GenSeg: On Generating Unified Adversary for Segmentation

Yuxuan Zhang[1], **Zhenbo Shi**[1,2], **Wei Yang**[1,2,3,*], **Shuchang Wang**[1,3], **Shaowei Wang**[4], **Yinxing Xue**[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] Institute of Artificial Intelligence and Blockchain, Guangzhou University

Abstract

Great advancements in semantic, instance, and panoptic segmentation have been made in recent years, yet the top-performing models remain vulnerable to imperceptible adversarial perturbation. Current attacks on segmentation primarily focus on a single task, and these methods typically rely on iterative instance-specific strategies, resulting in limited attack transferability and low efficiency. In this paper, we propose GenSeg, a **Gen**erative paradigm that creates unified adversaries for **Seg**mentation tasks. In particular, we propose an intermediate-level objective to enhance attack transferability, including a mutual agreement loss for feature deviation, and a prototype obfuscating loss to disrupt intra-class and inter-class relationships. Moreover, GenSeg crafts an adversary in a single forward pass, significantly boosting the attack efficiency. Besides, we unify multiple segmentation tasks to GenSeg in a novel category-and-mask view, which makes it possible to attack these segmentation tasks within this unified framework, and conduct cross-domain and cross-task attacks as well. Extensive experiments demonstrate the superiority of GenSeg in black-box attacks compared with state-of-the-art attacks. To our best knowledge, GenSeg is the first approach capable of conducting cross-domain and cross-task attacks on segmentation tasks, which are closer to real-world scenarios.

1 Introduction

Deep neural networks have excelled in diverse domains, however, their vulnerability to quasi-imperceptible adversarial perturbations remains a challenge [Goodfellow *et al.*, 2014b; Szegedy *et al.*, 2014]. Efforts to address this issue have predominantly focused on image classification tasks [Dong *et al.*, 2018; Ilyas *et al.*, 2018; Zhang *et al.*, 2022a], leaving a gap in understanding adversarial robustness in segmentation models. Given the broader applications of segmentation tasks (such as in autonomous driving and medical image analysis), there is an urgent need to develop effective strategies of adversarial attack and defense for them.

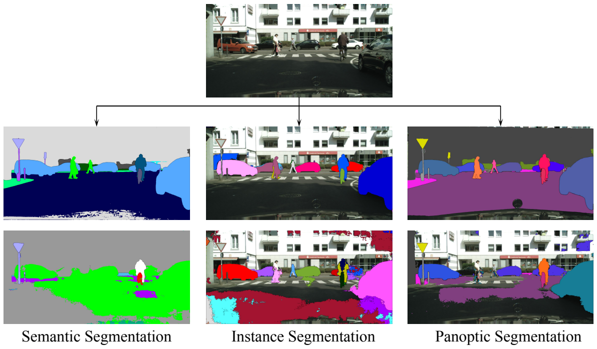

Figure 1: Semantic, instance, and panoptic segmentation models exhibit vulnerability to adversarial examples. The first row is the clean image, with its corresponding predictions exhibited in the second row. Upon the introduction of imperceptible adversarial noise, the predictions in the third row become totally different.

Predominant adversarial attacks on segmentation tasks, to date, have faced two main challenges: 1) *Transferability*: Dense-prediction tasks are more complicated than classification tasks, leading to the difficulty in designing a strong attack. Although progress has been made [Hendrik Metzen *et al.*, 2017; Xie *et al.*, 2017; Xu *et al.*, 2021], existing methods still exhibit limited transferability across different models due to their instance-specific strategies, relying too much on specific source models. 2) *Efficiency*: Current instance-specific attacks [Gu *et al.*, 2022; Agnihotri and Keuper, 2023; Rony *et al.*, 2023] typically employ an iterative gradient-based paradigm, which is time-consuming. However, insufficient iterations to reduce time consumption will decrease the effectiveness and further reduce attack transferability. Therefore, enhancing attack transferability and efficiency remains challenging at the same time. Additionally, existing approaches primarily focus on adversarial attacks in semantic segmentation (SS) [Bar *et al.*, 2020; Rossolini *et al.*, 2023], neglecting other segmentation tasks such as instance segmentation (IS) and panoptic segmentation (PS). However, these tasks are typically vulnerable to adversarial attacks, as illustrated in Figure 1.

The above challenges prompt two intriguing questions: 1)

[*] Corresponding author. Email: qubit@ustc.edu.cn
Code is available at: https://github.com/YXZhang979/GenSeg

How can we improve both attack transferability and efficiency of adversarial attacks at the same time? 2) *Is it possible to devise a unified attack paradigm that is applicable across diverse segmentation tasks*?

To address these two questions, we propose a pipeline to generate unified adversaries for segmentation tasks based on a generative model [Goodfellow *et al.*, 2014a], termed **GenSeg**. For the first question, contrasting to the iterative instance-specific methods, GenSeg crafts an adversary in a single forward pass, showcasing its high efficiency. To enhance attack transferability, we introduce an intermediate-level attack catering to segmentation tasks for optimizing the generator. Such an attack contains two objectives, i.e., a mutual agreement loss and a prototype obfuscating loss. The former aims to deviate the adversarial features from the benign features. As our goal is to deceive as many pixels as possible, we incorporate a dynamic weight strategy to punish the insufficient-perturbed pixels. The latter focuses on damaging the intra-class and inter-class prototype relationships [Zhang *et al.*, 2023b], which contributes to a stronger ambiguity in the latent space.

As to the second question, given the shared structure in different segmentation tasks (involving a backbone for feature extraction and a head network for prediction), we introduce a category-and-mask view to unify segmentation tasks and make it a uniform paradigm applicable to attack these tasks, including SS, IS, and PS.

To evaluate the effectiveness of GenSeg, we conduct comprehensive experiments on SS, IS, and PS, respectively. We employ 9 datasets and 15 models in total to validate our method. Experimental results show that GenSeg performs a stronger cross-model black-box attack transferability compared with state-of-the-art methods. To our best knowledge, GenSeg first extensively explores cross-domain and cross-task attacks in segmentation tasks, and sets a baseline.

We briefly summarize our contributions as follows:

- We propose GenSeg, a unified generative paradigm to attack diverse segmentation tasks, including semantic segmentation, instance segmentation, and panoptic segmentation.
- We design an intermediate-level objective for segmentation, including a mutual agreement attack to distance the feature space and a prototype obfuscating attack to disrupt the intra-class and inter-class relationships.
- Extensive experiments show GenSeg's superior attack capability on three tasks, encompassing 9 datasets and 15 models in total. GenSeg also pioneers the exploration of cross-domain and cross-task attacks in segmentation.

