# HPattack: An Effective Adversarial Attack for Human Parsing

Xin Dong[1,2], Rui Wang[1,2](✉), Sanyi Zhang[1,2], and Lihua Jing[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences,
No.19 Shucun Road, Haidian District, Beijing 100085, China
[2] School of Cyber Security, University of Chinese Academy of Sciences,
Beijing, China
{dongxin,wangrui,zhangsanyi,jinglihua}@iie.ac.cn

**Abstract.** Adversarial attacks on human parsing models aim to mislead deep neural networks by injecting imperceptible perturbations to input images. In general, different human parts are connected in a closed region. The attacks do not work well if we directly transfer current adversarial attacks on standard semantic segmentation models to human parsers. In this paper, we propose an effective adversarial attack method called HPattack, for human parsing from two perspectives, *i.e.*, sensitive pixel mining and prediction fooling. By analyzing the characteristics of human parsing tasks, we propose exploiting the human region and contour clues to improve the attack capability. To further fool the human parsers, we introduce a novel background target attack mechanism by leading the predictions away from the correct label to obtain high-quality adversarial examples. Comparative experiments on the human parsing benchmark dataset have shown that HPattack can produce more effective adversarial examples than other methods at the same number of iterations. Furthermore, HPattack also successfully attacks the Segment Anything Model (SAM) model.

**Keywords:** Adversarial attack · Human parsing · Sensitive pixel mining · Prediction fooling

## 1 Introduction

In the past few years, many adversarial attack methods have been proposed to generate adversarial examples for image classification tasks [8,15,19–21]. The core idea is implementing very small perturbations on the clean image to mislead the deep neural networks. While for dense prediction tasks, for example, the goal of the adversarial attacks on the semantic segmentation task [3–6] dedicates to segmenting each pixel as the wrong class, they are also vulnerable to adversarial attacks similar to the classification task [2,9–13,26,27]. Human parsing [17,22] is a specific semantic segmentation task that focuses on recognizing each pixel of the human regions with the correct human part, which has great potential

S. Rudinac et al. (Eds.): MMM 2024, LNCS 14555, pp. 28–41, 2024.
https://doi.org/10.1007/978-3-031-53308-2_3

in the areas of shopping platforms, human-computer interaction, image editing, and posture analysis in medical rehabilitation domain, and so on. Like semantic segmentation models, human parsers are still vulnerable to adversarial attacks, which in turn can cause applications that use human parsers to give incorrect information or be paralyzed, creating a great potential for life. Therefore, the study of adversarial examples against human parsers is an important research step to prevent it from being attacked.

Many existing adversarial attack methods designed for semantic segmentation can not achieve good attack results on the human parsing models, so there is still lots of room for improvement since they ignore the independent characteristics of the human parsing task. The core difference between human parsing and semantic segmentation is that human parsing focuses on segmenting the human region, and the pixels outside the human region are grouped into one class, *i.e.*, the background class. And the human parts are usually connected in a closed spatial region. An effective adversarial attack should meet the conditions of implementing imperceptible perturbations on the vulnerable pixels and misleading more pixels to be classified as the wrong classes. Motivated by these inherent properties and the attack demand, in this paper, we propose an effective adversarial attack method for human parsing called **HPattack**, it formulates two distinct mechanisms into a unified framework, *i.e.*, sensitive pixel mining and prediction fooling. For sensitive pixel mining, we find that the loss of pixels in the background region impedes improving the attack capability, so pixel selection should ignore this part. Apart from that, the pixels in the contour region [4,22] are wrongly recognized will lead to the whole segmentation accuracy decrease. Thus, the adversarial attack method should not only focus on the pixels inside the human region but also emphatically perturb the pixels in the contour region. Except for pixel mining, misleading the parsing predictions is also essential to further improve the attack capability. The dodging attack [7,24] and impersonation attack [1,23] are designed to keep the prediction results away from the ground truth, and close to a pre-defined target, respectively. To inherit the advantages of dodging and impersonation attacks, we introduce a novel background target attack fusing two types of attacks together to guide the generation of better adversarial examples, *i.e.*, the prediction results of each pixel are close to the background class. Extensive experiments on the human parsing benchmark dataset LIP [18] have shown the effectiveness of the proposed HPattack, it can achieve the best attack performance over state-of-the-art adversarial attack methods. In addition, we also conduct experiments on the Segment Anything model (SAM) [14], the proposed HPattack obtains better attack results than other methods. The main contributions can be summarized in the following:

- We propose an effective adversarial attack method for human parsing, HPattack, which can exploit and inject imperceptible perturbations on sensitive human body pixels and effectively fool human parsers into classifying more pixels as wrong labels.

- The human body and contour regions are introduced to exploit useful pixels for decreasing the accuracy of the whole human parsing. We find that pixels in the background region are harmful, and those in the contour region are conducive to the adversarial attack.
- A novel background target attack that combines dodging and impersonation attacks is proposed to further enhance the attack capability for generating high-quality adversarial examples.
- Extensive experimental results show that HPattack can generate effective adversarial examples. It achieves over 93.36% success rate using only three iterations. In addition, HPattack is also more capable of attacking SAM than other methods.

