Arezoo Rajabi*, Mahdieh Abbasi, Rakesh B. Bobba*, and Kimia Tajik

Adversarial Images Against Super-Resolution Convolutional Neural Networks for Free

Abstract: Super-Resolution Convolutional Neural Networks (SRCNNs) with their ability to generate high-resolution images from low-resolution counterparts, exacerbate the privacy concerns emerging from automated Convolutional Neural Networks (CNNs)-based image classifiers. In this work, we hypothesize and empirically show that adversarial examples learned over CNN image classifiers can *survive* processing by SRCNNs and lead them to generate poor quality images that are hard to classify correctly. We demonstrate that a user with a small CNN is able to learn adversarial noise without requiring any customization for SRCNNs and thwart the privacy threat posed by a pipeline of SRCNN and CNN classifiers (95.8% fooling rate for Fast Gradient Sign with $\epsilon = 0.03$). We evaluate the *survivability* of adversarial images generated in both black-box and white-box settings and show that black-box adversarial learning (when both CNN classifier and SRCNN are unknown) is at least as effective as white-box adversarial learning (when only CNN classifier is known). We also assess our hypothesis on adversarial robust CNNs and observe that the supper-resolved white-box adversarial examples can fool these CNNs more than 71.5% of the time.

Keywords: Survivability of Adversarial Example, Super-Resolution Convolutional Neural networks, Image Privacy

DOI 10.56553/popets-2022-0065
Received 2021-11-30; revised 2022-03-15; accepted 2022-03-16.

(a) LR **(b)** Super-resolved HR **(c)** Original HR

Fig. 1. Privacy threat posed by SRCNNs. Left: LR image of 16×16 pixels. Center: the super-resolved HR image (128×128 pixels), generated using Deep Face SRCNN ($\times 8$) [26]. Right: the original HR image. Both HR images are classified correctly by Clarifai.com's celebrity face recognition model, while the LR image is not.

1 Introduction

Single-image Super-Resolution Convolutional Neural Networks (SRCNNs) [27, 49, 64] are designed to generate/recover a High Resolution (HR) image from its Low Resolution (LR) counterpart. While SRCNNs have legitimate and beneficial applications in different domains [46, 54, 60], they raise serious privacy concerns. Previously, low resolution images were considered to provide some level of image privacy against recognition by humans and machine classifiers (*e.g.,* thumbnail-preserving image encryption methods [18, 29, 53, 58]). However, SRCNNs' ability to generate High Resolution (HR) images from Low Resolution (LR) ones to enable face recognition [5, 44, 66] poses a serious privacy threat (see Figure 1).

The privacy threat posed by face recognition CNNs, and AI-based face-recognition in general, is real as evidenced by the emergence of image search applications such as Clearview AI Facial Recognition App [16] with the ability to match photos with a database of more than 3 billion images scraped from Facebook, YouTube and millions of other websites. Due to the current challenges in learning large robust CNNs (with ability of thwarting all types of adversarial noise) [45], adversarial perturbation-based image privacy schemes have emerged as a potential defense against this threat of unauthorized automated face recognition [4, 11, 19, 40].

However, these schemes do not explicitly consider low-resolution images or take SRCNNs into account, and learn adversarial noise/perturbation(s) over CNN classifiers only. Regardless of vulnerability of SRCNNs to adversarial noise, adversarial noise learnt on SRC-

***Corresponding Author: Arezoo Rajabi:** University of Washington, E-mail: rajabia@uw.edu
Mahdieh Abbasi: Universite Laval, E-mail: mahdieh.abbasi.1@ulaval.ca
***Corresponding Author: Rakesh B. Bobba:** Oregon State University, E-mail: rakesh.bobba@oregonstate.edu
Kimia Tajik: Case Western Reserve University, E-mail: kxt328@case.edu

NNs alone leads only to quality degradation in super-resolved HR images [6] and the effectiveness of such noise against image classification by CNNs is unknown. While it is possible to learn adversarial noise through joint optimization over both SRCNNs and classification CNNs [61], this approach requires knowledge of both target SRCNN and CNN, which is not realistic.

Contributions: In this paper, we hypothesize and empirically show that adversarial perturbations learnt against CNN classifiers (alone) can *survive* the processing by unknown SRCNNs. Specifically, we introduce two approaches for generating adversarial LR images using local CNNs and evaluate the *survivability* of both white-box and black-box adversarial perturbations through the state-of-the-art SRCNNs. The first approach of down-scaling learned HR adversarial images is motivated by the observation that SRCNNs are optimized to recreate the original HR version of a given LR image as closely as possible. It investigates whether down-scaled adversarial examples carry sufficient amount of adversarial noise to make SRCNNs recreate adversarial images. The second approach of directly learning LR adversarial using small local CNNs is exploratory and investigates whether LR perturbations are translated into effective HR perturbations by SRCNNs.

Our findings demonstrate that adversarial perturbations against SRCNNs can come free with adversarial perturbations against classification CNNs (see Figure 2). Our evaluation of the proposed adversarial approaches against potential countermeasures such as robust CNNs [10, 28] and input filtering show that the adversarials were able to resist them reasonably. Briefly, we make the following contributions:

- This is the first work to study the impact of super-resolution on adversarial perturbations.
- We define and hypothesize the *survivability* of adversarial images learned only on CNN classifiers through unknown SRCNNs (see Section 5).
- We show that both (i) down-scaled HR adversarial images (learned on HR images) and (ii) LR perturbations directly learned using a small local CNN (trained on low resolution images), can survive through state-of-the-art SRCNNs and contribute to defense against the privacy threat posed by SRCNN-CNN pipelines (see Sections 6.2 and 6.3).
- We empirically evaluate how well HR adversarial examples that can fool robust CNN classifiers survive through SRCNNs. We found that such adversarial images cause robust CNNs to misclassify even after being down-scaled and super-resolved by SRCNNs.

The rest of this paper is organized as follows: We present some preliminaries in Section 2, and describe our assumptions and the system model in Section 3. We discuss potential solutions in Section 4. We define survivability of adversarial images and introduce our methodology for generating and evaluating LR adversarial images in Section 5. We present our evaluation and results in Section 6. We discuss the implications of our work and future directions in Section 7. We cover related work in Section 8, and conclude in Section 9.

2 Preliminaries

In this section, we first introduce convolutional neural network (CNN) classifiers and super-resolution convolutional neural networks (SRCNNs). We then briefly introduce the notion of adversarial examples and perturbations on these networks.

Convolutional neural network (CNN) classifiers: Convolutional Neural Networks (CNNs) as modern deep learning models are able to achieve nearly human-level performance on several computer vision tasks such as face detection [52], optical character recognition [14], object recognition [50], and object detection [41, 42]. A CNN consists of a series of convolution, pooling, and non-linear activation layers that are followed by a few fully connected layers. For a K-class classification task, usually a CNN is terminated by a softmax activation layer to map the logit layer (with K outputs) into conditional class probabilities. Concretely, a CNN can be denoted by a function $F(.;\theta) : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X} = [0,1]^D$ and $\mathcal{Y} = [0,1]^K$ are input and output spaces respectively, with θ compactly denoting parameters of the CNN. Upon feeding a CNN (*i.e.,* $F(.,\theta)$) with an input sample $x \in [0,1]^d$, it returns a vector of conditional class probabilities over K classes s.t. $\sum_{k=1}^{K} F_k(x;\theta) = 1$. To learn the parameter θ, cross-entropy loss function $J(.,.;\theta) : (\mathcal{X} \times \mathcal{Y}) \rightarrow R^+$ is minimized over N i.i.d. training samples, i.e. $(x^i, y^{*i})_{i=1}^{N}$, where $y^{*i} \in \{0,1\}^K$, as the true class associated with i-th input, is a binary vector with a single one bit at its k-th element. Formally, the cross-entropy is defined for a given sample (x, y^*)[1] as follows:

$$J(x, y^*; \theta) = -\sum_{k=1}^{K} y_k^* \log F_k(x;\theta) \qquad (1)$$

1 For simplicity in notation, the superscript i is dropped from (x^i, y^{*i})

Super-resolution convolutional neural networks (SRCNNs): Super-Resolution Convolutional Neural Networks (SRCNNs) are deep convolutional neural networks that generate HR images from their LR counterparts [9] with the aim of increasing the visual quality of HR images. In general, SRCNNs optimize Peak Signal to Noise Ratio (PSNR), Minimum Square Error (MSE), or Structural Similarity Index (SSIM) between original HR images and the super-resolved HR images as an objective function. Here we use $hr(.)$ to denote an SRCNN function which takes an LR image I (I_{lr}) and returns a super-resolved HR version.

Adversarial examples and perturbations: Adversarial generation methods find and add a small adversarial noise (δ) to an image that causes the target CNN to misclassify it.

$$\begin{aligned} &\min_{\delta} \|\delta\|_2 - J(F(x,\theta)) \\ &s.t. \;\; argmax \;\; F(x+\delta) \neq y^*, \;\; x+\delta \in [0,1]^d \end{aligned} \tag{2}$$

where δ, x, y^*, and $J(.)$ are the adversarial noise, given input image, the true label of the input image, and the loss function, respectively. Since this noise is imperceptible to the human eye, a benign image with adversarial noise is called an *adversarial example*. Many methods have been proposed to generate a small adversarial noise for a known CNN (*e.g.,* [3, 13, 31, 51]), but they suffer from low *transferability* to unknown CNNs. Transferable perturbation generation methods [24, 30] proposed to address this issue by producing adversarial noise that is typically perceptible, called *adversarial perturbation*. Images perturbed with adversarial perturbation are still recognizable to human eyes but can fool CNN classifiers (including unknown ones) with high probability. The detection adversarial example defenses can detect these adversarial images [1, 2, 39, 59], but these defenses cannot help the classifiers to classify them correctly.

3 System and Adversary Model

CNN classifiers do not perform well on LR images. For example, recent CNNs achieved more than 98% top-5 accuracy on ImageNet dataset containing images with resolution of 256×256 pixels [36], while top-5 accuracy for the images from the same dataset down-scaled to a resolution of 16×16 pixels is less than 65% [7]. In this paper, we consider a situation where an adversary has access only to LR images and aims to classify them using their convolutional neural network classifier. In order to improve their classification performance, the adversary first passes the low resolution images through a super resolution convolutional neural network (SRCNN) to generate a higher resolution image, and then uses a CNN classifier on the super-resolved image [66]. Figure 2 depicts such an adversary's classification pipeline. We assume that the adversary's image classification CNN is pre-trained (*e.g.,* using images scraped from the Internet; or using images previously stored or shared by users on online storage or sharing platforms).

