Frequency-driven Imperceptible Adversarial Attack on Semantic Similarity

Cheng Luo* Qinliang Lin* Weicheng Xie† Bizhu Wu Jinheng Xie Linlin Shen
[1]Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University
[2]Shenzhen Institute of Artificial Intelligence & Robotics for Society
[3]Guangdong Key Laboratory of Intelligent Information Processing
{luocheng2020, linqinliang2021}@email.szu.edu.cn, {wcxie, llshen}@szu.edu.cn

Abstract

Current adversarial attack research reveals the vulnerability of learning-based classifiers against carefully crafted perturbations. However, most existing attack methods have inherent limitations in cross-dataset generalization as they rely on a classification layer with a closed set of categories. Furthermore, the perturbations generated by these methods may appear in regions easily perceptible to the human visual system (HVS). To circumvent the former problem, we propose a novel algorithm that attacks semantic similarity on feature representations. In this way, we are able to fool classifiers without limiting attacks to a specific dataset. For imperceptibility, we introduce the low-frequency constraint to limit perturbations within high-frequency components, ensuring perceptual similarity between adversarial examples and originals. Extensive experiments on three datasets (CIFAR-10, CIFAR-100, and ImageNet-1K) and three public online platforms indicate that our attack can yield misleading and transferable adversarial examples across architectures and datasets. Additionally, visualization results and quantitative performance (in terms of four different metrics) show that the proposed algorithm generates more imperceptible perturbations than the state-of-the-art methods. Code is made available at https://github.com/LinQinLiang/SSAH-adversarial-attack.

1. Introduction

With the advent of deep learning, neural network models [10, 12, 16, 32] have demonstrated revolutionary performance in recognition tasks of real-world datasets. Nevertheless, the vulnerability of deep neural networks (DNNs) to image corruptions and adversarial examples has been unveiled [8, 35]. This problem hinders the applications of DNNs in security-critical domains and promotes research on understanding the robustness of DNNs, including adversarial attack [1, 8] and defense [22, 36, 40, 44].

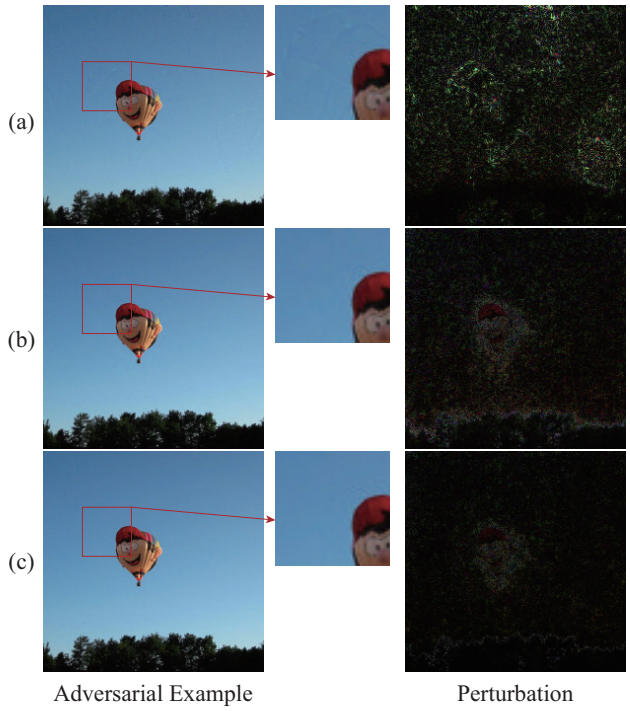

Figure 1. Comparison of the adversarial examples and perturbations generated by three different attack methods: (a) C&W, (b) Our SSA (semantic similarity attack), and (c) Our SSAH (semantic similarity attack on high-frequency components). For the visualization, we regularize the perturbation by taking its absolute value and multiplying it by 25.

The most intuitive approaches for white-box attacking are to increase the cost of the classification loss [8] to yield adversarial examples via gradient descent. Besides, they further apply ℓ_p distance to constrain the visual differences between benign and perturbed images. However, conventional approaches may suffer from the two open problems:

- **Inherent limitation in cross-dataset generalization.**

*Equal Contribution
†Corresponding Author

Due to the classification layer with learned weight vectors representing specific class proxies, current attack paradigms based on a white-box or surrogate classifier are limited to this setting, where images of the model training and attack domains are from the same set of categories. In real-world scenarios, however, an image from an open set [25] may belong to an unknown category to the classifier.

- **Poor imperceptibility to HVS.** Sharif *et al*. [30] have demonstrated that the ℓ_p distance metric is insufficient for assessing perceptual similarity. In other words, visual imperceptibility may not be explicitly reflected using only the perturbation intensity. For instance, C&W [1], a well-known attack method, generates easy-to-perceive perturbations on the smooth background, as shown in Fig. 1 (a).

Intuitively, a natural approach to circumvent the classification layer is to perform attacks in the feature space. In this work, we propose a general adversarial attack, namely semantic similarity attack (SSA), which builds on the similarity of feature representations. More specifically, we push apart the representations of adversarial and benign examples but pull that of adversarial and target (the most dissimilar) examples together. In this way, we can fool classifiers without the knowledge of the specific image category. The underlying assumption is that the high-level representation implies image discrimination and semantics. Hence, perturbing such representation can guide perturbations towards semantic regions within pixel space. As shown in Fig. 1 (b), SSA focuses on perturbing semantic regions such as objects in the scene while suppressing redundant perturbations on irrelevant regions.