## 2 Related Work

### 2.1 Adversarial Attacks on Segmentation Models

Adversarial attacks aim at adding small and imperceptible perturbations to the input of a deep neural network to generate adversarial examples that interfere with the model's capabilities. In semantic segmentation [4,6,30], most of the current work is dedicated to discovering the vulnerability of semantic segmentation models using existing adversarial attack methods [9,13,26,27]. In addition, the study [2] proposes a method for benchmarking the robustness of segmentation models, showing that segmentation models have different performances under FGSM [8] and BIM [15] attacks than classification methods, and demonstrating the robustness of segmentation models under multi-scale transformations. Hendrik *et.al.* [12] used universal adversarial perturbation to generate adversarial examples against semantic segmentation. MLAttack [11] adds the loss of the output of multiple intermediate layers of the model to avoid the impact of multi-scale analysis on attack performance. SegPGD [10] generates effective and efficient adversarial examples by analyzing the relationship between correctly and incorrectly classified pixels, calculated the loss function of both separately, and tested the capability to attack against robust semantic segmentation models, and found that these adversarially trained models are not resistant to the adversarial attack of SegPGD. All these works show that existing semantic segmentation models are highly against well-designed adversarial examples, and it is essential to study more effective attack methods and apply them to defense such as adversarial training. In this paper, the proposed approach is mainly improving the attack capability of white-box attacks on human parsing.

### 2.2 Human Parsing

Human parsing models aims to classify each image pixel as the correct human part, i.e., a pixel point at a given location is judged to belong to which of the categories of head, arm, leg, etc. The current solutions mainly exploit valuable clues to improve the parsing capability. Researchers have developed various useful clues, such as multi-scale context information is useful to solve various scale

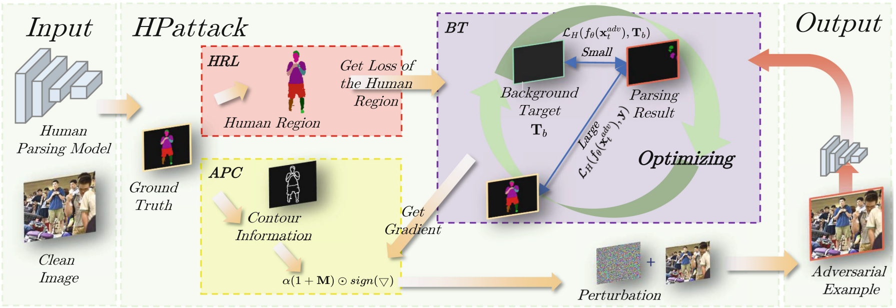
Input
Human Parsing Model
Clean Image
HPattack
Ground Truth
HRL
Human Region
Get Loss of the Human Region
APC
Contour Information
$\alpha(1+\mathbf{M}) \odot sign(\triangledown)$
Get Gradient
BT
$\mathcal{L}_H(f_\theta(\mathbf{x}_t^{adv}), \mathbf{T}_b)$
Small
Background Target $\mathbf{T}_b$
Parsing Result
Large $\mathcal{L}_H(f_\theta(\mathbf{x}_t^{adv}), \mathbf{y})$
Optimizing
Perturbation
+
Output
Adversarial Example

**Fig. 1.** The framework of HPattack. The input image is first fed into the pixel mining module, *i.e.*, the human region and the contour region (red and yellow background area), which leads the HPattack focusing on the sensitive pixels. Then the background target attack (purple background area) is further employed to optimize the model generating error prediction results. Finally, HPattack outputs a new adversarial image with imperceptible perturbation.

problems (*e.g.*, JPPNet [18] exploited the correlation between the two tasks of parsing and pose estimation, making them mutually reinforcing, CE2P [22] added edge information into the loss function to enhance the ability of human parsing, PGECNet [28] employed a Gather-Excite operation to accurately reflect relevant human parts of various scales), edge information across adjacent human classes helps to obtain better boundaries [22], fusing edge, human pose and parsing together *e.g.*, CorrPM [29], and human hierarchy structure [16,25]. In addition, some other mechanisms are also proposed, for example, self-correction mechanism SCHP [17] is proposed to utilize the pseudo label assistance and edge consistency, and so on.

## 3 HPattack: Adversarial Attack for Human Parsing

The proposed HPattack contains two main parts, *i.e.*, adversarial sensitive pixel mining (human body and contour regions) and prediction fooling (background target attack), the detailed HPattack is shown in Fig. 1.

