Fashion-Guided Adversarial Attack on Person Segmentation

Marc Treu*[1], Trung-Nghia Le*[2], Huy H. Nguyen*[2], Junichi Yamagishi[2,3], and Isao Echizen[2,3,4]

[1]TEHTRIS, France
[2]National Institute of Informatics, Japan
[3]The Graduate University for Advanced Studies, SOKENDAI, Japan
[4]University of Tokyo, Japan

Abstract

This paper presents the first adversarial example based method for attacking human instance segmentation networks, namely person segmentation networks in short, which are harder to fool than classification networks. We propose a novel Fashion-Guided Adversarial Attack (FashionAdv) framework to automatically identify attackable regions in the target image to minimize the effect on image quality. It generates adversarial textures learned from fashion style images and then overlays them on the clothing regions in the original image to make all persons in the image invisible to person segmentation networks. The synthesized adversarial textures are inconspicuous and appear natural to the human eye. The effectiveness of the proposed method is enhanced by robustness training and by jointly attacking multiple components of the target network. Extensive experiments demonstrated the effectiveness of FashionAdv in terms of robustness to image manipulations and storage in cyberspace as well as appearing natural to the human eye. The code and data are publicly released on our project page[1].

1. Introduction

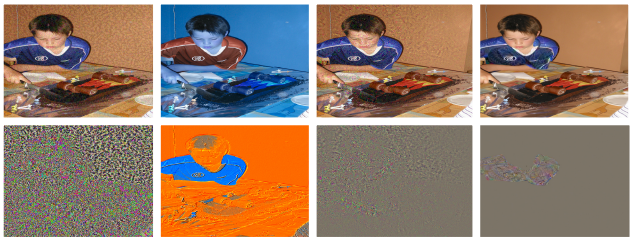
Powerful computer vision methods have been applied to object classification [14, 15], object detection [21, 35], semantic segmentation [19, 25], and instance segmentation [13, 20], enabling machines to perceive the world. With incremental learning using a massive amount of data obtained from the Internet and surveillance cameras, vision-based intelligent systems (*i.e.*, face recognition [39], and gait recognition [41]) can track and learn the behaviors of a large population. Such systems are typically trained using images and videos crawled from social networks without authorization, which raises a serious privacy issue requiring governments and companies to establish policies to prevent unauthorized surveillance and tracking on the basis of user data [39]. Moreover, social network users need a tool to protect their privacy. The tool should be robust against the transformations and compressions that occur during the uploading, sharing, and storing images and videos.

Figure 1. Overview of our proposed Fashion-Guided Adversarial Attack (FashionAdv). Top row: person is segmented successfully. Bottom row: given a target image and a guided-fashion style image, FashionAdv synthesizes natural clothing texture that can nevertheless fool a person segmentation network by making the person invisible to the network.

Several attack methods using adversarial examples have been developed to degrade the performance of deep neural networks in the digital world. Most such methods effectively fool object classification by directly attacking the digital image or video files [3, 36]. Adversarial noise-based attack methods were developed first and became well-established [11, 18, 28, 30]). Subsequently, attacks based on color [36] and texture [3, 8], which also target human perception, have been developed. A few attack methods have

*Equal contributions. This work was done when Marc Treu interned at NII, Japan. Contact email: marc.treu@tehtris.com.

[1]https://github.com/nii-yamagishilab/fashion_adv



(a) FGSM (b) ColorFool (c) tAdv (d) FashionAdv

Figure 2. Effectiveness of our proposed FashionAdv, compared with conventional digital-level adversarial attacks. From left to right: adversarial noise attack [11], adversarial color attack [36], adversarial texture attack [3], and our FashionAdv. Bottom row: corresponding adversarial perturbations. Changes made by FashionAdv are inconspicuous to the human eye.

been developed to target object detection and semantic segmentation [46]. To the best of our knowledge, a method for attacking instance segmentation, which is a combination of object detection and semantic segmentation, has not been developed. In this work, we develop a method for targeted adversarial attacks against human instance segmentation, namely person segmentation.

Naturalness and robustness are two essential properties of adversarial examples [3, 49]. Naturalness means that adversarial examples are unlikely to be noticed by people, while robustness means that they are not paralyzed by data compression, transformation, or image filter effects (*e.g.*, beauty applications). There is a trade-off between these properties. Adversarial examples created using traditional approaches can deal with naturalness [3, 36], but they are easily negated by image transformation [12] or compression [11]. *This work aims to balance this trade-off.*

To this end, we propose an attack method for use against person segmentation network (*i.e.*, YOLACT [4], one of the first networks attempting real-time instance segmentation). Our proposed Fashion-Guided Adversarial Attack (FashionAdv) is designed to synthesize ***adversarial textures*** *by blending generated adversarial noise transferred from customizable fashion style images into the original clothing regions*. The new image with adversarial textures, an adversarial image, is used to spoof the target person segmentation network by making the persons invisible to the network (*c.f.* Fig. 1). Directly optimizing the adversarial texture through infusion of texture from the fashion style image helps make the synthesized adversarial texture inconspicuous and appear natural to the human eye despite their large perturbations. Indeed, a quick visual comparison of the results with FashionAdv with those of conventional adversarial attacks in Fig. 2 indicates that FashionAdv is capable of generating highly inconspicuous adversarial textures. Unlike conventional adversarial attacks [3, 11, 36], which attack the entire image, FashionAdv automatically detects the attackable regions (*i.e.*, clothing regions) to minimize the impact

