Neural Networks 132 (2020) 416–427

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

On the robustness of skeleton detection against adversarial attacks

Xiuxiu Bai *, Ming Yang, Zhe Liu

School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Article history:
Received 21 June 2020
Received in revised form 12 September 2020
Accepted 21 September 2020
Available online 28 September 2020

Keywords:
Convolutional neural network
Skeleton detection
Adversarial attacks
Robustness

ABSTRACT

Human perception of an object's skeletal structure is particularly robust to diverse perturbations of shape. This skeleton representation possesses substantial advantages for parts-based and invariant shape encoding, which is essential for object recognition. Multiple deep learning-based skeleton detection models have been proposed, while their robustness to adversarial attacks remains unclear. (1) This paper is the first work to study the robustness of deep learning-based skeleton detection against adversarial attacks, which are only slightly unlike the original data but still imperceptible to humans. We systematically analyze the robustness of skeleton detection models through exhaustive adversarial attacking experiments. (2) We propose a novel Frequency attack, which can directly exploit the regular and interpretable perturbations to sharply disrupt skeleton detection models. Frequency attack consists of an excitatory-inhibition waveform with high frequency attribution, which confuses edge-sensitive convolutional filters due to the sudden contrast between crests and troughs. Our comprehensive results verify that skeleton detection models are also vulnerable to adversarial attacks. The meaningful findings will inspire researchers to explore more potential robust models by involving explicit skeleton features.

1. Introduction

In neuroscience, an increasing amount of evidence suggests that the skeleton representation exists in the V1 and inferotemporal (IT) visual cortices of the brain (Hung, Carlson, & Connor, 2012; Kovacs & Julesz, 1994; Lescroart & Biederman, 2013; Sing Lee, Mumford, Romero, & Lamme, 1998). The skeleton feature can well express the symmetry, connectivity and deformation of different object parts in low dimensions (Marr & Nishihara, 1978). People's perception of an object's skeleton is especially robust to diverse perturbations of shape (Ayzenberg, Chen, Yousif, & Lourenco, 2019). This skeleton representation possesses substantial advantages for parts-based and invariant shape encoding, therefore it can be used as a basic task to assist many computer vision tasks, including object recognition (Ayzenberg & Lourenco, 2019), text recognition (Lake, Salakhutdinov, & Tenenbaum, 2015; Xiao, Qin, & Yin, 2020), action detection (Liu, Wang, & Liu, 2020; Yu & Lee, 2015; Zhu, Zou, Zhu, & Hu, 2019) and road detection (Sironi, Lepetit, & Fua, 2014).

Constructing skeleton detection models with the generalization beyond the training data is a key step toward human learning abilities (Hofstadter, 1985). Multiple deep learning-based skeleton detection models (Shen et al., 2016; Wang et al., 2019; Zhao, Shen, Gao, Li, & Cheng, 2018) have been proposed, while their robustness to adversarial attacks remains unclear. The vulnerability of deep neural networks has recently garnered significant attention, because these state-of-the-art deep classification models are fooled by adversarial examples which are only slightly unlike the original data but still scarcely conspicuous to humans (Moosavi-Dezfooli, Fawzi, & Frossard, 2016; Szegedy et al., 2013). It causes suspicion that deep networks are applied in safety-critical fields like medical image processing or autonomous vehicles (Janai, Güney, Behl, & Geiger, 2017). This raises concern regarding skeleton detection, because the disrupted detection results can influence the accuracy and reliability of systems adopting skeleton detection as their principal modules.

This paper is the first work to rigorously study the robustness of skeleton detection against adversarial attacks so far as we know. The reason why we concern skeleton detection is that it is a noticeable object representation ability in the human visual cortex (Hung et al., 2012). The current convolutional neural networks (CNNs) are strongly biased toward recognizing local textures rather than shapes (Brendel & Bethge, 2019; Geirhos et al., 2019), while humans learn mainly by shape information (Ritter, Barrett, Santoro, & Botvinick, 2017). To some extent, it can explain why CNNs are much more vulnerable than humans. Shape-bias CNNs are more robust than texture-bias CNNs (Brochu, 2019; Geirhos et al., 2019). However, skeleton features are more robust than shape features in the human vision (Ayzenberg et al., 2019; Ayzenberg & Lourenco, 2019). We want to make sure if this

* Corresponding author.
E-mail address: xiubai@xjtu.edu.cn (X. Bai).

https://doi.org/10.1016/j.neunet.2020.09.018
0893-6080/

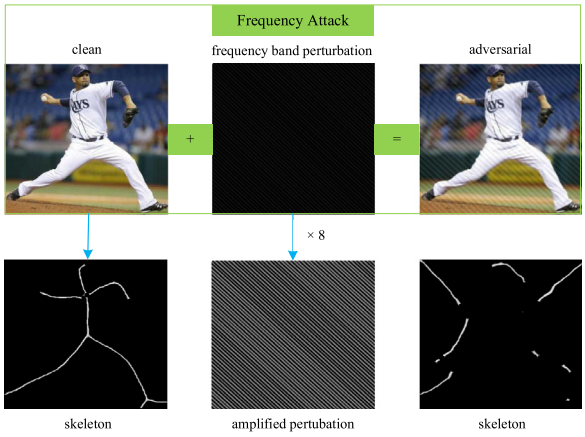


Fig. 1. Frequency attack. It can directly create regular and interpretable perturbations to disturb inputs. A state-of-the-art skeleton detection model named DeepFlux (Wang et al., 2019) is applied on a clean input and its adversarial disturbed versions. The perturbations are bounded with the $L\infty$ norm of 8. We amplify 8 times of perturbations to clearly display that. The perturbations with $\epsilon = 8$ become to be perceived by humans.

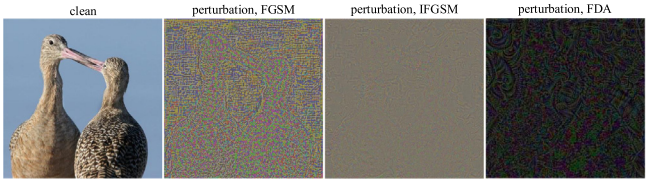


Fig. 2. Adversarial perturbations created by FGSM, IFGSM and FDA. Perturbations are bounded with the $L\infty$ norm of 32. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

intrinsic attribute of skeleton features can bring some resistance to adversarial attacks.

By analyzing adversarial perturbations of skeleton detection models, we discover that the shape of these perturbations is similar to the band structure in specific orientations (in Fig. 2). In addition, the shallow layers of CNNs are highly sensitive to the edges with the obvious gradients in diverse orientations. Inspired by these observations, we introduce a very simple but effective Frequency Attack, which can directly apply the regular perturbations to severely harm the skeleton prediction, as shown in Fig. 1. Our frequency attack composes of a positive–negative alternating frequency band, which confuses edge-sensitive convolutional filters of CNNs due to the sudden contrast between crests and troughs.

We summarize our contributions as follows:

- We are the first to use four adversarial attacks to disrupt skeleton detection models, including white-box, gray-box and black-box attacks, which are drawn from the attacks typically applied to classification tasks.
- We propose a novel Frequency attack, which can directly exploit regular perturbations to achieve a powerful fooling ability.
- We systematically analyze the robustness of skeleton detection models through detailed adversarial attacking comparative experimental results. Our meaningful findings will inspire researchers to explore more robust models by involving skeleton features.

The remainder of this paper is organized as follows: Section 2 reviews some related works. Section 3 describes the proposed method in detail. In Section 4, the experimental results are presented. Finally, we conclude this work in Section 5.

2. Related work

In this section, we present classical skeleton detection methods and adversarial attack approaches.

2.1. Skeleton detection

Traditional methods. Mignotte (2016) adopts a Hough-style voting method to calculate skeleton points. Tsogkas and Dickinson (2017) present an appearance medial axis transform by turning the skeleton extraction into a geometric set cover solution. Bai, Ye, Zhu, Zhu, and Komura (2020) introduce a zero-sum self-symmetric Skeleton Filter perceiving skeletons in the perturbed situations.

Deep learning-based methods. A wide range of deep learning-based skeleton detection models have been presented. Shen, Zhao et al. (2016) introduce Scale-related deep side-outputs (FSDS) to supervise learning the skeleton of different scales by leveraging scale-related skeleton labels. Ke, Chen, Jiao, Zhao, and Ye (2017) propose a Side-output Residual Network (SRN) by stacking the residual units in a deep-to-shallow fashion to minimize the loss. Liu, Ke, Qin, and Ye (2018) develop a Linear Span Network (LSN) by combining the feature and subspace linear span to effectively leverage multi-scale features. Hi-Fi (Zhao et al., 2018) possesses a hierarchical feature integration strategy by combining multi-scale features with bilateral cues, global semantics and local details.