In addition to ℓ_p norms [1,2,26,27], other measures such as CIEDE2000 [46], SSIM [9] and LPIPS [18] are applied to approximate perceptual similarity. In this work, we provide a different metric from the frequency domain perspective. Generally, the low-frequency component of an image contains the basic information, whereas the high-frequency components represent trivial details and noise. Inspired by it, we measure the variations of low-frequency components as the perceptual variations in image pixel space. We further build a low-frequency constraint to limit the perturbations within imperceptible high-frequency components. As depicted in Fig. 1 (c), the perturbations generated by the proposed framework, *i.e.*, SSAH, appear mostly on imperceptible regions such as object edges. Some works show that adversarial examples may be neither in high-frequency nor low-frequency components [23], and low-frequency perturbations with much perceptibility are especially effective for attacking defended models [31]. Nevertheless, we consider that developing attacks in high-frequency components is significant, as it helps improve perturbation imperceptibility to HVS and learn robust models that better align with human perception. Recent works [38, 41] also prove that these high-frequency signals are barely perceivable to HVS but can largely determine the prediction results of DNNs.

The main contributions can be summarized as follows:

- We propose a novel adversarial attack, SSA, which is applicable in wide settings by attacking the semantic similarity of images.
- We present a new perturbation constraint, the low-frequency constraint, into the joint optimization of SSA to limit perturbations within the imperceptible high-frequency components.
- We conduct extensive experiments on three datasets, *i.e.*, CIFAR-10, CIFAR-100, and ImageNet-1K, and the experiment results show that our proposed attack outperforms the state-of-the-art methods by significantly imperceptible perturbations.
- Experimental results demonstrate that adversarial perturbations generated by our SSAH are more transferable across different architectures and datasets.

2. Related Work

Feature Space Attack. Feature space attacks [13, 29] manipulate the image representation to appear remarkably similar to the target image from a different class. The same goal of these methods is to directly minimize the Euclidian distance between intermediate layer features of source and target images in the target DNN. Similar attacks recently applied to Person Re-identification [37, 39] or Image Retrieval [7, 19] generate adversarial perturbations or patterns by minimizing the distance of the inter-class pair while maximizing the distance of the intra-class pair. In this work, we perturb the class-specific representations of image instances from the perspective of feature similarity and design a more flexible optimization scheme.

Imperceptible Attack. A rich line of works [1, 8, 9, 14, 18,45,46] resort to devising perceptual similarity metrics to constrain perturbations during adversarial example generation. Among these metrics, ℓ_p norms of perturbations are generally employed [1]. However, recent works have revealed that ℓ_p norms do not well align with human perception. Thus, other perceptual distances in terms of similarity of object structures [9], edges [14], color [46], and Learned Perceptual Image Patch Similarity (LPIPS) [18,45] are proposed to improve the imperceptibility of perturbations. In this work, we decompose images into various frequency components by wavelets and measure image pair similarity via the distance of their low-frequency components.

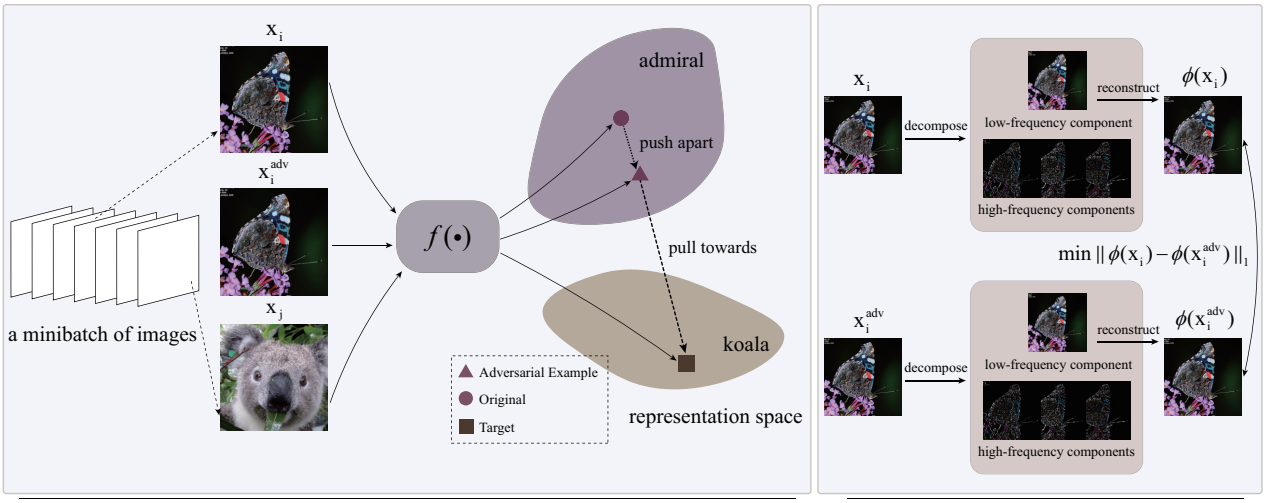

Figure 2. An overview of proposed SSAH. Left: Semantics Similarity Attack; Right: Low-frequency Constraint. $f(\cdot)$ is the mapping from an image to its embedding in representation space. $\phi(\cdot)$ is a shallow network that decomposes an image into different frequency components and reconstructs it using the low-frequency component.

Wavelets in Deep Learning. Wavelet is an effective tool for time-frequency analysis, and Discrete Wavelet Transform (DWT) is frequently used to decompose image data into various frequency components. Recent works [6, 21] explore implementing wavelet transform in deep learning for various visual tasks such as image segmentation. In particular, Li *et al*. [20] design a DWT/IDWT layer, making discrete wavelet transform easily applicable in DNNs.