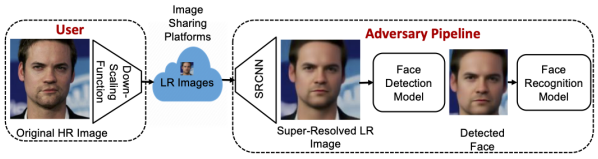

Fig. 2. Users only share their LR images. Adversary uses an SRCNN to super-resolve such LR images before applying face detection and recognition models.

Our goal is to help owners of the (low resolution) images (*e.g.,* users' of social networks or image sharing platforms) thwart these unauthorized automated classification pipelines using image perturbation. Here perturbations that cause SRCNNs to generate high resolution images that will be misclassified by the CNN classifier need to be learned and applied to the low resolution image. However, it is unlikely that the image owners have any prior knowledge of the SRCNNs and classification CNNs that may be used against their images in an adversary's pipeline. Thus, we assume that users do not have any knowledge about the CNN(s) that an adversary might use. For the image classification CNN, we consider the following two scenarios:

Black-box attacks: In this scenario,end users do not have any knowledge about the adversary's image classification CNN. This is a more likely case in real-world use. For example, in image sharing platforms, end users do not know what classifiers may be used by service providers or other curious users. This scenario requires perturbations for the images to be learned in a black-box setting. Since the adversary uses publicly shared images, the images used for target classes could be similar for both user and the adversary. These assumptions are consistent with the black-box adversarial setting used in machine learning literature [34].

White-box attacks: We also consider the scenario where there is access to the target CNN or the target CNN's function can be estimated by sending queries [35]. In this case, perturbations for images can be learned using white-box adversarial learning techniques.

While this scenario is not quite realistic, we consider this to study whether white-box perturbations have an advantage compared to black-box perturbations against SRCNNs.

4 Potential Solutions

Adversarial perturbations have been explored as a potential privacy defense against unauthorized image classification [4, 19, 40]. However, the previous work did not consider the use of SRCNNs in the adversary's pipeline (see Figure 2). In this section we discuss some plausible ways to extend previous work to address the privacy threat posed by SRCNNs.

Learning jointly over both the SRCNN and CNN classifier: One way to defeat an adversary's pipeline shown in Figure 2 is to learn image perturbations optimized jointly over both the target SRCNN and CNN. This approach was explored in [61], where a joint optimization was proposed to learn an imperceptible noise that leads a known SRCNN to generate adversarial HR images against a known target CNN. In this approach, full knowledge of both the target SRCNN and CNN is necessary which makes it impractical for use in our setting where end-users may not have such knowledge. To address this, one could potentially try to learn transferable adversarial perturbations (*e.g.,* using ensemble of models [24] or universal perturbation generation methods [30]) that are jointly optimized over both a local CNN(s) and a local SRCNN(s). This approach is likely to be computationally expensive for end users given the joint optimization over both CNN(s) and SRCNN(s).

Learning transferable perturbations against SRCNNs: Another potential approach to defend against unauthorized classification by neural network pipelines involving SRCNNs, is to learn and embed a perturbation just against the SRCNNs. Learning adversarial images to degrade the quality of super-resolved HR images generated by the SRCNN has been explored recently [6]. However the transferability of those adversarial images to other SRCNNs is undetermined. While this may be addressed by employing transferable adversarial perturbation generation methods proposed for CNN classifiers [24, 30], it is not clear how effective just degrading the quality of super-resolved images will be against the target CNNs employed for recognition. It has been shown that CNNs can be robust to noise [12, 40] and therefore simply degrading the quality of super resolved images is unlikely to be sufficient.

Learning transferable adversarial images over CNN(s): The last alternative approach is to try to subvert the image classification CNN in the adversary's pipeline by learning transferable adversarial images that can fool the CNN. There are several methods for learning transferable adversarial images against CNN classifiers [24, 30, 31, 40]. However, perturbations in such adversarial images have to *survive* the super-resolution processing by the SRCNN in the adversary's pipeline. To the best of our knowledge, there has been no work investigating the impact of super-resolution on adversarial perturbations. In this work we explore this third approach and investigate the *survivability* of adversarial images learned over CNNs through SRCNN processing. We discuss details of this approach and the intuition behind it in the following section.

5 Survivable Adversarial Images

Our goal is to learn adversarial images against the classification CNN in the adversary's pipeline (see Figure 2) that can *survive* the processing by the SRCNN. In this section, we first describe our approach for generating such adversarial images and the intuition behind the approach. We then define the notion of *survivability* for such adversarial images, and finally the metrics for assessing their survivability.

5.1 Generating Survivable Low Resolution Adversarial Images

There are several adversarial attack models for learning adversarial images against a CNN both in a black-box or a white-box setting [3, 13, 30, 31]. If a user has access to the adversary's target CNN or can estimate the target CNN [35] by sending queries, white-box adversarial approaches are a better choice as they can generate adversarial images with imperceptible noise. However, black-box attacks that generate perceptible perturbations are more suitable (compared to white-box adversarial noise) for image privacy, since they are designed to fool unknown CNNs with high probability. But either approach is only useful against an SRCNN-CNN pipeline shown in Figure 2 if the resulting adversarial images can survive the processing by SRCNNs. This is because end users can only modify their LR images and cannot directly access or perturb the super-resolved image generated by the adversary's SRCNN.

Down-scaling HR adversarial images: We propose to generate adversarial images that can survive SRCNN processing by using the success of SRCNNs against them. Specifically, SRCNNs are optimized to generate HR images from LR images that are as close to the original HR images as possible. Using this observation, we propose to learn adversarial images (using either white-box or black-box approaches) corresponding to the HR user images and down-scale them to generate the LR adversarial images. Specifically, we use block-averaging for down-scaling the HR adversarial image. Since the owners only share LR images, they can deploy any down-scaling approach that preserves adversarial noise. Here, we find that a simple block-averaging approach is sufficient and can preserve the adversarial noise. This approach is depicted in Figure 3. The intuition behind this approach is that when an adversary takes such a down-scaled LR adversarial image and feeds it into their SRCNN it will produce a super-resolved HR image that is as close to the original HR adversarial image as possible which can then fool the adversary's classification CNN.
Directly learning LR adversarial images: We also explore and evaluate an alternate and a more direct approach to generating potentially survivable LR adversarial images. As shown in Figure 3 (b), black-box adversarial perturbations are learned directly on the LR user images using transferable learning techniques. Such an approach is useful when end users only have access to LR images or do not have sufficient training data or computational resources [2] needed to learn adversarial images on HR images. However, it is not clear at the outset whether such perturbations can survive the processing by SRCNNs.

5.2 Survivability of Adversarial Images Through SRCNNs

Intuitively, super-resolving the LR adversarial images created by down-scaling HR adversarial images, as proposed in the preceding section, will lead to HR images that are very similar to the original HR adversarial images. However, perceptual similarity doesn't necessarily imply that the adversarial nature and CNN fooling efficacy of the image are preserved. In order to assess the effectiveness of proposed approaches for generating adversarial images that can survive SRCNNs We define the notion of *survivability* as follows:

Definition 5.1 (Survivability)**.** *An LR adversarial image is considered to have* survived *through an SRCNN if the corresponding super-resolved HR image (i) is able to fool the target CNN, and (ii) perceptually preserves the true class* [3]*, i.e., visually similar.*

Note that the second requirement of perceptually preserving the true class by the super-resolved image is a strong requirement and is not necessary for thwarting unauthorized classification by SRCNN-CNN pipelines. However, meeting this requirement demonstrates that adversarial images can truly *survive* processing by SRCNNs and is of independent interest. Our evaluation will demonstrate that imperceptible noise learned over just CNN(s) can indeed survive SRCNNs, *i.e.,* retains its adversarial nature while still preserving recognizability (perceptually belongs to true class) for humans.

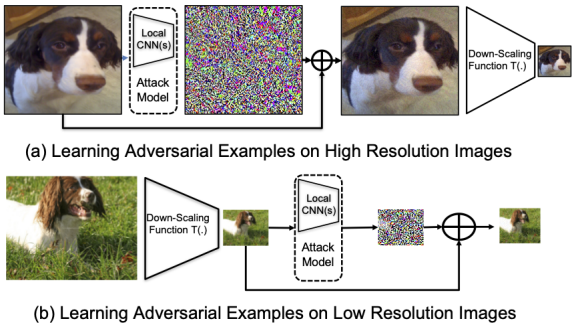

Fig. 3. Proposed approaches for generating LR adversarial examples for unknown SRCNNs using only a local CNN(s). a) Adversarial examples are learnt over clean HR images and down-scaled to produce LR adversarial examples. b) Adversarial examples are directly learnt on clean LR images. Adversarial learning can be either black-box (unknown target CNN) or white-box (known target CNN). No knowledge is assumed for the SRCNN (unknown).

5.3 Metrics

In this section, we introduce the metrics used for evaluating the survivability of adversarial images through SRCNNs. As defined, survivability has two properties. To assess the first property, namely, the ability to fool target CNNs, we measure transferability (see below).

2 Learning adversarial images on LR images is computationally cheaper.

3 The adversarial image belongs to the class of original (clean) HR image perceptually

To evaluate the second property, perceptually preserving the true-class, we use three image similarity metrics of (i) Peak Signal to Noise Ratio (PSNR), (ii) Structural Similarity Index(SSIM) [21], and (iii) Perceptual Similarity (PerSim) [63]. PSNR estimates the pixel-value similarity between two images, while the other two metrics focus on the visual similarity between images.

Transferability (TR): To measure the transferability of super-resolved HR images, we consider the misclassification rate which is defined as follows:

$$TR = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\text{argmax}F(hr_k(I'_i)) \neq Y^*_{I'_i})$$

where I'_i, $hr_k(I'_i)$, and $Y^*_{I'_i}$ are the i^{th} perturbed image, the super-resolved images by k^{th} super resolution convolutional neural network (SRCNN) and the true label of the perturbed image, respectively. $F(x)$ returns the probability vector generated by a CNN classifier for image x and $\mathbb{I}(t)$ is the identity function which returns one when t is true, and zero otherwise. In other words, TR measures the misclassification rate of the CNN on super-resolved adversarial images.

Peak signal to noise ratio (PSNR): This metric uses the normalized Minimum Square Error (MSE) between two images. Unlike MSE that depends on the scale of pixels' value of images, PSNR considers the ratio of the maximum possible value for a pixel to the difference between pixels as described below:

$$PSNR = 10\log_{10}\frac{R^2}{MSE}, \quad MSE = \frac{1}{M \times N}\|I_1 - I_2\|_2^2$$

where I_1 and I_2 are the images being compared, M and N are the dimensions (number of pixel rows and columns) of the image, and R is the maximum pixel value. For two identical images MSE converges to 0 and therefore PSNR value converges to infinity. In other words, if two image pixel values are very close, their PSNR would have a very large value.