2 Related Work

2.1 Adversarial Attack in Classification

The pioneered work [Goodfellow *et al.*, 2014b] highlights the vulnerability of neural networks, leading to numerous approaches using iterative gradient-based optimization for generating *instance-specific* adversaries [Madry *et al.*, 2017; Dong *et al.*, 2018; Xie *et al.*, 2019; Shi *et al.*, 2021]. However, these methods are time-consuming and lack transferability across models or domains due to their reliance on the output score of a specific source model [Liu *et al.*, 2016; Naseer *et al.*, 2021; Shi *et al.*, 2022; Multimedia, 2023]. In contrast, *universal perturbations* aim to deceive models using global patterns, yet their effectiveness is relatively limited [Moosavi-Dezfooli *et al.*, 2017; Li *et al.*, 2022]. More recently, *generative perturbations* have been proven to be more effective than directly optimizing the universal perturbations [Feng *et al.*, 2023; Yang *et al.*, 2022; Naseer *et al.*, 2021]. Our method follows the concept of generative perturbations, crafting adversaries in a single forward pass. Unlike previous approaches, we focus on designing a generator tailored to generating adversaries for segmentation tasks, targeting the deception of as many pixels as possible.

2.2 Adversarial Attack in Segmentation

Current adversarial attacks in segmentation primarily focus on SS. [Arnab *et al.*, 2018] transfers the attacks in classification to assess the adversarial robustness on segmentation models. [Hendrik Metzen *et al.*, 2017] optimizes a universal perturbation to fool the majority of test images in SS, while the performance is unsatisfactory. For deceiving as many pixels as possible, approaches like DAG [Xie *et al.*, 2017], SegPGD [Gu *et al.*, 2022], and CosPGD [Agnihotri and Keuper, 2023] pay more attention to attacking benign pixels. [Rony *et al.*, 2023] proposes a white-box attack to reduce the perturbation budget. In addition, there are a few works [Zhang *et al.*, 2022b; Li *et al.*, 2018; Lang *et al.*, 2022] on adversarial attacks in IS. However, most of these approaches treat IS as a detect-then-segment method, while one-stage and query-based approaches have become the mainstream paradigm. Furthermore, only [Daza *et al.*, 2022] delves into adversarial attack exploration in PS, to the best of our knowledge.

Despite their advancements, current methods are still limited by 1) their focus on a single task, 2) the poor transferability to other target models, and 3) low efficiency with the iterative strategies. To address these issues, we design GenSeg, a unified attack framework applicable to multiple segmentation tasks, motivated by MaskFormer series [Cheng *et al.*, 2021; Cheng *et al.*, 2022]. Moreover, GenSeg also performs higher attack transferability and efficiency compared with iterative instance-specific attacks.

3 Methodology

3.1 General Formulation of Segmentation

We start by providing a unified formulation for segmentation tasks in a category-and-mask view. Despite the apparent diversity of SS, IS, and PS, current solutions share a common structure, a backbone for feature extraction and a head network for prediction. Given an image $I \in \mathbb{R}^{H\times W\times 3}$, we derive the intermediate feature F using the backbone. Subsequently, F is input into the head network to predict an instance mask set $\hat{S}^c = \{\hat{y}_i^c\}_{i=1}^{\hat{N}_c}$ for each category $c \in \mathcal{C}$, approximating the ground-truth $S^c = \{y_i^c\}_{i=1}^{N_c}$. Here, $y_i^c \in \{0,1\}^{H\times W}$ denotes the binary mask of the i-th instance for category c and N_c is the total mask count for category c.

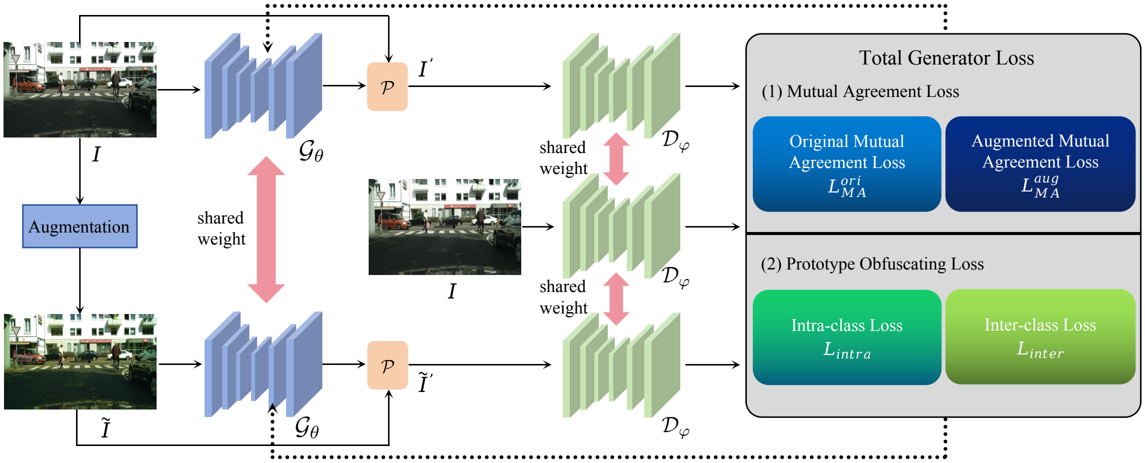

Figure 2: Overview of GenSeg. During training, GenSeg first employs $\mathcal{G}_\theta$ to generate the adversaries of the clean image and its augmentation. Then D_φ extracts the features of both adversaries and the clean image to establish the objective for training $\mathcal{G}_\theta$, which consists of a mutual agreement loss and a prototype obfuscating loss.

Note that, $\mathcal{C}$ varies across different tasks. As to the image segmentation tasks, $\mathcal{C}$ typically comprises two subsets: *stuff* and *thing*, which are denoted as $\mathcal{C}^{\mathrm{St}}$ and $\mathcal{C}^{\mathrm{Th}}$, respectively, and $\mathcal{C}^{\mathrm{St}} \cap \mathcal{C}^{\mathrm{Th}} = \emptyset$. *Stuff* segmentation involves segmenting uncountable regions that do not correspond to individual objects, e.g., sky, grass, and wall. *Thing* segmentation focuses on identifying each separate object, e.g., car, bus, and person.

In SS, we do not need to distinguish different *thing* objects. That is, we treat the countable *thing* equal to the uncountable *stuff*. Therefore, the segmented category $\mathcal{C} = \mathcal{C}^{\mathrm{St}} \cup \mathcal{C}^{\mathrm{Th}}$ and $N_c = \{0, 1\}$ for each category $c \in \mathcal{C}$. Given the ground-truth mask set $S^c = \{y^c_i\}^{N_c}_{i=1}$ for $c \in \mathcal{C}$, we have $\sum_{c\in\mathcal{C}} \sum^{N_c}_{i=1} ||y^c_i||_1 = H \times W$.