3. Methodology

In a white-box setup, an adversary can access details of the target classifier (*i.e.*, architectures, parameters, gradients of the loss with respect to (w.r.t.) the input) to craft an adversarial example $\boldsymbol{x}^{adv} = \boldsymbol{x} + \boldsymbol{\delta}$ with the image perturbation $\boldsymbol{\delta}$ to the benign example $\boldsymbol{x}$. Generally, a distance metric $\mathcal{D}$ is required to quantify the perceptual similarity (between an adversarial example and its original one) and is used as the constraint of the perturbation. We can formulate the adversarial examples in the untargeted attack scenario as a solution to the following problem:

$$\boldsymbol{x}^{adv} = \boldsymbol{x} + \arg\min_{\boldsymbol{\delta}} \{\mathcal{D}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\delta}) | \arg\max_{i} \{z'_i\} \neq y\}, \tag{1}$$

where $z'_i = \boldsymbol{w}_i^T f(\boldsymbol{x}+\boldsymbol{\delta})$ denotes the logit (*i.e.*, the similarity score between the embedding vector $f(\boldsymbol{x}+\boldsymbol{\delta})$ of the example and the weight vector $\boldsymbol{w}_i$ $(i = 1, 2, \ldots, C)$, y denotes the ground-truth label and C is the number of classes.

In this work, we propose a novel semantic similarity attack on high-frequency components (SSAH), and the framework is depicted in Fig. 2. SSAH is composed of an attack paradigm (semantic similarity attack) and a new perturbation constraint (low-frequency constraint). The semantic similarity attack does not require the classification layer but tends to change the similarity of pairwise feature representations. The low-frequency constraint preserves the basic information of objects and limits perturbations within imperceptible high-frequency components.

3.1. Semantics Similarity Attack

3.1.1 Attack Design

Conventional white-box attack methods solve the problem presented in Eq. (1) by maximizing the classification loss or changing logits. However, given a minibatch of N instances, *i.e.*, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]$, we instead optimize the representation of the i-th adversarial example $\boldsymbol{x}_i^{adv}$ as:

$$\boldsymbol{x}_i^{adv} = \arg\min_{\boldsymbol{x}'_i} [s'_{i,i} - \min\{s'_{i,j} | j \neq i\}]_+, \tag{2}$$

where $[\cdot]_+$ denotes $\max(\cdot, 0)$, $\boldsymbol{x}'_i$ is the optimization variable and initialized as $\boldsymbol{x}_i$, $s'_{i,i} = \mathrm{sim}(f(\boldsymbol{x}'_i), f(\boldsymbol{x}_i))$ and $s'_{i,j} = \mathrm{sim}(f(\boldsymbol{x}'_i), f(\boldsymbol{x}_j))$ are similarity scores. In our method, we use cosine similarity of embeddings, which is defined as:

$$s'_{i,j} = \frac{f(\boldsymbol{x}'_i)^T f(\boldsymbol{x}_j)}{\|f(\boldsymbol{x}'_i)\|_2 \|f(\boldsymbol{x}_j)\|_2}. \tag{3}$$

Likewise, we can define the attack in the targeted scenario as:

$$\boldsymbol{x}_i^{adv} = \arg\min_{\boldsymbol{x}_i'} [s'_{i,i} - s'_{i,t}]_+, \tag{4}$$

where t denotes the index of the target image in the mini-batch. Eq. (4) aims to encourage the adversarial example $\boldsymbol{x}_i'$ to be close to the target $\boldsymbol{x}_t$ in terms of the feature representation. Without loss of generality, we only discuss the case of untargeted attacks.

The attack objective in Eq. (1) changes logits of the classification layer. In other words, it pushes the embedding of an adversarial example apart from its ground-truth class centroid. By contrast, we directly change pair-wise similarity: reducing the similarity between the adversarial example and its original, while increasing the similarity between the adversarial example and its most dissimilar one in the mini-batch. In this way, our attack misleads a classifier to map the example representation into a different subspace.

3.1.2 Self-paced Weighting

To avoid redundant perturbations, we design a self-paced weighting scheme to improve the optimization flexibility. This scheme design is inspired by Circle loss [33] that uses it in metric learning. It aims at adjusting the optimization pace for each similarity score as:

$$\begin{aligned} \boldsymbol{x}_i^{adv} &= \arg\min_{\boldsymbol{x}_i'} \mathcal{L}_{SSA}(\boldsymbol{x}_i, \boldsymbol{x}_i') \\ &= \arg\min_{\boldsymbol{x}_i'} [\alpha_i s'_{i,i} - \beta_i \min\{s'_{i,j} | j \neq i\}]_+, \end{aligned} \tag{5}$$

where α_i and β_i are adjusted in a self-paced manner as:

$$\begin{cases} \alpha_i = [s'_{i,i} - m]_+, \\ \beta_i = [1 + m - \min\{s'_{i,j} | j \neq i\}]_+, \end{cases} \tag{6}$$

where $m \geq 0$ is a pre-defined margin. Eq. (6) is the weight factor setting for our attack. Compared to the optimization of $(s'_{i,i} - s'_{i,j})$ in Eq. (2), we introduce the adaptive weighting as $(\alpha_i s'_{i,i} - \beta_i s'_{i,j})$. During the optimization of variable $\boldsymbol{x}_i'$, the gradient with respect to $(\alpha_i s'_{i,i} - \beta_i s'_{i,j})$ is multiplied with α_i (β_i) when back-propagated to $s'_{i,i}$ $(s'_{i,j})$. Consequently, the similarity score close to its optimum is assigned with a smaller gradient, whereas the less optimized similarity score is assigned with a larger gradient.

3.2. Low-frequency Constraint

Although SSA yields perturbations in the representation space, there is still a risk that these perturbations may distribute in regions perceptible to HVS. Conventional constraints, on the other hand, may result in a random distribution of perturbation. Therefore, we seek a new constraint into the joint optimization of SSA, limiting perturbations into imperceptible details of objects.