Structural similarity index (SSIM) [21]: Unlike PSNR, SSIM is based on visible structures in the image, since human visual perception depends on structural information existing in an image and not just on pixel values. Therefore, this metric focuses on the local pattern of pixel intensities that have been normalized for luminance and contrast. SSIM metric returns a value in the range of $[0, 1]$ where two identical images will return an SSIM value of 1. Increasing the difference between the two images will cause their SSIM similarity value to decrease and approach zero.

Perceptual similarity (PerSim) [63]: Both PSNR and SSIM are susceptible to noise and even a small amount of noise could cause a significant degradation on these metrics, while the images would not change perceptually. To address this problem, the perceptual similarity was proposed in [63], which measures the similarity not in the image space (similarity between pixel values), but rather in a feature space. Feature space is the output of intermediate layers of a CNN. This metric considers two images similar if for a given CNN they have similar values in feature space. This metric returns a value in the range of $[0, 1]$. For two identical images, the PerSim value will be zero. We use all three metrics as they are all commonly used by designers to assess SRCNN performance, and since no one metric is perfect.

6 Evaluation

In this section, we evaluate the survivability of adversarial images through SRCNNs. We consider both white-box and black-box adversarial learning settings, and both methods of generating LR adversarial images described in Section 5.1 (see Figure 3). We describe our evaluation setup next including SRCNNs, CNN classifiers, datasets, and the adversarial learning approaches deployed.

6.1 Evaluation Setup

SRCNNs: We selected four state-of-the-art super-resolution convolutional neural networks, namely: (i) RCAN [22, 64], (ii) CAR [49] which obtained the best PSNR, (iii) SPSR [27] whose objective function minimizes PerSim, and (iv) DeepFace super-resolution [26] which is trained specifically for faces. SPSR is trained for a scale of 4 [4] and Deep Face SRCNN is trained for scales of 4 and 8. Both CAR and RCAN support scales of 2, 4 and 8.

Datasets: The different SRCNNs we selected impose different restrictions on the datasets that can be used. For example, DeepFace super-resolution can only work with CelebA for scaling up by a factor of 8. Keeping such constraints in mind we chose three different datasets: (i) Facescrub dataset [33], (ii) a CelebA dataset [25],

4 An SRCNN with a scale of 4 generates a HR image with resolution of $4M \times 4N$ from a LR image with resolution of $M \times N$.

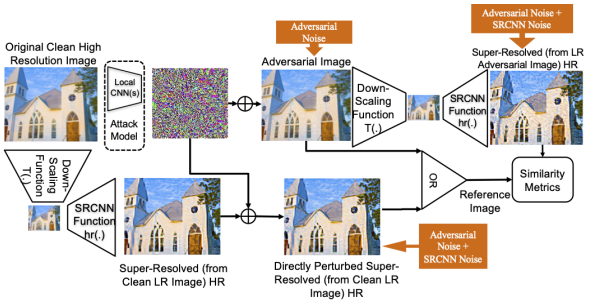

Fig. 4. Two approaches for selecting a reference image for similarity metrics: (i) original HR adversarial image, and (ii) perturbed super-resolved (from clean LR image) HR image.

and (iii) ImageNet dataset [43]. These datasets are well-known and commonly used for image classification tasks. However, they are very large datasets (*e.g.,* ImageNet dataset contains more than 1000 classes and $1M$ images). So we sampled a set of 1000 images at random from each of the three datasets to train local CNNs used for learning adversarial images. For ImageNet dataset sampled from 10 classes which are known to be less sensitive to noise and consequently more difficult to misclassify.

Classifiers: To measure the transferability of the adversarial noise and perturbation, we consider a state-of-the-art online face recognition model for celebrities called `clarifai.com` and pre-trained ImageNet [43] classifiers ResNet101 and ResNet152 with accuracy of 75.6% and 80.8%, respectively. ResNet101 and ResNet152 are the state-of-the-art CNN structures for large datasets [15].

To train LR adversarial images using small local CNNs, we used VGG-11 (a CNN with 11 layers) [48] on the 10 easiest classes of ImageNet dataset. VGG-11 could obtain high accuracy on CIFAR-10 dataset with 10 classes and it is suitable for small datasets [48]. We used 10 classes as it has been demonstrated that using small CNNs with 10 classes is sufficient to learn transferable perturbations [40]. We selected the 10 easiest classes since samples from these classes are less sensitive to noise and thus difficult to misclassify. We also used pre-trained robust CNNs from the robustness package of Python [10] trained for the large dataset of ImageNet [43].

Adversarial learning approaches: To train adversarial images with imperceptible noise we use Fast Gradient Sign (FGS) attack [13] when target CNN is known (white-box attacks). This attack is computationally inexpensive and fast. In addition, we evaluate the survivability of transferable (black-box attacks) adversarial perturbation trained using universal ensemble perturbation [40] and universal perturbation [30] methods (see Section A in Appendix for more details).

Evaluating survivability: We assess the ability of survivable adversarial images to fool the target classification CNN using the transferability metric. Also, we assess if the true class is preserved perceptually using the image similarity metrics discussed in Section 5.3. In other words, we assess whether super-resolved images are perceptually in the same class as their corresponding HR clean images. However, to be able to interpret the image similarity results, we need to establish a baseline and account for the noise introduced by the SRCNNs as discussed next.

Baseline: Similarity of original HR images to super-resolved clean LR images. First, as a baseline we evaluate the performance of SRCNNs in reconstructing original HR images from their clean LR counterparts *i.e.,* how similar super-resolved LR images are to their corresponding original HR images without any adversarial perturbations or noise. To this end, we measure the similarity between original HR images and super-resolved clean LR images ($\texttt{Sim}(I, hr_k(T(I))$ using the metrics discussed in Section 5.3.

Visual similarity: Similarity between original HR adversarial images and super-resolved LR adversarial images. We expect a super-resolved LR adversarial image to be similar to the original HR adversarial counterpart and belong to the true-class of HR clean image perceptually. To this end, we measure the similarity between super-resolved LR adversarial image and its original HR adversarial counterpart ($\texttt{Sim}(I+\delta, hr_k(T(I+\delta)))$, using image similarity metrics discussed in Section 5.3.

Noise adjusted visual similarity: Similarity between directly perturbed HR images super-resolved from clean LR images, and super-resolved LR adversarial images. SRCNNs are not able to reconstruct the exact original HR images of their LR inputs, and add a small amount of noise. As shown in Figure 4, to adjust for any degradation in the visual similarity introduced by the SRCNNs, we measure the similarity between directly perturbed super-resolved (from clean LR image) HR images and super-resolved LR adversarial images ($\texttt{Sim}(hr_k(T(I)) + \delta, hr_k(T(I+\delta)))$). In our evaluation, we show that this latter measurement gives a better estimation of adversarial images' survivability. Here, $\texttt{Sim}$, I, hr_k, T and δ denote a similarity metric function (PSNR, SSIM or PerSim), an original HR image, the k^{rh} SRCNN function, a down-scaling function, and an adversarial noise/perturbation respectively.

Evaluation scenarios: As shown in Table 1, we consider 3 different scenarios: i) down-scaling of HR ad-

	Learning LR Adversarial Images	Down-Scaling HR Adversarial Images
Black Box	FGS (Section 6.3) UP (Section 6.3)	UEP (Section 6.2) & FGS (Section 6.2)
White Box	NA	FGS (Section 6.2)

Table 1. Our 3 different evaluation scenarios. We consider two attack models of white-box and black-box with our two different approaches of generating LR adversarial images.

versarial images in white-box setting, ii) down-scaling of HR adversarial images in black-box setting, and iii) learning LR adversarial images on small CNNs in black-box setting. We do not consider a white-box setting for directly learning LR adversarial images as it is not applicable given the difference in size of the target and the local CNNs.

For learning HR adversarial images in a white-box attack, we use ImageNet dataset [43] with 1000 classes and two classifiers, ResNet101 and ResNet152, as local and target CNN classifiers. Also, we use the FGS attack approach to train HR adversarial images on the local CNN and test the super-resolved versions of down-scaled HR adversarial images on the target CNN (same as local in white-box setting and different in black-box setting).

We also use down-scaled UEP adversarial perturbation trained on 3 local CNNs [40] and test the super-resolved images (through different SRCNNs) on face recognition model of `clarifai.com` (black-box setting). We used FaceSrcub dataset for this scenario, since UEP is designed for image privacy against automated face recognition [40]. However, for the super-resolution with the scale of 8 we used CelebA dataset since the DeepFace SRCNN is trained on this dataset and can work well on this dataset for that scale. Evaluation of the down-scaled HR adversarial approach is discussed in Section 6.2, except for UEP evaluation with the scale 8 which is included in the Appendix B.

For directly learning LR adversarial images on a small local CNN, we use VGG-11 CNN with 10 classes and train adversarial images with imperceptible noise using FGS attack [13], and create transferable perturbations using the universal perturbation(UP) attack [30]. We use ImageNet dataset for both the attack approaches and pre-trained ImageNet CNNs as unknown target CNNs. This evaluation scenario is discussed in Section 6.3. Finally, we evaluate the survivability of these adversarial approaches on robust CNNs in Section 6.4.

6.2 Down-Scaled HR Adversarial Images

In this section, we evaluate the survivability of down-scaled HR adversarial images generated using universal ensemble perturbation (UEP) [40], designed for preserving image privacy (black-box setting), and Fast Gradient Sign [13], a fast and inexpensive adversarial attack model which generates adversarial images with imperceptible noise (both black-box and white-box).

Universal ensemble perturbation (UEP): To evaluate the survivability of UEP through SRCNNs, we use Facescrub dataset, since this perturbation was designed for image privacy and is learned on faces. We use `clarifai.com's` celebrity face recognition model as the adversary's CNN classifier since it provides large, highly accurate models for different classification tasks and the models and pre-processing steps are unknown. `clarifai.com's` models were previously used in the literature to evaluate the transferability of black-box adversarial perturbations [24, 40].

Transferability: We use the 1000 sampled face images from Facescrub dataset which are all recognizable by `clarifai.com` and provided the down-scaled images to SRCNNs. Table 2 presents the accuracy ($= 100 -$ `Transferability` for adversarials) of `clarifai.com` on super-resolved images by SRCNNs for scales of 2 and 4. SRCNNs we used could not reconstruct classifiable high resolution images for the larger scale 8 for this dataset [5]. As shown, 86.16% of the super-resolved (using CAR ($\times$4)) LR images are recognized by the celebrity face recognition model of clarifai.com. However, clarifai.com face recognition model could only recognize about 33.05% of HR images super-resolved from clean LR images using RCAN ($\times$4). We then perturbed the original HR images using UEP with $\beta = 3$ (as suggested in [40]). To generate perturbed LR images, we down-scaled the UEP perturbed HR images using block averaging. We then used the SRCNNs to super-resolve the resulting perturbed LR images. As shown in Table 2, at least $98\%(< 100\% - 1.79\%)$ of super-resolved UEP perturbed LR images can fool the face recognition model of `clarifai.com` successfully across all the SRCNNs considered.

