RP-PGD: Boosting Segmentation Robustness with a Region-and-Prototype Based Adversarial Attack

Yuxuan Zhang[1], **Zhenbo Shi**[1, 2, 5*], **Shuchang Wang**[1], **Wei Yang**[1, 2, 3*], **Shaowei Wang**[4], **Yinxing Xue**[1]

[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]Institute of Artificial Intelligence and Blockchain, Guangzhou University
[5]Laboratory for Advanced Computing and Intelligence Engineering, Wuxi, China

Abstract

Adversarial attack and defense have been extensively explored in classification tasks, but their study in semantic segmentation remains limited. Moreover, current attacks fail to act as strong underlying attacks for adversarial training (AT), making it difficult to achieve segmentation robustness against strong attacks. In this paper, we present **RP-PGD**, a novel **R**egion-and-**P**rototype based **P**rojected **G**radient **D**escent attack tailored to fool segmentation models. In particular, we propose a region-based attack, which leverages a spatial-temporal way to separate the pixels into three disjoint regions, and highlights the attack on the crucial True Region and Boundary Region. Moreover, we introduce a prototype-based attack to disrupt the feature space, further enhancing the attack capability. To boost the robustness of segmentation models, we inject adversaries generated by RP-PGD into the clean data and perform AT. Extensive experiments on multiple datasets showcase that RP-PGD generates adversaries with faster convergence and stronger attack effectiveness, surpassing state-of-the-art attacks by a large margin. Consequently, RP-PGD serves as a strong underlying attack for segmentation models to perform AT, assisting them in defending against a variety of strong attacks without incurring additional computational costs during inference.

Introduction

Deep neural networks are susceptible to quasi-imperceptible adversarial perturbations. To address this issue, great efforts (Szegedy et al. 2013; Zhang and Wang 2019; Liao et al. 2018; Tramer et al. 2020; Zhou et al. 2020) have been made to defend against adversarial attacks on classification tasks. As an extension task, semantic segmentation has broad applications in various fields. However, the study of adversarial attack and defense in semantic segmentation remains limited (Hendrik Metzen et al. 2017; Yang et al. 2020), resulting in ample room for enhancing segmentation robustness.

Previous methods (Arnab, Miksik, and Torr 2018; Hendrik Metzen et al. 2017; Poursaeed et al. 2018) primarily focus on transferring attack patterns from classification and evaluating the segmentation performance. However, unlike the classification task, merely maximizing the pixel-wise cross-entropy loss to fool the model is inappropriate for segmentation, as the gradient is dominated by the already-misclassified pixels. Attacking these pixels does not work well, since the attack target is to mislead as many pixels as possible to produce incorrect predictions, considering the inherent property of the dense-prediction task.

To defend against adversarial attacks, DDC-AT (Xu, Zhao, and Jia 2021) adopts adversarial training (AT) to enhance segmentation robustness. Subsequently, SegPGD (Gu et al. 2022) demonstrates that enhancing the attack capability can further boost the segmentation robustness with AT. However, these methods are still vulnerable to strong attacks like Projected Gradient Descent (PGD) (Madry et al. 2017) with 100 iterations. This vulnerability arises from their failure to fully consider the inherent properties of generating strong underlying adversaries tailored for segmentation. Consequently, AT with weak attacks gives a false sense of segmentation robustness.

To address this issue, on the one hand, we propose a novel **region**-based attack. Specifically, we employ a spatial-temporal way to classify the pixels into three disjoint regions: True Region, False Region, and Boundary Region. The True Region consists of pixels that consistently maintain accurate predictions during previous attack iterations. We focus on attacking the pixels in this region to increase the number of misclassified pixels, which can greatly enhance the attack capability. The False Region comprises pixels that are initially mispredicted and are difficult to transform to correct predictions even with successive iterations. Recognizing the limited potential for improvement, we allocate less attention to this region during the attack process. Finally, the Boundary Region includes pixels near the decision boundary, which initially produce correct predictions but can be misled afterwards. We consider this region as crucial because the pixels may revert to their benign state if they are not sufficiently targeted during the attack. Thus, we place significant emphasis on preventing the pixels in this region from becoming benign throughout the attack iterations.

On the other hand, we introduce a **prototype**-based approach to enhance the attack efficacy from the feature space. Motivated by (Zhou et al. 2022; Zhang, Yang, and Wang 2023), we leverage a prototype to represent the distinguishable feature of each category. Note that, the inter-class proto-

*Corresponding authors, E-mail: {zbshi,qubit}@ustc.edu.cn

types are separated from each other. Hence, we aim to maximize the similarity between different prototypes to make the feature space ambiguous, hindering the classifier from identifying the features of different categories. This prototype-based attack is only applied to the True Region, as this region remains clean information.

Finally, we combine the above two attacks and formulate a region-and-prototype based hybrid attack, termed RP-PGD, as shown in Figure 1. Such an attack exhibits superior performance in fooling various segmentation models. By injecting adversaries into clean data, the segmentation models achieve remarkable robustness with AT. We apply the multi-step attack paradigm PGD (Madry et al. 2017) to our method for crafting adversaries, and conduct evaluations on multiple datasets. Extensive experiments show that RP-PGD is effective and efficient in attacking segmentation models. AT with RP-PGD enhances segmentation robustness against various attacks without adding extra computation during inference, significantly outperforming state-of-the-art approaches.

In summary, our contributions are as follows:

- We formulate a stronger adversarial attack for semantic segmentation, termed RP-PGD. When segmentation models are adversarially trained with RP-PGD, their robustness will be substantially boosted.
- We introduce two attack objectives, i.e., region-based attack and prototype-based attack. The former uses a region-divided strategy, which focuses on fooling True Region and Boundary Region. The latter leverages a prototype view to perturb the feature space.
- Extensive experiments show that RP-PGD is an effective and efficient method for enhancing segmentation robustness, achieving top-performing results on various segmentation models and datasets against diverse attacks.

Related Work

Semantic Segmentation

Semantic segmentation is a fundamental task in computer vision area. Its target is to perform a pixel-wise classification into pre-defined semantic categories. The mainstream methods are based on the fully convolutional network (FCN) (Long, Shelhamer, and Darrell 2015). Such a network provides a paradigm for modern segmentation approaches (Gao 2023; Zhou et al. 2019; Ji et al. 2020; Strudel et al. 2021; Ge, Fu, and Zha 2022; Ge et al. 2024; Zhang and Yang 2022). However, these models are prone to be fooled by adversarial attacks. In this paper, we leverage the classic PSPNet (Zhao et al. 2017) and DeepLabv3 (Chen et al. 2017) models to assess the segmentation robustness.

