



Evaluating Robustness of Deep Image Super Resolution Against Adversarial Attacks

评估深度图像超分辨率对对抗性攻击的鲁棒性

摘要

单一图像的超分辨率旨在生成低分辨率图像的高分辨率版本，它是许多计算机视觉应用中的一个重要组成部分。本文研究了基于深度学习的超级分辨率方法对对抗性攻击的鲁棒性，对抗性攻击可以显著将超分辨率图像变差，而在被攻击的低分辨率图像中没有明显的失真。事实证明，目前最先进的深度超级分辨率方法非常容易受到对抗性攻击。从理论上和实验上分析了不同方法的不同稳健性水平。我们还提出了关于攻击的可转移性，以及目标攻击和普遍攻击的可行性的分析。

1.简介

单一图像的超分辨率，也就是对低分辨率图像生成一个高分辨率版本，是近年来的热门研究领域之一。虽然简单的插值方法，如双线性和双三次上采样已经被普遍使用，但是由一个简单的卷积网络模型引发，名为超分辨率卷积神经网络（SRCNN）[7]造成的，基于深度学习的方法的发展，提供了更好的上采样图像的质量。超分辨率技术的改进将其应用扩展到更广泛的领域，包括视频流、监控、医疗诊断和卫星摄影[23]。

虽然已经有了许多基于深度学习的超级分辨率方法，但它们对预定攻击的鲁棒性还没有被彻底研究。近年来，深度网络的脆弱性一直是一个重要的问题，因为各种调查报告显示，攻击可以愚弄深度分类模型，并会造成严重的安全问题[9, 17]。类似的问题也可以在超分辨率应用中提出，因为恶化的输出会直接影响采用超分辨率作为其关键组件的系统的可靠性和稳定性。

在本文中，我们研究了基于深度学习的超分辨率对对抗性攻击的鲁棒性，就我们所知，这是第一项工作。我们的攻击在输入图像中产生扰动，这些扰动在视觉上并不明显，但在很大程度上会使输出的质量恶化。本文的主要贡献可以归纳为以下几点：

我们提出了三种针对超分辨率的对抗性攻击方法，这些方法对给定的低分辨率图像进行轻微扰动，但会导致输出图像明显恶化，包括基本攻击、通用攻击和部分攻击。这些方法是基于广泛用于图像分类任务的方法，我们为超分辨率任务优化了这些方法。

我们通过提供使用对抗性攻击方法的实验结果，对超分辨率方法的鲁棒性进行了彻底的分析。我们采用了各种最先进的基于深度学习的超级分辨率方法，这些方法在模型结构、训练目标和模型大小方面具有不同的特点。

我们进一步研究稳健性与模型属性的关系，并衡量其可转移性。此外，我们还提供了三个高级课题，包括受攻击的攻击、与攻击无关的鲁棒性测量，以及攻击的简单防御方法。

2.相关工作

超分辨率。最近，超分辨率研究的趋势已经转移到基于深度学习的方法。其中一个值得注意的方法是增强型深度超分辨率 (EDSR) 模型[12]，它实现了性能的大幅提升。随后，Zhang等人[25]提出了一个更先进的网络模型，名为残差通道注意网络 (RCAN)，它应用注意机制来有效利用图像特征。

上述方法侧重于在峰值信噪比 (PSNR) 方面实现高性能，但一些研究人员认为，只考虑这样的失真度量不一定能提高图像的感知质量[5]。为了解决这个问题，人们提出了感知优化的超级分辨率方法，它采用生成对抗网络 (GANs) [8]。最先进的方法之一是增强型超分辨率生成式对抗网络 (ESRGAN) [19]，它比其他传统方法产生更有视觉吸引力的输出图像，尽管PSNR值较低。Choi等人[6]开发了四通道感知增强型上采样超分辨率 (4PP-EUSR) 方法，该方法同时考虑定量和感知质量，以获得更自然的上采样图像。

由于超级分辨率在移动应用中也是一个有用的组成部分，一些研究的重点是在保持合理性能的同时节约计算资源。例如，Ahn等人[3]提出了级联残差网络 (CARN) 及其移动版本 (CARN-M)，它们采用共享模型参数的级联残差块。

对抗性攻击。最近的研究表明，深度图像分类器很容易受到各种对抗性攻击。Szegedy等人[18]提出了一种基于优化的攻击方法，其目的是在改变分类器的分类结果的情况下尽量减少扰动量。Goodfellow等人[9]开发了快速梯度信号法 (FGSM)，它使用从分类器获得的梯度的信号。Kurakin等人[11]将其扩展为一种迭代方法 (I-FGSM)，该方法显示出比FGSM更高的攻击成功率。这些攻击被称为强攻击方法，可以骗过几乎所有最先进的图像分类器，而且成功率高[17]。