The prior deep learning-based skeleton detection methods always build on the architecture of HED (Xie & Tu, 2015), and regard the issue as a binary pixel classification. Departing from the above mechanism, DeepFlux (Wang et al., 2019) achieves the state-of-the-art skeleton detection performance by leveraging context flux as a proxy skeleton, which embodies the connection between skeleton pixels and relevant edge pixels.

2.2. Adversarial attack

Szegedy et al. (2013) first discovered adversarial examples for deep neural networks in the image classification task. Following that, many works (Athalye, Carlini, & Wagner, 2018; Dong et al., 2018; Kurakin, Goodfellow, & Bengio, 2016; Moosavi-Dezfooli et al., 2016; Mygdalis, Tefas, & Pitas, 2020; Oregi, Ser, Pérez, & Lozano, 2020; Vidnerová & Neruda, 2020) have demonstrated that deep classification networks are vulnerable to kinds of adversarial attacks.

Goodfellow, Shlens, and Szegedy (2015) present a fast gradient sign approach (FGSM) by utilizing the sign of gradients from neural networks. Based on FGSM, Kurakin, Goodfellow, and Bengio (2017) explore an iterative mode to generate adversarial samples, which can achieve stronger fooling capability.

Most previous attackers (Dong et al., 2018; Goodfellow et al., 2015; Kurakin et al., 2017) create perturbation examples with the optimization of the prediction layer. It can change the last output, while the inner feature extractions keep original object information with some noise (Xie, Wu, Maaten, Yuille, & He, 2019). Motivated by this observation, Ganeshan et al. introduce a Feature Disruptive Attack (FDA) (Ganeshan & Babu, 2019) to damage features at each layer. Guo et al. (2019) propose a simple black-box adversarial attack, which takes a relatively low number of queries in a low frequency space.

A few of works develop adversarial attacks for other visual problems, for instance semantic segmentation (Arnab, Miksik, & Torr, 2018; Metzen, Kumar, Brox, & Fischer, 2017; Xie et al., 2017), object detection (Xie et al., 2017) and super-resolution (Choi, Zhang, Kim, Hsieh, & Lee, 2019).

Adversarial attack on skeleton detection. Skeleton features exhibit considerable advantages for parts-based and invariant shape encoding, thus it can be used as a pretext task to aid many computer vision tasks (Ayzenberg & Lourenco, 2019; Lake et al., 2015). However, the robustness of skeleton detection against adversarial attacks has not been considered in previous works. To the best of our knowledge, our paper is the first work to research adversarial attacks on skeleton detection.

3. Attacks on skeleton detection

This section presents characteristics of the skeleton detection task, the previous white-box, gray-box and black-box attacks and our proposed frequency attack on the skeleton detection task.

3.1. The characteristics of skeleton detection

From the neuroscience perspective. The skeleton representation is an important object representation ability in the human visual cortex. Therefore, we want to check how about the robustness of the skeleton representation is in the human vision.

(1) In neuroscience, skeletons and the corresponding external object shapes are simultaneously activated in IT visual cortex (Hung et al., 2012).

(2) Humans perceiving object skeletons can be especially robust to diverse perturbations of shapes (Ayzenberg et al., 2019).

(3) Humans can robustly recognize objects by involving global shape cues. In the human vision, the global shape can be distilled down to a low-dimensional skeleton feature that are steady beyond various viewpoints (Ayzenberg & Lourenco, 2019).

From the deep learning model perspective. Deep learning-based skeleton detection models possess various characteristics including model architectures, ground-truth skeleton labels and loss functions.

Network architecture. DeepFlux uses an atrous spatial pyramid pooling to expand the receptive field. Other skeleton detection models build on a VGG16 backbone with side output layers connected to the convolutional layers, which can extract multi-scale skeletons. Hi-Fi employs a hierarchical feature combination strategy via bilateral cues, global semantics and local details.

Ground-truth skeleton label. DeepFlux introduces context flux as a proxy for skeleton ground-truth label, which makes it achieve the state-of-the-art skeleton detection performance. This flux embodies the connection between skeleton pixels and relevant edge pixels. Some models introduce scale-related ground-truth skeleton labels to implicitly express the relation between skeleton pixels and edge pixels.

Loss function. Because of an extreme imbalance in the number of skeleton and background pixels in the skeleton detection task, they all use a weighted balancing method in the loss function. DeepFlux adopts a balanced L_2 loss. Hi-Fi and FSDS use a balanced softmax loss.

In summary, a central challenge of deep learning-based skeleton detection is that skeletons are highly sensitive to spatial relationship and object scales.

3.2. White-box attack

The purpose of adversarial attacks on skeleton detection is to induce an imperceptible perturbation on the original image, which leads to unreasonable skeleton predictions. To conduct this, we attack skeleton detection by extending FGSM (Goodfellow et al., 2015), Iterative FGSM (Kurakin et al., 2017), which are some of the most well-known attackers for classification.

FGSM (Goodfellow et al., 2015) **attack**. FGSM generates perturbed samples by rising the loss of neural network on a given image x

$$x^{adv} = x + \epsilon sign(\nabla_x L(J(\theta, x, y))) \tag{1}$$

where $\nabla_x L()$ denotes gradients of the cost function. y is the target with x. $sign()$ is the sign function. FGSM minimizes the $L\infty$ norm of the perturbation bounded with ϵ.

Iterative FGSM (Kurakin et al., 2017) **attack**. Based on FGSM, Iterative FGSM (IFGSM) introduces an iterative mode, which can produce superior adversarial examples to attack the network

$$\begin{aligned} x_0^{adv} &= x \\ x_{n+1}^{adv} &= clip_\epsilon(x_n^{adv} + \alpha sign(\nabla_{x_n^{adv}} L(J(\theta, x_n^{adv}, y)))) \end{aligned} \tag{2}$$

where n denotes the number of iterations. $clip\ ()$ ensures results clipped by the same value ϵ. It can maintain that the perturbation of each iteration is less than ϵ, which prevents visible changes of attacked samples.

3.3. Gray-box attack

Many previous attackers (Goodfellow et al., 2015; Kurakin et al., 2017) create perturbed examples with the optimization of the last prediction layer. It can change the final output, while inner feature extractions keep original object information with some noises. Inspired from that, FDA (Ganeshan & Babu, 2019) is introduced to damage features at each layer. We extend FDA to attack skeleton detection task which is a feature-based task. To do this, we apply FDA attack for image classification task to generate adversarial samples of skeleton detection datasets.

The spirit of FDA is to weaken signals approving the current prediction, conversely strengthen signals against the current prediction (Ganeshan & Babu, 2019). It uses spatial mean across channels $C(h, w)$ as average signals. The layer objective L for a given layer l_i as

$$\begin{aligned} L(l_i) = &\ log(D(\{l_i(x^{adv})_{(h,w,c)} | l_i(x)_{(h,w,c)} < C_i(h, w)\})) \\ &- log(D(\{l_i(x^{adv})_{(h,w,c)} | l_i(x)_{(h,w,c)} \geq C_i(h, w)\})) \end{aligned} \tag{3}$$

where D means the l_2 norm of inputs $l_i(x^{adv})$. Then it sums the per-layer objective as

$$\begin{aligned} &O = -\sum_{i=1}^{K} L(l_i), \\ &\| x^{adv} - x \|_\infty < \epsilon \end{aligned} \tag{4}$$

It uses the $L\infty$ norm limit of the perturbation at each iteration.

3.4. Black-box attack

Guo et al. (2019) propose a simple black-box adversarial attack, which randomly chooses a vector from a set of candidate orthonormal discrete cosine transform (DCT) basis to add or subtract to the target image. It takes a relatively low number of queries in a low frequency space. For example, the average query count of SimBA-DCT (Guo et al., 2019) for untargeted attacks on ImageNet is 1283. Since it is very fast and effective, SimBA-DCT can serve as a strong baseline for future black-box attacks.