Adversarial Attack and Defense

Deep learning models are vulnerable to imperceptible adversarial perturbations, and numerous works (Duan et al. 2021; Wang et al. 2021; Sriramanan et al. 2020) have explored deceiving the networks, which mainly focus on the classification task. Most approaches aim to maximize the classification loss (Goodfellow, Shlens, and Szegedy 2014) so that a classifier will make mispredictions. To resist these attacks, diverse strategies (Goldblum et al. 2020; Athalye, Carlini, and Wagner 2018) are proposed to boost the model robustness. Among them, AT (Madry et al. 2017) has emerged as the dominant strategy for defending against adversaries, formulating the defending process as a min-max problem. However, the process is time-consuming and the effectiveness depends highly on the quality of the generated adversaries, as weak attacks will give a false sense of the model robustness. In this paper, we focus on developing potent adversaries tailored for semantic segmentation, and explore whether the attack can be employed to perform AT better.

Adversarial Defense in Segmentation

Previous works (Arnab, Miksik, and Torr 2018; Hendrik Metzen et al. 2017; Rossolini et al. 2023; Rony, Pesquet, and Ben Ayed 2023; Agnihotri and Keuper 2024; Zhang et al. 2024) find that semantic segmentation models are also vulnerable to imperceptible perturbations. To mitigate this challenge, (Xiao et al. 2018; Tran et al. 2021) introduce perturbation detection methods to resist adversaries. Moreover, pioneered by DDC-AT (Xu, Zhao, and Jia 2021), AT on segmentation model has become a prevailing strategy to defend against adversarial attacks. Subsequently, SegPGD (Gu et al. 2022) proposes a novel attack perspective to boost AT and achieve top-performing results against various attacks. However, the pixel-divided strategy of SegPGD is somewhat rough, resulting in a weak attack effectiveness. Different from previous attacks, we introduce a novel region-and-prototype based hybrid attack for semantic segmentation, which performs a stronger attack capability and further boosts segmentation robustness with AT.

Methodology

AT is an effective way to resist adversarial attacks, which is formulated as a min-max problem:

$$\min_{\theta} \max_{\delta \in S_\epsilon} \mathcal{L}(\theta, x+\delta, y) \tag{1}$$

where the inner maximum problem refers to the process of generating adversaries. Specifically, given the input data x and its ground-truth y, we need to find a perturbation δ that can lead the network to obtain a false prediction $\hat{y}$ ($\hat{y} \neq y$) to maximize the loss function $\mathcal{L}(\cdot)$. S_ϵ refers to an ϵ-ball and ϵ is the maximum allowed perturbation with the commonly used l_∞-norm. Besides, the outer minimum problem stands for the AT process, which injects the obtained adversaries into the clean data for training. The target is to obtain the parameter θ to minimize the total loss, which enhances the capability to defend against adversarial attacks.

As demonstrated by (Madry et al. 2017), AT with stronger adversaries can significantly enhance the model robustness, while weak attack methods provide a misleading sense of robustness. In semantic segmentation, previous gradient-based attack methods mainly maximize the pixel-wise cross-entropy loss to make the predictions deviate from the ground-truth, which is calculated as:

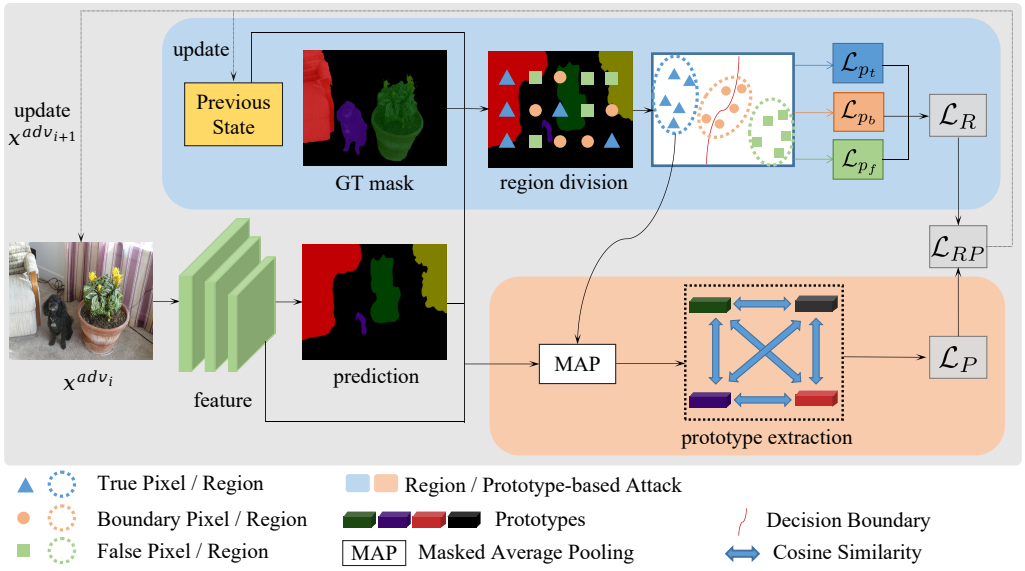

Figure 1: Overall pipeline of RP-PGD. In each attack iteration, RP-PGD calculates the region-based and prototype-based losses to perform a hybrid attack, which fools segmentation models in both output prediction and feature space.

$$\mathcal{L}_{CE} = \frac{1}{HW} \sum_{j=1}^{HW} CE(f(x_j), y_j) \tag{2}$$

where $x \in \mathbb{R}^{H \times W \times 3}$ and $y \in \mathbb{R}^{H \times W \times C}$ separately stand for the input RGB image and the ground-truth mask with C categories, H and W are the height and the width, $f(\cdot)$ denotes the segmentation function, and j is the pixel index.

Our goal is to maximize the ratio of mispredicted pixels, hence perturbing the already-misclassified pixels is redundant. Moreover, the cross-entropy loss is dominated by the misclassified pixels, with their gradients significantly outweighing those of truly-predicted pixels. This situation hinders the increase of the misprediction rate and results in a weak attack capability. To tackle this problem, we introduce our region-based approach, catering to the multi-step attacks for semantic segmentation.

Region-Based Attack

During the attack iteration, we partition the image into three disjoint regions: True Region, False Region, and Boundary Region. In the first attack iteration, pixels that are correctly predicted in the clean image x^{clean} are assigned to the True Region, while mispredicted pixels are assigned to the False Region. As discussed earlier, we believe that pixels initially misclassified are challenging to convert to a correctly predicted state, even after multiple attack iterations. To validate our thought, we conduct experiments using PSPNet (Zhao et al. 2017) on Pascal VOC (Everingham et al. 2010) dataset under PGD attack (Madry et al. 2017) with the typical cross-entropy loss. The results reveal that during 100 attack iterations, less than 1% of the initially mispredicted pixels could be randomly restored to the truly predicted state.

On the other hand, during the subsequent attack iterations, part of the pixels in True Region will be misled to obtain false predictions by the accumulated perturbations. These pixels have the boundary attribute, and thus we place them in the Boundary Region. Such pixels are susceptible to returning to their accurately predicted state if they are not adequately attended to. Hence, to increase the quantity of mispredicted pixels, we need to consistently focus on deceiving the pixels in True Region and prevent the pixels in Boundary Region from becoming benign again.