### 3.1 PGD for Human Parsing

In this section, we first make a formulaic statement on the adversarial attack for human parsing by the baseline PGD [19]. In human parsing, given a human parsing model $f_\theta$ parameterized by $\theta$, a clean input image $\mathbf{x}^{clean} \in \mathbb{R}^{H\times W\times 3}$ and its corresponding ground truth $\mathbf{y} \in \mathbb{R}^{H\times W}$, each value of $\mathbf{y}$ belongs to $C = \{0, 1, ..., N-1\}$, where $N$ is the total number of categories. The model classifies each pixel of the input image $f_\theta\left(\mathbf{x}^{clean}\right) \in \mathbb{R}^{H\times W\times N}$, where $(H, W)$ is the sizes of the input image. PGD for human parsing aims at finding an adversarial example $\mathbf{x}^{adv}$ that can allow the model to misclassify all pixels in the input image, which can be formulated as:

$$\arg\max_{\mathbf{x}^{adv}} \mathcal{L}^{CE}\left(f_\theta\left(\mathbf{x}^{adv}\right), \mathbf{y}\right), \quad ||\mathbf{x}^{adv} - \mathbf{x}^{clean}||_p < \epsilon, \tag{1}$$

where $\mathcal{L}^{CE}$ is the cross-entropy loss value, the difference between adversarial example $\mathbf{x}^{adv}$ and clean image $\mathbf{x}^{clean}$ needs to be $\epsilon$-constraint at the $L_p$ parametrization to ensure that the added perturbation remains imperceptible.

Specifically, PGD [19] creates adversarial examples for human parsing that can be represented as:

$$\mathbf{x}^{adv}_{t+1} = \mathrm{Clip}_{\epsilon,\mathbf{x}^{clean}}\left(\mathbf{x}^{adv}_t + \alpha \times sign\left(\bigtriangledown_{\mathbf{x}^{adv}_t}\mathcal{L}^{CE}\left(f_\theta\left(\mathbf{x}^{adv}_t\right), \mathbf{y}\right)\right)\right), \tag{2}$$

where $\alpha$ and $\epsilon$ represent each iteration's step size and perturbation range. $\mathbf{x}^{adv}_t$ is the adversarial example generated at the $t$-th iteration, and the initial value $\mathbf{x}^{adv}_0$ is set to $\mathbf{x}^{adv}_0 = \mathbf{x}^{clean} + \mathcal{U}(-\epsilon, +\epsilon)$, *i.e.*, random initialization of the perturbation on the input image. The $\mathrm{Clip}(\cdot)$ function constrains the perturbation to the $L_p$ parametrization under the $\epsilon$-constraint. $\mathcal{L}^{CE}$ is required to get gradually larger during the attacking process, *i.e.*, leading adversarial perturbation added according to the direction of the gradient.

### 3.2 Adversarial Pixel Mining of HPattack

***Human Region Loss of HPattack (HRL).*** For the adversarial attack on human parsing, the core goal is to misclassify as many pixels of the human region as possible, and misclassification of the background region is not necessary. However, the background pixels may take up a large ratio of the whole image, so the loss value of the background's pixels will affect the optimization direction of the pixel gradient in the human region and the attack capability of the adversarial example. To reduce the influence, we ignore the pixels of the background region computed in the loss function to guide the adversarial example generation. Thus, original loss $\mathcal{L}$ can be expressed as the sum of two separate components, *i.e.*,

$$\mathcal{L} = \mathcal{L}_{B+H} = \frac{1}{H \times W}\left(\sum_{i \in S_B} \mathcal{L}^{CE}_i + \sum_{j \in S_H} \mathcal{L}^{CE}_j\right), \tag{3}$$

where $\mathcal{L}^{CE}$ is the cross-entropy loss for each pixel, $S_B$ is the set of pixels in the background region, and $S_H$ is the set of pixels in the human region.

Based on the above considerations, we first use ground truth $\mathbf{y}$ to select pixels that are in the human region, then keep only the loss of these pixels to guide the generation of the adversarial example, as Eq. (4) shows.

$$\mathcal{L}_H = \frac{1}{H \times W}\sum_{j \in S_H} \mathcal{L}^{CE}_j. \tag{4}$$

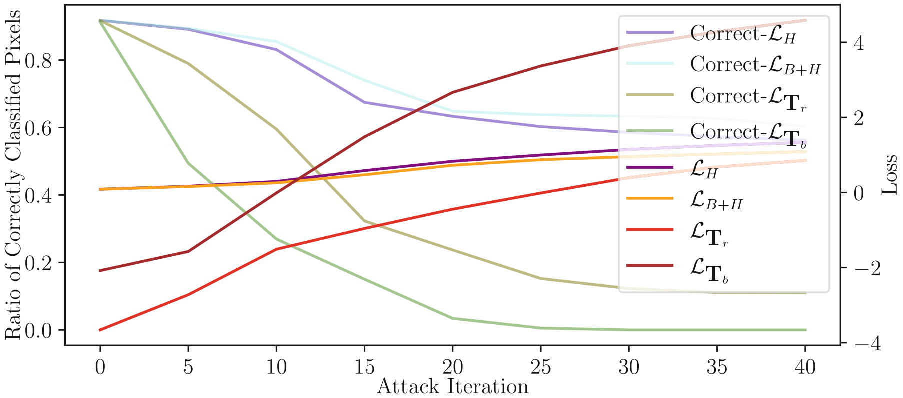

**Fig. 2.** The effectiveness validation of the loss design implemented on pixels only in the human region ($\mathcal{L}_H$) and background target attack ($\mathcal{L}_{\mathbf{T}_b}$) to improve the attack capability of PGD when attacking CE2P.