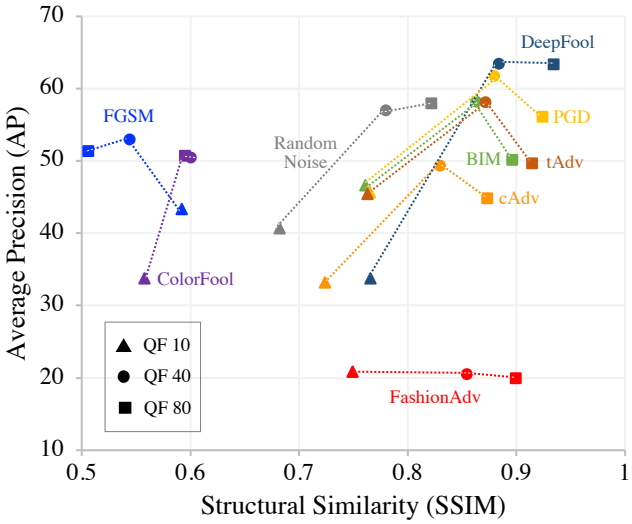


Figure 3. Naturalness (*i.e.*, SSIM [42]) v.s. robustness (*i.e.*, AP [24]) on MS-COCO dataset. FashionAdv significantly outperformed conventional adversarial example methods in terms of robustness against JPEG compression (quality factors 10, 40, and, 80) while maintaining naturalness competitive with that of the conventional methods. Hence, FashionAdv successfully spoofed the target person segmentation network (*i.e.*, YOLACT [4]), and the changes made were inconspicuous to the human eye.

of adversarial examples on image quality. Only clothing regions are identified as attackable regions because (1) these regions are suitable for embedding adversarial noise while maintaining image quality, and (2) human perception is less sensitive to these regions compared with other regions such as faces [22]. FashionAdv effectively works on all persons in the image, hiding them from deep instance segmentation networks. Moreover, we propose jointly attacking two network components, the classification and the segmentation sub-networks. Multiple cues are integrated into the loss function, making the adversarial attack robust against image manipulations and storage in cyberspace.

Extensive experiments on the MS-COCO dataset [24] demonstrated that the *changes made by FashionAdv are not only robust against image manipulations and storage in cyberspace* (*e.g.*, image filters and JPEG compression) *but are also inconspicuous and appear natural to the human eye despite having large perturbations* (*c.f.* Fig. 3). The code and data are made available on our project page[2].

Our contributions are as follows.

- We have developed a novel adversarial attack method for use against person segmentation networks that achieves both robustness and naturalness. Our Fashion-Guided Adversarial Attack (FashionAdv) is the first reported attack method for use against instance segmentation.
- FashionAdv flexibly synthesizes clothing textures

[2] https://github.com/nii-yamagishilab/fashion_adv

from guided-fashion style images that are inconspicuous and appear natural to the human eye despite their large perturbations.
- FashionAdv focuses on only the clothing regions in images, which minimizes the effect on image quality. It is also robust against image manipulations and storage in cyberspace, such as image filtering and JPEG compression, respectively.
- Extensive experiments demonstrated that FashionAdv significantly outperforms conventional methods in term of robustness while still maintaining naturalness.

2. Related Work

2.1. Adversarial Attacks

There are two types of adversarial attacks, including digital-level and physical-level. The former target object classification, while the latter target both object classification and object detection. In digital-level attacks, adversarial perturbations are added to digital images or videos [1]. In physical-level attacks, real adversarial objects are crafted [18]. For both levels, the adversarial perturbations can be noise [26], patterns [8, 16, 48], colors [36, 3], or patches [5, 40, 49].

Most **digital-level adversarial attacks** target object classification [1]. Adversarial noise-based attacks (*e.g*., gradient-based attacks) were developed first and became well-known, including FGSM [11], BIM [18], DeepFool [30], and PGD [28]). FGSM [11] uses gradients of the target neural network to craft adversarial examples. BIM [18] performs multiple FGSM attacks with small step size. DeepFool [30] estimates the minimal adversarial perturbation by using a simple iterative method. PGD [28] treats adversarial attacks as a constrained optimization problem.

Furthermore, attacks based on color and texture (*e.g*., unrestricted attacks), which also target human perception, have been developed. Shamsabadi *et al*. [36] devised ColorFool attack in which colors within a certain range are modified rather than adversarial noise or patches being crafted. Duan *et al*. [8] used both colors and patterns to generate photorealistic adversarial examples. Bhattad *et al*. [3] introduced two attacks that use realistic color perturbation (cAdv) and texture extracted from a source image (tAdv).

Most **physical adversarial attacks** target object detection. The Expectation Over Transformation (EOT) framework [2] was originally developed for object classification, and then was extended to attack object detection. Thys *et al*. [40] used adversarial patches to attack a human detection system using YOLO-v2 [33]. Zhao *et-al.* [49] proposed two types of adversarial patches (*i.e*., hiding attack and appearing attack) to attack Faster RCNN [35] and YOLO-v3 [34]. Zhang *et al*. [48] learned camouflaged adversarial patterns to attack Faster RCNN [35] and YOLO [32]. Following this, Huang *et al*. [16] made camouflage adversarial patterns universal. On the other hand, few adversarial patch-based methods were proposed for attacking object recognition in the real-world, especially road objects, which is important for autonomous driving systems [5, 8, 26]. In addition, few methods have been recently developed to attack person detectors in the real-world by printing adversarial patches on t-shirts [45, 47]. However, these methods look extremely unnatural to the human eye and adversarial patches must be manually attached to other objects (*i.e*. t-shirts).