Accordingly, we give formal definitions of the above three regions in a *spatial-temporal* way (We leverage a multi-step attack with N iterations, and the adversarial sample for the i-th attack iteration is denoted as x^{adv_i}):

Definition 1. *True Region $\mathcal{T}$: Given an attack iteration i ($i \in \{1, 2, ..., N\}$) and a pixel p_j ($j \in \{1, 2, ..., HW\}$), the predictions of p_j in $\{x^{\text{clean}}, x^{\text{adv}_1}, x^{\text{adv}_2}, ..., x^{\text{adv}_i}\}$ are always correct ($\hat{y}_{p_j}^{\text{clean}} = \hat{y}_{p_j}^{\text{adv}_1} = \hat{y}_{p_j}^{\text{adv}_2} = ... = \hat{y}_{p_j}^{\text{adv}_i} = y_{p_j}$). Then pixel $p_j \in \mathcal{T}$ in iterations $\{1, 2, ..., i\}$.*

Definition 2. *Boundary Region $\mathcal{B}$: Given an attack iteration i ($i \in \{1, 2, ..., N\}$) and a pixel p_j ($j \in \{1, 2, ..., HW\}$), it occurs that $\hat{y}_p^{\text{clean}} = \hat{y}_{p_j}^{\text{adv}_1} = \hat{y}_{p_j}^{\text{adv}_2} = ... = \hat{y}_{p_j}^{\text{adv}_{i-1}} = y_{p_j}$, but $\hat{y}_{p_j}^{\text{adv}_i} \neq y_{p_j}$. Then $p_j \in \mathcal{B}$ in iterations $\{i, i+1, ..., N\}$.*

Definition 3. *False Region $\mathcal{F}$: Consider a pixel $p_j (j \in \{1, 2, ..., HW\})$, if it is misclassified in the initial clean image x^{clean} ($\hat{y}_{p_j}^{\text{clean}} \neq y_{p_j}$), then $p_j \in \mathcal{F}$ throughout all attack iterations $\{1, 2, ..., N\}$.*

In short, $\mathcal{F}$ contains the pixels that are initially mispredicted, $\mathcal{T}$ includes the pixels that are consistently correctly predicted in previous iterations, and $\mathcal{B}$ consists of the pixels that are initially correct but later mispredicted. According to the definition of $\mathcal{B}$, once a pixel $p \in \mathcal{T}$ is mispredicted in

the i-th iteration, it will consistently belong to $\mathcal{B}$ in the rest of the iterations, even if it may restore to the benign state afterwards.

Formally, we have the following theorems:

Theorem 1. *In each attack iteration,* $\mathcal{T} \cap \mathcal{B} = \mathcal{T} \cap \mathcal{F} = \mathcal{B} \cap \mathcal{F} = \phi$.

Proof. A pixel can transition through multiple states during the attack iterations. Specifically, it may start off belonging to $\mathcal{T}$ initially, and subsequently transfers to belonging to $\mathcal{B}$. Given an attack iteration i ($i \in \{1, 2, ..., N\}$), the pixel $p_t \in \mathcal{T}$ is always correctly predicted during the iterations $\{1, 2, ..., i\}$ according to the definition of $\mathcal{T}$. However, as for the pixel $p_b \in \mathcal{B}$, we assume that it first becomes mispredicted in the j-th iteration, where $j \in \{1, 2, ..., i\}$. According to the definition of $\mathcal{B}$, this pixel belongs to $\mathcal{T}$ in the iterations $\{1, 2, ..., j-1\}$. Even though p_b may become benign in the iterations $\{j+1, j+2, ..., i\}$, it still belongs to the Boundary Region $\mathcal{B}$. Hence, in each specific iteration, $\mathcal{T} \cap \mathcal{B} = \phi$. On the other hand, pixel $p_f \in \mathcal{F}$ is mispredicted in the clean image x^{clean}, while p_t and p_b are both correctly predicted in x^{clean}, and thus $\mathcal{T} \cap \mathcal{F} = \mathcal{B} \cap \mathcal{F} = \phi$. □

Theorem 2. *In each attack iteration,* $|\mathcal{T} \cup \mathcal{B} \cup \mathcal{F}| = H \times W$.

Proof. Theorem 2 is equal to the statement that in each iteration i ($i \in \{1, 2, ..., N\}$), the pixels in the three regions cover the whole image. Assuming that $\exists p_e \in x$ (x represents the pixel set of the image), subject to $p_e \notin (\mathcal{T} \cup \mathcal{B} \cup \mathcal{F})$. We obtain that $p_e \notin \mathcal{T}$, thus p_e is mispredicted in the clean image x^{clean} or in an intermediate attack iteration j, where $j \in \{1, 2, ..., i\}$. Assuming that $p_e \notin \mathcal{B}$, p_e must be mispredicted in the clean image x^{clean}. Hence we draw that $p_e \in \mathcal{F}$, which contradicts the condition of $p_e \notin (\mathcal{T} \cup \mathcal{B} \cup \mathcal{F})$. Therefore, $\forall p_e \in x$, $p_e \in (\mathcal{T} \cup \mathcal{B} \cup \mathcal{F})$, which is equal to $x \subseteq (\mathcal{T} \cup \mathcal{B} \cup \mathcal{F})$. On the other hand, $\mathcal{T}$, $\mathcal{B}$, and $\mathcal{F}$ are derived from x, and thus $(\mathcal{T} \cup \mathcal{B} \cup \mathcal{F}) \subseteq x$. Therefore, we conclude that $|\mathcal{T} \cup \mathcal{B} \cup \mathcal{F}| = |x| = H \times W$. □

These theorems provide the following guarantees: 1) Ensuring non-intersecting regions prevents redundant pixel definition and attack; 2) Covering the entire image makes certain that all pixels can be attacked. These form the basis of our spatial-temporal region-based attack.

Accordingly, we formulate the region-based objective as:

$$\mathcal{L}_R = \frac{1}{HW}(\lambda_1 \sum_{p_t \in \mathcal{T}} \mathcal{L}_{p_t} + \lambda_2 \sum_{p_b \in \mathcal{B}} \mathcal{L}_{p_b} + \lambda_3 \sum_{p_f \in \mathcal{F}} \mathcal{L}_{p_f}) \tag{3}$$

where λ_1, λ_2 and λ_3 are the weights, and $\mathcal{L}_p$ represents the cross-entropy loss of the pixel p, , with subscripts t, b and f corresponding to True Region, Boundary Region and False Region, respectively. Based on the analysis above, we assign the weights in a linear way as follows:

$$\lambda_1 = \frac{2N - i}{2N}, \quad \lambda_2 = \frac{i - 1}{2N}, \quad \lambda_3 = \frac{1}{2N} \tag{4}$$

where i is the current step among a total of N attack steps.