一些研究对深度学习模型的鲁棒性进行了深入分析。Liu等人[13]测量了对抗性图像的可转移性，也就是找出为一个分类器发现的扰动是否对另一个分类器也有效。Moosavi-Dezfooli等人[15]研究了一种通用的扰动，可以应用于给定数据集中的所有图像。Weng等人[20]提出了一个理论上的鲁棒性措施，它不依赖于特定的攻击方法。

对超分辨率的对抗性攻击。最近，出现了将超分辨率任务与对抗性攻击相结合的情况。Mustafa等人[16]提出了一种采用超分辨率的方法来防御深度图像分类器的对抗性攻击。Yin等人[22]采用对超分辨率的对抗性攻击来欺骗后续的计算机视觉任务。然而，这些研究调查了对抗性攻击对其他任务的有效性，而不是超分辨率任务本身，包括图像分类、风格转移和图像字幕，其中超分辨率被用作主要任务之前的预处理步骤。因此，本文所研究的超分辨率本身对对抗性攻击的鲁棒性，以前还没有被解决。

3. 对超分辨率的攻击

3.1. 基本攻击

对超分辨率模型的对抗性攻击的目标是在给定的输入图像中注入少量的扰动，使扰动在视觉上无法察觉，但导致超分辨率输出的显著恶化。为此，我们开发了一种基于I-FGSM[11]思想的算法，它是分类模型最广泛使用的强攻击之一。

让 X_0 表示原始低分辨率输入图像， X 表示 X_0 的攻击版本。从这些图像中，我们通过一个给定的超分辨率模型 $f(\cdot)$ ，分别得到超分辨率的高分辨率图像 $f(X_0)$ 和 $f(X)$ 。我们的目标是使超分辨率输出的劣化量最大化，可以定义为：

$$L(X, X_0) = \|f(X) - f(X_0)\|_2. \quad (1)$$

为了使用有界的 $\ell_\infty - norm$ 约束($\|X - X_0\|_\infty \leq \alpha$)找到一个 X 来最小化(1)，我们采用I-FGSM更新规则，通过以下方式反复更新 X 。

$$\tilde{X}_{n+1} = clip_{0,1}(X_n + \frac{\alpha}{T} sgn(\nabla L(X_n, X_0))) \quad (2)$$

$$X_{n+1} = clip_{-\alpha, \alpha}(\tilde{X}_{n+1} - X_0 + X_0) \quad (3)$$

其中 T 是迭代次数， $sgn(\nabla L(X_n, X_0))$ 是(1)的梯度信号，以及

$$clip_{a,b}(X) = min(max(X, a), b) \quad (4)$$

期限 α 不仅控制了每次迭代时计算的梯度所提供的贡献量，而且还限制了最大的扰动量，以防止被攻击的输入图像发生明显的变化。最后的对抗性例子是由 $X = X_T$ 得到的。

3.2.通用攻击

虽然可以像第3.1节那样为每个图像找到一个对抗性图像，但也可以找到一个与图像无关的对抗性扰动，它可以影响到某个超级分辨率方法的任何输入图像[15]。在我们的研究中，我们通过改变基本攻击的表述来应用这一概念，具体如下。

假设数据集中有 K 张图像，其中第 k 张图像表示为 X_0^k 。有了一个普遍的扰动 Δ ，我们可以得到对抗性的图像，即：

$$X^k = \text{clip}_{0,1}(X_0^k + \Delta) \quad (5)$$

然后，我们计算恶化的平均数量为：

$$F(\Delta) = \frac{1}{K} \sum_{k=1}^K L(X^k, X_0^k) \quad (6)$$

从 $\Delta_0 = 0$ 开始，通用扰动被迭代更新为：

$$\Delta_{n+1} = \text{clip}_{-\alpha, \alpha}(\Delta_n + \frac{\alpha}{T} \text{sgn}(\nabla F(\Delta_n))) \quad (7)$$

最后的通用扰动是由 $\Delta = \Delta_T$ 得到的。

3.3.部分攻击

第3.1节中的基本攻击找到一个覆盖给定图像整个区域的扰动。我们进一步研究超分辨率方法的稳健性，只攻击图像的某些部分，但测量没有被攻击的区域的恶化程度。通过这个实验，我们可以研究在超分辨率过程中，扰动渗透到邻近区域的程度。

让 M 表示扰动 Δ 的二进制掩码，其中只有要被攻击的区域被设置为1。掩码扰动为 $\Delta \circ M$ ，其中 \circ 表示元素相乘。那么(2)可以修改为：

$$\tilde{X}_{n+1} = \text{clip}_{0,1}(X_n + \frac{\alpha}{T} \text{sgn}(\nabla L_M(X_n, X_0)) \circ M) \quad (8)$$