In IS, we only care about *thing* segmentation. Thus $\mathcal{C} = \mathcal{C}^{\mathrm{Th}}$. As IS aims to distinguish intra-class objects, we have $N_c \in \mathbb{N}$ for $c \in \mathcal{C}$. Note that, the different masks may have an intersection, since the initial masks are typically encoded as a polygon. Thus we cannot determine which one is greater between $\sum_{c\in\mathcal{C}} \sum^{N_c}_{i=1} ||y^c_i||_1$ and $H \times W$. However, this does not influence the unified attack objective.

In PS, we need to not only segment *stuff* category, but also identify different objects in *thing* category. Accordingly, we have $\mathcal{C} = \mathcal{C}^{\mathrm{St}} \cup \mathcal{C}^{\mathrm{Th}}$, $N_c = \{0, 1\}$ for $c \in \mathcal{C}^{\mathrm{St}}$, and $N_c \in \mathbb{N}$ for $c \in \mathcal{C}^{\mathrm{Th}}$. It is worth noting that the masks are not allowed to have intersections in PS, which is different from IS. Thus we draw that $\sum_{c\in\mathcal{C}} \sum^{N_c}_{i=1} ||y^c_i||_1 = H \times W$.

Despite their differences, we are able to unify them into this category-and-mask formulation. This formulation enables the development of a unified objective in the generative network, providing a seamless approach to attack various segmentation tasks.

3.2 Overview of GenSeg

The overall procedure of GenSeg is illustrated in Figure 2. Such a pipeline contains two key networks, i.e., a generator $\mathcal{G}_\theta$ and a discriminator $\mathcal{D}_\varphi$, which are parameterized by θ and φ, respectively. Given a clean image I, we first create an augmented copy $\tilde{I}$, aiming to enhance attack transferability [Naseer *et al.*, 2021]. After that, I and $\tilde{I}$ are input into $\mathcal{G}_\theta$ to produce unrestricted adversaries. As the perturbation should be imperceptible, we strictly constrain the adversary into an ϵ-ball with l_∞-norm. This is guaranteed by a differentiable projecting operation, which is formulated as:

$$I' = \mathrm{Clip}\{\min(I + \epsilon, \max(\mathcal{W} * (\mathcal{G}_\theta(I)), I - \epsilon)\} \tag{1}$$

where ϵ is the perturbation budget and $\mathcal{W}$ is a smoothing operator with fixed weights [Feng *et al.*, 2023]. Such a projection tightly bounds the output of $\mathcal{G}_\theta$ within l_∞-norm. Meanwhile, with the smooth operation $\mathcal{W}$, the generator is guided to avoid redundant high frequencies during training, making $\mathcal{G}_\theta$ converge to a reasonable solution. Clip is a clipping operation.

Subsequently, we put adversary I', augmented adversary $\tilde{I}'$, and clean image I into $\mathcal{D}_\varphi$ for feature extraction. Accordingly, we obtain the feature F', $\tilde{F}'$, and $F \in \mathbb{R}^{H\times W\times D}$, respectively, where D is the dimension of the latent space. Then the features are employed to calculate an intermediate-level objective for optimizing $\mathcal{G}_\theta$, which is formulated as:

$$L_{total} = \lambda_1 L_{MA} + \lambda_2 L_{PO} \tag{2}$$

where λ_1 and λ_2 are two balanced weights, and we set $\lambda_1 = \lambda_2 = 0.5$ empirically. The *mutual agreement* loss L_{MA} aims to distance the feature of the adversary and its benign state, and the *prototype obfuscating* loss L_{PO} further leverages the intra-class and inter-class prototype relationships to enhance the intermediate-level attack. We will present the details of both losses in the following subsections.

Accordingly, during the training period, we employ L_{total} to optimize the parameter θ of the generator, which guides $\mathcal{G}_\theta$ to produce an adversary whose feature is deviated from the expected one. During inference, we solely apply $\mathcal{G}_\theta$ and a clip operation to craft an adversary in a single forward pass.

3.3 Mutual Agreement Attack

Intuitively, if the adversarial feature F' is totally different from the benign feature F, it would mislead the subsequent classifier to conduct false predictions. Given this, our attack design is to pull the distribution of adversarial features away from the benign distribution. To achieve this goal in the dense-prediction task, we take Kullback Leibler (KL) divergence [Kullback and Leibler, 1951] to measure the pixel-wise mutual agreement between F' and F, which is calculated by:

$$D_{KL}(F'_j||F_j) = \sum_{d=1}^{D} \sigma(F'{}_j)_d \log \frac{\sigma(F'{}_j)_d}{\sigma(F_j)_d} \tag{3}$$

where j represents the 2D spatial location, d stands for the d-th dimension of the latent space, and σ denotes a softmax operation for normalization. In short, KL divergence measures the pixel-level differences between F and F'. It is worth noting that, KL divergence is asymmetric, i.e., $D_{KL}(F'_j||F_j) \neq D_{KL}(F_j||F'_j)$, and the value is not normalized. Thus we employ Jensen Shannon (JS) divergence [Goodfellow *et al.*, 2014a] to achieve both symmetry and normalization, which is formulated by:

$$D_{JS}(F'_j||F_j) = \frac{1}{2} D_{KL}(F'_j||\bar{F}_j) + \frac{1}{2} D_{KL}(F_j||\bar{F}_j) \tag{4}$$

where $\bar{F}_j = \frac{F_j + F'_j}{2}$ denotes the average of F_j and F'_j.

Inspired by [Gu *et al.*, 2022], a strong attack can deceive as many pixels as possible in segmentation tasks. Therefore, paying equal attention to all pixels is less effective. In particular, the pixels in F' with a higher similarity to F are supposed to be highlighted. Accordingly, we introduce a dynamic weight to each pixel, calculated by: $W_j = \frac{F_j \cdot F'_j}{||F_j|| \cdot ||F'_j||}$, and we focus on reducing the *original mutual agreement* loss between F and F' by:

$$L^{ori}_{MA} = \frac{1}{H \times W} \sum_{j=1}^{H \times W} W_j \cdot (1 - D_{JS}(F'_j||F_j)) \tag{5}$$

If F'_j is close to F_j, we have $D_{JS}(F'_j||F_j) \to 0$. Thus we assign a large weight W_j to punish this high alignment, which contributes to the deception of more pixels.

Besides, to enhance attack transferability, we employ an augmented copy, which guides the generator to produce an adversary that is robust to the input transformation. Similarly, we reduce the feature similarity between F and $\tilde{F}$ by the *augmented mutual agreement* loss:

$$L^{aug}_{MA} = \frac{1}{H \times W} \sum_{j=1}^{H \times W} \tilde{W}_j \cdot (1 - D_{JS}(\tilde{F'}_j||F_j)) \tag{6}$$

where $\tilde{W}_j = \frac{F_j \cdot \tilde{F}'_j}{||F_j|| \cdot ||\tilde{F}'_j||}$ is a dynamic weight.