We concentrate more on $\mathcal{T}$ at the beginning. With the progress of attack iterations, the number of pixels in $\mathcal{B}$ significantly increases, consequently gaining more attention. As to $\mathcal{F}$, we consistently assign it a fixed and small weight, as focusing on this region is unable to fool more pixels.

We will discuss more weight schedules in the experiment, and Eq. (4) is the best choice. Different from SegPGD, we distinguish $\mathcal{B}$ from $\mathcal{F}$ and assign different weights. Such a region-divided strategy focuses more on $\mathcal{B}$ and effectively keeps the pixels in $\mathcal{B}$ being misled.

Prototype-Based Attack

Inspired by the prototype view in segmentation (Zhou et al. 2022; Zhang, Yang, and Hu 2023), we further propose a prototype-based attack strategy, which enhances the attack capability in deceiving existing segmentation models.

Each specific category exhibits a high intra-class pixel-wise similarity in the feature space, allowing us to utilize a prototype to represent each class. Typically, the prototype is computed using the masked average pooling operation (Wang et al. 2019), considering all pixels available. However, in adversarial attacks, the perturbations also contaminate the latent space, causing the features extracted from $\mathcal{F}$ and $\mathcal{B}$ to be impure. Consequently, we only focus on the True Region $\mathcal{T}$ for pure feature extraction, and the prototype is calculated as follows:

$$P_c = \frac{\sum_{p_t \in \mathcal{T}} M_{p_t} \mathbb{1}[\hat{y_{p_t}} = c]}{\sum_{p_t \in \mathcal{T}} \mathbb{1}[\hat{y_{p_t}} = c]} \tag{5}$$

where $P_c \in \mathbb{R}^{1 \times 1 \times d}$ denotes the prototype of the category c, $M \in \mathbb{R}^{H \times W \times d}$ is the feature map, $\mathbb{1}$ stands for the indicator function, and d represents the dimension of the latent space.

In this way, assuming that there are N_P different classes $C_{\mathcal{T}} = \{c_i | i = 1, 2, ..., N_p\}$ in $\mathcal{T}$, we can obtain N_P prototypes $P = \{P_{c_i} | i = 1, 2, ..., N_p\}$ for the corresponding categories. As described in (Okazawa 2022), reducing the inter-class prototype similarity is beneficial to enhance the separation performance. Consequently, we can attack the segmentation models by obfuscating the inter-class prototype relationships. Once the feature space is totally perturbed, the final classifier will be difficult to conduct correct predictions.

Hence, the prototype-based objective is formulated as:

$$\mathcal{L}_P = \frac{1}{N_{Comb}} \sum_{P_{c_i} \in P} \sum_{P_{c_j} \in P} cos(P_{c_i}, P_{c_j}) \mathbb{1}[c_i \neq c_j] \tag{6}$$

where $N_{Comb} = \binom{N_P}{2}$ is the number of pairwise combinations of different prototypes, cos means cosine similarity, and $\mathbb{1}$ is an indicator function. Maximizing the prototype loss $\mathcal{L}_P$ contributes to obfuscating the latent space of the True Region, which enhances the attacking capability by making the inter-class representations ambiguous.

Overall RP-PGD

Combining the above two attack strategies, we formulate the hybrid attack objective of RP-PGD as follows:

Algorithm 1: RP-PGD: A Region-and-Prototype based multi-step attack for semantic segmentation models

Input: Clean image x^{clean}, label y, segmentation model $f(\cdot)$, attack iteration N, parameters β, step size α, maximum allowed perturbation ϵ
Output: Adversary x^{adv_N}

$x^{\text{adv}_0} = x^{\text{clean}} + \mathcal{U}(-\epsilon, +\epsilon)$
Initialize the previous state matrix S
for $i \leftarrow 1$ to N **do**
  $\hat{y} = f(x^{\text{adv}_{i-1}})$ // obtain prediction
  $\mathcal{T}, \mathcal{B}, \mathcal{F} \leftarrow \hat{y}, y, S$ // region division
  Using Eqs. (3) and (4) to calculate $\mathcal{L}_R$
  Using Eq. (5) to obtain prototypes $\{P_{c_i}|_{i=1}^{N_p}\}$ from $\mathcal{T}$
  Using Eq. (6) to calculate $\mathcal{L}_P$
  Using Eq. (7) to calculate $\mathcal{L}_{RP}$
  Update x^{adv_i} using Eq. (9) and update S
end for
return x^{adv_N}

$$\mathcal{L}_{RP} = \beta \mathcal{L}_R + (1-\beta)\mathcal{L}_P \tag{7}$$

where β is a balanced factor to adjust $\mathcal{L}_R$ and $\mathcal{L}_P$, and we set $\beta = 0.75$ according to the experimental results.

Multi-step Attack. We combine the attack objective with the multi-step attack paradigm PGD (Madry et al. 2017), and propose our RP-PGD for semantic segmentation models. We use an iterative manner to generate powerful yet imperceptible adversarial examples, which can be further employed to enhance segmentation robustness. This process can be recognized as a gradient-ascent step, which is formulated as:

$$x^{\text{adv}_0} = x^{\text{clean}} + \mathcal{U}(-\epsilon, +\epsilon) \tag{8}$$

$$x^{\text{adv}_{i+1}} = \phi_\epsilon(x^{\text{adv}_i} + \alpha \cdot \text{sign}(\bigtriangledown_{x^{\text{adv}_i}} \mathcal{L}_{RP}(x^{\text{adv}_i}, y))) \tag{9}$$

where x^{adv_i} denotes the adversary in the i-th iteration, $\mathcal{U}$ represents a uniform distribution, ϕ_ϵ stands for restricting the perturbation in the ϵ-ball with l_∞-norm, and α is the step size. The reason for incorporating a uniform distribution is to ensure that the adversarial noise is distributed within the sphere instead of being concentrated on the surface, which enhances the attack strength. The detailed process of RP-PGD is demonstrated in Algorithm 1.

Adversarial Training with RP-PGD

AT is an effective approach for training a robust model, which can effectively resist adversarial samples. However, the training process is time-consuming compared with training a non-robust model, as we need to first generate the adversarial examples in an iterative manner for training.

To handle this issue, we improve the training process by employing RP-PGD to generate adversaries. RP-PGD has a faster convergence than other gradient-based attacks (the analysis will be discussed in detail below), and thus not only the adversarial examples are more effective, but also the generating process is more efficient. By injecting the adversarial samples into the clean data, we train a segmentation model with great robustness and acceptable time cost. The detailed AT process is demonstrated in the appendix.

Experiment

Dataset

To evaluate the performance of RP-PGD, we conduct experiments on three widely-used segmentation datasets: Pascal VOC (Everingham et al. 2010), Cityscapes (Cordts et al. 2016), and ADE20K (Zhou et al. 2017). Pascal VOC consists of 20 foreground categories, with 1464, 1499, and 1456 images allocated for training, validation, and testing, respectively. Following (Hariharan et al. 2015), we augment the training set to include 10,582 images. Cityscapes comprises urban scene understanding images with 19 different categories. It contains 2975, 500, and 1525 pixel-wise annotated images for training, validation, and testing, respectively. Besides Pascal VOC and Cityscapes, we also employ the commonly used segmentation dataset ADE20K to evaluate the attack effectiveness of RP-PGD. As to the dataset description and corresponding experimental results of ADE20K, please refer to Appendix C for more details.