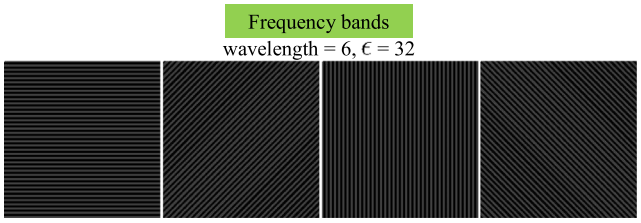


Fig. 3. Frequency bands. To clearly display that, perturbations are bounded with the $L\infty$ norm of 32.

Let $\alpha_i \in \{-\epsilon_{iter}, 0, \epsilon_{iter}\}$ represent the sign of search directions sampled at step n. ϵ_{iter} denotes the step size per iteration. q_n is a vector from a set of orthonormal basis Q. The perturbation δ is as follows:

$$\begin{aligned} \delta_{n+1} &= \delta_n + \alpha_n q_n \\ \delta_N &= \sum_{n=1}^{N} \alpha_n q_n \end{aligned} \tag{5}$$

where N means the number of steps. The final perturbation δ_N is a sum of these basis vectors. It can obtain the adversarial sample x^{adv} of input x by

$$x_N^{adv} = clip_\epsilon(x + \delta_N) \tag{6}$$

where *clip()* ensures results clipped by the same value ϵ.

We try to use SimBA-DCT (Guo et al., 2019) to attack skeleton detection models.

3.5. Frequency attack

We propose a novel Frequency attack, which can directly exploit the regular and interpretable perturbation to sharply disrupt skeleton detection models, as shown in Fig. 1. The proposed attacker is a black-box approach, which utilizes a pair of input and output to seek the optimization setting. Our attacker is an untargeted error-generic evasion method. It aims to cause an integrity and availability violation. The specific attack scenarios of our method are some skeleton-based visual tasks, such as road centerline detection in autonomous vehicle, text detection and blood vessel centerline in medical diagnosis.

Motivation. Fig. 2 shows adversarial perturbations created by FGSM, IFGSM and FDA at $\epsilon = 32$. Among them, FGSM and IFGSM directly utilize the gradient information of skeleton detection models, and FDA indirectly uses gradient signals of classification model and transfers to skeleton detection task. However, we discover that the shape of these perturbations is alike to the band structure in the specific orientation. It reveals that CNNs may be high sensitive to these band structures in the specific orientation.

Moreover, the shallow layers of CNNs are high sensitive to edges with obvious gradients in diverse orientations. Inspired by that, we conceive that the distribution of perturbations exactly fill with the color contrast signals, which are easily captured by orientation-like filters of CNNs. Therefore, we design an excitatory-inhibition alternating waveform, which is full of the contrast information between crests and troughs. These very regular perturbations probably disrupt prediction results. We will strictly verify our assumption.

Frequency bands. We introduce frequency bands to attack the given models. Frequency bands denote that the wave crest is filled with 1 and the wave trough is filled with -1. This waveform of frequency bands fills with the excitatory-inhibition signals. Frequency bands involve a group of N_{orient} orientations and N_{wave} wavelengths. Fig. 3 visualizes frequency bands. We first create the horizontal frequency band *frequency_band*[0]

$$\begin{aligned} &frequency\_band[0] = ones([w, h]) \\ &frequency\_band[0][k, :] = -1, \ \ if \ \ k \,\%\, \lambda < \frac{\lambda}{2} \end{aligned} \tag{7}$$

where *frequency_band* denotes frequency bands. *ones*() generates a new $w \times h$ array filled with ones. λ denotes the wavelength of frequency bands. By calculating the remainder operation, it can fill a half of wave with -1. $\frac{\lambda}{2}$ requires to be integer.

It can obtain a group of frequency bands with various orientations by rotating a horizontal band.

$$\begin{aligned} frequency\_band[n] &= rotate(frequency\_band[0], \beta) \\ \beta &= n \times \frac{\pi}{N_{orient}}, n \in [0, N_{orient}) \end{aligned} \tag{8}$$

where *rotate*() is the rotation function. β means the rotation angle. N_{orient} denotes the number of orientations.

To sum up, the key parameters of frequency bands are the wavelength λ and orientation β. The calculation formulas are Eqs. (7) and (8).

$$frequency\_band = f(\lambda, \beta) \tag{9}$$

We can obtain the adversarial sample x^{adv} of input x by

$$x^{adv} = x + \epsilon frequency\_band(\lambda, \beta) \tag{10}$$

where the perturbations are bounded with the $L\infty$ norm of ϵ. ϵ can also be treated as the amplitude of frequency bands. The adversarial samples are closely related to the orientation and wavelength of frequency bands.

Relation to orientation and wavelength. Fig. 4 visualizes the sensitivity to orientation and wavelength of frequency bands. With the same wavelength, the attack abilities of various orientations of frequency bands are significantly different. Similarly, with the same orientation, the corrupting effects of diverse wavelengths of frequency bands are substantially different. We will make an in-depth analysis as follows.

(1) Why are the short wavelength of frequency bands more aggressive? The intensity of adversarial perturbations requires a kind of L norm to constraint, which keeps imperceptible to humans. The proposed frequency bands are bounded with the $L\infty$ norm of ϵ. Therefore, the slight frequency band perturbations need to be successively enlarged through the network layers to disrupt last prediction results.

If the wavelength of frequency bands is very long, it expects a large receptive field in deeper layers of CNNs to detect the contrast signals between crests and troughs. Then, it will reduce the chance of aggravating perturbations, thus the last attack effect is much less.

(2) Is the shortest wavelength the most effective? The width of a single crest or trough demands to be integer, so the shortest wavelength of a frequency band is $\frac{\lambda}{2} = 1$. In vgg16, the receptive fields of the first two layers are 3 and 5. As long as $\frac{\lambda}{2}$ is in the range of 5, the contrast signals between crests and troughs will be detected by the first two layers. The attack effects of frequency bands are also related to the orientation of bands and the distribution of inputs. However, we can set a small number of wavelengths as a frequency band space and search for the optimal attack wavelength.

(3) Why does it require the specific orientation of frequency bands? In Fig. 4, the feature maps at the shallow layer reveal that frequency bands with diverse orientations can all be detected by skeleton models. As going to deeper layers, frequency bands with the specific orientation will be aggravated and seriously disrupt object features. It is relevant to the relationship between

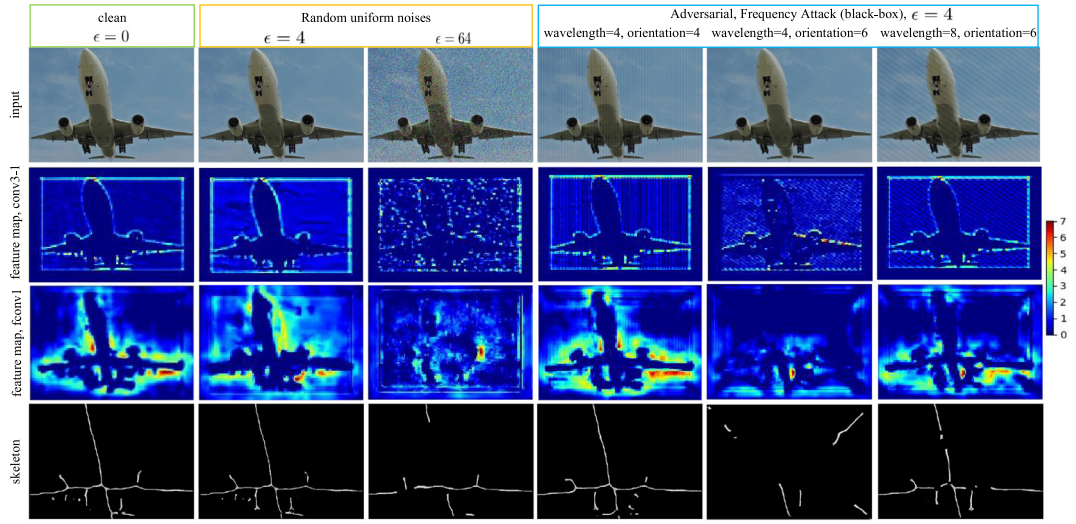


Fig. 4. Sensitive to the orientation and wavelength of frequency bands. The feature maps are from the same channel of conv3-1 and fconv1 layers in the DeepFlux (Wang et al., 2019) trained on clean inputs. The perturbations are bounded with the $L\infty$ norm of 4. We set the number of orientations $N_{orient} = 8$. $orientation = 4$ denotes that the chosen orientation is $frequency\_band[4]$. $wavelength = 4$ denotes the width of a wave is 4 line pixels.

distribution of frequency bands and data distribution. If they are consistent, it will be equivalent to amplifying the object signal. If not, it will bring some confusion to the distribution of the original object. Only in a certain orientation, it will sharply harm the predictions.