In this work, *we focus on attacks based on adversarial examples in cyberspace (digital level)*. According to the best of our knowledge, there has been no work on attacking instance segmentation. Therefore, in this work, we target adversarial attacks for use against person segmentation. Particularly, we target YOLACT [4], one of the first networks attempting real-time instance segmentation. YOLACT breaks instance segmentation into two parallel subtasks (*i.e*., generating a set of prototype masks and predicting per-instance mask coefficients) and then linearly combines the prototypes with the mask coefficients. We note that we target only persons segmented from YOLACT.

2.2. Improvements on Robustness and Naturalness

Easily created adversarial examples can fool several kinds of deep neural networks, including ones for object detection and semantic segmentation [1]. However, adversarial perturbations created using traditional approaches can be negated by image transformations [12, 27], or compression [11]. Recent work has been carried out to improve the robustness of adversarial perturbations, especially in the physical world. Physical adversarial objects crafted based on the basis of the EOT framework [2] are robust to a combination of noise, distortion, and affine transformation. The algorithm has thus been integrated into different attacks to improve robustness [5, 31, 44].

A comprehensive benchmark recently performed by Dong *et al*. [7] provides an overview of the robustness of several adversarial attacks. By considering the limitations of pixel-wise and global adversarial attack methods, Dong *et al*. [6] developed an attention-based attack that focuses only on salient objects to improve robustness. Luo *et al*. [26] introduced a human-perception-based metric for estimating the distance between pixels and a greedy optimization algorithm to maximize the noise tolerance of adversarial examples. Zhao *et al*. [49] used feature-interference reinforcement and enhanced realistic constraints generation to enhance the robustness of their proposed hiding attack.

Although naturalness is essential for human perception, this property has not been well explored in the literature. There has been little work dealing with the naturalness of

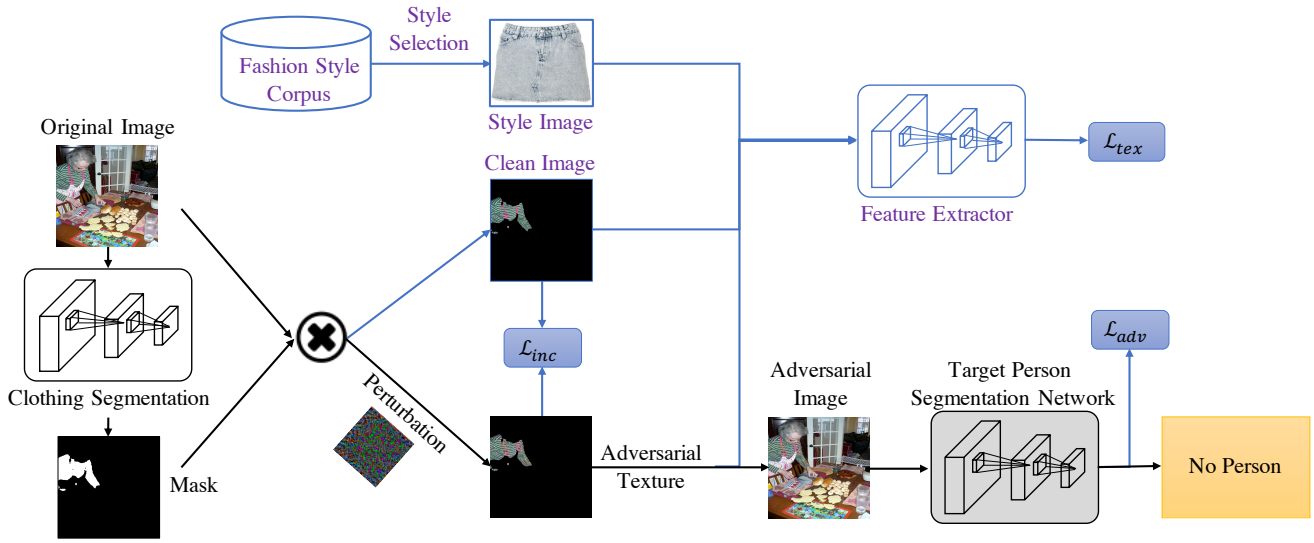


Figure 4. Workflow of proposed FashionAdv framework. Given a target image, our FashionAdv identifies attackable regions in the image and generates an adversarial texture from a guided-fashion style image, which is the image closest to the target image automatically obtained by searching our constructed fashion corpus. The optimized adversarial image is used to spoof a person segmentation network. Blue flow (*i.e.*, blue arrows and blue blocks) illustrates the learning process only.

adversarial attacks. Shamsabadi *et al*. [36] simply modified colors in the target image within ranges that are perceived as natural by humans. Bhattad *et al*. [3] searched texture sources in feature space to find the one most similar to the target image and used it to generate more natural-looking images. Duan *et al*. [8] transferred adversarial perturbations into natural-looking styles, namely camouflage styles that appear natural to human observers. Wu *et al*. [45] trained adversarial patterns from real images and printed them on posters and T-shirts for physical-level attacks.