Implementation Details

We employ the commonly used segmentation models PSPNet (Zhao et al. 2017) and DeepLabv3 (Chen et al. 2017) to evaluate our RP-PGD, and leverage ResNet-50 (He et al. 2016) as the backbone of both segmentation models. As to the metrics, we apply the widely used MIoU to evaluate the attack performance and the robustness of segmentation models, and we leverage MisRatio (Gu et al. 2022) to analyze the convergence of RP-PGD. MisRatio reflects the percentage of the mispredicted pixels, and has been proven to be more suitable for dense-prediction tasks. Moreover, we use the top-performing attack methods PGD (Madry et al. 2017), SegPGD (Gu et al. 2022), CosPGD (Agnihotri and Keuper 2024), and Proximal (Rony, Pesquet, and Ben Ayed 2023) as our baselines. We use PGDn (where n can be an arbitrary natural number) to represent PGD attack with n iterations, and PGDn-AT stands for AT with the adversaries produced by PGD. For example, CosPGD7-AT means that we adopt the adversaries generated by CosPGD with 7 iterations to conduct AT. For convenience, we use PSPNet-AT and DeepLabv3-AT to represent PGD3-AT PSPNet and PGD3-AT DeepLabv3, respectively. The attack is based on l_∞-norm perturbation, with the maximum value $\epsilon = 0.03$. The step size α is set to 0.01. In addition, to evaluate the robustness of the models trained with RP-PGD, we leverage multiple attack methods, i.e., Deepfool (DF) (Moosavi-Dezfooli, Fawzi, and Frossard 2016), CW (Carlini and Wagner 2017) and l_2-norm BIM (Kurakin, Goodfellow, and Bengio 2018). For more details, please refer to Appendix I.

Convergence Analysis

We conduct experiments on Pascal VOC and Cityscapes to study the convergence of RP-PGD, compared with PGD, SegPGD, CosPGD, and Proximal. As shown in Appendix B,

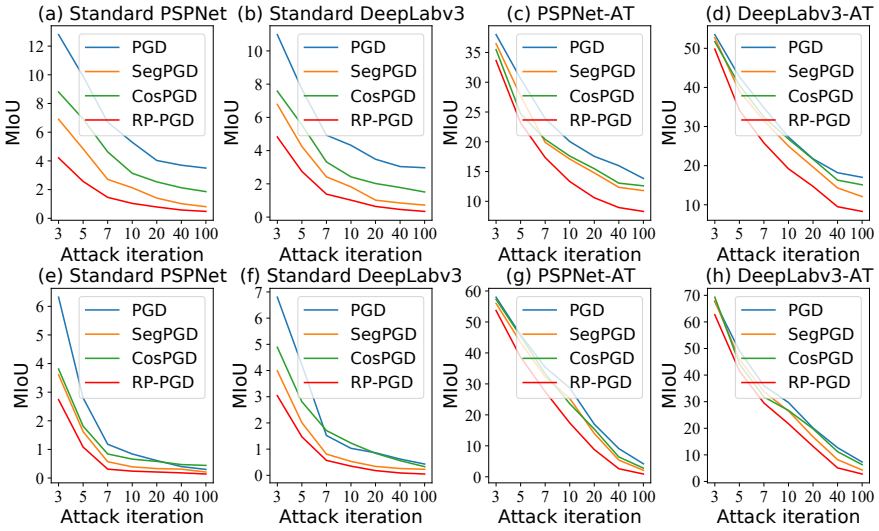

Figure 2: Attack effectiveness of RP-PGD compared with current top-performing attacks. Figures (a)-(d) show the MIoU results on Pascal VOC, and the results on Cityscapes are shown in figures (e)-(h). RP-PGD performs a stronger attack capability than other methods.

RP-PGD achieves a higher MisRatio with fewer attack iterations on both datasets. It converges faster than PGD in the first few steps by focusing on the True Region $\mathcal{T}$. Specifically, the MisRatio of RP-PGD under the $2nd$ iteration is nearly equal to that of PGD under the $10th$ iteration. RP-PGD outperforms the competitors in fooling segmentation models, with a significantly higher MisRatio in the last few iterations. The combination of reasonable region division and prototype-based strategy contributes to RP-PGD's superiority, making it an efficient and effective attack method. As a result, RP-PGD can generate stronger adversaries with fewer iterations, which significantly benefits the AT process. Note that, while Proximal exhibits the capability to generate potent attacks with smaller maximum perturbation, it suffers from a slower convergence, typically necessitating around 500 iterations. Additionally, its computational cost per iteration is high, diminishing its overall efficiency. Consequently, Proximal is deemed unsuitable for the application in AT.

Attack Effectiveness

To evaluate the attack efficacy, we conduct comprehensive comparisons between PGD, SegPGD, CosPGD, and RP-PGD. The experiments are performed on both standard models and PGD3-AT models. For fair comparisons, we report MIoU at iterations 3, 5, 7, 10, 20, 40, and 100, respectively. Figure 2 depicts the results, showing that RP-PGD not only has a faster convergence, but also exhibits a stronger attack capability compared to state-of-the-art attacks. The reasons may lie in two key aspects: 1) Focusing on the True Region and Boundary Region enables a more effective and efficient attack; 2) The hybrid attack strategy of RP-PGD deceives the model from multiple perspectives, further enhancing the overall attack effectiveness.

Moreover, we evaluate the black-box attack transferability to other models using a specific surrogate model. The settings and results are presented in Appendix E, which demonstrate that the transferability of RP-PGD surpasses other attacks. This superiority is attributed to the intermediate-level attack, which enhances the transferability by relying less on the specific parameters of the surrogate models.

Ablation Study

To fully analyze RP-PGD, we explore the impact of the region-based and prototype-based attacks. We also discuss the weight schedules for different regions, prototype extraction strategies, and the balanced factor of the hybrid attacks.

Attack Combinations. We conduct six groups of experiments, each comprising 100 attack iterations, to investigate the impact of different region combinations and the prototype factor on the attack capability. The results are presented in Table 1. Notably, the region-divided approach ($\mathcal{T}$, $\mathcal{B}$, and $\mathcal{F}$) coupled with the prototype attack strategy (group 6) achieves the highest attack performance.

Group 1 is similar to the previous work DAG (Xie et al. 2017), which only takes the True Region into account and can be viewed as a special case of RP-PGD. However, when considering the Boundary Region, Groups 2 and 4 exhibit stronger attack capabilities compared with Groups 1 and 3.