Baseline: As discussed in Section 6.1, to baseline the performance of SRCNNs on generating HR resolution images, we measure the similarity between the super-resolved images and original HR clean images. We use

5 SRCNNs usually do not work well at larger scales for images that they have not seen during training.

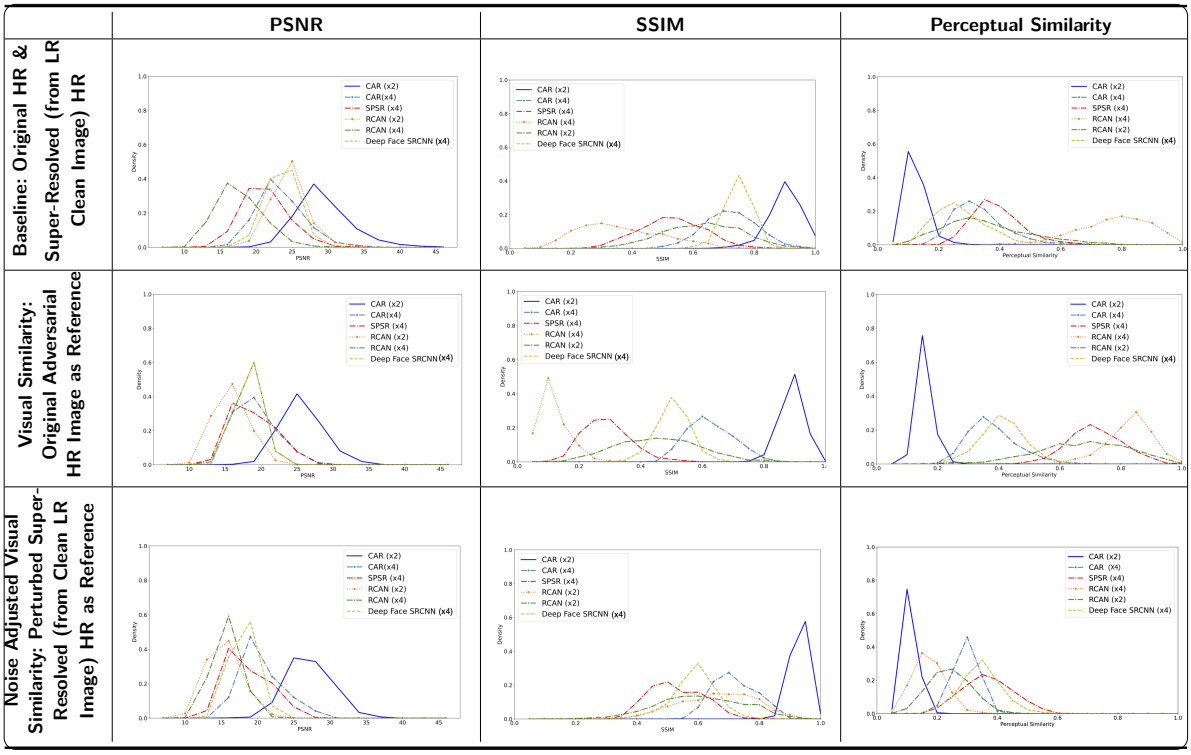

Fig. 5. The density distributions for a similarity metric to compare (i) Baseline: the super-resolved clean LR images and original HR images, (ii) Visual Similarity: super-resolved UEP perturbed LR images and original HR UEP perturbed images, and (iii) Noise Adjusted Visual Similarity: super-resolved UEP perturbed LR images and super-resolved clean LR images perturbed with UEP.

	Accuracy on Super-resolved HR Images from clean LR images		Accuracy on Super-resolved HR Images from UEP perturbed LR Images	
SRCNNs (scale)	**Face Detection**	**Face Recognition**	**Face Detection**	**Face Recognition**
CAR (x2)	**100%**	**87.44%**	**99.8%**	**1.72%**
CAR (x4)	**100%**	**86.16%**	**99.75%**	**0.5%**
RCAN (x2)	**99.26%**	**71.74%**	**99.8%**	**1.61%**
RCAN (x4)	**88.73%**	**33.05%**	**100%**	**0.4%**
SPSR (x4)	**100%**	**69.25%**	**99.65%**	**0.129%**
DeepFace SRCNN (x4)	**100%**	**71.82%**	**99.61%**	**1.66%**

Table 2. Accuracy of clarifai.com celebrity face detection and recognition models for super-resolved clean LR images (Baseline) and super-resolved UEP perturbed LR images.

the 3 similarity metrics of PSNR, SSIM and PerSim introduced in Section 5.3. As shown in Figure 5, we measure the density for each similarity metric. The first row in Figure 5 shows the similarity between an original HR image and the super-resolved HR image created from the clean LR counterpart. CAR (×2), Deep Face SRCNN and CAR (×4) SRCNNs reconstruct better than the others *i.e.,* the super-resolved images are more similar to their original counterparts. These results are compatible with the face recognition model's accuracy presented in Table 2.

As we expect, CAR (×2) which obtained highest accuracy on clarifai.com model has larger PSNR and SSIM values and smaller values for PerSim metric. For the scale of 8, none of CAR, RCAN and Deep Face SRCNNs could generate recognizable images for the face recognition model on FaceScrub dataset. Therefore, we use CelebA dataset with Deep Face SRCNN for evaluating the survivability of UEP for the larger scale of 8, as discussed later in the section.

Visual similarity: To assess how well survivable adversarial images preserve the true-class perceptually, we compare the similarity of the super-resolved LR perturbed images with directly perturbed HR images using metrics of PSNR, SSIM and PerSim. As shown in Figure 5, CAR (×2) and RCAN (×2) reconstruct origi-

nal adversarial images better. For example, the average value of PerSim between super-resolved LR UEP perturbed images by CAR (×2) and the original HR adversarial images (perturbed with UEP) is less than 0.15 for more than 75% of them (see graph in second row and the last column).

Noise adjusted visual similarity: recall that SRCNNs add noise and cannot reconstruct original HR images exactly. This noise leads similarity metric values to degrade. To adjust for any degradation in the visual similarity introduced by the SRCNN noise we consider a directly perturbed super-resolved HR image generated from a clean LR image as reference image for the visual similarity evaluation (see Figure 4). As shown in the last row of Figure 5, when using this noise adjusted measurement, we can see that the similarity metrics for survivable adversarial images have similar distributions as the baseline case. Note that since both reference images and super-resolved images have SRCNNs' noise, the similarity metrics have better values when compared to the case in which original HR adversarial images are used as reference images.

In summary, UEP survives through the SRCNN and the super-resolved images are not classifiable for the `clarifai.com` face recognition model, however, given the scaling level the super-resolved perturbed images are not perceptually similar to the original (See Figure 8 in Appendix). While without UEP, generated high resolution images are well-recognizable both for humans and automated classifiers. Loss of visual similarity however does not impact the privacy protections against SRCNNs provided by UEP as the super-resolved images are not accessed by end-users.

Fast gradient sign: To generate HR adversarial images on ImageNet dataset, we first select 1000 images classified correctly by the target CNN (ResNet101). Among those, we select images whose super-resolved HR images (from their LR counterpart) are classified correctly as well. Our experiments show that the super-resolved HR images by CAR (x2), RCAN (x2) and CAR (x4) have 86%, 69.9% and 39% accuracy on the ImageNet ResNet101 CNN classifier, respectively, and SPSR (x4) and RCAN (x4) could not generate classifiable images. We then learn adversarial examples (imperceptible noise) for these images against ResNet101 using Fast Gradient Sign method for different values of step size ($\epsilon \in \{0.001, 0.01, 0.02, 0.04, 0.05\}$).

Transferability: We measure transferability for both white-box settings in which the both local and target CNN are ImageNet classifier trained on ResNet101 and black-box setting in which the target CNN is ImageNet classifier trained on ResNet152. We also consider two cases (i) transferability of super-resolved LR adversarial examples (real-setting in which users do not have access to SRCNNs and can only perturb their LR image), and (ii) transferability of directly perturbed super-resolved LR clean images (idealized setting in which users can perturb the output of the SRCNNs). As shown in Table 3, for black-box setting the transferability is slightly lower (*i.e.,* values in column 3 are lower than values in column 5 on average) as adversarial example generation methods add less adversarial noise at cost of losing transferability. Also, the transferability of super-resolved LR adversarial images (when SRCNNs are not known) is actually better than the transferability of directly perturbed super-resolved LR clean images (*i.e.,* values in column 3(5) are better than values in column 4(6)). In other words, it shows having access to SRCNNs to directly perturb their output did not increase transferability. Our results in Table 3 show that the directly perturbed super-resolved images from clean LR images have lower transferabilities. Note that super-resolved images from LR images contain the SRCNNs' noise which may mitigate the transferability of the adversarial noise trained on clean HR images (i.e., SRCNN's noise cancels out some part of adversarial noise).

		Black-Box (ResNet152)		White-Box (ResNet101)	
	ϵ	$TR_R \uparrow$	$TR_I \uparrow$	$TR_R \uparrow$	$TR_I \uparrow$
CAR (x4)	0.01	29.2%	24.78 %	28.31%	19.76%
	0.02	43.95%	26.55%	42.3%	26.54%
	0.03	56.63%	28.02%	58.7%	28.61%
	0.04	66.37 %	29.45%	69.32%	31.56%
	0.05	75.51 %	31.27%	76.7%	37.46%
CAR (x2)	0.01	27.86%	16.9%	38.32%	23.96%
	0.02	39.05 %	23.96%	52.43 %	38.81 %
	0.03	50.85%	28.47%	59.97%	47.32%
	0.04	57.29%	35.89 %	64.84%	52.92%
	0.05	64.72%	41.24%	69.09%	57.42%
RCAN (x2)	0.01	50.33%	48.82%	55.03%	55.87%
	0.02	66.78 %	66.61%	68.79%	71.14%
	0.03	57.05 %	56.71%	57.88%	62.08%
	0.04	68.28 %	66.94%	70.80%	69.46%
	0.05	77.68 %	77.18%	77.18%	77.34%

Table 3. Transferability of super-resolved LR adversarial examples (TR_R) and transferability of directly perturbed super-resolved LR clean images(TR_I) for both black-box (ResNet152) and white-box setting (ResNet101)

As before, to evaluate perceptual preservation of the true class, we use similarity metrics using (i) the directly perturbed original HR adversarial examples (Visual Similarity), and (ii) directly perturbed images