3. Proposed Method

3.1. Overview

We overview our proposed FashionAdv method by describing how it can be used to attack a state-of-the-art instance segmentation network (*i.e.*, YOLACT [4]), particularly targeting person segmentation (*i.e.*, human instance segmentation). By generating adversarial textures and painting them on clothing regions, FashionAdv should cause YOLACT to fail in segmenting individual persons under diverse perturbation transformation conditions (*i.e.*, image compression, image blurring, noise addition, color variation, and perspective transformation).

Given an image X of size (w, h) consisting of K persons, FashionAdv decomposes the image into person regions P and non-person regions $\bar{P}$. It focuses on only the person regions:

$$P = \{P_k : P_k = X \cdot M_k + X \cdot \bar{M}_k\}_{k=1}^{K}, \quad (1)$$

where P_k indicates the k-th person, and $M_k, \bar{M}_k \in \{0, 1\}^{w,h}$ are binary masks of the attackable regions (*e.g.*, clothing) and unattackable regions (*e.g.*, face, hair, hand, and leg) of person P_k, respectively. These binary masks specify the locations to which the pixels of image X belong, and "$\cdot$" denotes pixel-wise multiplication.

FashionAdv manipulates only the attackable regions (*e.g.*, clothing), corresponding to masks M_k instead of the entire image, resulting in an adversarial image $\tilde{X}$. That is, it creates an adversarial texture for the attackable regions of each person $\tilde{P}_k$:

$$\tilde{P} = \{\tilde{P}_k : \tilde{P}_k = \tilde{X} \cdot M_k + X \cdot \bar{M}_k\}_{k=1}^{K}. \quad (2)$$

Our goal is to generate adversarial image $\tilde{X}$ with both robustness and naturalness. For *naturalness*, FashionAdv synthesizes adversarial textures $\tilde{P}_k$ from fashion style image S and minimizes the distance between $\tilde{P}_K$ with both P_K and S. This learning criterion should yield adversarial textures that look inconspicuous to the human eye and look natural as new fashion clothing:

$$\text{argmin}_{\tilde{X}} \sum_{k=1}^{K} (D_1(\tilde{P}_k, P_k) + D_2(\tilde{P}_k, S)). \quad (3)$$

For *robustness*, FashionAdv uses the expectation over transformation (EOT) framework to formalize the adversarial attack problem [2]. In particular, it optimizes the adversarial image $\tilde{X}$ to minimize the outputs $\Psi_t(\tilde{P}_k)$ of the attacked person segmentation network (*i.e.*, classification score and segmentation mask) in expectation:

$$\text{argmin}_{\tilde{X}} \mathbb{E}_{t\sim T} \sum_{k=1}^{K} \Psi_t(\tilde{P}_k), \quad (4)$$

where t denotes a perturbation transformation, T is the distribution over all possible transformations $\{t\}$, and $\Psi_t(\cdot)$ are the outputs of the attacked person segmentation network (*i.e.*, classification score and segmentation mask) over the transformed image.

3.2. Fashion-Guided Adversarial Attack Framework

Figure 4 depicts the FashionAdv workflow. Given an original target image, FashionAdv performs four basic steps: 1) segment the clothing regions to identify attackable regions in the image, 2) search for and select a fashion style image in a fashion corpus for use in generating an adversarial texture, 3) use the adversarial texture to create an adversarial image that is then optimized for use in an adversarial attack, and 4) use the image to spoof a person segmentation network.

FashionAdv first identifies regions in the target image in which the textures can be modified within an arbitrary range and still look natural to the human eye. The effect on image quality is minimized by focusing on only the clothing regions. These regions are segmented by using the Self-Correction for Human Parsing (SCHP) [23], which exhibited superior performance on the Look into Person (LIP) challenge [10]. We used the off-the-shelf model pre-trained on the LIP dataset [10], which was provided by the authors. FashionAdv combines semantic maps of the "upper-clothes, dress, coat, pants, skirt, and jumpsuits" labels, outputting a binary mask of the clothing regions. This mask is used to protect unattackable regions in the target image.

Next, FashionAdv automatically selects a fashion style image from a fashion style corpus we assembled. To improve the naturalness of the target image's texture, the fashion style image closest in texture to the target image is selected by minimizing the texture transfer cost in the feature space for the target image [3]. We assembled the fashion style corpus by manually selecting different raw images in the DeepFashion2 dataset [9]. We initially selected images of only clothing, without a person in the image. Then, we manually discarded images with patterns or materials similar to those of other images. We ended up with 140 style images in our corpus, each with a distinct style.

FashionAdv then initializes adversarial noise and overlays it on the mask-covered target image, outputting an adversarial texture. Using the adversarial texture and the fashion style image, FashionAdv performs robustness training (*c.f.* Section 3.4) to optimize the adversarial texture directly. In this training process, the adversarial texture is gradually modified to minimize both the classification score and segmentation mask produced by the attacked network while maintaining the texture's naturalness. The adversarial image following mask removal should be robust against various transformations in cyberspace. Finally, the optimized adversarial image is directly fed into the person segmentation network to be attacked.