其中

$$L_M(X, X_0) = \|(f(X) - f(X_0)) \circ (1 - M_H)\|_2 \quad (9)$$

在(9)中, M_H 是 M 的一个高分辨率对应。期限($1 - M_H$)确保恶化的数量只在未受扰动的区域计算。最后的对抗性例子是由 $X = X_T$ 得到的。

4. 实验结果

数据集。我们采用了三个广泛用于基准测试超分辨率方法的图像数据集。Set5 [4], Set14 [24], 和BSD100 [14]。每个数据集分别由5、14和100张图像组成。

超分辨率方法。我们考虑了八种基于深度学习的超分辨率方法, 它们具有不同的模型大小和属性, 包括EDSR[12]、EDSR-baseline[12]、RCAN[25]、4PP-EUSR[6]、ESRGAN[19]、RRDB[19]、CARN[3]和CARN-M[3]。表1显示了它们在模型参数数量、卷积层数量以及是否采用GANs进行训练方面的特点。

Method	# parameters	# layers	GAN-based
EDSR [12]	43.1M	69	-
EDSR-baseline [12]	1.5M	37	-
RCAN [25]	15.6M	815	-
4PP-EUSR [6]	6.3M	95	✓
ESRGAN [19]	16.7M	351	✓
RRDB [19]	16.7M	351	-
CARN [3]	1.1M	34	-
CARN-M [3]	0.3M	43	-

Table 1. Properties of the super-resolution methods.

EDSR-baseline是EDSR的一个较小的版本, RRDB是ESRGAN的一个替代版本, 在没有GAN的情况下进行训练, 而CARN-M是CARN在模型参数数量方面的一个轻量级版本。此外, 我们还考虑了双三次插值, 以比较其相对于基于深度学习的方法对对抗性攻击的鲁棒性。我们考虑所有超分辨率方法的缩放系数为4。此外, 我们还采用了原作者提供的预训练模型。

实施细节。我们的对抗性攻击方法是在TensorFlow框架上实现的[2]。对于所有的攻击方法, 我们设定 $\alpha \in 1/255, 2/255, 4/255, 8/255, 16/255, 32/255$ 和 $T = 50$ 。对于通用攻击, 需要一个具有固定空间分辨率的扰动 Δ , 以便将其应用于数据集中的所有图

像。因此，我们以固定的分辨率裁剪每个输入图像的中心区域。对于部分攻击，我们设置掩码 M ，以便攻击输入图像的中心部分，即

$$M_{(x,y)} = \begin{cases} 1 & \text{if } \frac{w}{4} \leq x < \frac{3w}{4}, \frac{h}{4} \leq y < \frac{3h}{4} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

其中 $M_{(x,y)}$ 是 M 在 (x,y) 处的值， w 和 h 分别是输入图像的宽度和高度。

性能衡量。我们用PSNR来衡量超分辨率方法对我们的对抗性攻击方法的稳健性。对于低分辨率(LR)图像，我们计算原始图像和被攻击图像之间的PSNR值，即 X_0 和 X 。对于超分辨率(SR)图像，PSNR是测量从原始和攻击的输入图像得到的输出图像，即 $f(X_0)$ 和 $f(X)$ 之间的关系。我们报告每个数据集的平均PSNR值。对于部分攻击，我们只计算输出图像的外部区域的PSNR值，该区域对应于攻击期间的屏蔽区域。

4.1. 基本攻击

图1比较了第3.1节中解释的I-GSM攻击的超分辨率方法在PSNR方面的表现。

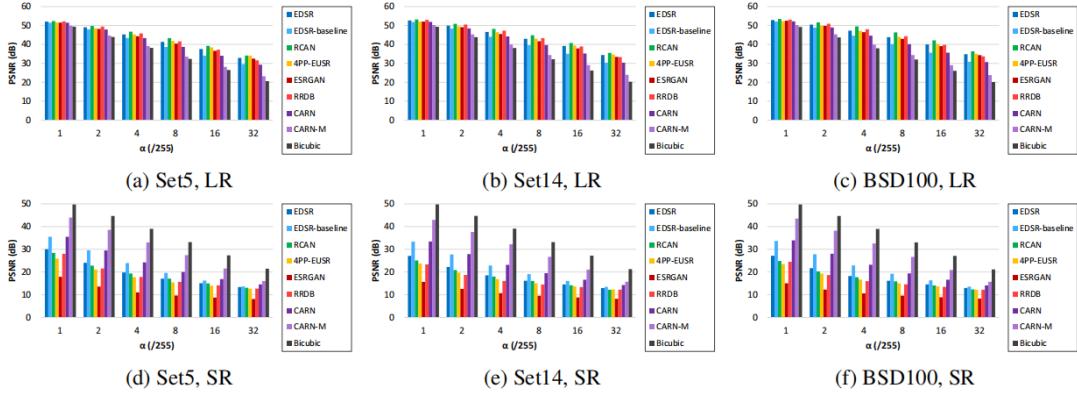


Figure 1. Comparison of the PSNR values of low-resolution (LR) and super-resolved (SR) images with respect to different α values for the basic attack on the Set5 [4], Set14 [24], and BSD100 [14] datasets.