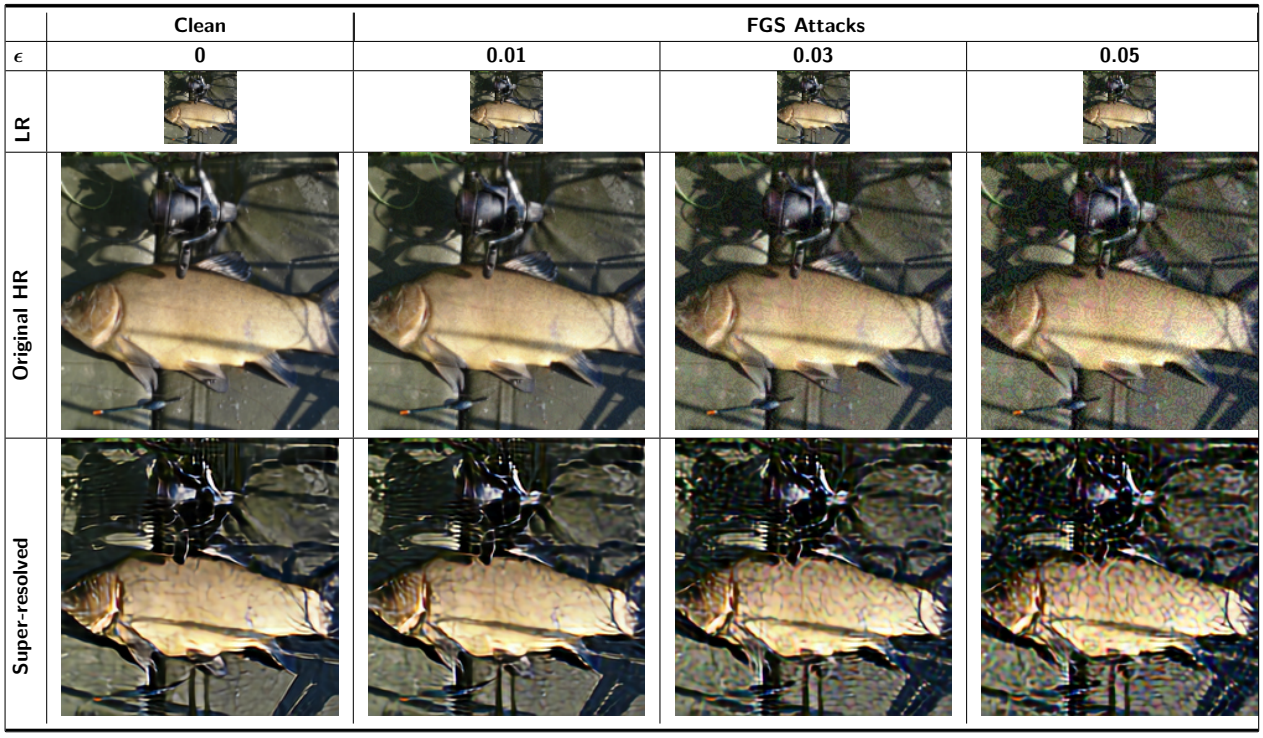

Fig. 6. The first row shows the LR image. The second row shows the original HR image. The third row shows the super-resolved HR image using CAR($\times 4$). The first column of each row shows the clean image and the rest of the columns show the adversarial examples generated for benign HR images with FGS attack and different ϵ values.

super-resolved from clean LR images (Noise Adjusted Visual Similarity) as reference images.

Visual similarity: As shown in Table 4, for the small values of ϵ, the adversarial noise does not transfer well. However, the larger ϵ values have higher transferability (misclassification rate) but lead to more distortion in super-resolved images even though the adversarial noise in low resolution images is imperceptible (see Figure 6). Consequently, larger values of ϵ lead the similarity metrics to degrade. For example, the LR adversarial images generated for $\epsilon = 0.01$ and $\epsilon = 0.03$ (the first image row (LR) in Figure 6) are perceptually similar but the super-resolved images (the third image row (Super-resolved) in Figure 6) are different. As shown in Table 4, the PerSim values of CAR (x4) for $\epsilon = 0.01$ and $\epsilon = 0.03$ are 0.276 and 0.446 on average.

Noise adjusted visual similarity: As expected, since both reference images and super-resolved LR adversarial examples ($hr(T(I + \delta))$) have SRCNN's noise, similarity metric values for PSNR and SSIM are higher and PerSim value is lower compared to the case above. For example, for $\epsilon = 0.03$ and CAR (x4), the average of PerSim reduces to 0.273 from 0.446.

6.3 Directly Learning LR Adversarial Images

As discussed in Section 5, we aim to evaluate if one can train a small CNN on low-resolution images to directly learn LR adversarial images which can survive through SRCNNs. To this end, we train a small CNN (VGG-11) on low resolution images of the 10 easiest classes of ImageNet. This local CNN with a few classes (small) is used by the end-user to learn adversarials directly on LR images. These LR adversarial images are then evaluated against the adversary's pipeline for effectiveness. We assume that the adversary uses publicly available images of the end-user to train their CNN and hence we allow for overlap in the training data for common classes but the adversary is assumed to have a much larger CNN. To learn a small CNN, we reduce the resolution of the images to 56×56 and use CAR (x4) as a target SRCNN. Here, we assume that an adversary has a large ImageNet classifier (the ResNet152 classifier learned on ImageNet dataset one with 1000 classes with input size of 224x224) and that she wants to classify the users' images. We use images from 10 easiest classes. Here we assume that the user uses her publicly shared images to train her local CNN. In other words,

		Visual Similarity: Original HR adversarial Image as Reference			Noise Adjusted Visual Similarity: Directly Perturbed Super-Resolved Clean LR as Reference		
	FGS(ϵ)	PSNR (avg (std)) ↑	SSIM (avg (std)) ↑	PerSim (avg (std)) ↓	PSNR (avg (std)) ↑	SSIM (avg (std)) ↑	PerSim (avg (std)) ↓
CAR (x2)	0.01	26.65 (2.97)	0.869 (0.034)	0.1 (0.024)	33.07 (2.09)	0.946 (SSIM)	0.027 (0.018)
	0.02	25.84 (2.34)	0.82 (0.03)	0.12 (0.024)	28.92 (1.46)	0.87 (0.04)	0.055 (0.024)
	0.03	24.77 (1.83)	0.78 (0.034)	0.143 (0.029)	26.48 (1.16)	0.82 (0.047)	0.09 (0.035)
	0.04	23.64 (1.45)	0.742 (0.038)	0.167 (0.034)	24.65 (0.98)	0.775 (0.05)	0.124 (0.042)
	0.05	22.53 (1.166)	0.712 (0.04)	0.189 (0.039)	23.16 (0.84)	0.7396 (0.05)	0.156 (0.047)
CAR (x4)	0.001	20.22 (3.23)	0.677 (0.105)	0.231 (0.059)	47.9 (2.76)	0.999 (0.0004)	0.0008 (0.0005)
	0.01	20.07 (3.12)	0.638 (0.086)	0.276 (0.056)	29.41 (2.23)	0.91 (0.023)	0.067 (0.041)
	0.02	19.62 (2.77)	0.564 (0.056)	0.366 (0.079)	325.22 (1.74)	0.785 (0.058)	0.1762 (0.089)
	0.03	19.02 (2.36)	0.50 (0.047)	0.446 (0.096)	22.88 (1.39)	0.688 (0.079)	0.273 (0.114)
	0.04	18.36 (1.98)	0.457 (0.047)	0.51 (0.104)	21.18 (1.146)	0.616 (0.088)	0.353 (0.12)
	0.05	17.69 (1.66)	0.422 (0.048)	0.562 (0.11)	19.83 (0.97)	0.56 (0.09)	0.418 (,0.132)
RCAN (x2)	0.01	18.73 (4.41)	0.61 (0.153)	0.356 (0.186)	24.44 (8.82)	0.85 (0.159)	0.072 (0.077)
	0.02	16.58 (4.22)	0.552 (0.164)	0.446 (0.2)	19.79 (6.47)	0.778 (0.15)	0.0958 (0.066)
	0.03	18.68 (4.69)	0.572 (0.164)	0.333 (0.2)	22.69 (6.3)	0.852 (0.094)	0.092 (0.07)
	0.04	17.63 (4.3)	0.55 (0.155)	0.35 (0.17)	19.1 (5.78)	0.7 (0.17)	0.147 (,0.09)
	0.05	15.39 (3.41)	0.47 (0.134)	0.41 (0.173)	16.64 (4.68)	0.649 (0.16)	0.183 (0.086)

Table 4. Similarity metrics for evaluating the survivability of adversarial examples through SRCNNs. We measure PSNR, SSIM and perceptual similarity for super-resolved LR adversarial examples for two different cases of (i) original adversarial examples as reference images, and (ii) directly perturbed super-resolved clean LR images as reference images.

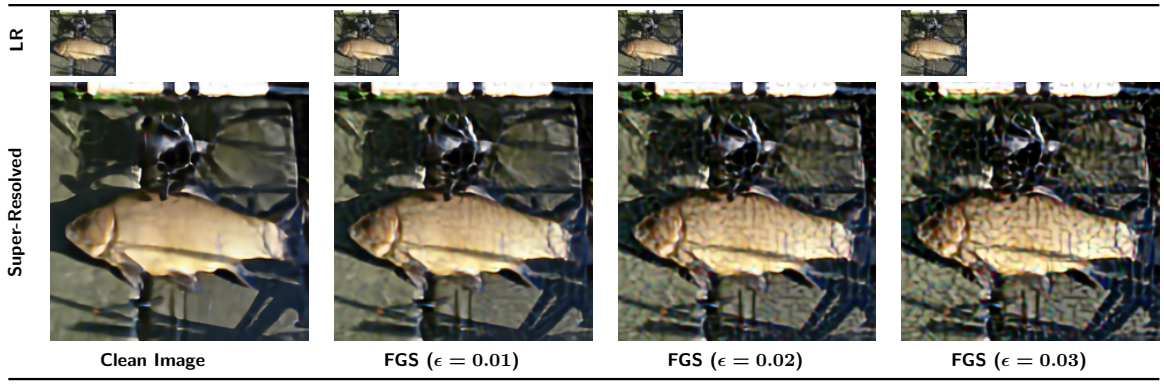

Fig. 7. The first column shows a clean low resolution image and its high resolution images by CAR (x4) and the rest columns show the adversarial examples generated on low resolution images and their high resolution images by CAR (x4). As shown by increasing the ϵ values leads the adversarial noise to increase and the quality of generated high resolution images to decrease.