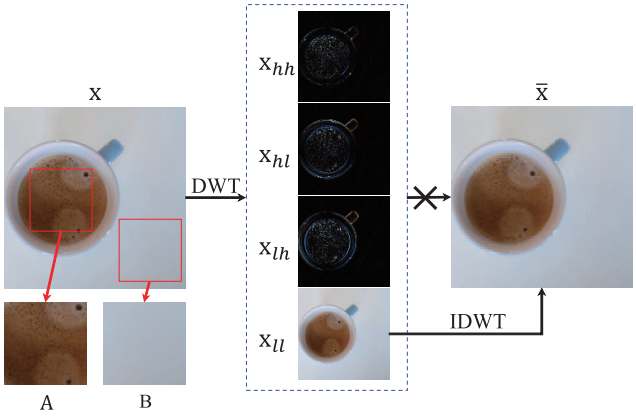

Figure 3. Illustration of our image decomposition and reconstruction by wavelet transforms. An image $\boldsymbol{x}$ with complex (*e.g.*, Part A) and smooth (*e.g.*, Part B) contexts can be decomposed into the low-frequency component ($\boldsymbol{x}_{ll}$) and high-frequency components ($\boldsymbol{x}_{lh}$, $\boldsymbol{x}_{hl}$ and $\boldsymbol{x}_{hh}$) by Discrete Wavelet Transform (DWT). The reconstructed image $\bar{\boldsymbol{x}}$ has the same fundamental shape and resolution as the original image $\boldsymbol{x}$.

We observe that HVS is more sensitive to object structures and smooth regions, whereas it is not easy to perceive object edges and complex textures. For example, the perturbations hidden in the dense bubbles of Part A (in Fig. 3) are more invisible than those in the smooth background like Part B. It motivates us to limit perturbations into regions less sensitive to HVS.

From a frequency domain perspective, the high-frequency components representing noise and textures are more imperceptible than the low-frequency component containing basic object structure. As a time-frequency analysis tool, discrete wavelet transform (DWT) can decompose an image $\boldsymbol{x}$ into one low-frequency and three high-frequency components, *i.e.*, $\boldsymbol{x}_{ll}$, $\boldsymbol{x}_{lh}$, $\boldsymbol{x}_{hl}$, $\boldsymbol{x}_{hh}$ as:

$$\begin{aligned} \boldsymbol{x}_{ll} = \boldsymbol{L}\boldsymbol{x}\boldsymbol{L}^T, \boldsymbol{x}_{lh} = \boldsymbol{H}\boldsymbol{x}\boldsymbol{L}^T, \\ \boldsymbol{x}_{hl} = \boldsymbol{L}\boldsymbol{x}\boldsymbol{H}^T, \boldsymbol{x}_{hh} = \boldsymbol{H}\boldsymbol{x}\boldsymbol{H}^T, \end{aligned} \tag{7}$$

where $\boldsymbol{L}$ and $\boldsymbol{H}$ are the low-pass and high-pass filters of an orthogonal wavelet, respectively. As shown in Fig. 3, $\boldsymbol{x}_{ll}$ preserves the low-frequency information of the original image, whereas $\boldsymbol{x}_{lh}$, $\boldsymbol{x}_{hl}$ and $\boldsymbol{x}_{hh}$ are associated with edges and drastic variations.

Normally, inverse DWT (IDWT) uses all four components to reconstruct the image. In this work, we drop the high-frequency components and reconstruct an image with only the low-frequency component as $\bar{\boldsymbol{x}} = \phi(\boldsymbol{x})$, where

$$\phi(\boldsymbol{x}) = \boldsymbol{L}^T \boldsymbol{x}_{ll} \boldsymbol{L} = \boldsymbol{L}^T (\boldsymbol{L}\boldsymbol{x}\boldsymbol{L}^T) \boldsymbol{L}. \tag{8}$$

Based on this process of image decomposition and reconstruction, we can obtain the main image information. It means that we can assess the perceptual similarity between

two images in terms of the main information. On that basis, we develop a new constraint between $\boldsymbol{x}$ and $\boldsymbol{x}'$:

$$\mathcal{D}_{lf}(\boldsymbol{x}, \boldsymbol{x}') = \|\phi(\boldsymbol{x}) - \phi(\boldsymbol{x}')\|_1. \tag{9}$$

Consequently, the loss of perceptual information specific to perturbed images is reduced by minimizing Eq. (9).

3.3. The Unified Attack

We define the objective of SSAH as the semantic similarity attack SSA under the new constraint $\mathcal{D}_{lf}$. The adversarial example $\boldsymbol{x}_i^{adv}$ can be obtained as:

$$\begin{aligned} \boldsymbol{x}_i^{adv} &= \arg\min_{\boldsymbol{x}_i'} \mathcal{L}_{SSAH}(\boldsymbol{x}_i, \boldsymbol{x}_i') \\ &= \lambda \mathcal{D}_{lf}(\boldsymbol{x}_i, \boldsymbol{x}_i') + \mathcal{L}_{SSA}(\boldsymbol{x}_i, \boldsymbol{x}_i'), \end{aligned} \tag{10}$$

where λ is a hyperparameter specific to the low-frequency constraint. In practice, we replace $\boldsymbol{x}_i'$ with a variable $\boldsymbol{r}_i = \arctanh(2\boldsymbol{x}_i' - 1)$ for optimization. For clarity, we present the pseudo-code in Algorithm 1 to outline the main procedures of our SSAH.

Algorithm 1 Adversarial attack with SSAH

Require: A minibatch of original images $\{\boldsymbol{x}_i\}_{i=1}^N$; the number of iterations K; the encoder $f(\cdot)$ of a classifier.