3.3. Loss Function

The total loss is a combination of adversarial loss $\mathcal{L}_{adv}$ and naturalness loss $\mathcal{L}_{nat}$ in the adversarial image:

$$\mathcal{L} = \alpha\mathcal{L}_{adv} + \mathcal{L}_{nat}. \tag{5}$$

Adversarial loss $\mathcal{L}_{adv}$ is directly obtained from the target person segmentation network (*i.e.*, YOLACT) and is defined as the combination of classification loss $\mathcal{L}_{cls}$ (*i.e.*, softmax loss) and segmentation loss $\mathcal{L}_{mask}$ (*i.e.*, pixel-wise binary cross entropy). This is described in detail elsewhere [4].

Naturalness loss $\mathcal{L}_{nat}$ as introduced here consists of inconspicuous loss $\mathcal{L}_{inc}$ and texture transfer loss $\mathcal{L}_{tex}$. Inconspicuous loss $\mathcal{L}_{inc}$ helps make the adversarial texture inconspicuous to the human eye while texture loss $\mathcal{L}_{tex}$ helps create a new fashion texture for clothing.

$$\mathcal{L}_{nat} = \mathcal{L}_{inc} + \beta\mathcal{L}_{tex}, \tag{6}$$

Inconspicuous loss is used to make the adversarial image $\tilde{X}$ similar to the original image X so that the changes are inconspicuous to the human eye:

$$\mathcal{L}_{inc} = \lambda_1\mathcal{L}_{sim}(X, \tilde{X}) + \lambda_2\mathcal{L}_{tv}(\tilde{X}), \tag{7}$$

where $\mathcal{L}_{sim}$ is computed as the multi-scale structural similarity index measure (MS-SSIM) between $\tilde{X}$ and X [43], and $\mathcal{L}_{tv}$ stands for the total variation loss, which is helpful for reducing noise in the adversarial texture [29].

Inspired by the efficient performance of style transfer [3, 8], we use texture loss to generate a new clothing texture with large perturbations that still looks natural to the human eye. It consists of content loss $\mathcal{L}_c$, which helps preserve the content of the original image X, and style loss $\mathcal{L}_s$, which helps to generate a new style from the fashion style image S:

$$\mathcal{L}_{tex} = \sum_{l \in L_c} w_l\mathcal{L}_c^l(\tilde{X}, X) + \sum_{l \in L_s} w_l\mathcal{L}_s^l(\tilde{X}, S), \tag{8}$$

where l is the l-th feature (*e.g.*, the l-th layer of the network) in the sets of content layers L_c and style layers L_s, and $w_l = 1/C_{l+1}^2$ indicates normalized weights used to configure the layer preferences, with C_l being the number of filter maps in layer l.

The content of the adversarial image $\tilde{X}$ may appear different from that of the original image X. The content loss is used to ensure that the adversarial image can preserve the content of the original image in the deep representation space through mean squared error minimization:

$$\mathcal{L}_c^l(\tilde{X}, X) = \left|\left|\phi_l(\tilde{X}) - \phi_l(X)\right|\right|_2^2, \tag{9}$$

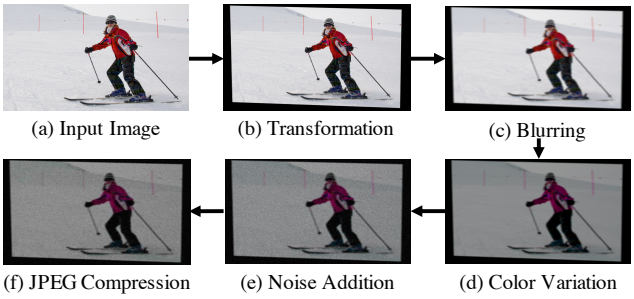


Figure 5. Image perturbation pipeline.

where $\phi_l(\cdot)$ denotes the feature extracted from the l-th layer in a set of content layers L_c of feature extractor ϕ.

The style of the fashion style image is transferred to the adversarial image by using the style loss, which is defined by the difference in their style representations:

$$\mathcal{L}_s^l(\tilde{X}, S) = \frac{\left|\left|\mathcal{G}(\phi_l(\tilde{X}), \phi_{l+1}(\tilde{X})) - \mathcal{G}(\phi_l(S), \phi_{l+1}(S))\right|\right|_2^2}{std(\mathcal{G}(\phi_l(\tilde{X}), \phi_{l+1}(\tilde{X})))}, \tag{10}$$

where $\mathcal{G}$ is the Gram matrix of features extracted from a set of style layers L_s of feature extractor ϕ.

3.4. Robustness Training

The performance of a person segmentation network is affected by fluctuations such as a shift in the viewpoint and by other transformations in the real world. In early work [2], the EOT framework was used to conduct adversarial attacks on classifiers by adding random distortions in the optimization to make the perturbations more robust. In this work, we adapt the EOT framework to train a robust adversarial attack for use against a target person segmentation network. The training was done by simulating realistic situations with a series of random image transformations in cyberspace. We included camera-related transformations for two reasons: (1) they have been proven to be effective in making adversarial perturbations robust [2], and (2) they make our method extendable to the physical domain. From Eqs. (3), (4), and (5), our EOT-based training is defined as:

$$\text{argmin}_{\tilde{X}} \mathbb{E}_{t\sim T} \alpha \mathcal{L}_{adv} + \mathcal{L}_{nat}. \tag{11}$$

Minimizing Eq. 11 results in the optimized adversarial image substantially degrading the performance of the detector, thereby resulting in a high spoofing rate. To improve the generation of adversarial images under various conditions, we used a series of perturbation transformations to simulate the cyberspace condition fluctuations. Figure 5 depicts the image perturbation pipeline.