		Super-Resolved LR Adversarial Image & Super-Resolved Clean LR Image			Clean LR Image & Adversarial LR Image		
ϵ	TR	PSNR (avg (std)) ↑	SSIM (avg (std)) ↑	PerSim (avg (std)) ↓	PSNR (avg (std)) ↑	SSIM (avg (std)) ↑	PerSim (avg (std)) ↓
0.01	88.8%	30.34(1.37)	0.86(0.04)	0.145 (0.079)	40.07(0.21)	0.988 (0.013)	0.003 (0.006)
0.15	90%	27.24 (1.05)	0.744 (0.07)	0.24 (0.10)	34.07 (0.23)	0.96 (0.04)	0.016 (0.03)
0.02	92.8%	25.02 (0.85)	0.64 (0.1)	0.32 (0.11)	34.07 (0.23)	0.96 (0.04)	0.016 (0.03)
0.03	95.8%	21.83 (0.67)	0.51 (0.11)	0.42 (0.11)	30.57 (0.25)	0.92 (0.0.7)	0.039 (0.05)
0.04	99.4%	19.57 (0.63)	0.413 (0.11)	0.48 (0.1)	28.09 (0.26)	0.88 (0.09)	0.069 (0.08)
0.05	99.7%	17.85(0.65)	0.35(0.1)	0.52 (0.09)	26.18 (0.28)	0.84 (0.1)	0.10 (0.1)

Table 5. Evaluating the survivability of black-box FGS attacks learned on low resolution images using a VGG-11 CNN with 10 classes. CAR (x4) SRCNN is used to scale up LR adversarial images and ResNet101 used for measuring the transferability. The first column shows the ϵ value for FGS attack. The second column shows transferability of high resolution images generated from low resolution adversarial examples. We measure similarity between the HR generated from clean LR image and the generated HR image from its LR resolution adversarial example as well as between two clean perturbed LR images.

the adversary has used the same images for the same classes but the user does not have access to any images from other classes. We select 300 LR images for which the target ImageNet classifier (ResNet101) classifies super-resolved images using CAR (×4) correctly. Then, we learn LR adversarial images using FGS and universal perturbation methods.

Fast gradient sign: To evaluate survivability of adversarial images with imperceptible noise, we learn FGS noise for different $\epsilon \in (\ 0.01, 0.015, 0.02, 0.03, 0.04, 0.05)$ values on those LR images. To measure the survivability, we evaluate the misclassification rate on ImageNet classifiers (ResNet101, ResNet152) and measure similarity between super-resolved clean LR images (as reference images) and super-resolved LR adversarial examples. As shown in Table 5, for even small values of $\epsilon(= 0.01)$, the adversarial images learned on small CNN are transferred well through CAR (x4) and fooled the ResNet101 ImageNet classifier for $> 88\%$ of the time. Figure 7 shows that the super-resolved LR adversarial image for $\epsilon = 0.01$ has imperceptible noise. In Table 5, the super-resolved LR adversarial example for $\epsilon = 0.01$ has large values for PNSR and SSIM metrics, and a low value for PerSim metric which demonstrates that the super-resolved LR adversarial images are perceptually close to super-resolved HR images without noise/perturbation. For larger values of ϵ, *e.g.,* $\epsilon = 0.03$, the PerSim value when comparing clean and perturbed LR image is 0.039 on average, while the perceptual similarity between original HR images and super-resolved LR adversarial images is 0.42 on average. Therefore, while increasing the value of ϵ leads to super-resolved LR adversarial images to be unclassifiable, they also have lower perceptual quality. We measure the similarity between the clean and perturbed LR images. The low values of PerSim metric (in Table 5) demonstrate that the clean and perturbed LR images are perceptually similar and consequently if clean LR images are recognizable for humans, then their adversarial LR images are expected to be recognizable for humans (also seen from Figure 7).

Universal perturbation: To evaluate the survivability of transferable LR adversarial images learned on a small CNN, we use universal adversarial perturbation. We train a universal perturbation with $\epsilon = 0.03$ which was able to fool the small CNN for 80.1% of images. Then we use CAR (x4) to generate high-resolution images for ImageNet classifiers (both ResNet101 and ResNet152). The super-resolved LR adversarial images could fool ImageNet classifiers of ResNet101 and ResNet152 at least 92.4% and 91.8% of the times, respectively. Since we learn an adversarial perturbation for LR images, we do not have original HR adversarial images for comparison. Hence, we use super-resolved (from clean LR images) HR images as reference images. The average (std) of PSNR, SSIM and PerSim metric values are 31.69(1.74), 0.915(0.19) and 0.06(0.04), respectively. The high values of PSNR and SSIM similarity metrics and low value for PerSim indicate that the super-resolved LR perturbed images will be perceptually recognizable. Also, the average (std) of PSNR, SSIM and PerSim metric values between LR adversarial images and LR clean images are 43.48(0.238), 0.995(0.005) and 0.0005(0.001), respectively. This indicates that LR adversarial images are perceptually similar to clean LR images which is more critical as the end-users only interact with LR images in our setting.

6.4 Impact of Defense Knowledge

An adversary with knowledge of the users' defense, may use i) robust CNNs and ii) filtering techniques to counter users' defense. In this section, we show that the adversarial noise trained to fool robust CNNs can survive well through SRCNNs. We also show the impact of filtering on the defense.

Robust CNNs: To evaluate super-resolved adversarial LR images against robust CNNs, we use the pre-trained networks from robustness packages from Python [10]. We select CNNs trained on ImageNet dataset [43] with robustness $L_2 = 3.0$ and $L_\infty = 8$ as our local and target CNNs, respectively [6]. Note that robust training against adversarials causes the accuracy of the CNN classifiers on clean data to drop. For example, our ImageNet CNN without any robustness can achieve 76.13% accuracy while $L_2 = 3$ and $L_{inf} = 8$ robustness caused the accuracy to drop to 57.90% and 47.91%, respectively. We train adversarial images on a local robust CNN, since the HR adversarial images trained on naive CNNs with a simple attack like FGS, cannot fool robust CNNs. To train HR adversarial examples, we select 1000 images at random such that both original HR and the corresponding super-resolved LR images are classified correctly by robust CNNs. To learn LR adversarial images, we learn untargeted adversarial HR images on the CNN with the robustness of $L_2 = 3$ by setting the attack's parameters of ϵ, step size, and the number of iterations to $\{3, 6\}$, 0.5

6 The pre-trained networks and their specifications are available at https://github.com/MadryLab/robustness

ϵ	SRCNN	White-box		Black-box	
		Original HR Advesarial	Super-resolved Adversarial LR	Original HR Advesarial	Super-resolved Adversarial LR
3	CAR(x2)	33.5%	31.4%	12.7%	27.4%
	CAR(x4)	33.5%	27.6%	12.1%	22.3%
6	CAR(x2)	71.5%	71.6%	42.8%	47.7%
	CAR(x4)	71.5%	63.6%	45.4%	52.2%

Table 6. Transferability of adversarial examples on robust CNNs. The local CNN (white-box) has the robustness of $L_2 = 3.0$ and the target CNN (Black-box) attack has the robustness of $L_\infty = 8$. The attacker's lower values of ϵ cause the adversarial examples to have lower transferability. SRCNNs can blow up the adversarial noise in the images and increase the transferability.

and 500, respectively. We use CAR (x2) and CAR (x4) to generate HR images from LR images.

White-box attacks: We first evaluate the adversarial examples trained on the CNN with robustness $L_2 = 3.0$ on itself. Only 33.5% of adversarial examples could cause the robust CNNs to misclassify them. As shown in Table 6, using a higher value of ϵ causes the adversarial to fool CNNs more successfully. For example, $\epsilon = 6.0$ causes the CNN to misclassify 71.5% of the generated adversarial examples. Note that increasing ϵ value causes the amount of perturbation added to images to increase. For example, PSNR similarity between clean HR images and original HR adversarial images for $\epsilon = 3$ and $\epsilon = 6$ are 42.06 and 36.169 on average, respectively. We trained the adversarial noise on HR images (224×224). After down-scaling HR adversarial images to resolutions of 112×122 and 56×56, we use SRCNNs of CAR (x2) and CAR (x4) to super-resolve these images. Our empirical study shows that the transferability of the adversarial images supper-resolved by CAR (x2) and CAR (x4) only drops from 33.5% to 31.4% and 27.6% respectively for $\epsilon = 3$ and from 71.5% to 71.6% and 63.6% respectively for $\epsilon = 6$.

Black-box attacks: To assess the survivability of adversarial examples for black-box attacks, we assess the transferability of adversarial examples trained on the CNN with robustness $L_2 = 3.0$ on the CNN with the robustness of $L_{inf} = 8$ (target robust CNN). As shown in Table 6, the original HR resolution images have low transferability. While SRCNNs can blow up the adversarial noise in the images which leads the transferability to improve. Also, using higher values of ϵ could improve the transferability of the super-resolved adversarial examples at the cost of increasing the adversarial noise.

Filtering: Filtering techniques were shown to reduce the impact of adversarial image perturbations and increase the classification accuracy [59]. Therefore, we applied those filtering techniques on super-resolved images before passing them to the classifiers to evaluate whether such filters can effectively thwart the impact of our perturbations. To this end, we studied the impact of applying average2D, blur, median blur, and bilateral filters to the output of SRCNNs. Our experiments show that for larger values of ϵ, median blur and bilateral filters may decrease the transferability but only insignificantly. For example, for $\epsilon = 6$ the bilateral filter (median blur) causes the transferability of CAR (x2) and CAR (x4) to drop to 41.8% (44%) and 43.9% (46.3%), respectively. For $\epsilon = 3$ none of these filtering techniques could reduce the transferability [7].

7 Discussion

Unauthorized image classification is a serious privacy threat that is exacerbated by the advent of SRCNNs. Adversarial perturbation is being proposed as a way to address privacy concerns with scalable automated CNN-based face recognition [4, 11, 19, 40]. In this work, we considered adversarial learning as a potential defense against unauthorized image classification pipelines that use both an SRCNN and a CNN.

7.1 Findings and Implications

Adversarial images against SRCNNs are indeed free: To the best of our knowledge, this is the first work to investigate the effectiveness of adversarial images learned using only CNN classifiers in fooling SRCNNs. SRCNNs are optimized to return the original HR of a given LR. Therefore, we expect that super-resolved images from LR adversarial images are close to their true corresponding HR images i.e., the original HR adversarial examples. Our empirical study shows

7 Our code is available at https://github.com/rajabia/Adversarial-Images-Against-Super-Resolution-Convolutional-Neural-Networks-SRCNNs-for-Free

that down-scaled HR adversarial images carry adversarial noise which leads SRCNNs to generate adversarial images, and simply learning adversarial images against CNNs is sufficient to effectively counter unauthorized classification of images by ML pipelines with SRCNNs without requiring any knowledge of such SRCNNs.

Both down-scaling perturbed HR images and directly perturbing LR images are effective: We investigated two approaches for creating LR adversarial images to counter SRCNNs: (i) down-scaling HR adversarial images, and (ii) directly learning LR adversarial images. Our findings show that regardless of the approach used for generating the LR adversarial images, these adversarial images can survive through SRCNNs and fool the target CNN effectively.