Initialize $\{\boldsymbol{x}_i'\}_{i=1}^N$ with $\{\boldsymbol{x}_i\}_{i=1}^N$;
for $i = 1$ to N **do**
  Initialize the variable $\boldsymbol{r}_i$ as $\arctanh(2\boldsymbol{x}_i' - 1)$;
  for $k = 1$ to K **do**
    Calculate the cosine similarity scores $\{s_{i,j}'\}_{j=1}^N$ as in Eq. (3) and use $s_{i,i}'$ and the lowest similarity score $\min\{s_{i,j}' | j \neq i\}$ in Eq. (5);
    Calculate the constraint loss $\mathcal{D}_{lf}(\boldsymbol{x}_i, \boldsymbol{x}_i')$ as in Eq. (9);
    Optimize the variable $\boldsymbol{r}_i$ by minimizing $\mathcal{L}_{SSAH}(\boldsymbol{x}_i, \boldsymbol{x}_i')$ as in Eq. (10) and obtain $\boldsymbol{x}_i'$ through $\boldsymbol{r}_i$;
  end for
end for
return $\{\boldsymbol{x}_i'\}_{i=1}^N$.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate the performance of our method on three general datasets, namely CIFAR-10 [15], CIFAR-100 [15], and ImageNet-1K [28]. In particular, CIFAR-10 contains 50K training samples and 10K testing samples with the size of 32×32 from 10 classes; CIFAR-100 has 100 classes, containing the same number of training (testing) samples as CIFAR-10; ImageNet-1K has 1K classes, containing about 1.3M images for training and 50K images for validation.

Implementation details. Adam optimizers with the learning rates of 0.01, 0.01, and 0.001 are used for C&W, PerC-AL [46] and our SSAH, respectively. The default values for the hyperparameters m in Eq. (6) and λ in Eq. (10) are 0.2 and 0.1, respectively. The perturbation budget (ϵ) is set to 8/255 under the ℓ_∞ for BIM [17], PGD [22], AA (AutoAttack) [3] and MIM [4], respectively. This budget is specified with the iterative step size $\alpha = 1/255$. We use ResNet-20 models that achieve the 7.4% and 30.4% top-1 test errors on CIFAR-10 and CIFAR-100, respectively, as the white-box model for these two datasets. For ImageNet-1K, pre-trained ResNet-50 that achieves the 23.85% top-1 error is employed. For the DWT/IDWT layer in our low-frequency constraint, Haar wavelet is used. All our experiments are conducted on a NVIDIA A100 GPU with 40GB memory.

Evaluation metrics. For the performance evaluation and comparison, we use the attack success rate (ASR) and four different metrics, including conventional average ℓ_2 distortion, maximum perturbation intensity (ℓ_∞), Fréchet Inception Distance (FID) [11] and a newly introduced metric, *i.e.*, average distortion of low-frequency components (LF) based on 2D DWT, for approximating the perceptual similarity. LF (LF $= \frac{1}{N}\sum_{i=1}^N \|\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_i^{adv})\|_2$) is employed to quantify average variations of the basic structure information between the original and adversarial examples.

4.2. White-box Attacks

In this section, we evaluate the adversarial strength and imperceptibility of the examples generated by different approaches in a white-box scenario, where the knowledge of the target system is fully accessible.

Tab. 1 shows the performances of nine attack approaches in terms of five different metrics. It demonstrates that our attack, with the lowest ℓ_p (*i.e.*, ℓ_2 or ℓ_∞) norm of perturbations, is successful on all three datasets. More importantly, our semantic similarity attack (SSA), without tightly constraining the ℓ_p norms or other perceptual distances, can generate perturbations that are imperceptible.

Generally, FID is consistent with human judgment and well reflects the level of disturbance. It calculates the distance between the benign and perturbed images in the feature space of an Inception-v3 [34] network. The proposed SSAH achieves the FID of 3.90 on Imagenet-1K, which outperforms the state-of-the-art models like PerC-AL (11.56 FID) by a large margin. Such improvement suggests that our feature-oriented attack generates adversarial examples with more realistic visual effects in pixel space but fewer variations in feature space.

Tab. 1 also shows that our attack significantly outperforms the other methods among all cases in terms of LF. It implies that our attack can effectively preserve the ob-

Dataset	Attack	Iteration	RunTime (s) ↓	ASR (%) ↑	ℓ_2 ↓	ℓ_∞ ↓	FID ↓	LF ↓
CIFAR-10	BIM [17]	10	35	100	0.85	0.03	14.85	0.23
	PGD [22]	10	37	100	1.28	0.03	27.86	0.34
	MIM [4]	10	46	100	1.90	0.03	26.00	0.48
	AA ℓ_∞ [3]	100	184	100	1.91	0.03	34.93	0.61
	AdvDrop [5]	150	392	99.92	0.90	0.07	16.34	0.34
	C&W ℓ_2 [1]	1000	991	100	0.39	0.06	8.23	0.11
	PerC-AL [46]	1000	1221	98.29	0.86	0.18	9.58	0.15
	SSA (ours)	150	192	99.96	0.29	**0.02**	5.73	0.07
	SSAH (ours)	150	198	99.94	**0.26**	**0.02**	**5.03**	**0.03**
CIFAR-100	BIM [17]	10	34	99.99	0.85	**0.03**	15.26	0.32
	PGD [22]	10	31	99.99	1.29	**0.03**	27.74	0.42
	MIM [4]	10	30	99.99	1.87	**0.03**	26.04	0.65
	AA ℓ_∞ [3]	100	184	100	1.91	**0.03**	33.86	0.61
	AdvDrop [5]	150	332	99.93	0.80	0.07	15.59	0.31
	C&W ℓ_2 [1]	1000	751	100	0.52	0.07	11.04	0.19
	PerC-AL [46]	1000	919	99.61	1.41	0.21	12.83	0.37
	SSA (ours)	150	150	99.90	0.48	**0.03**	9.68	0.17
	SSAH (ours)	150	149	99.80	**0.45**	**0.03**	**9.20**	**0.13**
ImageNet-1K	BIM [17]	10	3998	99.98	26.85	0.03	51.92	11.18
	PGD [22]	10	3451	99.98	54.97	0.03	45.51	17.41
	MIM [4]	10	7847	99.98	91.78	0.03	101.88	39.42
	AA ℓ_∞ [3]	100	27312	96.97	71.62	0.03	77.49	30.45
	AdvDrop [5]	150	48355	99.76	14.95	0.06	11.28	5.67
	C&W ℓ_2 [1]	1000	$>$ 100000	99.27	**1.51**	0.04	12.14	0.67
	PerC-AL [46]	1000	$>$ 100000	98.78	4.35	0.12	11.56	1.59
	SSA (ours)	200	35414	98.56	2.34	**0.01**	4.63	1.05
	SSAH (ours)	200	38018	98.01	1.81	**0.01**	**3.90**	**0.06**