Moreover, the utilization of the prototype-based attack in Group 6 further enhances the attack capability by obfuscating the feature space, making it harder for the classifier to give correct predictions. Consequently, dividing the regions into $\mathcal{T}$, $\mathcal{B}$, and $\mathcal{F}$ and integrating the prototype-based attack strategy yields the most potent attack effectiveness.

Weights Schedules of Different Regions. In pursuit of the optimal weights assignment for the region-based attack, we explore both fixed and dynamic schedules, and conduct five groups of experiments. The results are shown in Table 2.

From Groups 1 and 2, it is evident that $\mathcal{B}$ carries greater significance than $\mathcal{F}$. Despite having similar targets (prevent-

Region $\mathcal{T}$	Region $\mathcal{B}$	Region $\mathcal{F}$	Proto	PSPNet	PSPNet-AT	DeepLabv3	DeepLabv3-AT
✓				2.65	13.37	3.26	13.59
✓	✓			0.65	10.06	0.54	10.47
✓		✓		2.97	12.84	3.02	13.38
✓	✓	✓		0.61	9.83	0.51	10.35
			✓	8.74	26.17	7.57	25.09
✓	✓	✓	✓	**0.48**	**7.69**	**0.34**	**8.30**

Table 1: MIoU results of different combinations of region selections and prototype-based attack (Proto) on Pascal VOC. Our region division strategy combined with prototype attack achieves the best attack.

λ_1	λ_2	λ_3	PSPNet	PSPNet-AT	DeepLabv3	DeepLabv3-AT
0.4	0.3	0.3	3.60	14.31	3.47	13.92
0.4	0.4	0.2	1.64	12.38	1.49	12.09
0.6	0.2	0.2	1.17	10.96	1.04	10.80
0.2	0.6	0.2	6.84	20.27	6.71	19.54
Eq. (4)			**0.61**	**9.83**	**0.51**	**10.35**

Table 2: MIoU results of weight schedules for different regions on Pascal VOC. Dynamic weight has a stronger attack.

ing pixels from reversing to the truly-predicted states), the boundary pixels are more susceptible to becoming benign again. Hence, a larger weight is assigned to $\mathcal{B}$ to prioritize its protection, so as to maintain these vulnerable mispredicted pixels. Especially, in SegPGD, there always exists $\lambda_2 = \lambda_3$, which can be regarded as a special case of RP-PGD.

For Groups 2 to 4, none of the performances surpass the dynamic schedule in Group 5. In the early attack iterations, the focus should be on attacking pixels in $\mathcal{T}$ to maximize the MisRatio. However, as the attack iteration progresses, the number of pixels in $\mathcal{B}$ increases significantly, necessitating an increased weight, while the main focus is still on $\mathcal{T}$. The dynamic setting in Eq. (4) is verified to be the most effective, outperforming other fixed settings. Moreover, we explore several different dynamic schedules to seek the optimal choice, and the details are available in Appendix D.

Prototype Extraction Strategy. To determine which region contributes to establishing reasonable prototypes in the prototype-based attack, we hypothesize that extracting features from $\mathcal{B}$ and $\mathcal{F}$ might contaminate the prototypes. To validate this hypothesis, we conduct several ablation experiments focusing on different regions, and the results are presented in Table 3. As expected, the results confirm our hypothesis, showing that solely extracting features from $\mathcal{T}$ yields the best attack performance. This finding suggests that the information from $\mathcal{T}$ plays a crucial role in generating effective and meaningful prototypes for the attack, while features from $\mathcal{B}$ and $\mathcal{F}$ might introduce noise or irrelevant information, reducing the quality of the prototypes. Therefore, by concentrating on $\mathcal{T}$, RP-PGD better obfuscates the feature space and maximizes its attack effectiveness. It is worth mentioning that this region selection is solely for prototype establishment, not for region-based attacks.

Balance of Hybrid Attacks. We explore the optimal value of the balanced factor β. The detailed results, available in Appendix D, indicate that $\beta = 0.75$ yields the best performance. On the one hand, the region-based attack offers

$\mathcal{T}$	$\mathcal{B}$	$\mathcal{F}$	PSPNet	PSPNet-AT	DeepLabv3	DeepLabv3-AT
✓			**8.74**	**26.17**	**7.57**	**25.09**
✓	✓		10.75	28.52	9.56	26.41
✓		✓	18.40	31.71	16.68	29.62
✓	✓	✓	17.51	30.15	15.70	27.87

Table 3: MIoU results of prototype extraction strategies on Pascal VOC. Focusing on $\mathcal{T}$ achieves stronger attacks.

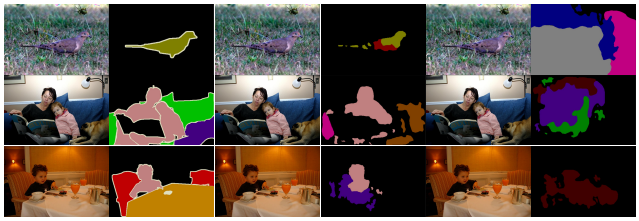

Figure 3: Visualization of the predictions under RP-PGD attack. (a)-(f) represent clean image, GT mask, adversary of SegPGD, prediction of SegPGD, adversary of RP-PGD, and prediction of RP-PGD, respectively.

a direct way to deceive the network and produce mispredictions by targeting specific regions. On the other hand, the prototype-based attack focuses on obfuscating the latent space, indirectly leading to false predictions. By combining the both strategies, the attack capability is significantly enhanced. The chosen value of $\beta = 0.75$ represents a well-balanced trade-off between these two attacks, ensuring powerful adversarial examples in RP-PGD. The attack effectiveness of RP-PGD is shown in Figure 3. As shown in the figure, RP-PGD exhibits a stronger attack capability.

Adversarial Training with RP-PGD

We explore whether the underlying attack RP-PGD can boost segmentation robustness through AT. To ensure a fair comparison, we adopt the same setting in DDC-AT (Xu, Zhao, and Jia 2021) for AT. The baseline approaches employ PGD, SegPGD, and CosPGD as their underlying attacks for AT. In contrast, we employ RP-PGD as the underlying attack method for AT to evaluate the segmentation robustness. To guarantee the reliability, each experiment is conducted 7 times, and the average results are reported for evaluation.

Against white-box attack. To comprehensively evaluate the robustness of the segmentation models against white-box attacks, we leverage several commonly used methods to attack them. The main results are presented in Table 4 and Table 5, respectively. Standard PSPNet and DeepLabv3 models are found to be vulnerable to adversarial attacks, with the MIoU reduced to near-zero, indicating their susceptibility to being easily fooled. AT with the underlying attack provides some tolerance to adversarial perturbations, yet still shows unsatisfactory performance against strong attacks like PGD100. In contrast, RP-PGD-AT demonstrates significant improvements in MIoU performance against various attack methods compared with PGD-AT, SegPGD-AT, and CosPGD-AT on different datasets and segmentation models.