Black-box learning is equally effective: While white-box adversarial learning was found to exhibit more success rate on the known target CNNs with less adversarial noise, black-box adversarial learning was equally effective on unknown CNNs when learning adversarial perturbations. This shows it is possible to defend against adversary pipelines without any detailed knowledge of either the SRCNN or the CNN used.

SRCNNs might even make adversarial learning based defense even more practical: We showed that users only need a small local CNN (10 classes vs. 1000 classes for the target CNN) to learn LR adversarials and do not need large training datasets. Using a small CNN to learn adversarial images is computationally inexpensive and consequently makes adversarial-based image privacy more accessible than previously thought.

Practical viability: Despite CNNs' proven vulnerability to adversarial images, it is a long way before adversarial-based image privacy schemes can be reliably used for individual privacy. The development of robust CNNs with the ability of classifying adversarial images correctly is a potential threat, and developing such CNNs is currently a very active research topic [45]. While, our experiments show that learning transferable adversarial examples with the ability of fooling current state-of-the-art robust CNNs is possible, their effectiveness does reduce and fooling future robust CNNs is an open question. Also, while our experiments demonstrate current filtering methods cannot counter users' defense significantly newer filtering techniques may emerge. Given the lack of forward guarantees, unlike with cryptographic techniques, perturbation-based defenses may be more suitable for blunting the effectiveness of unauthorized classification (*e.g.,* across a population) rather than for individual privacy protections.

7.2 Open Questions

Better understanding survivability: This is the first work to study the impact of super-resolution on adversarial perturbations. While the first approach for generating survivable adversarials (*i.e.*, down-scaling perturbed HR images) is well supported by intuition, the second approach (*i.e.*, directly learning LR adversarials) was exploratory and understanding the reasoning for its effectiveness remains an interesting open problem. Emerging work on scale-invariance of adversarials [23] may provide a good starting point.

Survivability of adversarial images through robust SRCNNs: Several methods have proposed to improve performance of SRCNNs on generating HR images from poor quality LR images (*e.g.,* blurred, noisy images, etc.,) [47, 57] but they have not considered LR adversarial images with imperceptible noise which aim to degrade the quality of super-resolved HR images. It has been shown that learning CNN classifiers on both clean and adversarial images can improve their robustness to adversarial images [28, 37]. Even though these approaches do not guarantee robustness against all kinds of adversarial examples [45], it is worth investigating robust SRCNNs trained using enriched dataset with adversarial images.

Survivability of other adversarial attack models: In this paper, to evaluate the survivability of adversarial images, we considered three different adversarial attack models; (i) UEP, a transferable perturbation designed for image for image privacy, iii) UP, a universal perturbation approach and (iii) Fast Gradient Sign (FGS), a simple and fast adversarial images generation model. Assessing the survivability of other attack models [3, 31, 56] is an open problem. For example, adaptive adversarial learning is proposed to bypass the state-of-the-art adversarial detection methods [56], and consequently is a good candidate for evaluation against adversary countermeasures. Investigating the survivability of such adversarial perturbation can be an interesting future research direction.

Evaluating other down-scaling approaches: Even when using the same adversarial learning models, it is not clear how effective they would be if other "down-scaling" methods (*i.e.,* other than block-averaging) are used. That is, further evaluating the impact of combining different adversarial learning methods with different down-scaling approaches remains an open problem.

Impact of the accuracy of SRCNNs: Our approach relies on the fact that SRCNNs try to super-resolve as close to the original image as possible. How much impact

does the accuracy of the SRCNNs have on the success of the proposed approaches? Clearly reducing the accuracy of SRCNNs will impact accuracy of the adversary's pipeline, but is there a sweet spot for the adversary?

Survivability of random noise: Adversarial noise is trained to fool CNNs with a minimum amount of noise without provable privacy guarantees. Unlike this adversarial noise, using Laplacian noise may provide a way to argue for formal differential privacy guarantees for images. Therefore, an interesting direction is to investigate whether such perturbations can survive SRCNNs.

8 Related Work

Image privacy: Solutions for image privacy like blurring, mosaicing or redaction have been leveraged for a long time (*e.g.,* [20, 62, 65]). These approaches however are typically not reversible and therefore they are not practical in real-life applications [17]. To address this issue, two reversible schemes are proposed which thwart classifiers using false-color [8, 67]. Unfortunately these schemes do not provide recognizable images. Cryptographic image privacy techniques have been proposed (*e.g.,* [38, 55]), but they obfuscate the entire image and as a consequence introduce usability barriers. Recently, reversible image obfuscation approaches like thumbnail-preserving encryption (TPE) [29, 53, 58] which only leak pixelated version of the images, since it was thought LR images are not usable for classifiers. However, advent of single image deep super-resolution neural networks (SRCNNs) with their ability to generate high quality HR images from LR images threaten the security (or usability if the size of leaked thumbnail is reduced to compensate) of such schemes. Recently, due to the demonstrated vulnerability of CNN classifiers to adversarial noise, adversarial-based image privacy schemes received attention in the research community [4, 19, 40]. These schemes can fool unknown CNNs and provide recognizable images for end-users. It has been shown that such perturbations can be reversible [40]. Therefore, here we focus on adversarial-based image privacy methods.

Adversarial examples for CNNs: Adversarial perturbation techniques have succeeded in fooling convolutional neural network classifiers [3, 13, 31] and due to the transferability of adversarial noise to other CNNs, their are getting for privacy applications *e.g.,* image privacy [4, 11, 19, 40], membership privacy [32], etc. Recently, it has been shown that adversarial attack models can be extended for other types of convolutional neural networks *e.g.,* SRCNNs. Here, we briefly discuss adversarial learning approaches proposed for SRCNNs. Recently, in [6], an attack model has been proposed for generating adversarial noise on low resolution images to decrease the quality of HR images generated by SRCNNs. Similar to FGS attack for CNN classifiers, this method tends to learn a minimum adversarial noise on LR resolution image on SRCNNs in order to maximize the difference between a generated high resolution image and its original high resolution counterpart.

This method assumes that the target SRCNN is known to users. Moreover, the generated HR images from perturbed low resolution images are not necessarily able to fool a CNN, since their goal is degradation of super-resolved image's quality (not fooling a CNN classifier). To address this problem, a joint optimization has been proposed for learning adversarial noise on low resolution images in [61], such that it will mislead the well-trained and known image caption system (which is composed of a CNN-based encoder and an RNN-based decoder) to generate wrong captions for the super-resolved high resolution images by a given SRCNN. This method finds an optimum adversarial noise for fooling a known image caption system in the adversary pipeline, which makes it impractical for real-world applications.

9 Conclusion

In this paper, we hypothesized and empirically showed that adversarial images learned on CNN classifiers can lead super resolution convolutional neural networks to generate adversarial high resolution images. We evaluated the survivability of adversarial images, generated using well-known adversarial learning methods in both white-box and black-box settings, through state-of-the-art SRCNNs. We showed that even for imperceptible white-box noise, the quality of generated high resolution images by SRCNNs downgrades significantly, and that one can learn adversarial examples effective against SRCNNs using a small local CNN classifier trained on a limited number of images. We also found that our proposed approaches can be effectively used against robust CNN targets by using local robust CNNs for learning the adversarials. Our work shows that while SRCNNs pose a serious privacy threat, perturbation-based defenses developed against image classification CNNs may be leveraged to counter this threat.

Acknowledgements

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] M. Abbasi, A. Rajabi, C. Gagné, and R. B. Bobba. Towards dependable deep convolutional neural networks (cnns) with out-distribution learning. 2018.

[2] M. Abbasi, A. Rajabi, C. Gagné, and R. B. Bobba. Toward adversarial robustness by diversity in an ensemble of specialized deep neural networks. In *Advances in Artificial Intelligence*, pages 1–14. Springer International Publishing, 2020.

[3] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 39–57. IEEE, 2017.

[4] V. Chandrasekaran, C. Gao, B. Tang, K. Fawaz, S. Jha, and S. Banerjee. Face-off: Adversarial face obfuscation. *Proceedings on Privacy Enhancing Technologies*, 2:369–390, 2021.

[5] Z. Cheng, X. Zhu, and S. Gong. Low-resolution face recognition. In *Asian Conference on Computer Vision*, pages 605–621. Springer, 2018.

[6] J.-H. Choi, H. Zhang, J.-H. Kim, C.-J. Hsieh, and J.-S. Lee. Evaluating robustness of deep image super-resolution against adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 303–311, 2019.

[7] P. Chrabaszcz, I. Loshchilov, and F. Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

[8] S. Çiftçi, P. Korshunov, A. O. Akyüz, and T. Ebrahimi. Using false colors to protect visual privacy of sensitive content. In *Human Vision and Electronic Imaging Xx*, volume 9394, page 93941L. International Society for Optics and Photonics, 2015.

[9] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks, 2015.

[10] L. Engstrom, A. Ilyas, H. Salman, S. Santurkar, and D. Tsipras. Robustness (python library), 2019.

[11] I. Evtimov, P. Sturmfels, and T. Kohno. Foggysight: A scheme for facial lookup privacy. *arXiv preprint arXiv:2012.08588*, 2020.

[12] I. Goodfellow, Y. Bengio, and A. Courville. Deep learning (adaptive computation and machine learning series), 2016.

[13] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[14] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*, 2013.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] K. Hill. *The Secretive Company That Might End Privacy as We Know It*, 1/182020.

[17] S. Hill, Z. Zhou, L. Saul, and H. Shacham. On the (in) effectiveness of mosaicing and blurring as tools for document redaction. *Proceedings on Privacy Enhancing Technologies*, 2016(4):403–417, 2016.

[18] S. Hill, Z. Zhou, L. Saul, and H. Shacham. On the (in)effectiveness of mosaicing and blurring as tools for document redaction. In *Proceedings on Privacy Enhancing Technologies*, volume 2016, pages 403–417. Sciendo, 2016.

[19] S. Joon Oh, M. Fritz, and B. Schiele. Adversarial image perturbation for privacy protection – a game theory perspective. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[20] K. Lander, V. Bruce, and H. Hill. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology*, 15(1):101–116, 2001.

[21] C. Li and A. C. Bovik. Content-partitioned structural similarity index for image quality assessment. *Signal Processing: Image Communication*, 25(7):517–526, 2010.

[22] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[23] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft. Nesterov accelerated gradient and scale invariance for improving transferability of adversarial examples. *CoRR*, abs/1908.06281, 2019.

[24] Y. Liu, X. Chen, C. Liu, and D. Song. Delving into transferable adversarial examples and black-box attacks. *International Conference on Learning Representations (ICLR)*, 2017.

[25] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[26] C. Ma, Z. Jiang, Y. Rao, J. Lu, and J. Zhou. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[27] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou. Structure-preserving super resolution with gradient guidance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[28] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations(ICLR)*, 2018.