随着 α 的增加，LR和SR图像的质量下降都变得很严重。然而，除了双三次插值外，它在SR图像中比LR图像要显著得多（即PSNR值较低）。例如，在Set5数据集上，当 $\alpha = 8/255$ 时，EDSR模型的LR和SR图像的PSNR值分别为41.37dB和17.05dB。请注意，两个PSNR值高于30dB的图像可以被视为视觉上相同的图像[10]。

图2显示了 $\alpha = 8/255$ 的LR和SR图像示例。



Figure 2. Visual comparison of the super-resolved outputs for the inputs attacked with $\alpha = 8/255$. In each case, (top-left) is the original input in Set5 [4], (top-right) is the adversarial input, and (bottom) is the output obtained from the adversarial input. The input images are enlarged two times for better visualization.

总的来说，对于所有的超分辨率方法来说，原始图像和扰动的输入图像之间没有明显区别。然而，在所有方法的SR图像中都可以观察到明显的质量下降。ESRGAN显示出最差的视觉质量，SR图像的所有部分都有退化，这也可以从图1d、1e和1f中的最低PSNR值观察到。对于其他超分辨率模型，观察到的是类似指纹的图案。这证明了所有基于深度学习的超级分辨率方法都非常容易受到对抗性攻击。相比之下，双三次方法虽然在干净的数据上具有较低的超分辨率质量，但与基于深度学习的方法相比，它要鲁棒得多。

与模型目标的关系。 ESRGAN和4PP-EUSR采用GANs来考虑感知质量的改善，比其他方法产生更明显的退化输出。由于ESRGAN与RRDB具有完全相同的结构，但以不同的目标（即考虑感知质量）进行训练，ESRGAN比RRDB更明显的脆弱性意味着训练目标的不同影响了对对抗性攻击的稳健性。众所周知，采用GANs的方法往往比其他方法产生更清晰的纹理，以确保升级后的图像具有自然吸引人的质量[5]。因此，这些方法会显著放大小的扰动，并产生不理想的纹理，这使得它们比没有GANs的方法更容易受到对抗性攻击。

与模型尺寸的关系。 据观察，超分辨率模型的脆弱性与它们的模型尺寸有关。例如，EDSR-baseline是EDSR的一个较小版本，它对SR图像的PSNR值比EDSR高，如图1d、1e和1f所示。这在图3中得到了证实，我们在图中比较了与模型大小有关的稳健性。

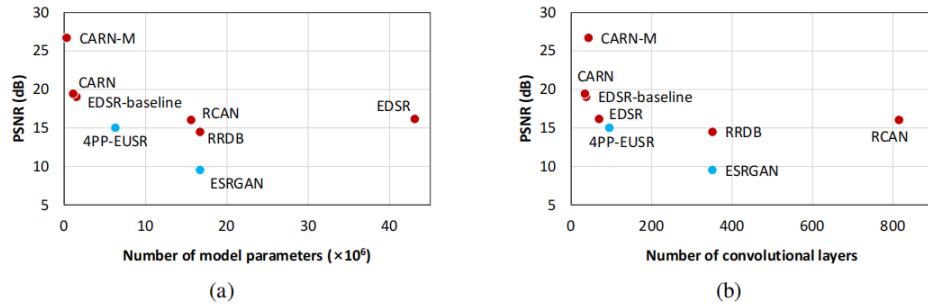


Figure 3. Comparison of the PSNR values of SR images for BSD100 [14] with respect to the model sizes in terms of (a) the number of model parameters and (b) the number of convolutional layers ($\alpha = 8/255$). Blue and red colors indicate the models trained with and without GANs, respectively.