AT Strategy	Pascal VOC								Cityscapes							
	Clean	CW	DF	BIM	PGD10	PGD20	PGD40	PGD100	Clean	CW	DF	BIM	PGD10	PGD20	PGD40	PGD100
Standard	76.64	4.72	14.2	15.32	5.21	4.09	3.64	3.37	73.98	5.94	12.68	12.36	0.96	0.61	0.42	0.27
DDC-AT	76.44	55.37	57.16	50.58	22.94	19.17	16.40	13.62	73.31	34.59	36.82	33.04	31.17	19.29	12.60	5.32
PGD3-AT	74.51	52.23	55.46	51.56	20.04	17.34	15.84	13.89	71.28	35.21	36.84	32.22	28.79	17.30	9.29	3.95
SegPGD3-AT	75.38	56.52	59.47	50.17	26.60	20.69	17.19	14.49	71.01	36.30	38.27	35.34	33.52	25.23	19.22	13.04
CosPGD3-AT	74.04	53.37	57.19	52.04	25.33	19.75	17.82	13.64	72.51	34.70	34.18	31.24	30.76	22.37	18.74	11.98
RP-PGD3-AT	75.17	**59.34**	**62.37**	51.19	**30.25**	**24.86**	**20.37**	**18.69**	71.58	**38.64**	**39.70**	**38.44**	**35.19**	**29.63**	**24.34**	**18.57**
PGD7-AT	74.99	42.30	45.05	47.21	21.79	19.39	17.99	16.97	69.85	27.78	28.44	27.87	26.00	24.75	23.86	22.80
SegPGD7-AT	74.45	48.79	51.44	45.15	25.73	22.05	20.61	19.23	70.21	29.59	30.68	32.55	27.13	25.56	24.29	23.13
CosPGD7-AT	75.12	47.43	49.86	47.81	26.10	24.97	18.52	19.74	69.18	29.32	30.87	32.16	26.69	26.20	23.77	20.63
RP-PGD7-AT	74.93	**55.65**	**60.32**	**48.10**	**29.15**	**25.58**	**23.91**	**23.14**	70.08	**31.47**	**33.82**	**35.94**	**30.74**	**28.80**	**27.49**	**26.04**

Table 4: MIoU results of different AT strategies on PSPNet against several white-box attacks. AT with RP-PGD presents the most robust performances, especially on strong attacks like PGD100.

AT Strategy	Pascal VOC								Cityscapes							
	Clean	CW	DF	BIM	PGD10	PGD20	PGD40	PGD100	Clean	CW	DF	BIM	PGD10	PGD20	PGD40	PGD100
Standard	77.36	5.24	13.57	14.76	4.36	3.46	3.05	2.85	73.82	8.24	14.26	13.86	1.07	0.84	0.62	0.44
DDC-AT	75.81	56.46	60.72	37.03	26.18	20.56	18.94	18.07	71.59	34.85	38.13	35.15	30.77	21.06	12.69	6.58
PGD3-AT	75.03	57.10	60.23	36.83	28.16	20.77	18.12	16.91	71.45	36.72	38.98	36.78	29.52	20.23	12.22	6.74
SegPGD3-AT	75.01	59.55	62.12	39.46	26.29	20.92	19.10	18.24	71.04	37.93	37.63	34.54	32.11	25.49	17.67	15.23
CosPGD3-AT	74.60	58.01	62.32	37.19	28.26	21.04	18.57	18.02	70.74	36.83	39.07	34.29	31.81	24.73	16.60	13.90
RP-PGD3-AT	75.20	**61.21**	**64.40**	**39.81**	**31.22**	**25.45**	**23.73**	**22.84**	71.25	**38.67**	38.39	**37.30**	**36.26**	**30.97**	**24.43**	**22.70**
PGD7-AT	73.45	48.51	48.87	43.13	26.23	21.15	20.06	19.10	69.91	28.87	29.63	30.58	25.64	24.48	22.87	21.24
SegPGD7-AT	74.46	51.42	51.47	42.91	30.95	26.68	24.32	23.09	69.93	29.73	31.30	32.35	30.43	28.78	26.73	25.31
CosPGD7-AT	72.88	50.66	52.37	42.51	28.70	24.84	23.35	20.33	69.26	28.38	30.70	32.14	31.08	28.45	25.66	22.16
RP-PGD7-AT	74.01	**57.58**	**59.60**	**45.08**	**33.52**	**32.35**	**30.56**	**28.10**	69.67	**30.33**	**34.07**	**34.55**	**36.16**	**33.20**	**31.74**	**29.87**

Table 5: MIoU results of different AT strategies on DeepLabv3 against several white-box attacks. AT with RP-PGD presents the most robust performances, especially on strong attacks like PGD100.

For instance, RP-PGD3-AT boosts the robustness against the strong adversaries PGD100, which is 4.2 and 5.05 MIoU higher than SegPGD3-AT and CosPGD3-AT on PSPNet for Pascal VOC, respectively. RP-PGD-AT also demonstrates superior robustness on Cityscapes. Note that, as AT is performed in the training period, these improvements in MIoU do not bring extra computational cost during inference.

Moreover, we also apply strong attacks, e.g., SegPGD and RP-PGD, to attack the models PGD-AT, SegPGD-AT, CosPGD-AT, and RP-PGD-AT, respectively. The experimental results are shown in Appendix E. RP-PGD-AT maintains superior performances, while other models are fragile against strong adversarial attacks.

Additionally, as shown in Tables 4, 5, and Appendix E, RP-PGD3-AT achieves a comparable or even superior performance compared to SegPGD7-AT, indicating that RP-PGD3-AT requires less than half the iterations for generating adversaries, making it a more efficient way to achieve relatively robust models. Overall, RP-PGD not only improves the segmentation robustness, but also provides an efficient option to obtain a relatively robust segmentation model.

Against black-box attack. We investigate the robustness of defending black-box attacks using a transfer-based setting, following the approach in DDC-AT (Xu, Zhao, and Jia 2021). In this experiment, we generate adversaries using PSPNet-AT (surrogate model) and use these adversaries to attack DeepLabv3 with various AT strategies. The results, displayed in Table 6, showcase that AT with RP-PGD achieves the best performance against transfer-based strong attacks on both Pascal VOC and Cityscapes. This demonstrates the effectiveness of RP-PGD in enhancing the robustness of segmentation models against black-box attacks.

AT Strategy	Pascal VOC		Cityscapes	
	SegPGD100	RP-PGD100	SegPGD100	RP-PGD100
DDC-AT	11.94	8.92	13.32	11.35
PGD3-AT	12.38	8.73	14.26	10.10
SegPGD3-AT	13.34	9.17	20.11	12.55
CosPGD3-AT	12.81	9.35	17.62	11.88
RP-PGD3-AT	**17.79**	**15.10**	**23.56**	**16.73**

Table 6: MIoU results of defending black-box attack. Adversaries are generated by PGD3-AT PSPNet under SegPGD100 and RP-PGD100 attacks, which are further applied to attack DeepLabv3 with various AT strategies.