To demonstrate the validity of Eq. (4), we compare the attack capability of Eq. (4) and Eq. (3) as loss functions in PGD, respectively. We calculate the ratio of pixels still correctly classified by CE2P [22] and the value of the loss function. As shown in Fig. 2, Eq. (4) has fewer correctly classified pixels and larger loss values compared to Eq. (3). This indicates that the loss of pixels in the background region affects the direction of the gradient and blocks the improvement of attack capability.

***Aggravated Perturbation on Contour of HPattack (APC)***. The correct classification of the contour regions can sharpen and refine the segmentation results and thus further help the pixels within each class to be correctly classified, it has been widely verified for improving the performance of human parsing and semantic segmentation models [6,22]. Inspired by this, we add one extra perturbation to the pixels belonging to the contour region of the clean image. Specifically, we extract contour mask $\mathbf{M} \in \mathbb{R}^{H \times W}$ for the ground truth of the input image as shown in Fig. 1 via

$$\mathbf{M} = g\left(\mathbf{y}\right), \tag{5}$$

where $g(\cdot)$ is a contour extraction algorithm used in CE2P [22]. The value $\mathbf{M}_i \in \{0, 1\}$ for each position, 1 means that the point at that position lies on the contour, and 0 means that it does not. Then we add one more perturbation value to the points on the contour via

$$\mathbf{x}_{t+1}^{adv} = \mathrm{Clip}_{\epsilon,\mathbf{x}^{clean}}\left(\mathbf{x}_t^{adv} + \alpha\left(1 + \mathbf{M}\right) \odot sign\left(\bigtriangledown_{\mathbf{x}_t^{adv}} \mathcal{L}_H\left(f_\theta\left(\mathbf{x}_t^{adv}\right), \mathbf{y}\right)\right)\right). \tag{6}$$

### 3.3 Background Target Attack of HPattack

The prediction quality of human parsing directly relates to attack capability. Thus, we introduce the background target attack to mislead the human parsers predicting wrong classes. There are two types of adversarial attacks, *i.e.*, untargeted attack (dodging attack) and targeted attack (impersonation attack). The PGD attack only implements the dodging attack, which makes the prediction result of each pixel as far away from the correct label as possible. Still, it will cause a lack of attack capability because the distance away from the correct label is insufficient each time. Thus, it is also necessary to add the impersonation attack to make the prediction result close to a pre-defined target, which is as far away as possible from the correct label. By forcing the prediction result to be as far away as possible from the correct label, the attack capability will be significantly enhanced.

Since dense human parsing focuses on classifying each pixel as the correct label, we only set a target value for each pixel for convenience. We have stated that the pixels in the background region limit the attack ability. Thus, we set the target for each pixel in the human region to be 0, *i.e.*, leading the model to classify pixels of the human region as background class. We define it as background target $\mathbf{T}_b \in \mathbb{R}^{H \times W}$, where each value is 0. So the loss function $\mathcal{L}_H$ in Eq. (6) can be replaced as

$$\mathcal{L}_{\mathbf{T}_b} = \mathcal{L}_H\left(f_\theta\left(\mathbf{x}_t^{adv}\right), \mathbf{y}\right) - \mathcal{L}_H\left(f_\theta\left(\mathbf{x}_t^{adv}\right), \mathbf{T}_b\right), \tag{7}$$

where $\mathcal{L}_H(f_\theta(\mathbf{x}_t^{adv}), \mathbf{y})$ represents the loss of the dodging attack, which has to become larger to move away from the correct labels. $\mathcal{L}_H(f_\theta(\mathbf{x}_t^{adv}), \mathbf{T}_b)$ represents the loss of impersonation attack, which has to keep getting smaller as to move closer to the background target, we define this module as **BT**.

Regarding the different choices of the target label, in addition to choosing the background class 0, it is also possible to randomly select any class other than the correct label for each pixel. Specifically, we design a comparison that randomly chooses another class as the target label. For a pixel $j \in S_H$ with a correct class $\mathbf{y}_j$, then randomly choose from $C \setminus \mathbf{y}_j$. Combining the results of random selection per pixel, we denote this random target as $\mathbf{T}_r \in \mathbb{R}^{H \times W}$.

Again, we compared the two choices of the target in Fig. 2 and find that the loss function of $\mathbf{T}_b$ increases more than twice as much as that of $\mathbf{T}_r$. The ratio of pixels correctly classified by $\mathbf{T}_b$ decreases faster than that of $\mathbf{T}_r$ and is already close to 0% at the 40-th iteration, which shows that $\mathbf{T}_b$ will be more capable of attacking than $\mathbf{T}_r$. After that, a detailed comparison experiment of their attacking capability is conducted in Subsect. 4.3.