Finally, the total mutual agreement loss is calculated by:

$$L_{MA} = L^{ori}_{MA} + L^{aug}_{MA} \tag{7}$$

3.4 Prototype Obfuscating Attack

With segmentation masks, the intermediate feature map offers a greater interpretability compared with classification tasks. This allows us to improve intermediate-level attacks guided by the ground-truth masks. Specifically, we employ a global descriptor known as a *prototype* to represent the average feature of each category. The prototype is widely used in numerous segmentation tasks [Wang *et al.*, 2019; Zhang *et al.*, 2023a; Xu *et al.*, 2022]. Damaging the prototype relationship for each category will enhance the intermediate-level attack capability.

On the one hand, the intra-class features share a high similarity, which is the foundation for utilizing a prototype to represent each specific category. Given an image I and its intermediate feature F, we divide F into category-specific regions based on the ground-truth mask set $S^c = \{y^c_i\}^{N_c}_{i=1}$ for $\forall c \in \mathcal{C}_{seg}$, where $\mathcal{C}_{seg} = \{c | N^c \geq 1, \forall c \in \mathcal{C}\}$ denotes the set of categories that appears in the ground-truth mask. Then, the prototype $p^c \in \mathbb{R}^d$ of category c is calculated by a masked average pooling operation:

$$p^c = \frac{\sum_{i=1}^{N_c} \sum_{j=1}^{H \times W} (F \odot \mathbb{1}[\tau(y^c_i) = 1])_j}{\sum_{i=1}^{N_c} \sum_{j=1}^{H \times W} \mathbb{1}[(y^c_i)_j = 1]} \tag{8}$$

where $\odot$ and $\mathbb{1}$ denote a Hadamard product and an indicator function, respectively. i represents the i-th mask of category c, and j is the 2D spatial index. Besides, $\tau(y^c_i)$ stands for the dimension expansion of y^c_i (from $\mathbb{R}^{H \times W}$ to $\mathbb{R}^{H \times W \times D}$), which aims to align the dimension of F.

After obtaining the prototype for each segmented category, we make all the pixel-wise representations deviated from their original cluster center. Therefore, the intra-class attack objective of category c can be formulated as:

$$L^c_{intra} = \frac{\sum_{i=1}^{N_c} \sum_{j=1}^{H \times W} cosSim(F_j, p^c) \mathbb{1}[(y^c_i)_j = 1]}{\sum_{i=1}^{N_c} \sum_{j=1}^{H \times W} \mathbb{1}[(y^c_i)_j = 1]} \tag{9}$$

where $cosSim(\cdot, \cdot)$ denotes cosine similarity. The total intra-class objective considering all categories is computed by:

$$L_{intra} = \frac{1}{|\mathcal{C}_{seg}|} \sum_{c \in \mathcal{C}_{seg}} L^c_{intra} \tag{10}$$

Our attack goal is to minimize L_{intra}, thereby reducing the similarity between the pixel-wise representation and its corresponding prototype. This approach dilates the feature space of each segmented category, consequently misleading the head network to conduct false predictions.

On the other hand, the inter-class feature spaces is supposed to be discernible. In particular, a substantial margin between feature clusters of distinct categories enhances the separation performance, resulting in a distinct decision boundary. Hence, our attack is aimed at blurring the decision boundary, and the objective for inter-class attack is formulated as:

$$L_{inter} = 1 - \frac{1}{K} \sum_{c_1, c_2 \in \mathcal{C}_{seg}} cosSim(p^{c_1}, p^{c_2}) \mathbb{1}[c_1 \neq c_2] \tag{11}$$

where $K = 2\binom{|\mathcal{C}_{seg}|}{2}$ denotes the number of pair-wise prototype permutations.

After obtaining the above two attack objectives, we formulate the overall prototype obfuscating loss by:

$$L_{PO} = L_{intra} + L_{inter} \tag{12}$$

4 Experiments

Dataset. To evaluate the attack effectiveness of GenSeg, we employ the commonly used Pascal VOC (20 classes) [Everingham *et al.*, 2010], Cityscapes (19 classes) [Cordts *et al.*, 2016], and ADE20K (150 classes) [Zhou *et al.*, 2017] for SS. For IS, we use the widely used CityScapes (8 *things*), COCO (80 *things*) [Lin *et al.*, 2014], and ADE20K (100 *things*). As to PS, we adopt COCO (80 *things* and 53 *stuff*), Cityscapes (8 *things* and 11 *stuff*), and ADE20K (100 *things* and 50 *stuff*). Please see Appendix B for detailed descriptions of all datasets used in experiments.

Evaluation metrics. For SS, we use the metric of mean Intersection-over-Union (**mIoU**) for evaluation. For IS, we report the standard average precision (**AP**) metric. For PS, we employ the panoptic quality (**PQ**) metric. The **drop** score of each metric correlates directly with the attack capability.

4.1 Implementation Detail

We utilize the widely-adopted ResNet-based model [He *et al.*, 2016] in [Naseer *et al.*, 2019] for generator $\mathcal{G}_\theta$, producing adversaries that match the input size. The discriminator $\mathcal{D}_\varphi$ employs a pre-trained ResNet50 to generate a multi-scale hybrid feature map. During training, $\mathcal{D}_\varphi$ remains frozen while optimizing $\mathcal{G}_\theta$. We use the Adam optimizer with a learning rate of 5e-3 ($\beta_1 = .5$, $\beta_2 = .999$) for 100 epochs. For each segmentation task, we train an individual generator on each separate dataset, allowing the evaluation of both intra-domain and cross-domain attacks. We set the perturbation budget ϵ to the typical value of $8/255$. Additionally, our smooth operator $\mathcal{W}$ is a differentiable Gaussian kernel specified in [Naseer *et al.*, 2021]. After training, the generator creates adversaries without any augmentation. Subsequently, we conduct attacks across various models and domains to thoroughly evaluate the effectiveness of GenSeg.

4.2 Attack Scenarios and Results

We assess untargeted black-box attacks of GenSeg in the following two scenarios: 1) *Cross-model Intra-domain*: The attacker lacks access to the target model, while both the source and target models are trained within the same domain. 2) *Cross-model Cross-domain*: The attacker is unaware of the target model's architecture. Moreover, the source and target models are trained on different domains. This scenario more closely reflects real-world conditions.

Cross-model Intra-domain Attack

In this scenario, GenSeg is trained on the same dataset as the target models. Then we assess the attack transferability of GenSeg in comparison to other transfer-based attacks.