First, we simulated **perspective transformations** of different camera views by generating a random homography matrix in the range $[-0.2, 0.2]$ to transform the original image. Next, we simulated **image blurring** by using a Gaussian blur with kernel size and standard deviation randomly sampled in $\{1, 3, 5, 7, 9, 11\}$ and $[0.1, 3]$, respectively. We then approximated **color variation** by shifting the color values of each channel in terms of hue, saturation, brightness, and contrast in the range $[-0.02, 0.02]$ for hue and $[-0.2, 0.2]$ for the others. We also took into account **image noise** by using a uniform noise model in the range $[-0.02, 0.02]$. Finally, we applied an **approximation of JPEG compression** [37] with a quality factor (QF) generated uniformly in the range $[18, 22]$. We note that the original image color values were normalized in the range $[0, 1]$, and then the color values after transformations were clipped to $[0, 1]$.

3.5. Implementation Details

This section describes the implementation details of our approach. We implemented FashionAdv in PyTorch and conducted experiments on a computer with 32-GB RAM and a Tesla P100 GPU. We employed the ImageNet pre-trained VGG-19 network [38] as the feature extractor. We used $L_c = \{conv4\_2\}$ to present the content, and $L_s = \{conv1\_1, conv2\_1, conv3\_1, conv4\_1, conv5\_1\}$ to present the style, similar to previous usage [3].

We set the weights of the loss function empirically: $\alpha = 0.2, \beta = 5 * 10^3, \lambda_1 = 0.75$, and $\lambda_2 = 10^{-6}$. Each target image was trained using the EOT framework for 200 iterations, with the Adam optimization [17] and a learning rate of 0.02.

As our attack target, we selected the YOLACT, one of the first networks aimed at real-time instance segmentation. We used the pre-trained model on the MS-COCO dataset [24], which was provided by the authors.

4. Experiments

4.1. Benchmark Dataset and Evaluation Criteria

To evaluate the proposed method, we selected images containing at least one person from the MS-COCO dataset [24] and ended up with 1000 images. We evaluated robustness using Average Precision (AP) [24] and naturalness using Structural Similarity Index Measure (SSIM) [42]. To evaluate the effectiveness of methods, we computed AP by comparing the results of person segmentation before and after an attack.

4.2. Comparison with State-of-the-Arts

We compared the performance of FashionAdv with those of state-of-the-art adversarial example methods: FGSM [11], BIM [18], PGD [28], DeepFool [30], ColorFool [36], cAdv [3], and tAdv [3]. These methods were originally developed to attack classifiers; we thus adapted them to attack the classification sub-network in the person

Table 1. Comparison of performance between FashionAdv and conventional methods in terms of robustness as represented by average precision (AP) against JPEG compression for various quality factors (QFs). The two best results are shown in **blue** and **red**, respectively. FashionAdv significantly outperformed the conventional methods.

Method	Conference/Year	No Compression	JPEG Compression					
			QF 100	QF 80	QF 60	QF 40	QF 20	QF 10
Random Noise	-	58.37	58.36	57.89	57.62	56.96	54.32	40.65
FGSM [11]	ICLR 2014	49.71	49.93	51.33	52.63	52.90	51.97	43.30
BIM [18]	ICLRW 2016	40.87	41.63	50.11	54.82	58.13	59.16	46.65
DeepFool [30]	CVPR 2016	42.23	49.99	63.33	64.00	63.34	60.00	46.35
PGD [28]	ICLR 2018	3.74	4.70	56.05	61.12	61.66	45.65	45.65
ColorFool [36]	CVPR 2020	46.14	48.66	50.65	50.92	50.41	47.27	33.77
cAdv [3]	ICLR 2020	26.16	29.70	44.74	48.05	49.24	46.66	33.14
tAdv [3]	ICLR 2020	29.05	30.77	49.62	55.27	58.07	56.98	45.44
FashionAdv (rnd)	**CVPRW 2021**	**21.30**	**21.83**	**24.36**	**25.70**	**26.05**	**25.63**	**22.20**
FashionAdv	**CVPRW 2021**	**18.40**	**18.52**	**19.90**	**20.36**	**20.49**	**20.59**	**20.82**

Figure 6. Comparison of the resulting images of different methods. From left to right: the original image, FGSM [11], ColorFool [36], cAdv [3], tAdv [3], and our FashionAdv. Changes in image with FashionAdv are inconspicuous. (Best viewed online in color with zoom-in.)

segmentation network (*i.e.*, YOLACT). Instance segmentation networks are harder to fool than classifiers [45], and our empirical experiments revealed that the conventional methods with their default parameter settings are even weaker than a random noise attack, which is supposedly the weakest type of attack. We thus determined that it was pointless to use the default parameter settings to attack person segmentation networks. Therefore, we adjusted their parameter settings by using a gird search to make the compared methods more reliable than random noise methods in terms of AP score. The parameter settings and source codes are publicly released on our project page.

Figure 3 illustrates the trade-off between naturalness (*i.e.*, SSIM) and robustness (*i.e.*, AP) for the compared methods. FashionAdv significantly outperformed the other methods on robustness against JPEG compression with different QFs (*i.e.*, 10, 40, and 80): the AP scores of FashionAdv were around 20 while those of the other methods were higher than 30. FashionAdv also achieved competitive naturalness and was a top method in achieving the highest SSIM scores. As shown in Fig. 6, the adversarial images generated by FashionAdv looked the most natural.