该图说明，当采用更多的模型参数或更多的卷积层时，SR图像的PSNR值往往下降。关于这一现象的进一步分析将在第4.3节给出。

可转移性。在分类任务中，“可转移性”是指一个被错误分类的对抗性例子也被另一个分类器错误分类的可能性[13]。我们还研究了超级分辨率中对抗性攻击的可转移性。换句话说，一个为“源”超分辨率模型找到的对抗性例子被输入到另一个“目标”模型，并测量输出图像的PSNR值。

图4总结了BSD100数据集上基于深度学习的超级分辨率模型的可转移性，其中 $\alpha=8/255$ 。

PSNR (dB)		Target model							
		EDSR	EDSR -baseline	RCAN	4PP -EUSR	ESRGAN	RRDB	CARN	CARN-M
Source model	EDSR	16.14	32.88	25.82	23.86	16.26	23.57	32.80	37.74
	EDSR-baseline	24.44	19.19	23.65	21.23	15.15	22.29	26.82	33.62
	RCAN	30.57	35.49	15.89	26.60	18.94	29.74	35.57	40.46
	4PP-EUSR	27.25	32.76	26.83	15.02	16.16	24.71	32.87	37.97
	ESRGAN	28.64	33.11	28.59	24.28	9.57	24.46	33.30	36.56
	RRDB	25.55	33.09	25.31	23.77	15.86	14.59	32.91	38.11
	CARN	24.12	26.05	23.83	21.45	15.24	22.15	19.40	33.51
	CARN-M	27.34	28.20	27.20	23.49	16.27	26.77	28.20	26.66

Figure 4. Comparison of the transferability in terms of PSNR for the BSD100 dataset [14] when $\alpha = 8/255$. Red and blue colors indicate the lowest and highest PSNR values (except the diagonal cells) for each target model, respectively.

图中显示，对抗性例子在不同的模型之间有一定程度的可转移性，而可转移性的程度则因源模型和目标模型的组合而不同。在CARN和EDSR-baseline中发现的对抗性例子具有高度的可转移性，而在RCAN中发现的对抗性例子的可转移性最差。这一结果意味着RCAN在从输入图像中恢复纹理方面有其自身的特点，这使得与这种特点相关的扰动在其他超分辨率方法中不那么有效。

4.2. 通用攻击

图5比较了在应用普遍攻击时，BSD100数据集的超分辨率方法在不同 α 值方面的表现。

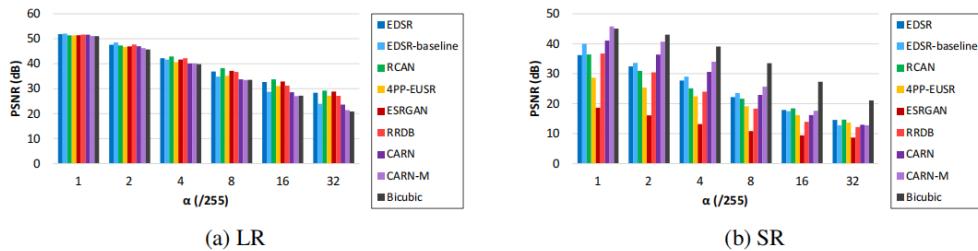


Figure 5. Comparison of the PSNR values of LR and SR images with respect to different α values for the universal attack on the BSD100 dataset [14].

该图证实了超分辨率模型也容易受到图像无关的通用攻击，尽管通用攻击需要对输入图像进行更大的扰动（即图5a中的PSNR值比图1c中的略低），并且比特定图像的攻击略低（即图5b中的PSNR值比图1f中的略高）。与基本攻击的结果相比（图1），观察到同样的趋势：ESRGAN和4PP-EUSR都是最脆弱的，双三次插值是最稳健的。

图6显示了RCAN的普遍攻击的视觉例子，其中 α 是4/255。

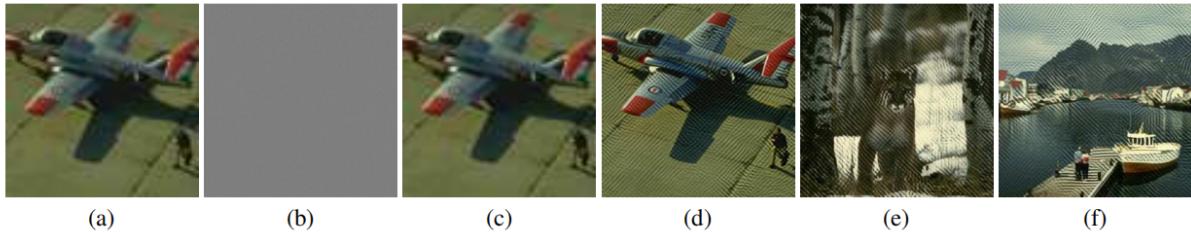


Figure 6. Visual examples of the universal attack with $\alpha = 4/255$ on the BSD100 dataset [14] for the RCAN model. (a) LR (original) (b) Perturbation (c) LR (attacked) (d) SR (e-f) Other examples obtained from the images attacked with the same perturbation

从BSD100数据集的所有图像中，我们的攻击方法找到了一个普遍的扰动（图6b），它将图6a所示的输入图像变为图6c中的图像。虽然被攻击的LR图像与原始图像几乎没有明显的差异，但它的放大版本却含有明显的伪影，如图6d所示。如图6e和6f所示，在其他用相同扰动攻击的SR图像中也可以观察到类似的伪影。这表明，使用深度学习的最先进的超级分辨率方法也容易受到普遍扰动的影响。