However, the objects have diverse shapes and textures, but the contrast distribution of frequency bands is very regular. In most cases, frequency bands are inconsistent with the object distribution. Due to the complexity of data distribution, it is hard to determine precise relations. Anyway, we can set a group of orientations as a frequency band space to search for the optimal attack orientation.

In addition, we compare regular frequency band perturbations with random uniform noises. With the same $\epsilon = 4$, our regular frequency bands are much more aggressive than random uniform noises. When ϵ climbs to 64, the random uniform noises become to disrupt skeleton features. It can also demonstrate that frequency bands filled with contrast signals are more effective.

Optimization strategy. We adopt the combination optimization strategy to search the optimization setting. There are $N_{freq} = N_{orient} \times N_{wave}$ settings of the frequency band space

$$\begin{aligned} y^k_{score} &= g(x^{adv_k}) \\ &= g(x + \epsilon frequency\_band(\lambda_m, \beta_n)), \\ m \in [0, N_{wave}), n \in [0, N_{orient}), k \in [0, N_{freq}) \end{aligned} \tag{11}$$

where $g()$ represents the target model. y_{score} denotes the evaluation metric of a given task. In the skeleton detection task, y_{score} uses F-measure metric explained in the experiment section.

We can directly seek the smallest y^k_{score} in the N_{freq} settings to obtain the optimal frequency band for a given input.

$$(\lambda_{opt}, \beta_{opt}) = argmin(y^k_{score}), k \in [0, N_{freq}) \tag{12}$$

The optimization time T_{optim} is calculated by

$$T_{optim} = N_{freq} T_{infer} \tag{13}$$

where T_{infer} is the inference time of a given model. Generally, because the smaller wavelengths are more effective, we can set $N_{wave} = 5$ and $\frac{\lambda}{2} \in [1, 5]$. N_{orient} can be set as 8, $\beta = n \times \frac{\pi}{N_{orient}}, n \in [0, 8)$. So, the frequency band space N_{freq} is 40.

In summary, we introduce a regular and explainable frequency band perturbation to realize a powerful adversarial attack capability, which provides a unique perspective to understand the vulnerability of CNNs.

4. Experiments

We use five representative adversarial attackers to perturb three state-of-the-art skeleton detection models on four popular datasets. Adversarial perturbations are limited with $L\infty$ norm of ϵ from {4, 8, 16}. Thus, we conduct 180 ($5 \times 3 \times 4 \times 3$) groups of adversarial attacking experiments to rigorously evaluate and analyze the robustness of skeleton detection task.

4.1. Experimental setup

Datasets. We use four widely skeleton detection datasets, including SK-LARGE (Shen et al., 2017), SYM-PASCAL (Ke et al., 2017), SYMMAX 300 (Tsogkas & Kokkinos, 2012) and WH-SYMMAX (Shen, Bai, Hu and Zhang, 2016). SK506 is a subset of SK-LARGE (Shen et al., 2017), thus we choose SK-LARGE to evaluate.

SK-LARGE (Shen et al., 2017) is stemmed from the MS COCO dataset (Chen et al., 2015). It consists of 746 training images and 745 testing images.

SYM-PASCAL (Ke et al., 2017) is built on the PASCAL VOC dataset (Everingham et al., 2010). It comprises 647 training images and 788 testing images.

SYMMAX300 (Tsogkas & Kokkinos, 2012) is derived from the BSDS300 dataset (Martin, Fowlkes, Tal, Malik, et al., 2001). It comprises 200 training images and 100 testing images.

WH-SYMMAX (Shen, Bai et al., 2016) is built on the Weizmann Horse dataset (Borenstein & Ullman, 2002). It comprises 228 training images and 100 testing images.

The images of these skeleton datasets are built on other semantic segmentation datasets, and the skeleton ground-truths are extracted from human annotated object segmentation labels. In the SK-LARGE and WH-SYMMAX datasets, images are cropped such that the object occupies the most space of images. In addition, all the objects in the WH-SYMMAX dataset are horses. While the SYM-PASCAL and SYMMAX300 datasets contain various complex backgrounds and multiple objects with different scales in the same image.

Skeleton detection methods. We employ these state-of-the-art deep learning-based skeleton detection methods, including DeepFlux (Wang et al., 2019), Hi-Fi (Zhao et al., 2018) and FSDS (Shen, Zhao et al., 2016). We use public code of original papers and trained models if they provided. For unprovided models, we

retrain to achieve nearly the same performance as the original paper by using the same hyper-parameters.

DeepFlux (Wang et al., 2019) achieves the state-of-the-art skeleton performance by utilizing context flux to represent object skeletons. The backbone network of DeepFlux is Deeplab v2 VGG16 (Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2017).

Hi-Fi (Zhao et al., 2018) possesses a feature integration strategy to combine multi-scale features with bilateral cues, including global semantics and local details. The backbone network of Hi-Fi is VGG16.

FSDS (Shen, Zhao et al., 2016) introduces scale-related skeleton ground-truth labels to supervise learning the skeleton of different scales. FSDS is built on the baseline architecture of HED (Xie & Tu, 2015) with the VGG16 backbone network.

Adversarial attacks. We employ FGSM (Goodfellow et al., 2015), IFGSM (Kurakin et al., 2017), FDA (Ganeshan & Babu, 2019) and SimBA-DCT (Guo et al., 2019) attacks. We use the public code of original papers. In our experiments, FGSM and IFGSM are used as white-box attacking methods, which can leverage the information of skeleton detection models. FDA is used as a gray-box attacking method. We use FDA for classification task to directly attack skeleton detection models without any skeleton task-specific details. The classification model uses a VGG16 backbone trained on ImageNet. SimBA-DCT and our Frequency attack are black-box methods. For a fair comparison, we mainly compare with SimBA-DCT.

In the Frequency attack, we set $N_{wave} = 5$ and the wavelength $\frac{\lambda}{2} \in [1, 5]$. N_{orient} can be set as 8, the orientation is $\beta = n \times \frac{\pi}{N_{orient}}, n \in [0, 8)$. The frequency band space $N_{freq} = N_{orient} \times N_{wave} = 40$. It means that the number of queries is $N_{freq} = 40$.

The parameters of SimBA-DCT (Guo et al., 2019) are the number of queries Q and step size per iteration ϵ_{iter}. ϵ_{iter} is set to 0.2. For a fair comparison with our method, Q is also set to 40. We adopt the $L\infty$ norm.

Based on Kurakin et al. (2017), the parameter of IFGSM attacks is the number of iterations, which is set to $min(\epsilon+4, \lceil 1.25\epsilon \rceil)$ and $\alpha = 1$ denotes the increment of each iteration.

The parameters of FDA (Ganeshan & Babu, 2019) are the tuple (ϵ, nb_{iter}, ϵ_{iter}). nb_{iter} denotes the number of iterations and ϵ_{iter} denotes the increment at each iteration. The used parameters of FDA are ($\epsilon = 4$, $nb_{iter} = 5$, $\epsilon_{iter} = 1$), ($\epsilon = 8$, $nb_{iter} = 10$, $\epsilon_{iter} = 1$) and ($\epsilon = 16$, $nb_{iter} = 10$, $\epsilon_{iter} = 2$).

Evaluation metric. The F-measure and precision–recall (PR) curves are the principal metrics used in evaluating skeleton detection performance. Applying different thresholds to predicted skeleton maps, it can obtain a series of precision/recall pairs to draw the PR curves. According to the formula F-measure $= \frac{2Precision \times Recall}{Precision + Recall}$, the highest F-measure can be obtained under the optimal threshold. The higher value of F-measure is better.

4.2. White-box attack

Tables 1–3 show the adversarial robustness of skeleton detection models for FGSM, IFGSM, FDA and our Frequency attack on four skeleton datasets at $\epsilon = 4, 8, 16$. With ϵ increasing, skeleton detection performance deteriorates moderately. Compared with FGSM attack, skeleton detection performance against IFGSM attack dramatically gets worse as ϵ growing. For different skeleton detection models, Hi-Fi exhibits a relatively better robustness to adversarial perturbation. The detailed analysis is in Section 4.6.

4.3. Gray-box attack

We introduce FDA to directly attack skeleton detection models without any skeleton task-related details. The purpose of attacking across different tasks is to further check the robustness of skeleton detection. On the other hand, it can also demonstrate whether these attackers possess strong attacking capability. Overall, we can obviously notice that the skeleton detection performance against FDA gets worse with ϵ increasing, particularly on the SYM-PASCAL dataset.