Table 2. Contributions of loss components to final result.

$\mathcal{L}_{adv}$	$\mathcal{L}_{tex}$	$\mathcal{L}_{sim}$	$\mathcal{L}_{tv}$	AP	SSIM
✓				18.68	0.954
✓	✓			18.62	0.954
✓	✓	✓		18.60	0.955
✓	✓		✓	18.63	0.956
✓	✓	✓	✓	**18.40**	**0.958**

Table 1 compares the results for robustness in detail. FashionAdv (rnd) is our proposed method using fashion style random selection strategy (see section 4.3 for more detail). Our method significantly outperformed state-of-the-art methods and achieved the highest robustness against image compression. The conventional methods were not robust even for uncompressed images (except for PGD) and JPEG compressed images. Their AP scores were higher than 30 and rapidly increased for the JPEG compressed images. In contrast, the scores for FashionAdv were lower than 20 on uncompressed images and remained stable at around 20 for the JPEG compressed images.

4.3. Ablation Study

We investigated the effect of different components of our proposed FashionAdv, such as targeted attack, components of loss function, fashion style selection, and robustness analysis.

Importance of Attacked Targets. Previous attack methods target only the classification module in deep networks, so they work only on image classification. For person segmentation, both the classification and segmentation modules must be targeted. Indeed, an experiment we conducted on attacked targets revealed that targeting both the classification and segmentation modules significantly reduced AP by 18.35, from 36.75 (corresponding to target only classification module by setting $\mathcal{L}_{mask} = 0$) to 18.40.

Contributions of Loss Components. Table 2 shows the contributions of the loss components to the final result. $\mathcal{L}_{tex}$ helps to generate strong adversarial textures while $\mathcal{L}_{sim}$ and

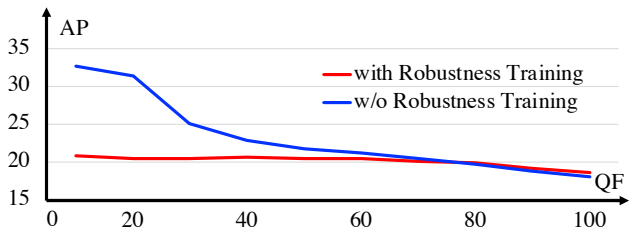


Figure 7. Robustness against JPEG compression for different quality factors QFs.

$\mathcal{L}_{tv}$ improve the naturalness and inconspicuousness of the generated textures. Integrating these loss components is thus useful in synthesizing robust and natural adversarial textures.

Fashion Style Selection. We employed two strategies for fashion style selection. In the first one, a fashion style image is randomly selected from our style corpus for each target image, denoted by *FashionAdv (rnd)*. In the second one, denoted by FashionAdv, fashion style image that minimized the texture transfer cost in the feature space is selected for each target image, similar to Bhattad et al. [3]. Using the optimal selection strategy effectively reduced AP by 3 to 5 compared with using the random selection strategy (*c.f*. Table 1). With this optimized strategy, the fashion style image closest to the target image is selected, resulting in more natural adversarial examples. This result also indicates the crucial role of using guided-fashion style images for synthesizing adversarial textures.

Robustness Against Image Storage. To evaluate the robustness of FashionAdv against image storage, *i.e*., against compression and saving in JPEG format, we investigated using JPEG cues in the robustness training. We compared training with and without JPEG cue for different JPEG quality factors (QFs). As shown in Fig. 7, training adversarial examples with JPEG cues kept the AP score around 20 for all QFs. In contrast, the AP score of training without JPEG cues increased to 33 at small QFs. Hence, integrating JPEG cues into the robustness training effectively defends against JPEG compression for various QFs.

Robustness Against Image Manipulations. To further exhibit the robustness of FashionAdv against image manipulations, we investigated the performance of person segmentation (in terms of AP) on several image editing operations: scaling, blurring, color jitter, and noise addition. We conducted manipulation attacks in two modes, easy and hard. In easy mode, we used randomly scaled images (0.8 to 1.2 of original size), Gaussian blur with a 5×5 kernel, histogram equalization for color jitter, and uniform noise of 0.05. In hard mode, we used randomly scaled images (0.5 to 1.5 of original size), Gaussian blur with a 9×9 kernel, randomly changing values of each color channel in the range [-0.2, 0.2], and uniform noise of 0.15. We remark that image color values were normalized and clipped to $[0, 1]$.

Table 3. Robustness against image manipulations in terms of AP score.

Attack Mode	Robustness Training	Scaling	Blurring	Color Jitter	Noise Addition
Easy	✓	19.93	20.75	19.12	18.04
	×	44.94	50.45	21.88	23.72
Hard	✓	23.50	23.66	22.93	19.23
	×	48.10	51.86	29.35	49.06

The effectiveness of the robustness training, which included perspective transform, image blurring, color variation, and image noise (Section 3.4) was demonstrated for each attack mode. Table 3 shows that the robustness training significantly outperformed the standard training (*i.e*., non-robustness training) for both attack modes, reducing the AP scores by 30. Utilizing our image perturbation pipeline in the robustness training process was effective against all image editing operations.