Semantic Segmentation: We evaluate the black-box attack capability of GenSeg by comparing it with several top-performing attacks, i.e., DAG [Xie *et al.*, 2017], SegPGD [Gu *et al.*, 2022], CosPGD [Agnihotri and Keuper, 2023], and Prox [Rony *et al.*, 2023]. Since most attacks are iterative instance-specific methods, we designate three source models, i.e., PSPNet-R50 (PSP-R50) [Zhao *et al.*, 2017; He *et al.*, 2016], DeepLabv3-R101 (DLv3-R101) [Chen *et al.*, 2017], and Mask2Former-Swin-S (M2F-Swin-S) [Cheng *et al.*, 2022; Liu *et al.*, 2021] for these attacks to generate adversaries and conduct transfer-based attacks. As to the victim target models, we use PSPNet-R50 (PSP-R50), PSPNet-R101 (PSP-R101), DeepLabv3-R50 (DLv3-R50), DeepLabv3-R101 (DLv3-R101), Mask2Former-R101 (M2F-R101), Mask2Former-Swin-T (M2F-Swin-T), and Mask2Former-Swin-S (M2F-Swin-S) for comprehensive evaluations. Both source and target models are well-trained within the same domain.

To ensure a fair comparison across these attacks, we reimplement them using the setting of the step size $3/255$ and the perturbation budget $8/255$. Besides, we set attack iteration to 5, striking a balance between attack capability and efficiency. Note that, unless otherwise specified, this setting generally applies to the iterative attacks mentioned later in this paper.

The results on Pascal VOC are summarized in Table 1. GenSeg demonstrates a strong attack capability across multiple SS models, showcasing its effective transferability. Although existing adversaries successfully deceive the source model (results in red in the table), their transferability to other target models remains limited. This limitation stems from instance-specific attacks heavily relying on the decision boundary of the specific source model. In contrast, GenSeg operates independently of the source model and employs an intermediate-level attack, which helps to enhance its attack transferability. Furthermore, the efficiency of GenSeg stands out as it generates adversaries in a single forward pass, presenting a more streamlined approach compared with the iterative attacks. For more details, please refer to Appendix F.

We also conduct experiments on Cityscapes and ADE20K, respectively. For detailed results, please refer to Appendix C. The results show that the adversaries generated by GenSeg effectively deceive a range of top-performing SS models, exhibiting the highlighted attack transferability of GenSeg.

Instance Segmentation: We explore the attack efficacy of GenSeg on IS, compared with the popular methods MIM [Dong *et al.*, 2018], Improved PGD (I-PGD) [Zhang *et al.*, 2022b], and DIM [Xie *et al.*, 2019]. As to the selection of source and target models, we choose Mask-RCNN (M-RCNN) [He *et al.*, 2017], Yolact [Bolya *et al.*, 2019], PolarMask [Xie *et al.*, 2020], and Mask2Former (M2F) [Cheng *et al.*, 2022]. These models represent four diverse solutions for IS: top-down, one-stage, contour-based, and query-based, respectively. As their structures are totally different, we use the same backbone ResNet50 for each model.

The results on Cityscapes are presented in Table 2, and the results on COCO and ADE20K are provided in Appendix C. Similar to SS, although iterative attacks own high efficacy in white-box attack cases, their transferability to different target models remains deficient. Note that, this defi-

Source Model	Attack	Target Model						
		PSP-R50	PSP-R101	DLv3-R50	DLv3-R101	M2F-R101	M2F-Swin-T	M2F-Swin-S
N/A	Clean	76.7	78.4	76.1	78.7	80.4	82.1	82.5
PSP-R50	DAG	64.1	18.4	23.5	15.7	16.2	11.9	8.2
	SegPGD	68.0	21.9	26.3	16.8	19.7	12.9	11.4
	CosPGD	66.7	23.5	24.0	18.1	23.5	15.0	14.4
	Prox	48.2	11.7	16.3	12.2	10.3	8.6	9.5
DLv3-R101	DAG	22.7	26.7	24.2	67.4	20.5	12.0	10.9
	SegPGD	27.4	30.4	30.5	71.1	18.1	14.7	11.5
	CosPGD	26.5	24.1	25.4	70.7	22.9	15.5	14.8
	Prox	10.4	12.1	10.2	44.3	7.3	8.2	8.7
M2F-Swin-S	DAG	28.7	17.6	22.9	18.0	25.7	29.5	72.5
	SegPGD	27.1	23.3	28.2	26.1	27.7	35.9	74.8
	CosPGD	32.8	34.0	29.9	23.6	28.6	31.6	71.6
	Prox	14.6	13.8	16.1	13.0	16.1	18.3	50.8
N/A	**GenSeg**	53.5	**47.9**	**53.3**	51.8	**46.0**	**56.2**	50.1

Table 1: Performances of cross-model intra-domain attacks on Pascal VOC for SS. Source models and target models are both well trained on Pascal VOC. The row in green represents the **original mIoU** for each target model without any attack. Subsequent rows exhibit **mIoU drop** after attacks. Red regions indicate the results of white-box attacks. The most effective attack is highlighted in bold, while results inferior only to white-box attack are underlined. GenSeg shows a strong capability to attack a wide range of SS models.

Source Model	Attack	Target Model			
		M-RCNN	Yolact	PolarMask	M2F
N/A	Clean	26.6	24.9	24.7	37.4
M-RCNN	MIM	19.8	4.7	3.0	7.4
	I-PGD	22.4	5.3	5.9	6.2
	DIM	19.2	6.2	5.7	8.2
Yolact	MIM	5.6	20.1	4.3	7.5
	I-PGD	5.1	21.4	6.0	7.1
	DIM	7.9	18.3	5.8	10.2
PolarMask	MIM	1.7	0.9	19.6	3.1
	I-PGD	3.2	3.1	21.4	4.8
	DIM	1.4	2.3	22.6	3.7
M2F	MIM	4.5	6.0	4.0	35.1
	I-PGD	5.7	7.1	3.4	33.2
	DIM	5.6	5.2	4.3	28.0
N/A	**GenSeg**	17.1	17.6	16.3	24.9

Table 2: Performance of cross-model intra-domain attacks on Cityscapes for IS. Source model and target model are well trained on Cityscapes. Results in green represent the **original AP** for each target model without attack, while others stand for **AP drop** after attacks. Red regions indicate the results of white-box attacks. GenSeg demonstrates a strong capability to attack a wide range of IS models.

ciency is more pronounced in comparison to SS, as the huge divergence among different IS models exacerbates the challenge. In contrast, GenSeg demonstrates robust transferability across diverse models as it is detached from any source model. Moreover, the intermediate-level attack is applicable to all solutions. Once the feature space is corrupted, all the subsequent head networks will be misled.

Panoptic Segmentation: We also explore the attack transferability of GenSeg on PS. To compare GenSeg with other top-performing iterative attacks (PGD [Madry *et al.*, 2017] and AutoPGD [Daza *et al.*, 2022]), we select four end-to-end PS models, i.e., DETR [Carion *et al.*, 2020], MaskFormer (MF) [Cheng *et al.*, 2021], Mask2Former (M2F) [Cheng *et al.*, 2022], and kMax-DeepLab (kMDL) [Yu *et al.*, 2022]. The backbone is selected to be ResNet50 for each model.