FDA is used as a gray-box method. Our Frequency attack performs better than FDA in most cases, particularly on the SK-LARGE, SYMMAX300 and WH-SYMMAX datasets. Compared with other datasets, SYM-PASCAL displays weaker than FDA, because this dataset includes multiple objects with different scales in the same image. Frequency attack is sensitive to the orientations and wavelengths. If the data distributions are so complicated, it is relatively difficult to seek the optimal orientations and wavelengths for multiple objects. However, Frequency attack can still effectively fool the skeleton predictions.

Compared with white-box attacks, our Frequency attack significantly outperforms FGSM, while it overall performs weaker than IFGSM. However, the attack effects of our method are better than IFGSM with ϵ increasing, particularly on the SYMMAX300 dataset at $\epsilon = 16$.

4.4. Black-box attack

SimBA-DCT (Guo et al., 2019) is a simple black-box adversarial attack for classification task. Since it is very fast and effective, SimBA-DCT can serve as a strong baseline for future black-box attacks. But SimBA-DCT cannot effectively attack skeleton detection models, as shown in Tables 1, 2, 3.

The possible reason is as follows: (1) Skeleton feature is a kind of relatively global feature. It needs a large receptive field to detect skeleton features in the large-scale part of the object. For the low frequency perturbations, it also needs a large receptive field to detect. However, in the same region of large receptive field, when the object and the perturbations exist at the same time, the intensity of perturbations is much weaker than that of the object itself, so the low frequency perturbations have limited ability to attack the skeleton detection models. (2) Through the test of our Frequency attack, the skeleton detection model is sensitive to the orientation of frequency bands. Specifically, in some orientations, the frequency band with any wavelength cannot effectively attack skeleton detection models; in some specific orientations, the frequency band with the specific wavelength is very effective to attack skeleton detection models. In SimBA-DCT, the used DCT basis includes only horizontal and vertical orthogonal orientations.

We explore high frequency bands to attack object skeleton detection models. The slight high frequency bands possess the opportunity to be successively enlarged to disrupt last prediction results. Moreover, we observe that the orientation of frequency bands is the key factor for the attacking ability, while SimBA-DCT has not involved this point. The experiments demonstrate that the effectiveness of our attacker. Therefore, our method is the orthogonal complement to SimBA-DCT.

4.5. Feature map analysis of adversarial samples

We visualize characteristics of feature maps for adversarial samples. Fig. 5 shows feature maps for clean inputs and perturbed counterparts generated by different attackers, including Frequency attack, FGSM, IFGSM, FDA and SimBA-DCT. The feature maps of each group are from the same channel of different layers in DeepFlux. These feature map flows expose that the initially tiny perturbation on the inputs are involved to corrupt the true signal of target objects. The attacking principle of FDA yields the signal of target objects weaken and the signal of background strengthen. In particular, the Frequency attack involves the regular tiny frequency band perturbation, which can be amplified in the deeper layers to disrupt skeleton features.

Table 1
Clean and adversarial performance. The performance metric is F-measure. $\epsilon = \mathbf{8}$.

	Attacker	SK-LARGE			SYM-PASCAL			SYMMAX300			WH-SYMMAX		
		DeepFlux	Hi-Fi	FSDS	DeepFlux	Hi-Fi	FSDS	DeepFlux	Hi-Fi	FSDS	DeepFlux	Hi-Fi	FSDS
	Clean	0.732	0.724	0.633	0.502	0.454	0.418	0.491	0.486	0.467	0.840	0.805	0.769
White-box	FGSM (Goodfellow et al., 2015)	0.505	0.638	0.547	0.273	0.358	0.293	0.364	0.459	0.440	0.582	0.714	0.663
	IFGSM (Kurakin et al., 2017)	0.196	0.644	0.352	0.149	0.324	0.108	0.256	0.450	0.282	0.242	0.716	0.509
Gray-box	FDA (Ganeshan & Babu, 2019)	0.524	0.543	0.426	0.201	0.180	0.139	0.367	0.366	0.338	0.586	0.668	0.600
Black-box	SimBA-DCT (Guo et al., 2019)	0.716	0.693	0.622	0.456	0.398	0.370	0.483	0.467	0.455	0.831	0.791	0.753
	Frequency (our)	**0.423**	**0.512**	0.457	**0.200**	0.229	0.201	**0.347**	**0.322**	**0.315**	0.621	**0.606**	**0.595**

Table 2
Clean and adversarial performance. The performance metric is F-measure. $\epsilon = \mathbf{4}$.

	Attacker	SK-LARGE			SYM-PASCAL			SYMMAX300			WH-SYMMAX		
		DeepFlux	Hi-Fi	FSDS	DeepFlux	Hi-Fi	FSDS	DeepFlux	Hi-Fi	FSDS	DeepFlux	Hi-Fi	FSDS
	Clean	0.732	0.724	0.633	0.502	0.454	0.418	0.491	0.486	0.467	0.840	0.805	0.769
White-box	FGSM (Goodfellow et al., 2015)	0.543	0.680	0.578	0.295	0.397	0.296	0.370	0.477	0.450	0.654	0.773	0.720
	IFGSM (Kurakin et al., 2017)	0.352	0.677	0.491	0.202	0.373	0.163	0.304	0.469	0.345	0.470	0.766	0.657
Gray-box	FDA (Ganeshan & Babu, 2019)	0.636	0.617	0.536	0.301	0.263	0.231	0.425	0.400	0.412	0.774	0.742	0.694
Black-box	SimBA-DCT (Guo et al., 2019)	0.716	0.693	0.622	0.456	0.399	0.370	0.483	0.467	0.454	0.837	0.790	0.753
	Frequency (our)	**0.546**	**0.607**	0.536	0.302	0.307	0.290	**0.420**	0.404	**0.408**	**0.746**	**0.708**	**0.689**

Table 3
Clean and adversarial performance. The performance metric is F-measure. $\epsilon = \mathbf{16}$.

	Attacker	SK-LARGE			SYM-PASCAL			SYMMAX300			WH-SYMMAX		
		DeepFlux	Hi-Fi	FSDS	DeepFlux	Hi-Fi	FSDS	DeepFlux	Hi-Fi	FSDS	DeepFlux	Hi-Fi	FSDS
	Clean	0.732	0.724	0.633	0.502	0.454	0.418	0.491	0.486	0.467	0.840	0.805	0.769
White-box	FGSM (Goodfellow et al., 2015)	0.450	0.557	0.483	0.177	0.287	0.251	0.345	0.417	0.399	0.490	0.608	0.503
	IFGSM (Kurakin et al., 2017)	0.129	0.606	0.245	0.124	0.259	0.091	0.223	0.431	0.258	0.134	0.648	0.377
Gray-box	FDA (Ganeshan & Babu, 2019)	0.377	0.461	0.287	0.106	0.105	0.087	0.277	0.303	0.229	0.331	0.573	0.397
Black-box	SimBA-DCT (Guo et al., 2019)	0.716	0.693	0.622	0.457	0.399	0.370	0.483	0.467	0.455	0.837	0.790	0.753
	Frequency (our)	**0.253**	**0.363**	0.330	**0.086**	**0.098**	**0.067**	**0.222**	**0.181**	**0.163**	0.377	**0.399**	0.417

Table 4
Clean and adversarial performance. The average adversarial performance indicates the average of all datasets with $\epsilon = 4, 8, 16$. The used performance metric is F-measure.

Model	scale skeleton	bilateral cue	multi-scale	flux skeleton	Average for all the datasets					
					Clean	FGSM (Goodfellow et al., 2015)	IFGSM (Kurakin et al., 2017)	FDA (Ganeshan & Babu, 2019)	SimBA-DCT (Guo et al., 2019)	Frequency
FSDS (Shen, Zhao et al., 2016)	✓				0.572	0.469	0.323	0.365	0.550	0.372
Hi-Fi (Zhao et al., 2018)	✓	✓	✓		0.617	**0.530**	**0.530**	**0.435**	0.587	**0.395**
DeepFlux (Wang et al., 2019)				✓	**0.641**	0.421	0.232	0.409	**0.623**	0.378

4.6. Robustness analysis of skeleton detection

We summarize all the average results to further analyze the robustness of skeleton detection task. Table 4 lists the clean and the average adversarial accuracy, which indicates the average of all datasets with $\epsilon = 4, 8, 16$.