Table 3 also shows that the completed FashionAdv (with robustness training) was significantly robust against color jitter and noise addition even in hard mode, in which the AP scores increased by less than 5. We remark that the AP for the original images was 18.04 (*c.f*. Table 1). FashionAdv also well defended against two more difficult image editing operations (*i.e*., scaling and blurring). Image scaling eliminates small details (including generated adversarial noise) for downscaled images or highlights adversarial noise for upscaled images. Image blurring can wash away adversarial noise in the images. For these two image editing operators, AP was smaller than 24, indicating that FashionAdv also well defends against these difficult attacks.

5. Conclusion

Our proposed adversarial example based method is aimed at spoofing person segmentation networks. FashionAdv automatically detects and attacks the clothing regions by synthesizing robust yet natural adversarial textures from guided-fashion style images, resulting in all persons in the image being invisible to the networks. Extensive testing showed that FashionAdv is not only robust against digital image filtering and compression but also produces images that appear natural to the human eye.

FashionAdv relies on SCHP for generating binary masks of the clothing regions. If the clothing regions cannot be segmented, the generated adversarial images will be unable to spoof person segmentation networks. If the binary masks are over-segmented, manipulations can be easily perceived. Future work includes investigating better clothing segmentation approaches. It also includes further exploration of adversarial attacks on general instance segmentation.

Acknowledgements. This work was partially supported by JSPS KAKENHI Grants JP16H06302, JP18H04120, JP21H04907, JP20K23355, and JP21K18023, and by JST CREST Grants JPMJCR18A6 and JPMJCR20D3, including the AIP challenge program, Japan.

References

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

[2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *ICML*, pages 284–293, 2018.

[3] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and D. A. Forsyth. Unrestricted adversarial examples via semantic manipulation. In *ICLR*, 2020.

[4] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *ICCV*, 2019.

[5] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. In *NeurIPS Workshop*, 2017.

[6] Xiaoyi Dong, Jiangfan Han, Dongdong Chen, Jiayang Liu, Huanyu Bian, Zehua Ma, Hongsheng Li, Xiaogang Wang, Weiming Zhang, and Nenghai Yu. Robust superpixel-guided attentional adversarial attack. In *CVPR*, pages 12895–12904, 2020.

[7] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *CVPR*, pages 321–331, 2020.

[8] Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A. K. Qin, and Yun Yang. Adversarial camouflage: Hiding physical-world attacks with natural styles. In *CVPR*, June 2020.

[9] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *CVPR*, 2019.

[10] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, pages 932–940, 2017.

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2014.

[12] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *ICLR*, 2018.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988, 2017.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, June 2016.

[15] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[16] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *CVPR*, pages 720–729, 2020.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[18] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLR Workshop*, 2016.

[19] Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *CVIU*, 184:45–56, 2019.

[20] Trung-Nghia Le and Akihiro Sugimoto. Semantic instance meets salient object: Study on video semantic salient instance segmentation. In *WACV*, pages 1779–1788, 2019.

[21] Trung-Nghia Le, Akihiro Sugimoto, Shintaro Ono, and Hiroshi Kawasaki. Toward interactive self-annotation for video object bounding box: Recurrent self-learning and hierarchical annotation based framework. In *WACV*, 2020.

[22] David A. Leopold and Gillian Rhodes. A comparative view of face perception. *Journal of Comparative Psychology*, 124(3):233–251, 2010.

[23] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *arXiv preprint arXiv:1910.09777*, 2019.

[24] 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, 2014.

[25] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, June 2015.

[26] Bo Luo, Yannan Liu, Lingxiao Wei, and Qiang Xu. Towards imperceptible and robust adversarial example attacks against neural networks. In *AAAI*, 2018.

[27] Yan Luo, Xavier Boix, Gemma Roig, Tomaso Poggio, and Qi Zhao. Foveation-based mechanisms alleviate adversarial examples. *arXiv preprint arXiv:1511.06292*, 2015.

[28] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[29] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *CVPR*, pages 5188–5196, 2015.

[30] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, pages 2574–2582, 2016.

[31] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *ICML*, pages 5231–5240, 2019.

[32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[33] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, pages 7263–7271, 2017.

[34] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015.

[36] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *CVPR*, pages 1151–1160, 2020.

[37] Richard Shin and Dawn Song. Jpeg-resistant adversarial images. In *NIPS Workshops*, volume 1, 2017.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ILSVRC*, 2014.

[39] Benjamin Sobel. Hiq v. linkedin, clearview ai, and a new common law of web scraping. *LinkedIn, Clearview AI, and a New Common Law of Web Scraping*, April 21, 2020.

[40] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *CVPR Workshops*, pages 0–0, 2019.

[41] Changsheng Wan, Li Wang, and Vir V Phoha. A survey on gait recognition. *ACM Computing Surveys*, 51(5):1–35, 2018.

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE T-IP*, 13(4):600–612, 2004.

[43] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *ACSSC*, volume 2, pages 1398–1402, 2003.

[44] Rey Reza Wiyatno and Anqi Xu. Physical adversarial textures that fool visual object tracking. In *ICCV*, pages 4822–4831, 2019.

[45] Zuxuan Wu, Ser-Nam Lim, Larry Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *ECCV*, 2020.

[46] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV*, 2017.

[47] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *ECCV*, 2020.

[48] Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *ICLR*, 2019.

[49] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors. In *ACM CCS*, pages 1989–2004, 2019.