The results on Cityscapes are shown in Table 3, and the results of COCO and ADE20K are depicted in Appendix C. GenSeg performs a good transferability to attack a diversity of models across different datasets, which is superior to existing instance-specific approaches.

Source Model	Attack	Target Model			
		DETR	MF	M2F	kMDL
N/A	Clean	54.7	58.9	62.1	64.3
DETR	PGD	38.5	8.7	11.0	6.5
	AutoPGD	41.4	7.3	10.9	9.7
MF	PGD	7.8	43.8	12.1	10.5
	AutoPGD	7.9	50.1	9.4	7.1
M2F	PGD	7.7	9.3	47.0	15.2
	AutoPGD	8.0	13.9	53.6	11.2
kMDL	PGD	6.6	8.9	8.4	47.3
	AutoPGD	7.8	10.5	7.6	50.5
N/A	**GenSeg**	31.7	39.1	37.6	42.9

Table 3: Performance of cross-model intra-domain attacks on Cityscapes for PS. Source models and target models are both well trained on Cityscapes with panoptic masks. Results in green represent the **original PQ** for each target model without any attack, while others stand for **PQ drop** after attacks. Red regions indicate the results of white-box attacks. GenSeg demonstrates a strong capability to attack a wide range of PS models.

Cross-model Cross-domain Attack

This scenario is more practical in the real world, as we have no access to either the deployed model structure or the target domain. Considering that we have no knowledge about the target domain and the guidance of the ground-truth mask, current iterative instance-specific attacks are unable to generate adversaries for unknown domains, as the misalignment of output dimension and ground-truth mask missing. However, as only one input image is required for GenSeg, it can be technically applied to generate adversaries for any unknown target domain.

We evaluate the cross-model cross-domain attack capability of GenSeg. For SS, we leverage GenSeg trained on source domains to generate adversaries to attack victim models that are trained on different target domains. We utilize Pascal VOC, Cityscapes, and ADE20K datasets to explore the cross-domain attack transferability. As to the victim target models, we select PSPNet-R50 (PSP-R50) [Zhao *et al.*, 2017; He *et al.*, 2016] and DeepLabv3-R50 (DLv3-R50) [Li *et al.*, 2018] to test their mIoU results. The results of mIoU drop are

Source Domain	Victim Model	Target Domain		
		Pascal VOC	Cityscapes	ADE20K
Pascal VOC	PSP-R50	-	23.3 (↓ 29.9%)	13.1 (↓ 31.8%)
	DLv3-R50	-	26.1 (↓ 33.0%)	14.4 (↓ 33.9%)
Cityscapes	PSP-R50	21.9(↓ 28.5%)	-	11.7(↓ 28.4%)
	DLv3-R50	27.3(↓ 35.8%)	-	12.5(↓ 29.4%)
ADE20K	PSP-R50	34.4(↓ 44.9%)	35.6(↓ 45.7%)	-
	DLv3-R50	29.7(↓ 39.0%)	33.8(↓ 42.7%)	-

Table 4: Performance of **mIoU drop** under cross-domain attacks in SS. The adversaries are generated by GenSeg, which is well trained on the source domain, and attack the victim models that are trained on other target domains.

Group	L_{MA}^{ori}	L_{MA}^{aug}	L_{intra}	L_{inter}	mIoU↓	AP↓	PQ↓
(1)	✓				40.4	14.3	31.5
(2)	✓	✓			47.1	15.4	34.8
(3)	✓	✓	✓		47.7	15.7	35.5
(4)	✓	✓		✓	49.5	16.3	35.1
(5)			✓	✓	33.0	9.8	20.7
(6)	✓	✓	✓	✓	**55.4**	**17.6**	**37.6**

Table 5: Ablation results on different combinations of losses.

reported in Table 4. As evident from the table, in the cross-domain attack scenario, GenSeg induces mIoU drops ranging from 28.4% to 45.7% for the target models. While this effectiveness is comparatively lower than that in the intra-domain attack, the substantial absolute value of mIoU drop remains noteworthy. This indicates that GenSeg possesses the capability to transfer and attack victim models trained on unknown domains without overly relying on the source domain.

Moreover, the settings and results of IS and PS are detailedly demonstrated in Appendix D. It can be observed that the adversaries produced by GenSeg can decrease the performances of victim models trained on other domains by 23%-47%, which is sufficient to bring potential risks to deceive the *unknown* deployed models trained on *unknown* domains.

4.3 Ablation Study

We conduct thorough ablation experiments to determine the optimal setting for GenSeg and analyze the potential reasons. Utilizing GenSeg trained on Cityscapes with SS, IS, and PS masks, we conduct attacks on corresponding target victim models. Specifically, we employ each corresponding GenSeg to attack PSPNet for SS, Yolact for IS, and Mask2Former for PS. The attack efficacy is measured in terms of the **metrics drops** in **mIoU**, **AP**, and **PQ** scores, respectively.

Loss combination: Table 5 illustrates the attack capability of GenSeg when trained on various loss combinations. Basically, the amalgamation of L_{MA}^{ori}, L_{MA}^{aug}, L_{intra}, and L_{inter} shows the most effective attack across all segmentation tasks, namely Group (6). By contrast, Groups (1) and (2) indicate the efficacy of the augmentation in enhancing attack transferability; Groups (3) and (4) highlight the advantageous use of L_{intra} and L_{inter}, respectively, as they obscure the representation space in a category-specific manner, significantly bolstering the overall attack capability; Group (5) underscores the pivotal role of L_{MA}^{ori} in ensuring a strong attack.

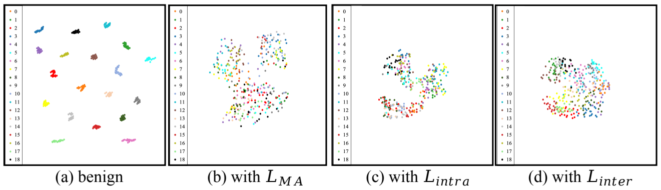
(a) benign (b) with L_{MA} (c) with L_{intra} (d) with L_{inter}

Figure 3: Visualization results of the prototypes on each category by *t*-SNE with different losses trained on GenSeg.

Augmentation	mIoU↓	AP↓	PQ↓
N/A	45.7	13.5	30.5
Gamma Transformation	52.5	17.1	35.4
Pepper Noise	54.7	16.5	**37.9**
Gaussian Noise	**55.4**	**17.6**	37.6

Table 6: Ablation results on augmentation-based training of GenSeg.