Relation to models. From the clean accuracy perspective, DeepFlux outperforms other skeleton detection models. It is because the flux skeleton introduced by DeepFlux can explicitly leverage the connection between skeletons and edges, which can effectively extract object skeletons.

From the adversarial robustness perspective, Hi-Fi performs better than other skeleton detection models. The reason is that Hi-Fi leverages scale skeleton, bilateral cues and multi-scale. The scale of skeleton pixel represents its distance to the nearest neighbor boundary pixel. Bilateral cues indicate local details and global semantic information. In addition, Hi-Fi adopts the average of multi-scale prediction by resizing inputs to various resolutions (0.5, 1, 1.5). The results show that scale skeleton, bilateral cue and multi-scale can ease the disruption by attackers to a certain extent.

Relation to attackers. FGSM and IFGSM corrupt features at the final prediction layer such that the higher semantic activation are perturbed. Hence, Hi-Fi can improve the robustness against these attackers by using bilateral cues.

The upgraded FDA destroys features at each layer so that both lower details and higher semantic signals are damaged. Although we employ FDA in gray-box setting to attack, it also can significantly corrupt skeleton features. Moreover, the bilateral cues of Hi-Fi against FDA are less effective than against FGSM and IFGSM.

Our Frequency attack introduces the very regular and interpretable perturbation, which is the big different from prior adversarial attacks.

Fig. 6 plots PR-curves for DeepFlux against adversarial attacks on the SK-LARGE dataset. Adversarial attacks include FGSM, IFGSM, FDA and Frequency attack. On the whole, IFGSM in white-box setting can significantly destroy skeleton feature extraction. The corrupting effect of FDA in gray-box setting is similar as

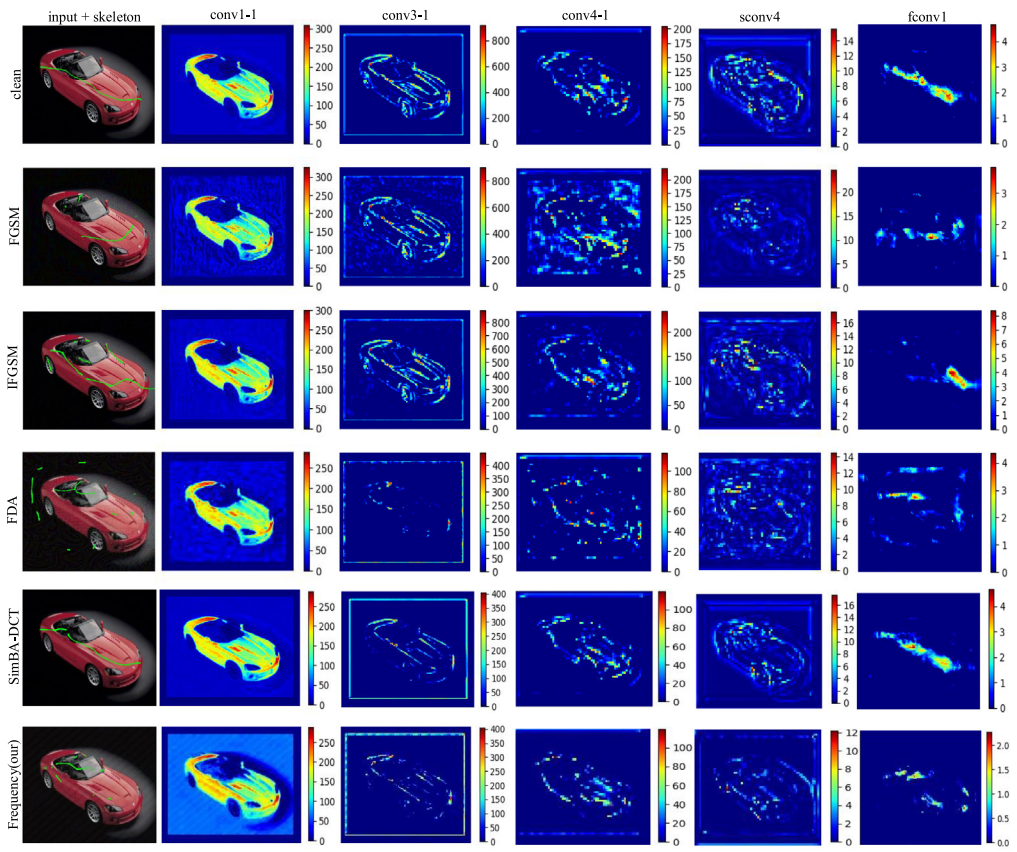
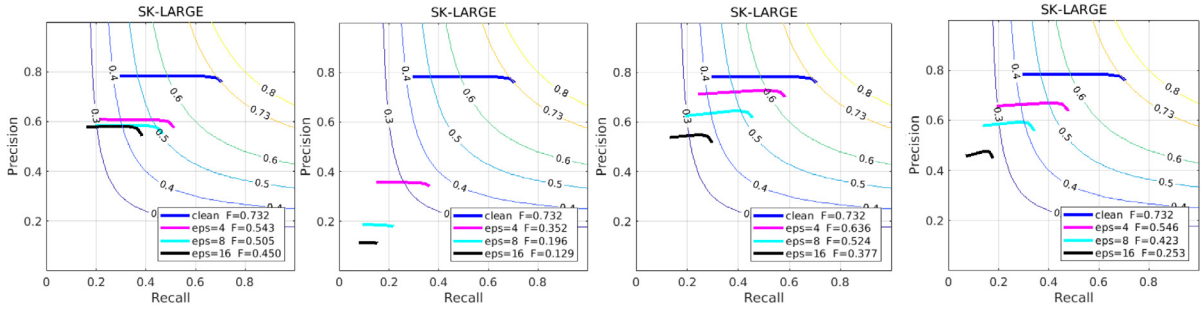


Fig. 5. Feature maps for a clean input and adversarial perturbed counterparts generated by different attackers, including FGSM, IFGSM, FDA, SimBA-DCT and our Frequency attack. The feature maps of each group are from the same channel of different layers in DeepFlux (Wang et al., 2019) trained on clean inputs. The perturbations are bounded with the $L\infty$ norm of 8.



(a) FGSM attack (b) IFGSM attack (c) FDA attack (d) Frequency attack

Fig. 6. PR curves for DeepFlux against adversarial attacks on the SK-LARGE dataset. Adversarial attacks include Frequency attack, FGSM, IFGSM and FDA attacks. Baseline means the original clean testing dataset. eps means ϵ.

single-step FGSM attacker in white-box setting. The Frequency attack outperforms FDA and white-box FGSM methods.

Fig. 7 presents visual comparison of skeleton detection results for input images perturbed by three attacks with $\epsilon = 8$. The used skeleton detection model is DeepFlux. We can observe that skeleton results of adversarial samples created by FGSM and IFGSM are deviated from the original position due to the perturbation. Moreover, the Frequency and FDA attacks yield the signal of target objects weaken and the signal of background strengthen.

4.7. Discussion

We first conduct an ablation study of our Frequency attack. Table 5 lists quantitative results of Frequency attack with different orientations and wavelengths. Fig. 8 shows that the relationship between the adversarial performance and the wavelength of Frequency bands. Fig. 9 shows that the relationship between the adversarial performance and the orientation of Frequency bands.

With the same orientation, the corrupting effects of diverse wavelengths of frequency bands are substantially different. Similarly, with the same wavelength, the attack abilities of various orientations of frequency bands are significantly different. Furthermore, only using the same orientation and wavelength of frequency bands for the whole dataset, it can effectively attack skeleton detection models.

Why can frequency bands attack effectively? The detailed analysis is in Section 3.5. The frequency band is a very simple positive–negative alternating waveform, which is filled with the contrast signals between crests and troughs. It would be easily

Fig. 7. Visual comparison of skeleton detection results for input images perturbed by four attacks with $\epsilon = 8$. The used skeleton detection model is DeepFlux. The perturbations with $\epsilon = 8$ become to be perceived by humans. In addition, because there is orientation selectivity in the V1 cortex of the brain (Jones & Palmer, 1987), frequency bands are also sensitive to humans. However, people can still precisely perceive skeleton features of objects under these perturbations, while the deep learning-based skeleton detection methods perform poorly against the frequency attack.

detected by the orientation-like filters at the shallow layers of CNNs. Further, the slight frequency bands possess the opportunity to be successively enlarged to disrupt the last prediction results. Fig. 8 and Table 5 can demonstrate that the high frequency perturbations are more aggressive for skeleton detection models. Fig. 9 and Table 5 can demonstrate that performance of skeleton detection models is much sensitive to the orientations of perturbations. We first notice that CNNs are so sensitive to the perturbation's orientation, which provides a new perspective to further study the generalization behavior of CNNs.