Conclusion

In this paper, we proposed RP-PGD, a novel region-and-prototype based hybrid adversarial attack tailored for semantic segmentation. On the one hand, RP-PGD involves a spatial-temporal region-divided strategy. We focused on deceiving the True Region while continuously misleading the Boundary Region, leading to a notable rise in the proportion of mispredicted pixels. On the other hand, we leveraged a prototype-based attack to introduce ambiguity into the latent space, effectively enhancing the attack efficacy. Extensive experiments reveal that RP-PGD exhibits faster convergence and stronger attack capabilities compared with state-of-the-art methods. Most importantly, we utilized RP-PGD as the underlying attack for segmentation AT, significantly enhancing model robustness against various strong attack methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62172385), the Natural Science Foundation of Jiangsu Province (BK20241819), the Inno-

vation Program for Quantum Science and Technology (No. 2021ZD0302900), the Laboratory for Advanced Computing and Intelligence Engineering Fund, the Jiangsu Province Science Foundation for Youths (BK20240463), the Xiaomi Young Talents Program, the Fundamental Research Funds for the Central Universities, the China Postdoctoral Science Foundation (2024M753115), and the Anhui Provincial Department of Science and Technology under Grant 202103a05020009.

References

Agnihotri, S.; and Keuper, M. 2024. CosPGD: a unified white-box adversarial attack for pixel-wise prediction tasks. In *ICML*.

Arnab, A.; Miksik, O.; and Torr, P. H. 2018. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 888–897.

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, 274–283. PMLR.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (sp)*, 39–57. IEEE.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Duan, R.; Chen, Y.; Niu, D.; Yang, Y.; Qin, A. K.; and He, Y. 2021. Advdrop: Adversarial attack to dnns by dropping information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7506–7515.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Gao, R. 2023. Rethinking Dilated Convolution for Real-Time Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4674–4683.

Ge, C.; Fu, X.; He, P.; Wang, K.; Cao, C.; and Zha, Z.-J. 2024. Neuromorphic Event Signal-Driven Network for Video De-raining. In *AAAI*, volume 38, 1878–1886.

Ge, C.; Fu, X.; and Zha, Z.-J. 2022. Learning Dual Convolutional Dictionaries for Image De-raining. In *ACM MM*, 6636–6644.

Goldblum, M.; Fowl, L.; Feizi, S.; and Goldstein, T. 2020. Adversarially robust distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 3996–4003.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gu, J.; Zhao, H.; Tresp, V.; and Torr, P. H. 2022. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *European Conference on Computer Vision*, 308–325. Springer.

Hariharan, B.; Arbeláez, P.; Girshick, R.; and Malik, J. 2015. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 447–456.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrik Metzen, J.; Chaithanya Kumar, M.; Brox, T.; and Fischer, V. 2017. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, 2755–2764.

Ji, J.; Shi, R.; Li, S.; Chen, P.; and Miao, Q. 2020. Encoder-decoder with cascaded CRFs for semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5): 1926–1938.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, 99–112. Chapman and Hall/CRC.

Liao, F.; Liang, M.; Dong, Y.; Pang, T.; Hu, X.; and Zhu, J. 2018. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1778–1787.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2574–2582.

Okazawa, A. 2022. Interclass prototype relation for few-shot segmentation. In *European Conference on Computer Vision*, 362–378. Springer.

Poursaeed, O.; Katsman, I.; Gao, B.; and Belongie, S. 2018. Generative adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4422–4431.

Rony, J.; Pesquet, J.-C.; and Ben Ayed, I. 2023. Proximal Splitting Adversarial Attack for Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20524–20533.

Rossolini, G.; Nesti, F.; D'Amico, G.; Nair, S.; Biondi, A.; and Buttazzo, G. 2023. On the real-world adversarial robustness of real-time semantic segmentation models for au-

tonomous driving. *IEEE Transactions on Neural Networks and Learning Systems*.

Sriramanan, G.; Addepalli, S.; Baburaj, A.; et al. 2020. Guided adversarial attack for evaluating and enhancing adversarial defenses. *Advances in Neural Information Processing Systems*, 33: 20297–20308.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7262–7272.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tramer, F.; Carlini, N.; Brendel, W.; and Madry, A. 2020. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33: 1633–1645.

Tran, H.-D.; Pal, N.; Musau, P.; Lopez, D. M.; Hamilton, N.; Yang, X.; Bak, S.; and Johnson, T. T. 2021. Robustness verification of semantic segmentation neural networks using relaxed reachability. In *Computer Aided Verification: 33rd International Conference, CAV 2021, Virtual Event, July 20–23, 2021, Proceedings, Part I 33*, 263–286. Springer.

Wang, K.; Liew, J. H.; Zou, Y.; Zhou, D.; and Feng, J. 2019. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, 9197–9206.

Wang, Z.; Guo, H.; Zhang, Z.; Liu, W.; Qin, Z.; and Ren, K. 2021. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7639–7648.

Xiao, C.; Deng, R.; Li, B.; Yu, F.; Liu, M.; and Song, D. 2018. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 217–234.

Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; and Yuille, A. 2017. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, 1369–1378.

Xu, X.; Zhao, H.; and Jia, J. 2021. Dynamic divide-and-conquer adversarial training for robust semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7486–7495.

Yang, J.; Xu, R.; Li, R.; Qi, X.; Shen, X.; Li, G.; and Lin, L. 2020. An adversarial perturbation oriented domain adaptation approach for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12613–12620.

Zhang, H.; and Wang, J. 2019. Defense against adversarial attacks using feature scattering-based adversarial training. *Advances in Neural Information Processing Systems*, 32.

Zhang, Y.; Shi, Z.; Yang, W.; Wang, S.; Wang, S.; and Xue, Y. 2024. GenSeg: On Generating Unified Adversary for Segmentation. In *IJCAI*.

Zhang, Y.; and Yang, W. 2022. BSOLO: Boundary-Aware One-Stage Instance Segmentation SOLO. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2594–2598. IEEE.

Zhang, Y.; Yang, W.; and Hu, R. 2023. BAProto: Boundary-Aware Prototype for High-quality Instance Segmentation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2333–2338. IEEE.

Zhang, Y.; Yang, W.; and Wang, S. 2023. FGNet: Towards Filling the Intra-class and Inter-class Gaps for Few-shot Segmentation. In *IJCAI*, 1749–1758.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 633–641.

Zhou, M.; Niu, Z.; Wang, L.; Zhang, Q.; and Hua, G. 2020. Adversarial ranking attack and defense. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 781–799. Springer.

Zhou, S.; Nie, D.; Adeli, E.; Yin, J.; Lian, J.; and Shen, D. 2019. High-resolution encoder–decoder networks for low-contrast medical image segmentation. *IEEE Transactions on Image Processing*, 29: 461–475.

Zhou, T.; Wang, W.; Konukoglu, E.; and Van Gool, L. 2022. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2582–2593.