To further analyze the impact of each loss, we illustrate the variations of representation space in Figure 3. These prototypes are obtained based on PSPNet with the guidance of ground-truth masks on Cityscapes for SS. As shown in the figure, the intra-class and inter-class prototypes of benign features are compacted and separate, respectively. With the introduction of the mutual agreement loss, the feature space tends to be obfuscated without any regular pattern. However, with the usage of L_{intra}, the prototypes within each specific category tend to be uncompacted, which leads to an ambiguous decision boundary. On the other hand, L_{inter} mixes the prototypes of different categories together, significantly reducing the margin of the decision boundary. Therefore, the mutual agreement loss brings a random ambiguity to the latent space, while the intra-class and inter-class losses give a directional perturbation to damage the category relationships. We take the advantages of them to formulate the total objective for GenSeg with a stronger effectiveness. Note that, L_{intra} and L_{inter} act as auxiliary objectives to further enhance attack ability. As to the reasons, we believe that the attack directions of L_{intra} and L_{inter} are related to the prototype space of D_{φ}, whereas the omnidirectional attack L_{MA} is model-agnostic. However, none of these attacks are domain-agnostic, leading to relatively inferior results of cross-domain attacks.

Augmentation: We investigate the efficacy of the augmentation-based training strategy. To guarantee pixel-wise alignment for segmentation tasks, we employ augmentations without spatial transformation, including Gamma Transformation, Pepper Noise, and Gaussian Noise, compared with the standard training without any augmentation. As demonstrated in Table 6, Gaussian Noise augmentation achieves the best results. The augmentations can make GenSeg robust to slight input transformations and focus specifically on perturbing the latent space, which significantly enhances the attack effectiveness.

Mutual agreement: According to Eq. (6), we use JS divergence to distinguish between the features of clean images and adversaries. We assess alternative metrics like cosine similarity and Euclidean distance as mutual agreement measures. Table 7 displays the results, indicating JS divergence as the

Mutual Agreement	mIoU↓	AP↓	PQ↓
Euclidean Distance	49.5	14.5	32.3
Cosine Similarity	48.0	15.2	30.9
JS Divergence	**55.4**	**17.6**	**37.6**

Table 7: Ablation results on different mutual agreement measures.

Weight	mIoU ↓	AP ↓	PQ ↓
Equal weight	43.5	12.2	31.1
Dynamic weight	**55.4**	**17.6**	**37.6**

Table 8: Ablation results on the usage of dynamic weight in L_{MA}.

superior metric. Compared with Euclidean distance, JS divergence's normalized range of 0-1 aids in optimizing the total loss. Furthermore, compared with cosine similarity, JS divergence avoids redundancy in attacks, as the prototype obfuscating loss also employs cosine similarity for intra-class and inter-class attacks. Moreover, JS divergence, following the softmax operation, highlights the salient features. Targeting these features can further deceive the segmentation model, as they encode crucial image patterns.

Dynamic weight: We explore the results of using dynamic weight in the mutual agreement loss, compared with the equal weight on all pixels, and the results are depicted in Table 8. As shown in the table, the dynamic weight strategy causes a larger metrics drop compared with the equal weight assignment. By assigning larger weights to the inadequate-perturbative representations, we are able to lead numerous pixel-wise representations apart from their benign states. This will impose obstacles on the following head network and indirectly impact the final predictions.

Discriminator: We use well-trained ResNet50 and Swin-S as discriminators for feature extraction. For each model, we extract intermediate features from the output of stages 2, 3, 4, and a hybrid output [Xie *et al.*, 2021] that incorporates them together, respectively, as the output of $\mathcal{D}_\varphi$. Bilinear interpolation is then applied to match the output size to the input image. Table 9 shows the results, highlighting the superior attack based on the hybrid feature. Moreover, the results of ResNet50 are close to Swin-S. On the one hand, the mutual agreement loss does not heavily rely on the feature quality, as it concentrates more on expanding mutual distances. On the other hand, the prototype establishment depends on semantic features, and the hybrid feature leverages multi-scale information, enhancing the prototype obfuscating attack. Considering the efficiency, we apply ResNet50 as $\mathcal{D}_\varphi$ for feature extraction.

4.4 Generalized Cross-task Attack

The above attacks necessitate training GenSeg for each task and each dataset separately. However, our ultimate goal is to train *a single GenSeg capable of generating adversaries across all segmentation tasks*. Unlike MaskFormer, achieving this objective with GenSeg is technically feasible, as it only requires outputting the common feature map and employing a unified mask formulation for optimization.

Structure	Stage	mIoU ↓	AP ↓	PQ ↓
ResNet50	2	44.0	14.2	31.6
	3	48.8	17.0	33.2
	4	47.3	15.5	33.0
	hybrid	**55.4**	**17.6**	**37.6**
Swin-S	2	48.3	16.8	33.8
	3	49.8	16.4	35.1
	4	44.8	17.3	36.2
	hybrid	**56.1**	**17.7**	**38.8**

Table 9: Ablation results on different output features of $\mathcal{D}_\varphi$.

Task	Victim Model	Pascal	ADE20K
SS	PSP-R50	20.2(↓ 26.3%)	6.6 (↓ 16.0%)
	DLv3-R50	16.7 (↓ 21.9%)	7.4 (↓ 17.4%)
Task	**Victim Model**	**COCO**	**ADE20K**
IS	Yolact	6.9 (↓ 23.9%)	3.7 (↓ 20.7%)
	PolarMask	5.7 (↓ 19.5%)	3.4 (↓ 18.3%)

Table 10: **MIoU drop** for SS and **AP drop** for IS (along with **percentage drop** compared with clean results), when using GenSeg trained on PS to attack SS and IS models on different datasets.

Considering PS as a fusion task of SS and IS, we explore whether GenSeg trained on PS can generate adversaries to deceive SS and IS models. Specifically, we train GenSeg on Cityscapes with panoptic masks. Subsequently, GenSeg generates adversaries to attack PSPNet-R50 and DeepLabv3-R50 in SS, as well as Yolact and PolarMask in IS. To simulate a general scenario, we employ Pascal VOC and ADE20K datasets for SS, as well as COCO and ADE20K for IS. The results, shown in Table 10, reveal that the attacks cause approximately 16%-26% metric drops. As segmentation models share similar semantic features, attacks targeting the latent space affect all subsequent head networks, despite variations in models, domains, and tasks. Of course, there is still ample room for enhancing attack efficacy in the cross-task manner.

5 Conclusion

In this paper, we introduced GenSeg, a unified adversarial attack paradigm tailored for multiple segmentation tasks. GenSeg employs mutual agreement loss and prototype obfuscating loss, showcasing robust attack transferability across diverse models, domains, and tasks. Moreover, the effective adversary is generated with a single forward pass. Extensive experiments demonstrate that the black-box attack capability of GenSeg outperforms the top-performing iterative instance-specific attacks across 3 tasks, 9 datasets and 15 models in total.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62172385), the Innovation Program for Quantum Science and Technology (No. 2021ZD0302900), the Anhui Provincial Department of Science and Technology under Grant 202103a05020009, and the Laboratory for Advanced Computing and Intelligence Engineering Fund.