In summary, HPattack first use ground truth to select pixels belonging to the human region and get the contour mask, then calculates the loss of the pixels in the human region with background target and ground truth, finally finding the gradient of the loss function $\mathcal{L}_{\mathbf{T}_b}$ with respect to $\mathbf{x}_t^{adv}$ and adding one more perturbation at the contour region. The overall flow is shown in Alg. 1.

**Algorithm 1.** HPattack

```
Input: The human parsing model $f_\theta$, the loss function $\mathcal{L}_{\mathbf{T}_b}$, a clean image $\mathbf{x}^{clean}$, the number of iteration $T$, the corresponding ground truth $\mathbf{y}$.
Output: The adversarial example $\mathbf{x}^{adv}$.
$\mathbf{x}_0^{adv} \leftarrow \mathbf{x}^{clean} + \mathcal{U}(-\epsilon, +\epsilon)$;
for $t = 0$ to $T - 1$ do
    Get pixels of human region and calculate the $\mathcal{L}_{\mathbf{T}_b}$ via $\mathcal{L}_{\mathbf{T}_b} = \mathcal{L}_H\left(f_\theta\left(\mathbf{x}_t^{adv}\right), \mathbf{y}\right) - \mathcal{L}_H\left(f_\theta\left(\mathbf{x}_t^{adv}\right), \mathbf{T}_b\right)$;
    Compute the gradient $\bigtriangledown_{\mathbf{x}_t^{adv}} \mathcal{L}_{\mathbf{T}_b}$;
    Get contour mask $\mathbf{M}$ and generate adversarial example $\mathbf{x}_{t+1}^{adv}$ via $\mathbf{x}_{t+1}^{adv} = \mathrm{Clip}_{\epsilon,\mathbf{x}^{clean}}\left(\mathbf{x}_t^{adv} + \alpha(1 + \mathbf{M}) \odot sign\left(\bigtriangledown_{\mathbf{x}_t^{adv}} \mathcal{L}_{\mathbf{T}_b}\right)\right)$;
end for
return $\mathbf{x}^{adv}$.
```

# 4 Experiments

In this section, we first introduce the experimental setup, and then divide experiments into evaluation of attack capability, ablation study, attack segment anything model, and performance against defense methods, all of which indicate that our HPattack has high performance.

## 4.1 Experimental Settings

We choose the large-scale human parsing benchmark dataset LIP [18] to verify the effectiveness of the proposed HPattack. There are 50,462 images in total, 30,462/10,000/10,000 images are chosen for training, validation, and testing, respectively. All images are annotated with categories $C = \{0, 1, ..., 19\}$, *i.e.*, a background class 0, and 19 categories belonging to the human region. Specifically, we implement adversarial attacks on the LIP validation set.

To verify the effectiveness of the attack, we choose two state-of-the-art models, CE2P [22] and SCHP [17], as the two attacked models for human parsing. The evaluation metric is the standard mIoU [22], which is commonly used in human parsing, *i.e.*, IoU is first calculated for each category, followed by the average value. We use mIoU for $C \setminus 0$ to reflect the attack success rate, the smaller the value of mIoU, the higher the attack success rate. The comparison methods we chose include two traditional adversarial attack methods, BIM [15] and PGD [19], and also two adversarial attack methods, MLAttack [11] and SegPGD [10] for semantic segmentation.

For the setting of hyperparameters, we choose a maximum perturbation value of $\epsilon = \frac{8}{255}$, a step size of $\frac{\epsilon}{T}$, and a constraint of $L_\infty$ according to the setting of SegPGD [10]. We test the attack capability from the number of iterations $T$ starting from 3 and up to 100.

**Table 1.** Comparison of the attacking capability, the smaller the mIoU score means the stronger the attack capability, the best results are shown in bold.

| Model | | mIoU | | | | | | |
|---|---|---|---|---|---|---|---|---|
| CE2P [22] | Clean | 52.86 | | | | | | |
| | Method/iter | 3 | 5 | 7 | 10 | 20 | 40 | 100 |
| | BIM [15] | 15.34 | 12.73 | 12.15 | 10.01 | 9.23 | 8.47 | 7.82 |
| | PGD [19] | 14.70 | 11.85 | 11.47 | 9.92 | 8.63 | 7.77 | 7.33 |
| | MLAttack [11] | 13.11 | 10.70 | 10.31 | 8.64 | 7.48 | 6.13 | 5.79 |
| | SegPGD [10] | 10.06 | 7.80 | 7.13 | 5.93 | 5.07 | 4.57 | 4.03 |
| | HPattack (Ours) | **3.20** | **1.43** | **0.64** | **0.37** | **0.16** | **0.11** | **0.09** |
| SCHP [17] | Clean | 57.06 | | | | | | |
| | Method/iter | 3 | 5 | 7 | 10 | 20 | 40 | 100 |
| | BIM [15] | 19.03 | 17.98 | 16.90 | 14.83 | 13.99 | 12.58 | 12.10 |
| | PGD [19] | 18.17 | 15.48 | 14.26 | 13.36 | 12.36 | 11.29 | 11.05 |
| | MLAttack [11] | 16.74 | 13.14 | 12.79 | 11.46 | 10.93 | 10.57 | 9.91 |
| | SegPGD [10] | 10.10 | 8.59 | 8.03 | 7.62 | 6.90 | 6.37 | 5.86 |
| | HPattack (Ours) | **4.12** | **2.90** | **2.37** | **2.05** | **1.51** | **1.22** | **1.18** |

### 4.2 Evaluation of Attack Capability

In this section, we conduct comparative experiments to show the effectiveness of our HPattack. Specifically, we first test the models' parsing capability on the clean images and then compare our HPattack with four state-of-the-art adversarial attack methods.

