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. Ieee (2017)
6. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
7. Cho, S., Jun, T.J., Oh, B., Kim, D.: Dapas: Denoising autoencoder to prevent adversarial attack in semantic segmentation. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2020)

8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
9. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
10. Dong, N., Xing, E.P.: Few-shot semantic segmentation with prototype learning. In: BMVC. vol. 3 (2018)
11. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010)
12. Gu, J., Zhao, H., Tresp, V., Torr, P.H.: Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In: European Conference on Computer Vision. pp. 308–325. Springer (2022)
13. Halmosi, L., Mohos, B., Jelasity, M.: Evaluating the adversarial robustness of semantic segmentation: Trying harder pays off. In: European Conference on Computer Vision. pp. 1–18. Springer (2024)
14. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 447–456 (2015)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. Hendrik Metzen, J., Chaithanya Kumar, M., Brox, T., Fischer, V.: Universal adversarial perturbations against semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 2755–2764 (2017)
17. Jia, X., Gu, J., Huang, Y., Qin, S., Guo, Q., Liu, Y., Cao, X.: Transegpgd: Improving transferability of adversarial examples on semantic segmentation. arXiv preprint arXiv:2312.02207 (2023)
18. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in neural information processing systems **33**, 21464–21475 (2020)
19. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
20. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
21. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2574–2582 (2016)
22. Nesti, F., Rossolini, G., Nair, S., Biondi, A., Buttazzo, G.: Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2280–2289 (2022)
23. Park, E.S., Park, M., Park, S., Shin, Y.G.: Fspgd: Rethinking black-box attacks on semantic segmentation. arXiv preprint arXiv:2502.01262 (2025)
24. Rossolini, G., Nesti, F., D'Amico, G., Nair, S., Biondi, A., Buttazzo, G.: On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving. IEEE Transactions on Neural Networks and Learning Systems (2023)
25. Wong, E., Schmidt, F., Kolter, Z.: Wasserstein adversarial examples via projected sinkhorn iterations. In: International conference on machine learning. pp. 6808–6817. PMLR (2019)

26. Xiao, C., Deng, R., Li, B., Yu, F., Liu, M., Song, D.: Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 217–234 (2018)
27. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 1369–1378 (2017)
28. Xu, X., Zhao, H., Jia, J.: Dynamic divide-and-conquer adversarial training for robust semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7486–7495 (2021)
29. Ye, S., Xu, K., Liu, S., Cheng, H., Lambrechts, J.H., Zhang, H., Zhou, A., Ma, K., Wang, Y., Lin, X.: Adversarial robustness vs. model compression, or both? In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 111–120 (2019)
30. Zhang, Y., Shi, Z., Wang, S., Yang, W., Wang, S., Xue, Y.: Rp-pgd: Boosting segmentation robustness with a region-and-prototype based adversarial attack. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10338–10347 (2025)
31. Zhang, Y., Shi, Z., Wang, S., Yu, Z., Wang, S., Yang, W., et al.: Leaving no ood instance behind: Instance-level ood fine-tuning for anomaly segmentation. Advances in Neural Information Processing Systems **38**, 120578–120611 (2026)
32. Zhang, Y., Shi, Z., Yang, W., Wang, S., Wang, S., Xue, Y.: Genseg: on generating unified adversary for segmentation. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. pp. 1733–1742 (2024)
33. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
34. Zhou, T., Wang, W., Konukoglu, E., Van Gool, L.: Rethinking semantic segmentation: A prototype view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2582–2593 (2022)
35. Zhu, M., Gupta, S.: To prune, or not to prune: exploring the efficacy of pruning for model compression. arXiv preprint arXiv:1710.01878 (2017)