Table 1. Results of the attack success rate (ASR) and three metrics related with perceptual similarity by nine attack approaches in the untargeted scenario. The best results are marked in bold.

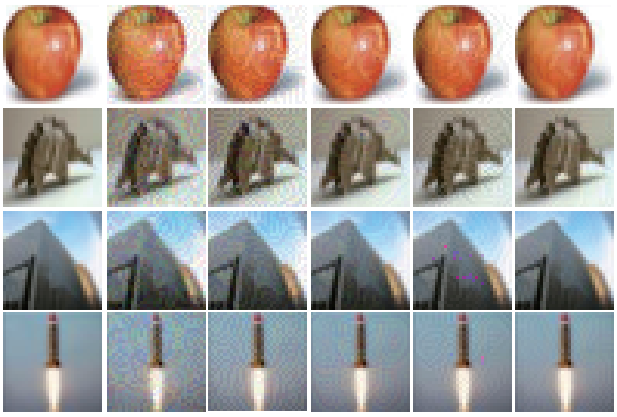

Figure 4. Adversarial examples generated by five different attack approaches on CIFAR-100.

ject structure and the low-frequency component. The results specific to the targeted attack scenario follow a similar pattern and can be found in the supplementary material.

Fig. 4 shows the adversarial examples generated by five approaches on CIFAR-100. Meanwhile, Fig. 5 displays adversarial examples and perturbations (with the same regularization as Fig. 1) of higher resolution images from ImageNet-1K. It is observed that images produced by our SSAH appear more natural to HVS.

4.3. Robustness

Defense	Attack	CIFAR-10	CIFAR-100
FSAT [43]	No Attack	89.98	74.11
	AdvDrop [5]	70.26	40.83
	C&W [1]	60.60	25.04
	PerC-AL [46]	89.80	74.00
	SSAH (ours)	**60.43**	**4.85**
TRADES [44]	No Attack	84.92	56.94
	AdvDrop [5]	84.42	56.37
	C&W [1]	81.24	**48.51**
	PerC-AL [46]	84.70	56.90
	SSAH (ours)	**78.68**	49.23

Table 2. Recognition accuracy (%) of two defense methods under four white-box attacks.

To study the robustness of the proposed attack, we com-

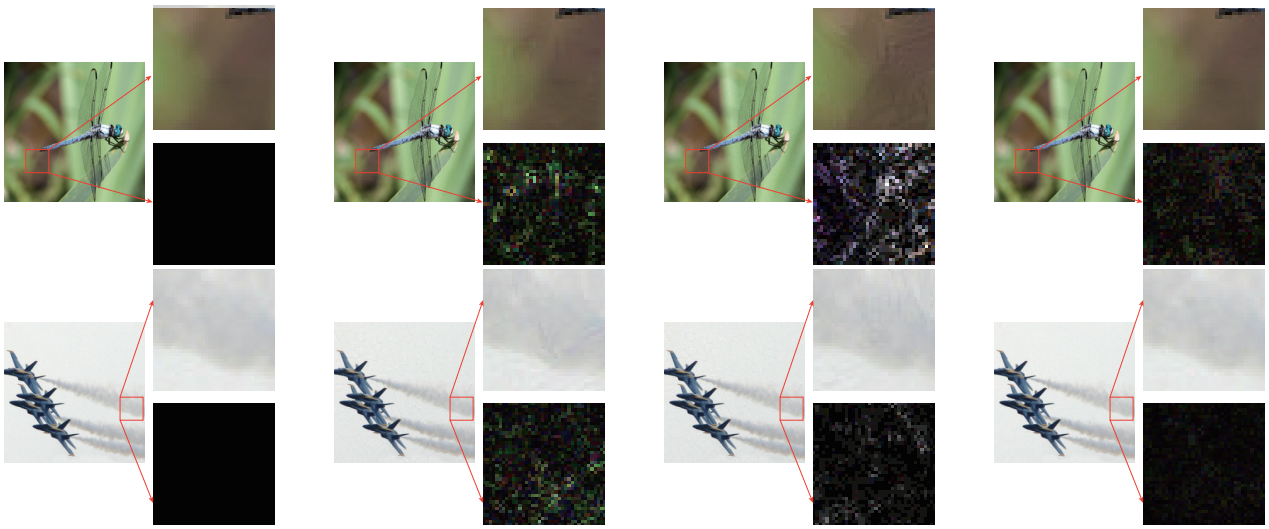

Figure 5. Adversarial examples and perturbations generated by three attack approaches on two high-resolution images from ImageNet-1K. This figure is best viewed in color/screen.

pare the attack success rates of four attack approaches against two defense schemes (FSAT [43] and TRADES [44]). The same network architecture as [44], *i.e.*, WRN-34-10 introduced in [42], is used to generate adversarial perturbations.

Based on the network trained with the defense method FSAT, our attack decreases the model accuracy by a large margin, *i.e.*, 29.55% on CIFAR-10 and 69.26% on CIFAR-100, and largely outperforms other approaches on CIFAR-100, as shown in Tab. 2. Against the more robust defense method (*i.e.*, TRADES), SSAH can still achieve competitive results. For CIFAR-10, an improvement of 2.56% is achieved by SSAH compared with C&W.