4.3. 部分攻击

图7显示了部分攻击的SR图像在不同 α 值下的PSNR值。

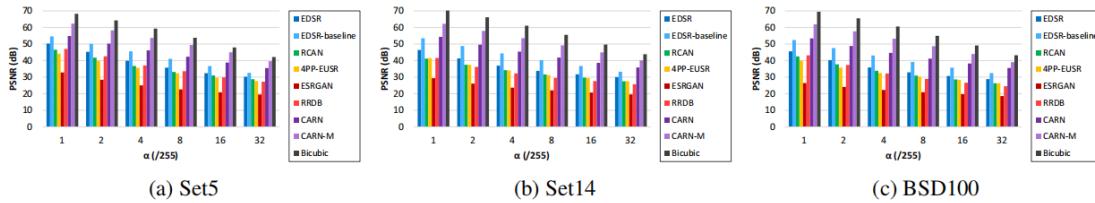


Figure 7. Comparison of the PSNR values of SR images with respect to different α values for the partial attack.

超分辨率方法在PSNR方面的排名与基本攻击相同，只是部分攻击的PSNR值远高于基本攻击，因为测量PSNR的区域在LR图像中没有被直接扰动。这表明，在上采样过程中，扰动相邻像素的传播说明了不同的超分辨率模型具有不同程度的脆弱性。例如，除了Set5中的 $\alpha=1$ ，ESRGAN的所有PSNR值都由于部分攻击而低于30dB。

图8显示了从Set14数据集中的图像中获得的SR图像的例子，这些图像被部分攻击， $\alpha=8/255$ 。

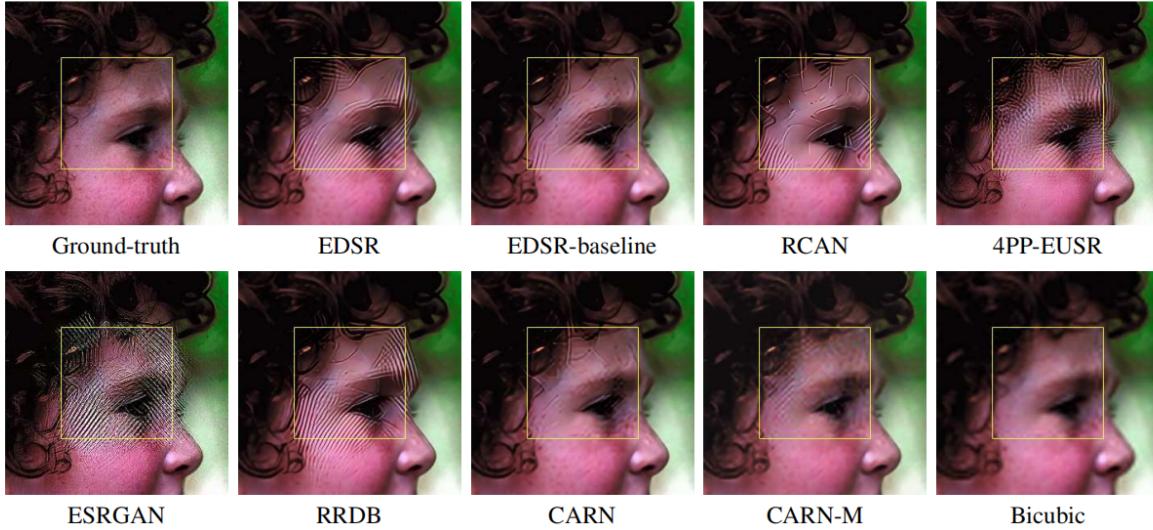


Figure 8. Visual comparison of the SR images for the partial adversarial attack with $\alpha = 8/255$ on an image of Set14 [24]. The regions marked with yellow boxes correspond to the regions where the attack is applied in the LR images.

攻击造成的退化会传播到被攻击区域之外，这对ESRGAN和RRDB来说特别明显。这是因为卷积层的核不仅对目标位置的像素进行操作，也对其相邻的像素进行操作。此外，这种操作引起的扰动的传播通过多个卷积层进一步扩展，这也是图3b所示结果的原因。

5. 高级专题

5.1. 有针对性的攻击

在分类任务的情况下，有可能攻击图像，使分类器错误地将图像分类为特定的目标类别。我们提出了一个展示，证明这个概念也可以应用于超分辨率方法。换句话说，有针对性的攻击不是降低输出图像的质量，而是使超分辨率方法生成的图像与目标图像的相似度高于原始的参考图像。对此，我们将(2)描述为：

$$\tilde{X}_{n+1} = clip_{0,1}(X_n - \frac{\alpha}{T} sgn(\nabla L(X_n, X^*))) \quad (11)$$