As shown in Table 1, the results for the clean images of CE2P [22] and SCHP [17] are 52.86 and 57.06, respectively. It achieves the worst attack capability for the BIM [15] attack, while PGD [19] has a better attack effect than BIM. MLAttack [11] reduces to about 5.79 at the 100-th iteration tested on the CE2P model, but SegPGD [10] method can reduce to about 5.93 just with 10 iterations. If we adopt the attack mechanism of our HPattack, it just needs three iterations which attack the CE2P [22] and the SCHP [17] both drop too much less than 5. This indicates that our HPattack method can generate adversarial examples with strong attack capability in a short time. In addition, if we increase more iterations implemented with our HPattack, the attack success rate can close to 100%, *i.e.*, let mIoU be almost 0. This means that the human parsing model is almost invalid.

Besides, we also provide qualitative comparison results in Fig. 3, where segmentation results are obtained by testing adversarial examples on the CE2P and SCHP model with 3 attack iterations. In particular, we choose the SegPGD as a comparison, which performed better in quantitative experiments than other comparison methods. It can be observed that the human parsing results of the HPattack misclassified almost all pixels of the whole image region compared to the SegPGD method.

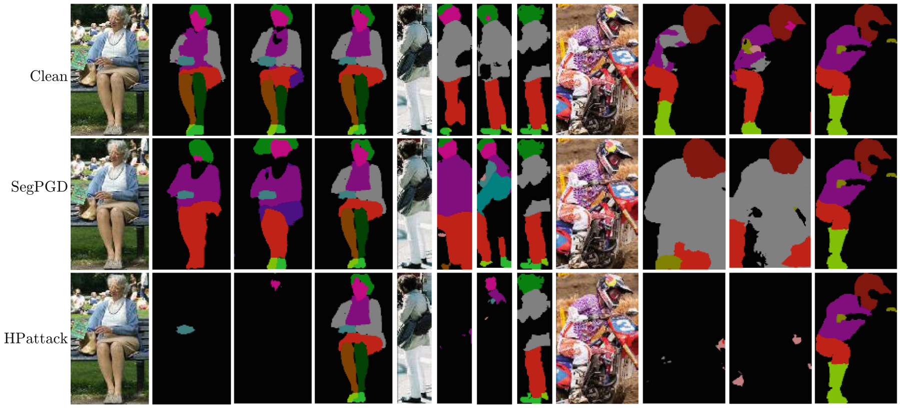

**Fig. 3.** The qualitative analysis of the adversarial attack. For each colorful image, the first column is itself, the second is the prediction result of CE2P on it, the third is the prediction result of SCHP on it, and the forth is the ground truth.

Furthermore, since we add the additional perturbations to the contour region in our HPattack, there is a risk of making the perturbation more prominent and thus resulting in the adversarial examples being vulnerable to human perception. Still, as shown in Fig. 3, the generated adversarial examples have no significant perturbations compared to the clean images and the adversarial examples generated by SegPGD. Overall, it can be seen that the attack capability of HPattack is significantly better than SegPGD.

## 4.3 Ablation Study

**Table 2.** Ablation study for HPattack, the smaller the mIoU, the stronger the attack capability, the best results are shown in bold.

| Model | Method | mIoU | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 3 | 5 | 7 | 10 | 20 | 40 | 100 |
| CE2P [22] | PGD [19] | 14.70 | 11.85 | 11.47 | 9.92 | 8.63 | 7.77 | 7.33 |
| | +HRL | 10.49 | 8.11 | 7.05 | 5.09 | 4.29 | 3.37 | 3.05 |
| | +HRL+APC | 10.02 | 7.74 | 6.16 | 5.00 | 4.03 | 3.34 | 3.01 |
| | +HRL+APC+RT | 7.16 | 5.10 | 3.42 | 3.14 | 1.76 | 1.33 | 0.91 |
| | +HRL+APC+IMA | 4.77 | 2.13 | 1.35 | 1.07 | 0.66 | 0.53 | 0.49 |
| | +HRL+APC+BT (HPattack) | **3.20** | **1.43** | **0.64** | **0.37** | **0.16** | **0.11** | **0.09** |
| SCHP [17] | PGD [19] | 18.17 | 15.48 | 14.26 | 13.36 | 12.36 | 11.29 | 11.05 |
| | +HRL | 11.98 | 9.34 | 8.51 | 7.69 | 6.81 | 5.97 | 5.58 |
| | +HRL+APC | 11.23 | 8.66 | 7.85 | 7.03 | 6.13 | 5.23 | 4.87 |
| | +HRL+APC+RT | 7.58 | 6.83 | 5.00 | 4.08 | 3.82 | 3.47 | 3.18 |
| | +HRL+APC+IMA | 6.59 | 3.67 | 2.91 | 2.58 | 1.92 | 1.69 | 1.55 |
| | +HRL+APC+BT (HPattack) | **4.12** | **2.90** | **2.37** | **2.05** | **1.51** | **1.22** | **1.18** |