4.4. Transferability

To study the transferability of the proposed algorithm in an open-set setting, we evaluate adversarial examples transferred across both architectures and datasets. That is, without the knowledge of the training set and architecture of a black-box model (*e.g.*, ResNet-18), we study to which extent attack approaches, based on another architecture (*e.g.*, ResNet-20) trained on another dataset (*e.g.*, CIFAR-10), affect the classification of this black-box model on the validation set (*e.g.*, ImageNet-1K). We use Gaussian noise and input-agnostic perturbations (*i.e.*, GD-UAP [24]) as the baselines and the ℓ_∞-norm bound of 10/255 for the perturbation generation. The experimental results in Tab. 3 show that our SSAH significantly outperforms these two baselines for eight out of eight cases.

To test attack effectiveness in real-world scenarios, we conduct experiments of attacking the online models on Microsoft Azure[1], Tencent Cloud[2] and Baidu AI Cloud[3]. The model and training data used in their platforms are completely unknown to us. We randomly sample 200 images from the ImageNet-1K validation set (the image names are listed in our supplementary material) and perturb them using four attack approaches on ResNet-152. Tab. 4 shows the attack success rates of these attacks against the online models. In Tab. 4, the proposed attack, without any classification query, achieves an attack success rate of 37.98% against the online model on Tencent Cloud, which outper-

Surrogate	Training set	Attack	ResNet-18	VGG-16
-	-	Gaussian Noise	9.18	10.20
ResNet-20	CIFAR-10	GD-UAP [24]	14.09	11.44
		SSAH (ours)	**17.66**	**16.87**
ResNet-20	CIFAR-100	GD-UAP [24]	12.72	10.22
		SSAH (ours)	**18.31**	**18.68**
VGG-11	CIFAR-10	GD-UAP [24]	14.63	12.38
		SSAH (ours)	**16.92**	**17.52**
VGG-11	CIFAR-100	GD-UAP [24]	12.93	10.39
		SSAH (ours)	**17.92**	**19.14**

Table 3. The attack success rates (%) of transferring adversarial examples across different architectures and datasets. The first column (Surrogate) and the second column (Training set) represent the surrogate's architecture and training set, respectively. The target classifier (*i.e.*, ResNet-18 or VGG-16) is trained on a different dataset (*i.e.*, ImageNet-1K) and the validation set of ImageNet-1K is used for the testing.

[1]https://azure.microsoft.com/
[2]https://cloud.tencent.com/
[3]https://cloud.baidu.com/

Attack	Microsoft Azure	Tencent Cloud	Baidu AI Cloud
AdvDrop [5]	16.26	15.83	17.82
C&W [1]	13.82	21.71	27.72
PerC-AL [46]	15.45	18.61	18.81
SSAH (ours)	**18.70**	**37.98**	**36.63**

Table 4. The attack success rates (%) of transferring adversarial examples to three online models.

forms the other approaches by a large margin.

4.5. Analysis

In this section, we give insight into the working mechanism of the proposed attack and study the behavior of each component in the attack.

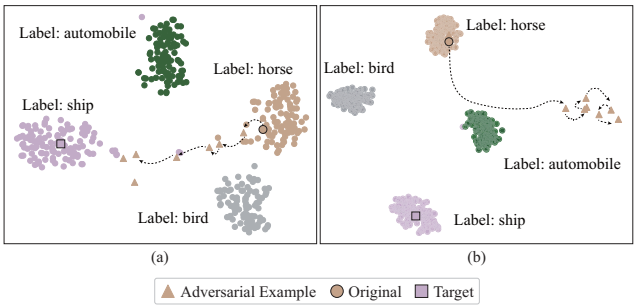

Figure 6. The 2D feature representation of the adversarial examples using the t-SNE algorithm under (a) SSAH and (b) C&W. An adversarial example representation gradually updates from its original class (horse) to the selected target class (ship). The results in the iteration of 10, 15, 20, 30, and 40 are presented.

To study the adversarial example generation of the proposed semantic similarity attack, we visualized the iterative adversarial examples on the 2D plane in Fig. 6. In this analysis, a subset of random instances of four classes from CIFAR-10 is used for visualization. Fig. 6 shows our attack can iteratively push the adversarial example away from the benign example, and gradually guide it toward the target class in the feature representation space. Compared with C&W, our semantic similarity attack is more effective at misleading a white-box network into mapping an image to the target class subspace.

To study the performance of the proposed low-frequency constraint, normalized perturbations by SSA and SSAH in different iterations are visualized in Fig. 7. This figure shows that the perturbations generated by SSA are prone to distribute on object foreground as well as some smooth background regions, whereas the perturbations by SSAH gradually appear in edges or complex textures.

To quantify the contribution of each component in SSAH, we conduct an ablation study in Tab. 5. The results in the 1st and 2nd rows of Tab. 5 show SSA, with the adjustment of SPW, significantly reduces FID by a margin of 1.20 compared to the variant without SPW. The results in

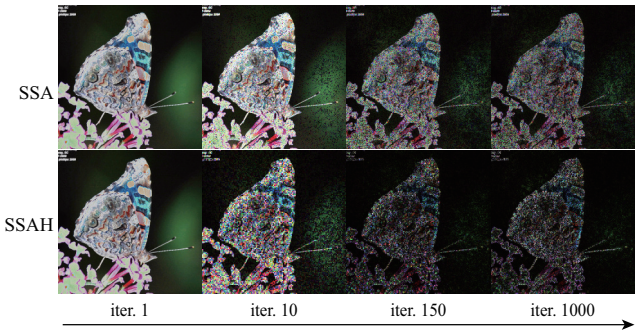

Figure 7. Normalized perturbations generated by SSA and SSAH in different iterations.