Besides, our proposed frequency bands are also sensitive to humans due to orientation selectivity existing in the V1 cortex of the brain (Jones & Palmer, 1987). However, people can still precisely perceive object skeleton features, while the deep learning-based skeleton detection methods perform poorly against the frequency attack. It exposes that the structures of shallow layers in CNNs are similar as that of human brains, but there is a large gap between the mechanisms of deeper layers in CNNs with that of human brains.

Moreover, we compare regular frequency bands with random uniform noises. With the same $\epsilon = 8$, the random uniform noises have a very limited attack effect to skeleton detection models. It can demonstrate that regular frequency bands filled with contrast signals are more effective.

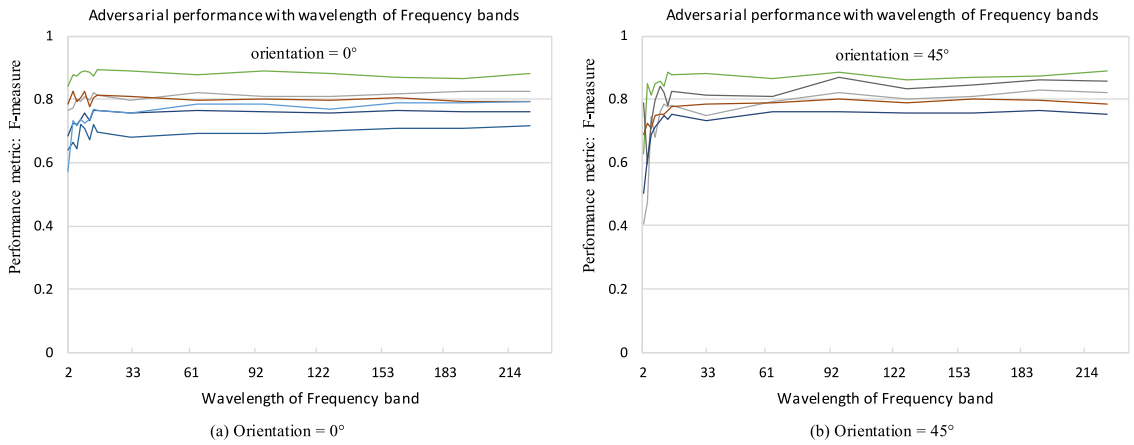


Fig. 8. Relationship between the adversarial performance and the wavelength of Frequency bands. The skeleton detection performance for input images perturbed by our Frequency attack with $\epsilon = 8$. The used skeleton detection model is DeepFlux. The lines with different colors denote results of different input images. The results of the whole dataset are in Table 5. With the same orientation, the attack abilities of high frequency perturbations are more effective for the skeleton detection task. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

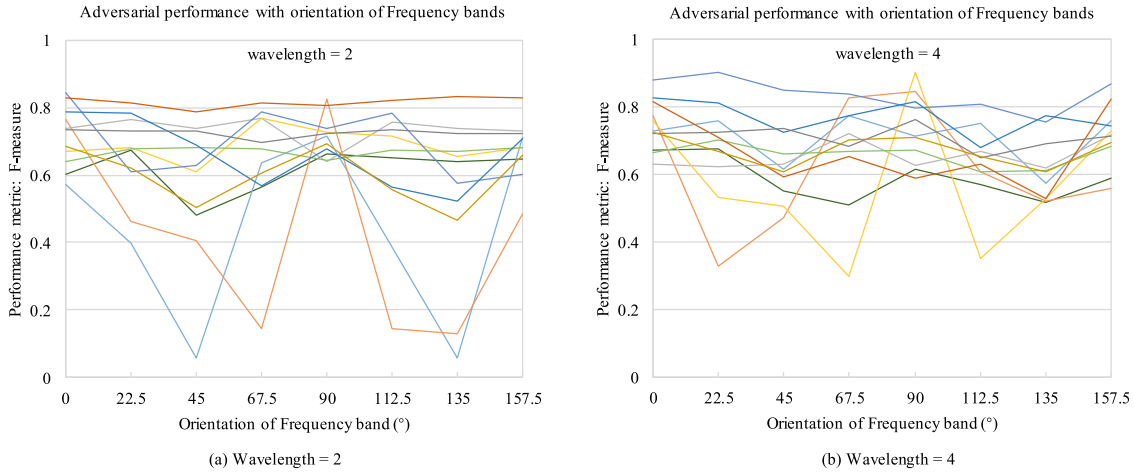


Fig. 9. Relationship between the adversarial performance and the orientation of Frequency bands. The skeleton detection performance for input images perturbed by our Frequency attack with $\epsilon = 8$. The used skeleton detection model is DeepFlux. The lines with different colors denote results of different input images. The results of the whole dataset are in Table 5. With the same wavelength, the attack abilities of various orientations of frequency bands are significantly different. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Furthermore, compared with previous adversarial attack methods, the uniqueness of our Frequency attack is that we propose a regular and interpretable perturbation distribution with a powerful adversarial attack ability. The existing adversarial perturbation distributions are hard to be clearly explained, although they directly or indirectly utilize the gradient information of the target model to generate.

Finally, although skeleton detection models are also vulnerable to adversarial attacks. The adversarial attacks dramatically harm the performance of classification (Ganeshan & Babu, 2019; Goodfellow et al., 2015; Kurakin et al., 2017) and semantic segmentation (Arnab et al., 2018), while they moderately decrease the performance of skeleton detection models. As the white-box attack methods, FGSM and IFGSM produce 22.0% and 40.9% performance degradation of the state-of-the-art skeleton detection model DeepFlux, respectively. As the gray-box attack method, FDA obtains 23.2% performance degradation of this DeepFlux model. As the black-box attack methods, SimBA-DCT and Frequency attack yield 1.8% and 26.3% performance degradation of this DeepFlux model, respectively. However, skeleton features are more robust than shape features in the human vision (Ayzenberg et al., 2019; Ayzenberg & Lourenco, 2019). This intrinsic attribute of skeleton features may bring some resistance to adversarial attacks. Anyway, it is hard to absolutely fairly compare the robustness of different tasks due to many uncontrolled factors. Our observations can only suggest to explore other visual tasks by introducing explicit skeleton features with the relatively robust attributes.

What inspiration can it bring us by studying the robustness of skeleton detection against adversarial attacks?

Classification task. The current CNNs are strongly biased toward recognizing local textures rather than shapes (Brendel & Bethge, 2019; Geirhos et al., 2019). The texture feature is a kind of relatively local and high frequency signals. For the classification task, the perturbations in low frequency space possess stronger attack capability (Guo et al., 2019).

Skeleton detection task. The skeleton feature is a kind of relatively global and low frequency signals. Our experiments demonstrate that the high frequency perturbations are more aggressive for skeleton detection models.

Speculation. The model detecting the high frequency features is more likely to be attacked by the low frequency perturbations, while the model detecting the low frequency features is more likely to be attacked by the high frequency perturbations. The possible reason is as follows. The filters in the same layer of the network can detect signals with similar frequencies due to

Table 5
The effect of Frequency attack with different orientations and wavelengths. The used skeleton detection model is DeepFlux on the SK-LARGE dataset. λ denotes the wavelength. β denotes the orientation.

Method	Setting		F-measure
Clean	$\epsilon = 0$		0.732
Random uniform Noises	$\epsilon = 8$		**0.712**
	$\epsilon = 64$		0.517
Frequency(our)	$\epsilon = 8$, $\beta = 2 \times \pi/8$	$\lambda/2 = 1$	0.686
		$\lambda/2 = 2$	0.601
		$\lambda/2 = 3$	**0.574**
		$\lambda/2 = 4$	0.593
		$\lambda/2 = 5$	0.615
		$\lambda/2 = 14$	0.678
		$\lambda/2 = 28$	0.691
		$\lambda/2 = 56$	0.703
		$\lambda/2 = 112$	0.712
	$\epsilon = 8$, $\lambda/2 = 3$	$\beta = 0$	0.668
		$\beta = 1 \times \pi/8$	0.617
		$\beta = 2 \times \pi/8$	0.574
		$\beta = 3 \times \pi/8$	0.602
		$\beta = 4 \times \pi/8$	0.663
		$\beta = 5 \times \pi/8$	0.584
		$\beta = 6 \times \pi/8$	**0.551**
		$\beta = 7 \times \pi/8$	0.600
	$\epsilon = 8$, using our optimization strategy		**0.423**

the same size of receptive fields. The adversarial perturbations are usually limited by L norm, which are much weaker than signals of object itself. (1) When object signals and perturbations with similar frequencies pass through the same layer, object signals are stronger than perturbations. Therefore, more object signals can be reserved through the maximum pooling layers. (2) When object signals and perturbations with large different frequencies pass through the same layer, the perturbations may dominate due to lack of object signals with similar frequencies. It makes perturbations transmit to the output layer to destroy the final prediction results. These observations inspire researchers to explore more robust classification models by involving explicit skeleton features, which will make CNNs to depend on both local and global features.