Our HPattack method consists of three components: the loss of human region (HRL), aggravated perturbation on contour region (APC), and background target attack (BT). Thus, we conduct ablation experiments on these three components to test the attack capability for attacking CE2P [22] and SCHP [17]. The baseline method is implemented with the standard PGD attack method. Then we add three modules to the PGD inch by inch, and the detailed comparison results are shown in Table 2. To further explain the effectiveness of the background attack mechanism, we provide a comparison that replaces the background target $\mathbf{T}_b$ with a random target $\mathbf{T}_r$, and we denote it as (+HRL+APC+RT). In addition, to test the effectiveness of using the minus sign to combine impersonation attacks, we tested the attack capability of the second part in Eq. 7 ($\mathcal{L}_{\mathbf{T}_b}$), *i.e.*, $-\mathcal{L}_H(f_\theta(\mathbf{x}_t^{adv}), \mathbf{T}_b)$, denoted as (+HRL+APC+IMA). If (+HRL+APC+IMA) leads to a very significant improvement, then it shows that using the minus sign for combining does not affect performance.

As shown in Table 2, the human region loss (HRL) module reduces the mIoU to two-thirds of PGD, and aggravated perturbation on contour (APC) only reduces the mIoU by about 0.5 compared to (+HRL) because it only adds one more perturbation to the pixels in the contour region. In contrast, adding the background target (BT) can obtain a performance drop of about 80% than the PGD result tested on the CE2P model with 3 attack iterations, which is the most significant improvement of the attack capability. From Table 2, we can observe that the random target (RT) will play some role in reducing mIoU relative to human region loss and aggravated perturbation on contour. However, the performance reduction is not significant enough relative to the background target attack, which verifies the analysis in Sect. 3.3 that the random target is not as effective as the background target in improving the attack capability. The mIoU scores of (PGD+HRL+APC+IMA) show that using the minus sign as the combination strategy can reduce mIoU to below 5 in many cases, *i.e.*, adopting the minus sign mechanism is effective.

### 4.4 Attack Segment Anything Model

We have shown the effectiveness of the HPattack on the standard human parsing models, an interesting discussion is whether the HPattack mechanism can be transferred to the large-scale pre-trained segmentation model, such as Segment Anything Model (SAM) [14]. SAM is a state-of-the-art image segmentation model trained with 11 million images and 1.1 billion masks, and SAM has very high accuracy in various semantic segmentation tasks. Specifically, we choose the pre-trained ViT-Base model of SAM to attack and use the SegPGD [10] as a comparison method, which performed relatively better in quantitative experiments. We should note that the current SAM can only output the segmentation result of the whole image, but the actual class of each pixel is not given.

The detailed comparison results are shown in Fig. 4. The highlighted regions with red circles show that HPattack fools the SAM model with worse segmentation results than SegPGD. For example, the first, second, and fourth columns of the image after being attacked by our HPattack compared to SegPGD, will