Attack	$\ell_2 \downarrow$	$\ell_\infty \downarrow$	FID $\downarrow$	LF $\downarrow$
SSA w/o SPW	3.12	**0.01**	5.83	1.40
SSA	2.34	**0.01**	4.63	1.05
SSAH	**1.81**	**0.01**	**3.90**	**0.06**

Table 5. Ablation study of the proposed attack on ImageNet-1K based on different modules, *i.e.*, self-paced weighting (SPW), low-frequency constraint $\mathcal{D}_{lf}$. w/o SPW means deleting the self-spaced weighting.

the 2nd and 3rd rows show that the proposed low-frequency constraint largely improves LF, *i.e.*, from 1.05 to 0.06.

5. Conclusion

We propose a novel framework, SSAH, for adversarial attack. It aims to perturb images by attacking their semantic similarity in representation space. Such an approach of attacking images in the feature space works well in a variety of settings. The proposed framework, in particular, could be used in more general and practical black-box settings, such as generating transferable adversarial examples across architectures and datasets, and misleading actual online models on various platforms while retaining high imperceptibility. It is more common in practice to attack in the open set scenario. For this reason, developing more effective algorithms in this scenario is worthwhile. Furthermore, the low-frequency constraint is introduced to limit adversarial perturbations within high-frequency components. Extensive experiments show that such constrained perturbations improve imperceptibility, particularly in smooth regions.

6. Acknowledgement

The work was supported by the National Natural Science Foundation of China under grants no. 61602315, 91959108, the Science and Technology Project of Guangdong Province under grant no. 2020A1515010707, the Science and Technology Innovation Commission of Shenzhen under grant no. JCYJ20190808165203670.

References

[1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, pages 39–57, 2017. 1, 2, 6, 8

[2] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: Elastic-net attacks to deep neural networks via adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2

[3] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, pages 2206–2216, 2020. 5, 6

[4] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193, 2018. 5, 6

[5] Ranjie Duan, Yuefeng Chen, Dantong Niu, Yun Yang, AK Qin, and Yuan He. Advdrop: Adversarial attack to dnns by dropping information. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7506–7515, 2021. 6, 8

[6] Yiping Duan, Fang Liu, Licheng Jiao, Peng Zhao, and Lu Zhang. Sar image segmentation based on convolutional-wavelet neural network and markov random field. *Pattern Recognition*, 64:255–267, 2017. 3

[7] Yan Feng, Bin Chen, Tao Dai, and Shu-Tao Xia. Adversarial attack on deep product quantization network for image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 10786–10793, 2020. 2

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 1, 2

[9] Muhammad Zaid Hameed and Andras Gyorgy. Perceptually constrained adversarial attacks. *arXiv preprint arXiv:2102.07140*, 2021. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017. 5

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 1

[13] Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7066–7074, 2019. 2

[14] Uyeong Jang, Xi Wu, and Somesh Jha. Objective metrics and gradient descent algorithms for adversarial examples in machine learning. In *Annual Computer Security Applications Conference (ACSAC)*, pages 262–277, 2017. 2

[15] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Tech report, Department of Computer Science, University of Toronto, Toronto, ON, Canada, 2009. 5

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 25:1097–1105, 2012. 1

[17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR) Workshops*, 2017. 5, 6

[18] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[19] Jie Li, Rongrong Ji, Hong Liu, Xiaopeng Hong, Yue Gao, and Qi Tian. Universal perturbation attack against image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4899–4908, 2019. 2

[20] Qiufu Li, Linlin Shen, Sheng Guo, and Zhihui Lai. Wavelet integrated cnns for noise-robust image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7245–7254, 2020. 3

[21] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) workshops*, pages 773–782, 2018. 3

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 5, 6

[23] Shishira R Maiya, Max Ehrlich, Vatsal Agarwal, Ser-Nam Lim, Tom Goldstein, and Abhinav Shrivastava. A frequency perspective of adversarial robustness. *arXiv preprint arXiv:2111.00861*, 2021. 2

[24] Konda Reddy Mopuri, Aditya Ganeshan, and R Venkatesh Babu. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2452–2465, 2018. 7

[25] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 754–763, 2017. 2

[26] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387, 2016. 2

[27] Jérôme Rony, Luiz G Hafemann, Luiz S Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direc-

tion and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 5

[29] Sara Sabour, Yanshuai Cao, Fartash Faghri, and David J Fleet. Adversarial manipulation of deep representations. In *International Conference on Learning Representations (ICLR)*, 2016. 2

[30] Mahmood Sharif, Lujo Bauer, and Michael K. Reiter. On the suitability of lp-norms for creating and preventing adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 2

[31] Yash Sharma, Gavin Weiguang Ding, and Marcus A Brubaker. On the effectiveness of low frequency perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3389–3396, 2019. 2

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 1

[33] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6398–6407, 2020. 4

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 5

[35] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 1

[36] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018. 1

[37] Hongjun Wang, Guangrun Wang, Ya Li, Dongyu Zhang, and Liang Lin. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep misranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 342–351, 2020. 2

[38] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8684–8694, 2020. 2

[39] Zhibo Wang, Siyan Zheng, Mengkai Song, Qian Wang, Alireza Rahimpour, and Hairong Qi. advpattern: Physical-world attacks on deep person re-identification via adversarially transformable patterns. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8341–8350, 2019. 2

[40] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning (ICML)*, pages 5286–5295, 2018. 1

[41] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems (NIPS)*, volume 32, 2019. 2

[42] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12, September 2016. 7

[43] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. *Advances in Neural Information Processing Systems (NIPS)*, 32:1831–1841, 2019. 6, 7

[44] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, pages 7472–7482, 2019. 1, 6, 7

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 2

[46] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1039–1048, 2020. 2, 5, 6, 8