5. Conclusion

In this paper, we are the first to study the robustness of deep learning-based skeleton detection models against adversarial attacks. The comprehensive experimental results verify that skeleton detection models are also vulnerable to adversarial attacks. Moreover, our Frequency attack applies regular perturbations to achieve a powerful adversarial attack effect. Frequency bands provide a unique perspective to understand the vulnerability of CNNs. The compelling findings will inspire researchers to develop more robust models by involving skeleton features.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under grants (No. 61802297).

References

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

Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. arXiv preprint arXiv:1802.00420.

Ayzenberg, V., Chen, Y., Yousif, S. R., & Lourenco, S. F. (2019). Skeletal representations of shape in human vision: Evidence for a pruned medial axis model. *Journal of Vision*, *19*(6), 1–21.

Ayzenberg, V., & Lourenco, S. F. (2019). Skeletal descriptions of shape provide unique perceptual information for object recognition. *Scientific Reports*, *9*(1), 9359.

Bai, X., Ye, L., Zhu, J., Zhu, L., & Komura, T. (2020). Skeleton filter: A self-symmetric filter for skeletonization in noisy text images. *IEEE Transactions on Image Processing*, *29*, 1815–1826.

Borenstein, E., & Ullman, S. (2002). Class-specific, top-down segmentation. In *European conference on computer vision* (pp. 109–122). Springer.

Brendel, W., & Bethge, M. (2019). Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *International conference on learning representations*.

Brochu, F. (2019). Increasing shape bias in imagenet-trained networks using transfer learning and domain-adversarial methods.. *arXiv: Computer Vision and Pattern Recognition*.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., et al. (2015). Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848.

Choi, J.-H., Zhang, H., Kim, J.-H., Hsieh, C.-J., & Lee, J.-S. (2019). Evaluating robustness of deep image super-resolution against adversarial attacks. In: *Proceedings of the IEEE international conference on computer vision* (pp. 303–311).

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., et al. (2018). Boosting adversarial attacks with momentum. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9185–9193).

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Ganeshan, A., & Babu, R. V. (2019). FDA: Feature disruptive attack. In: *Proceedings of the IEEE international conference on computer vision* (pp. 8069–8079).

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *ICLR*.

Guo, C., Gardner, J. R., You, Y., Wilson, A. G., & Weinberger, K. Q. (2019). Simple black-box adversarial attacks. In *International conference on machine learning* (pp. 2484–2493).

Hofstadter, D. R. (1985). *Metamagical themas: Questing for the essence of mind and pattern*. Basic Books.

Hung, C.-C., Carlson, E. T., & Connor, C. E. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron*, *74*(6), 1099–1113.

Janai, J., Güney, F., Behl, A., & Geiger, A. (2017). Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. arXiv preprint arXiv:1704.05519.

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*(6), 1233–1258.

Ke, W., Chen, J., Jiao, J., Zhao, G., & Ye, Q. (2017). SRN: side-output residual network for object symmetry detection in the wild. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1068–1076).

Kovacs, I., & Julesz, B. (1994). Perceptual sensitivity maps within globally defined visual shapes. *Nature*, *370*(6491), 644–646.

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533.

Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial machine learning at scale. *ICLR*.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lescroart, M. D., & Biederman, I. (2013). Cortical representation of medial axis structure. *Cerebral Cortex*, *23 3*, 629–637.

Liu, C., Ke, W., Qin, F., & Ye, Q. (2018). Linear span network for object skeleton detection. In: *Proceedings of the European conference on computer vision (ECCV)* (pp. 133–148).

Liu, J., Wang, Z., & Liu, H. (2020). Hds-sp: A novel descriptor for skeleton-based human action recognition. *Neurocomputing*, *385*, 22–32.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B: Biological Sciences*, *200*(1140), 269–294.

Martin, D., Fowlkes, C., Tal, D., Malik, J., et al. (2001). *A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics*. Iccv Vancouver.

Metzen, J. H., Kumar, M. C., Brox, T., & Fischer, V. (2017). Universal adversarial perturbations against semantic image segmentation. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 2774–2783). IEEE.

Mignotte, M. (2016). Symmetry detection based on multiscale pairwise texture boundary segment interactions. *Pattern Recognition Letters*, *74*, 53–60.

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

Mygdalis, V., Tefas, A., & Pitas, I. (2020). K-anonymity inspired adversarial attack and multiple one-class classification defense. *Neural Networks*, *124*, 296–307.

Oregi, I., Ser, J. D., Pérez, A., & Lozano, J. A. (2020). Robust image classification against adversarial attacks using elastic similarity measures between edge count sequences. *Neural Networks*, *128*, 61–72.

Ritter, S., Barrett, D. G. T., Santoro, A., & Botvinick, M. (2017). Cognitive psychology for deep neural networks: a shape bias case study. In *International conference on machine learning* (pp. 2940–2949).

Shen, W., Bai, X., Hu, Z., & Zhang, Z. (2016). Multiple instance subspace learning via partial random projection tree for local reflection symmetry in natural images. *Pattern Recognition*, *52*, 306–316.

Shen, W., Zhao, K., Jiang, Y., Wang, Y., Bai, X., & Yuille, A. (2017). Deepskeleton: Learning multi-task scale-associated deep side outputs for object skeleton extraction in natural images. *IEEE Transactions on Image Processing*, *26*(11), 5298–5311.

Shen, W., Zhao, K., Jiang, Y., Wang, Y., Zhang, Z., & Bai, X. (2016). Object skeleton extraction in natural images by fusing scale-associated deep side outputs. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 222–230).

Sing Lee, T., Mumford, D., Romero, R., & Lamme, V. (1998). The role of the primary cortex in higher level vision. *Vision Research*, *38*, 2429–2454. http://dx.doi.org/10.1016/S0042-6989(97)00464-1.

Sironi, A., Lepetit, V., & Fua, P. (2014). Multiscale centerline detection by learning a scale-space distance transform. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2697–2704).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

Tsogkas, S., & Dickinson, S. (2017). Amat: Medial axis transform for natural images. In *Computer vision (ICCV), 2017 IEEE international conference on* (pp. 2727–2736). IEEE.

Tsogkas, S., & Kokkinos, I. (2012). Learning-based symmetry detection in natural images. In *European conference on computer vision* (pp. 41–54). Springer.

Vidnerová, P., & Neruda, R. (2020). Vulnerability of classifiers to evolutionary generated adversarial examples. *Neural Networks*, *127*, 168–181.

Wang, Y., Xu, Y., Tsogkas, S., Bai, X., Dickinson, S., & Siddiqi, K. (2019). DeepFlux for Skeletons in the Wild. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5287–5296).

Xiao, Q., Qin, M., & Yin, Y. (2020). Skeleton-based chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural Networks*, *125*, 41–55.

Xie, S., & Tu, Z. (2015). Holistically-nested edge detection. In: *Proceedings of the IEEE international conference on computer vision* (pp. 1395–1403).

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

Xie, C., Wu, Y., Maaten, L. v. d., Yuille, A. L., & He, K. (2019). Feature denoising for improving adversarial robustness. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 501–509).

Yu, Z., & Lee, M. (2015). Real-time human action classification using a dynamic neural model. *Neural Networks*, *69*, 29–43.

Zhao, K., Shen, W., Gao, S., Li, D., & Cheng, M.-M. (2018). Hi-fi: hierarchical feature integration for skeleton detection. In *Proceedings of the 27th international joint conference on artificial intelligence* (pp. 1191–1197). AAAI Press.

Zhu, J., Zou, W., Zhu, Z., & Hu, Y. (2019). Convolutional relation network for skeleton-based action recognition. *Neurocomputing*, *370*, 109–117.