References

[Agnihotri and Keuper, 2023] Shashank Agnihotri and Margret Keuper. Cospgd: a unified white-box adversarial attack for pixel-wise prediction tasks. *arXiv preprint arXiv:2302.02213*, 2023.

[Arnab *et al.*, 2018] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *CVPR*, pages 888–897, 2018.

[Bar *et al.*, 2020] Andreas Bar, Jonas Lohdefink, Nikhil Kapoor, Serin John Varghese, Fabian Huger, Peter Schlicht, and Tim Fingscheidt. The vulnerability of semantic segmentation networks to adversarial attacks in autonomous driving: Enhancing extensive environment sensing. *IEEE Signal Processing Magazine*, 38(1):42–52, 2020.

[Bolya *et al.*, 2019] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *ICCV*, pages 9157–9166, 2019.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[Cheng *et al.*, 2021] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, 34:17864–17875, 2021.

[Cheng *et al.*, 2022] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022.

[Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[Daza *et al.*, 2022] Laura Daza, Jordi Pont-Tuset, and Pablo Arbeláez. Adversarially robust panoptic segmentation (arpas) benchmark. In *ECCV*, pages 378–395. Springer, 2022.

[Dong *et al.*, 2018] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, pages 9185–9193, 2018.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88:303–338, 2010.

[Feng *et al.*, 2023] Weiwei Feng, Nanqing Xu, Tianzhu Zhang, and Yongdong Zhang. Dynamic generative targeted attacks with pattern injection. In *CVPR*, pages 16404–16414, 2023.

[Goodfellow *et al.*, 2014a] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014.

[Goodfellow *et al.*, 2014b] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2014.

[Gu *et al.*, 2022] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *ECCV*, pages 308–325. Springer, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[Hendrik Metzen *et al.*, 2017] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *ICCV*, pages 2755–2764, 2017.

[Ilyas *et al.*, 2018] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, pages 2137–2146. PMLR, 2018.

[Kullback and Leibler, 1951] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[Lang *et al.*, 2022] Dapeng Lang, Deyun Chen, Sizhao Li, and Yongjun He. An adversarial attack method against specified objects based on instance segmentation. *Information*, 13(10):465, 2022.

[Li *et al.*, 2018] Yuezun Li, Daniel Tian, Ming-Ching Chang, Xiao Bian, and Siwei Lyu. Robust adversarial perturbation on deep proposal-based models. *arXiv preprint arXiv:1809.05962*, 2018.

[Li *et al.*, 2022] Maosen Li, Yanhua Yang, Kun Wei, Xu Yang, and Heng Huang. Learning universal adversarial perturbation by adversarial example. In *AAAI*, volume 36, pages 1350–1358, 2022.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[Liu *et al.*, 2016] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo.

Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Moosavi-Dezfooli *et al.*, 2017] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, pages 1765–1773, 2017.

[Multimedia, 2023] ACM Multimedia. Reinforcement learning-based adversarial attacks on object detectors using reward shaping. pages 8424–8432, 2023.

[Naseer *et al.*, 2019] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. *NeurIPS*, 32, 2019.

[Naseer *et al.*, 2021] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. On generating transferable targeted perturbations. In *ICCV*, pages 7708–7717, 2021.

[Rony *et al.*, 2023] Jérôme Rony, Jean-Christophe Pesquet, and Ismail Ben Ayed. Proximal splitting adversarial attack for semantic segmentation. In *CVPR*, pages 20524–20533, 2023.

[Rossolini *et al.*, 2023] Giulio Rossolini, Federico Nesti, Gianluca D'Amico, Saasha Nair, Alessandro Biondi, and Giorgio Buttazzo. On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Shi *et al.*, 2021] Zhenbo Shi, Wei Yang, Zhenbo Xu, Zhi Chen, Yingjie Li, Haoran Zhu, and Liusheng Huang. Adversarial attacks on object detectors with limited perturbations. In *ICASSP*, pages 1375–1379. IEEE, 2021.

[Shi *et al.*, 2022] Zhenbo Shi, Zhi Chen, Zhenbo Xu, Wei Yang, Zhidong Yu, and Liusheng Huang. Shape prior guided attack: Sparser perturbations on 3d point clouds. In *AAAI*, volume 36, pages 8277–8285, 2022.

[Szegedy *et al.*, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. 2014.

[Wang *et al.*, 2019] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*, pages 9197–9206, 2019.

[Xie *et al.*, 2017] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV*, pages 1369–1378, 2017.

[Xie *et al.*, 2019] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, pages 2730–2739, 2019.

[Xie *et al.*, 2020] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *CVPR*, pages 12193–12202, 2020.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Mingyu Ding, Ruimao Zhang, and Ping Luo. Polarmask++: Enhanced polar representation for single-shot instance segmentation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5385–5400, 2021.

[Xu *et al.*, 2021] Xiaogang Xu, Hengshuang Zhao, and Jiaya Jia. Dynamic divide-and-conquer adversarial training for robust semantic segmentation. In *ICCV*, pages 7486–7495, 2021.

[Xu *et al.*, 2022] Haiming Xu, Lingqiao Liu, Qiuchen Bian, and Zhen Yang. Semi-supervised semantic segmentation with prototype-based consistency regularization. *NeurIPS*, 35:26007–26020, 2022.

[Yang *et al.*, 2022] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Boosting transferability of targeted adversarial examples via hierarchical generative networks. In *ECCV*, pages 725–742. Springer, 2022.

[Yu *et al.*, 2022] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *ECCV*, 2022.

[Zhang *et al.*, 2022a] Jie Zhang, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Lei Zhang, and Chao Wu. Towards efficient data free black-box adversarial attack. In *CVPR*, pages 15115–15125, 2022.

[Zhang *et al.*, 2022b] Zhaoxin Zhang, Shize Huang, Xiaowen Liu, Bingjie Zhang, and Decun Dong. Adversarial attacks on yolact instance segmentation. *Computers & Security*, 116:102682, 2022.

[Zhang *et al.*, 2023a] Yuxuan Zhang, Wei Yang, and Rong Hu. Baproto: Boundary-aware prototype for high-quality instance segmentation. In *ICME*, pages 2333–2338. IEEE, 2023.

[Zhang *et al.*, 2023b] Yuxuan Zhang, Wei Yang, and Shaowei Wang. Fgnet: towards filling the intra-class and inter-class gaps for few-shot segmentation. In *IJCAI*, pages 1749–1758, 2023.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017.

[Zhou *et al.*, 2017] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017.