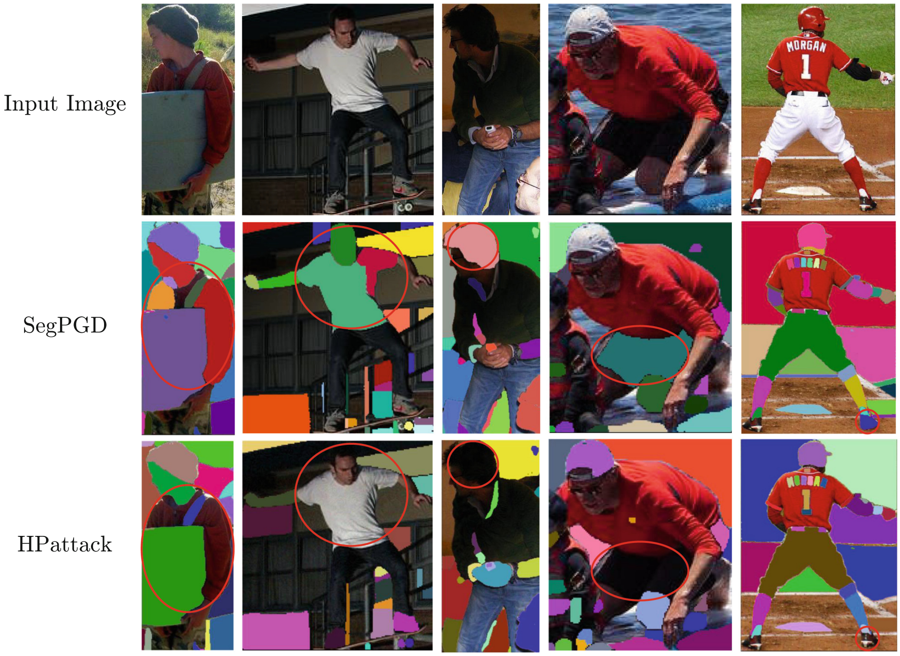

**Fig. 4.** Comparison of attacking SAM. HPattack performs better than SegPGD.

make SAM unable to segment the upper clothes or pants; the third and fifth columns will make the SAM fail to segment hair and shoes, respectively.

## 5 Conclusion

In this paper, we propose an effective adversarial attack method for human parsing, which boosts the adversarial attack capability by progressively combining the loss of human region, the aggravated perturbation on contour, and the background target attack. HPattack owns good attack ability on the human parsing models and applies to large-scale pre-trained segmentation model SAM. In conclusion, our approach can provide a meaningful reference for subsequent research on human parsing and even semantic segmentation.

**Acknowledgements.** This work is supported in part by the National Natural Science Foundation of China Under Grants No.62176253, No.62202461 and China Postdoctoral Science Foundation No.2022M723364.

## References

1. Akhtar, N., Mian, A., Kardan, N., Shah, M.: Advances in adversarial attacks and defenses in computer vision: a survey. IEEE Access **9**, 155161–155196 (2021)
2. Arnab, A., Miksik, O., Torr, P.H.: On the robustness of semantic segmentation models to adversarial attacks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 888–897 (2018)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs (2014). arXiv preprint arXiv:1412.7062
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
5. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation (2017). arXiv preprint arXiv:1706.05587
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
7. Dong, Y., et al.: Efficient decision-based black-box adversarial attacks on face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7714–7722 (2019)
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2014). arXiv preprint arXiv:1412.6572
9. Gu, J., Zhao, H., Tresp, V., Torr, P.: Adversarial examples on segmentation models can be easy to transfer (2021). arXiv preprint arXiv:2111.11368
10. Gu, J., Zhao, H., Tresp, V., Torr, P.H.S.: SegPGD: an effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. ECCV 2022. LNCS, vol. 13689. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19818-2_18
11. Gupta, P., Rahtu, E.: MLAttack: fooling semantic segmentation networks by multi-layer attacks. In: Fink, G.A., Frintrop, S., Jiang, X. (eds.) DAGM GCPR 2019. LNCS, vol. 11824, pp. 401–413. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33676-9_28
12. Hendrik Metzen, J., Chaithanya Kumar, M., Brox, T., Fischer, V.: Universal adversarial perturbations against semantic image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2755–2764 (2017)
13. Kang, X., Song, B., Du, X., Guizani, M.: Adversarial attacks for image segmentation on multiple lightweight models. IEEE Access **8**, 31359–31370 (2020)
14. Kirillov, A., et al.: Segment anything (2023). arXiv preprint arXiv:2304.02643
15. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv e-prints pp. arXiv-1607 (2016)
16. Li, L., Zhou, T., Wang, W., Li, J., Yang, Y.: Deep hierarchical semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1246–1257 (2022)
17. Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 3260–3271 (2022)
18. Liang, X., Gong, K., Shen, X., Lin, L.: Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE Trans. Pattern Anal. Mach. Intell. **41**(4), 871–885 (2018)

19. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks (2017). arXiv preprint arXiv:1706.06083
20. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574–2582 (2016)
21. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 372–387. IEEE (2016)
22. Ruan, T., Liu, T., Huang, Z., Wei, Y., Wei, S., Zhao, Y.: Devil in the details: towards accurate single and multiple human parsing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 4814–4821 (2019)
23. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: A general framework for adversarial examples with objectives. ACM Trans. Priv. Secur. (TOPS) **22**(3), 1–30 (2019)
24. Sun, L., Tan, M., Zhou, Z.: A survey of practical adversarial example attacks. Cybersecurity **1**, 1–9 (2018)
25. Wang, W., Zhou, T., Qi, S., Shen, J., Zhu, S.C.: Hierarchical human semantic parsing with comprehensive part-relation modeling. IEEE Trans. Pattern Anal. Mach. Intell. **44**(7), 3508–3522 (2021)
26. Xiao, C., Deng, R., Li, B., Yu, F., Liu, M., Song, D.: Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 217–234 (2018)
27. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1369–1378 (2017)
28. Zhang, S., Qi, G.J., Cao, X., Song, Z., Zhou, J.: Human parsing with pyramidical gather-excite context. IEEE Trans. Circuits Syst. Video Technol. **31**(3), 1016–1030 (2020)
29. Zhang, Z., Su, C., Zheng, L., Xie, X.: Correlating edge, pose with parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8900–8909 (2020)
30. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)