其中 X^* 是目标图像。

为了演示，我们使用了名为 "foreman"[1]的视频的两个相邻帧。图9显示了4PP-EUSR的结果，其中 $\alpha=16/255$ ， $T=50$ 。

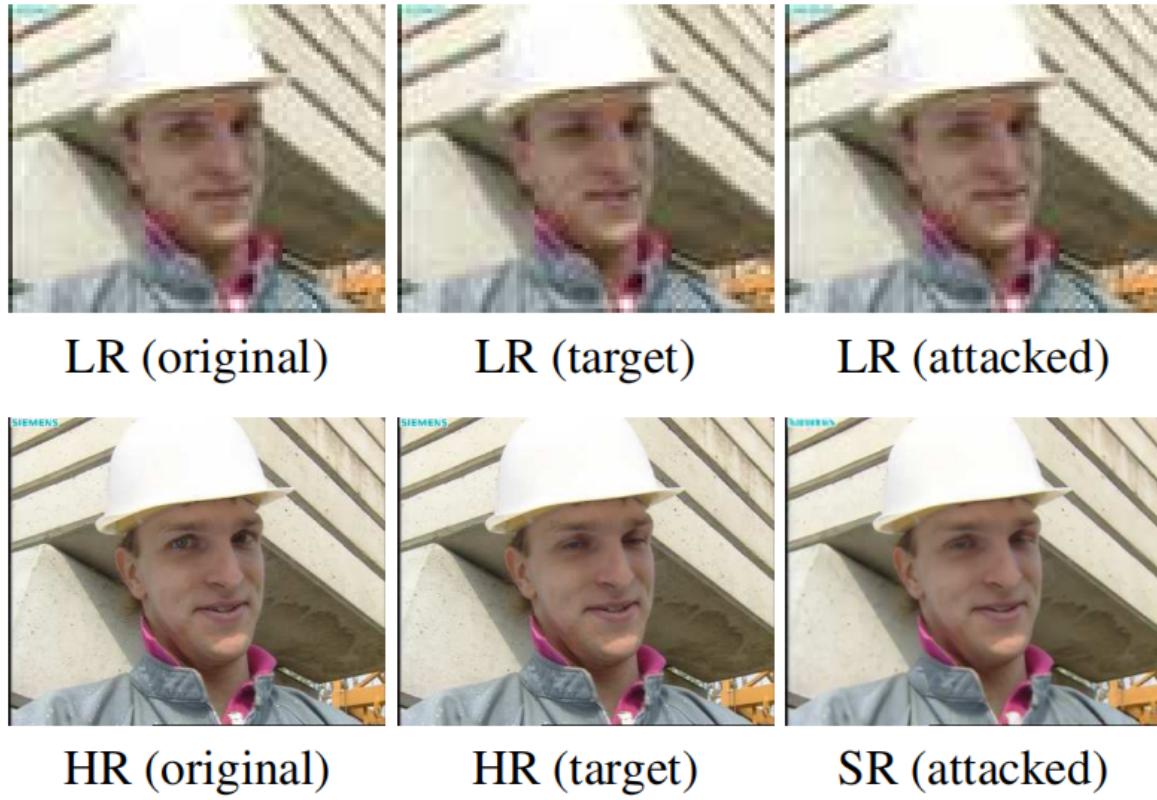


Figure 9. Result of the targeted attack with $\alpha = 16/255$ using two frames of a video “foreman” [1] for 4PP-EUSR [6].

图中显示，目标攻击是成功的。扰动的产生使超分辨率方法产生了放大的输出（“SR（被攻击）”），它看起来与半闭着眼睛的目标高分辨率（HR）图像（“HR（目标）”）比睁着眼睛的原始真实图像（“HR（原始）”）更相似，而被攻击的输入图像（“LR（被攻击）”）看起来仍比目标图像的低分辨率版本（“LR（目标）”）更接近原始图像。此外，我们对20名人类观察者进行了主观测试，其中10人将被攻击的输出（“SR（被攻击）”）识别为闭眼。此外，我们对20名人类观察者进行了主观测试，其中10人将被攻击的输出（“SR（被攻击）”）识别为闭眼。这些结果具有严重的安全意义：对超分辨率的攻击不仅会损害超分辨率的基本目标（即提高图像质量），而且还会危及对超分辨率图像的进一步人工或自动检查（例如，识别监控摄像头中的人或物体，识别图像中的文字等）。

5.2. 稳健性测量

最近，Weng等人[20]提出了一个与攻击无关的分类模型鲁棒性措施，称为网络鲁棒性的交叉Lipschitz极值（CLEVER），它不依赖于特定的攻击方法。它使用基于极值理论的

交叉Lipschitz常数来估计鲁棒性的下限。我们将这一方法的核心思想应用于超分辨率任务，以便从理论上验证第4节中的实验结果。

用 X_0 表示原始输入图像。我们首先得到 N_s 个随机扰动，每个像素都在 $[-\alpha, \alpha]$ 之内。用 $\Delta(i)$ 表示第*i*个随机扰动。然后，我们对所有的扰动计算 $b_i = \|\nabla L(X_0 + \Delta^{(i)}, X_0)\|_1$ ，其中 L 的定义见(1)。最后，我们将最大的 b_i 看作是稳健性指数；稳健性指数大，说明脆弱性高。我们将 N_s 和 α 分别设定为1024和1/255。

图10显示了八种基于深度学习的超分辨率方法对BSD100数据集的SR图像的PSNR值和鲁棒性指数，其中PSNR值是由相同 α 值的基本攻击得到的（4.1节）。

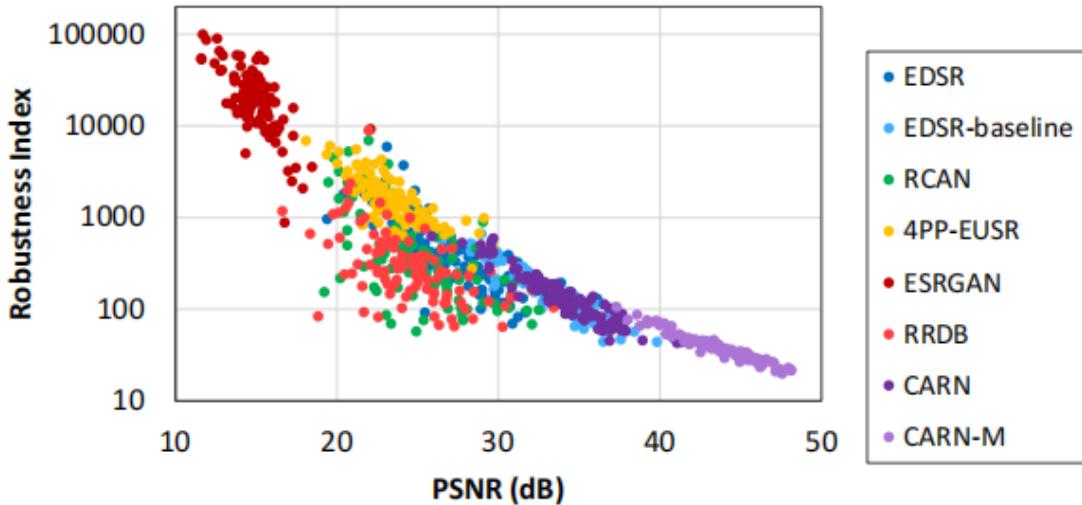


Figure 10. PSNR vs. the robustness index for the BSD100 dataset [14] when $\alpha = 1/255$. Each point corresponds to each image in the dataset.

结果表明，鲁棒性指数与PSNR密切相关。例如，ESRGAN具有最大的鲁棒性指数，它显示了最低的PSNR值。在PSNR和鲁棒性指数方面，EDSR-baseline模型具有与CARN模型相似的鲁棒性。此外，在每种方法中，鲁棒性指数成功地解释了不同图像的相对脆弱性。CLEVER方法适用于解释超分辨率方法的鲁棒性，这意味着对抗性攻击的基本机制在分类和超分辨率任务之间具有相似性。

5.3. 防御

我们展示了两种简单的防御攻击的方法。首先，我们采用调整大小的方法[21]，将被攻击的输入图像的大小减少一个像素，然后将其调整回原始分辨率，再输入到SR模型。这样一来， $\alpha = 8/255$ 的EDSR的PSNR从16.14dB增加到25.01dB。其次，我们采用了EDSR模型中使用的几何自组装方法[12]。这样一来， $\alpha = 8/255$ 的EDSR的PSNR从16.14dB增加到23.47dB。更先进的防御方法可以在未来的工作中进行研究。

6.总结

我们研究了基于深度学习的超分辨率方法对对抗性攻击的鲁棒性，其中分类任务的攻击方法针对我们的目标进行了优化。我们的结果表明，最先进的基于深度学习的超分辨率方法非常容易受到对抗性攻击，这主要是由于通过卷积操作的扰动传播造成的。使用攻击诊断鲁棒性措施来衡量不同方法的不同鲁棒性水平是非常有可能的。我们还展示了生成通用攻击和对于超分辨率方法的转移攻击的可行性。此外，事实表明，有针对性的攻击可以在超分辨率期间改变图像的内容。