[29] B. Marohn, C. V. Wright, W.-c. Feng, M. Rosulek, and R. B. Bobba. Approximate thumbnail preserving encryption. In *Proceedings of the 2017 on Multimedia Privacy and Security*, MPS '17, pages 33–43, New York, NY, USA, 2017. ACM.

[30] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. *arXiv*

preprint, 2017.

[31] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, June 2016.

[32] M. Nasr, R. Shokri, and A. Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 634–646, 2018.

[33] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE international conference on image processing (ICIP)*, pages 343–347. IEEE, 2014.

[34] N. Papernot, P. McDaniel, and I. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[35] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519. ACM, 2017.

[36] H. Pham, Z. Dai, Q. Xie, M.-T. Luong, and Q. V. Le. Meta pseudo labels. *arXiv preprint arXiv:2003.10580*, 2020.

[37] H. Phan, M. T. Thai, H. Hu, R. Jin, T. Sun, and D. Dou. Scalable differential privacy with certified robustness in adversarial learning. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7683–7694. PMLR, 13–18 Jul 2020.

[38] M.-R. Ra, R. Govindan, and A. Ortega. P3: Toward privacy-preserving photo sharing. In *Presented as part of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*, pages 515–528, Lombard, IL, 2013. USENIX.

[39] A. Rajabi and R. B. Bobba. Adversarial profiles: Detecting out-distribution & adversarial samples in pre-trained cnns. 2020.

[40] A. Rajabi, R. B. Bobba, M. Rosulek, C. V. Wright, and W.-c. Feng. On the (im) practicality of adversarial perturbation for image privacy. *Proceedings on Privacy Enhancing Technologies*, 2021(1):85–106, 2021.

[41] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[42] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[44] M. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[45] Y. Sharma and P.-Y. Chen. Attacking the madry defense model with l_1-based adversarial examples. *arXiv preprint arXiv:1710.10733*, 2017.

[46] W. Shi, J. Caballero, C. Ledig, X. Zhuang, W. Bai, K. Bhatia, A. M. S. M. de Marvao, T. Dawes, D. O'Regan, and D. Rueckert. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In *International conference on medical image computing and computer-assisted intervention*, pages 9–16. Springer, 2013.

[47] G. Shim, J. Park, and I. S. Kweon. Robust reference-based super-resolution with similarity-aware deformable convolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8422–8431, 2020.

[48] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015.

[49] W. Sun and Z. Chen. Learned image downscaling for upscaling using content adaptive resampler. *IEEE Transactions on Image Processing*, 29:4027–4040, 2020.

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[51] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[52] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[53] K. Tajik, A. Gunasekaran, R. Dutta, B. Ellis, R. B. Bobba, M. Rosulek, C. V. Wright, and W. Feng. Balancing image privacy and usability with thumbnail-preserving encryption. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*, 2019.

[54] M. W. Thornton, P. M. Atkinson, and D. Holland. Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. *International Journal of Remote Sensing*, 27(3):473–491, 2006.

[55] M. Tierney, I. Spiro, C. Bregler, and L. Subramanian. Cryptagram: Photo privacy for online social media. In *Proceedings of the First ACM Conference on Online Social Networks*, COSN '13, pages 75–88, New York, NY, USA, 2013. ACM.

[56] F. Tramer, N. Carlini, W. Brendel, and A. Madry. On adaptive attacks to adversarial example defenses. *Conference on Neural Information Processing Systems (NeurIPS)*, 33, 2020.

[57] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[58] C. V. Wright, W.-c. Feng, and F. Liu. Thumbnail-preserving encryption for jpeg. In *Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security*, pages 141–146, 2015.

[59] W. Xu, D. Evans, and Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *Network and*

Distributed System Security Symposium (NDSS), 2018.

[60] Y. Yang, P. Bi, and Y. Liu. License plate image super-resolution based on convolutional neural network. In *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*, pages 723–727, 2018.

[61] M. Yin, Y. Zhang, X. Li, and S. Wang. When deep fool meets deep prior: Adversarial attack on super-resolution network. MM '18, page 1930–1938, New York, NY, USA, 2018. Association for Computing Machinery.

[62] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan. iprivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Trans. Information Forensics and Security*, 12(5):1005–1016, 2017.

[63] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[64] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018.

[65] Q. A. Zhao and J. T. Stasko. The awareness-privacy tradeoff in video supported informal awareness: A study of image-filtering based techniques. Technical report, Goergia Tech, http://hdl.handle.net/1853/3452, 1998.

[66] W. W. W. Zou and P. C. Yuen. Very low resolution face recognition problem. *IEEE Transactions on Image Processing*, 21(1):327–340, 2012.

[67] S. Çiftçi, A. O. Akyüz, and T. Ebrahimi. A reliable and reversible image privacy protection based on false colors. *IEEE Transactions on Multimedia*, 20(1):68–81, 2018.

A Adversarial Learning Approaches

We used the following three adversarial learning methods:

Fast gradient sign (FGS) [13]: This method aims to minimize the maximum changes in each pixel for an untargeted attack. More precisely, it tends to minimize L_∞. Therefore, it uses the sign of gradient of loss function as follows:

$$x_{new} = x + \epsilon \times sign(\nabla J(F(x, \theta), y^*)) \tag{3}$$

Here, y^* is the true label and ϵ is a small constant coefficient which controls the maximum changes of pixels.

Universal perturbation: Moosavi-Dezfooli *et al.* [30], showed that it is possible to find a single (universal) perturbation that can be applied to multiple images to successfully fool a CNN into misclassifying all of them. Specifically, universal perturbation is defined as a noise pattern δ, which when added to input images leads a CNN to misclassify the input images (x's) with probability of p. Universal perturbations are shown to be transferable to other CNNs with different structures, but trained on the same dataset as the original.

Universal ensemble perturbation (UEP): This method was proposed in [40] to generate a universal transferable perturbation using only a few small local CNNs.

This approach allows parameterized control over the level of perturbation to increase the transferability as follows:

$$x_{i,pert} = \frac{1}{2}(\tanh\left(\operatorname{arctanh}(2 \times (\mathrm{x_i} - 0.5)) + \beta \times \delta\right)) + 0.5 \tag{4}$$

where $x_{i,pert}$ is the perturbed version of the image x_i and δ is the learned perturbation. Here β is a weighting factor that tunes transferability versus perturbation amount. UEP perturbations have been shown to thwart a state of the art face recognition model, called `clarifai.com`, more than 85% of the time.

B UEP for the Scale of 8

To evaluate the survivability of very low resolution adversarial images, we used the DeepFace super-resolution [26] trained on CelebA dataset [25] and were able to scale up 16×16 pixel LR images to 128×128 pixel HR images (*i.e.,* scale-up of 8). We used 1000 of the test images and perturbed them with UEP.

Transferability: As before, to evaluate the transferability, we utilize `clarifai.com's` celebrity face recognition model. We submit clean LR resolution images and their super-resolved images by Deep Face SRCNN to this model. The accuracy of those images were 0% and 100%, respectively. In other words, this model could not recognize any faces in the very LR images. Then we downscaled the HR UEP perturbed images (to generate LR adversarial images). After super-resolving these UEP perturbed images, we submit them to `clarifai.com's` celebrity face recognition model. This model could not recognize any of the super-resolved images generated from perturbed images leading to a 100% transferability or fooling rate.

Baseline: The first row in Table 7 presents similarity metric values for baseline evaluation in which we compare a HR image super-resolved from a clean LR image with its original HR counterpart. As shown, SSIM and PSNR metrics have large value and PerSim metric has low value on average which shows that Deep Face super-resolution can reconstruct the original high resolution images well and is consistent with baseline face recognition accuracy discussed above. Both baselines for Deep

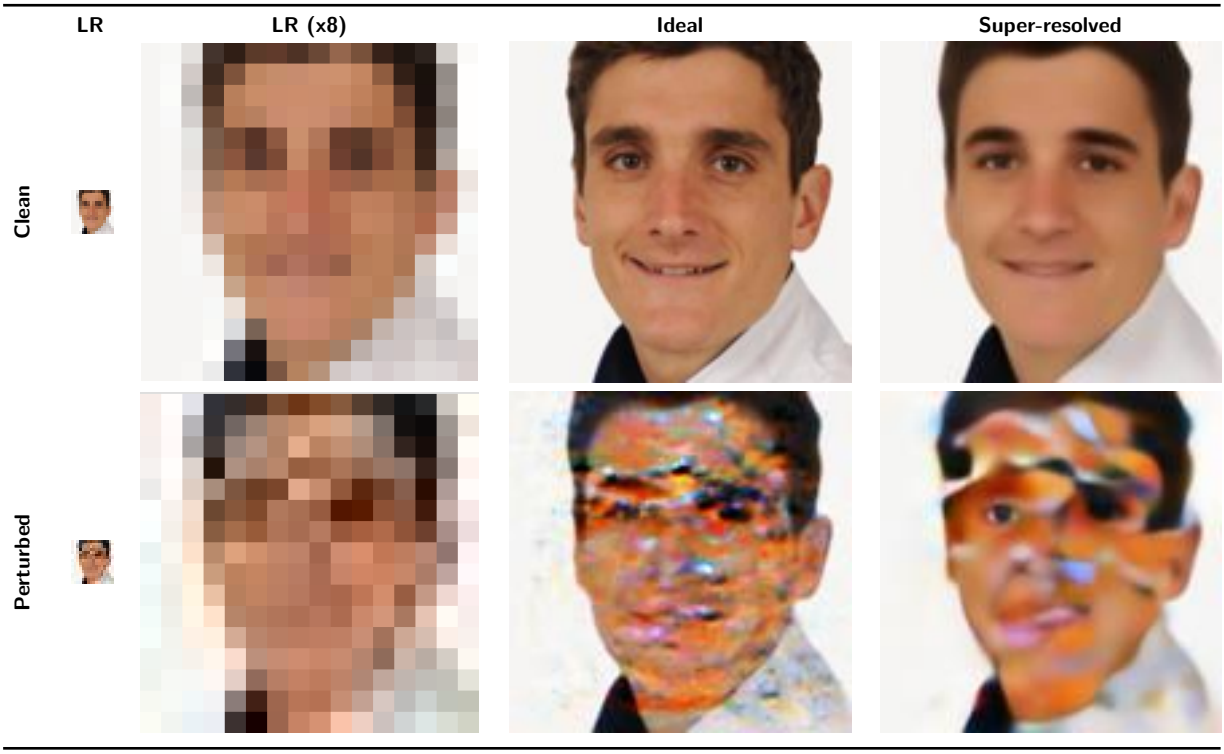

Fig. 8. The first column (LR) shows clean and UEP perturbed LR images, respectively. The second column (LR $\times 8$) shows the enlarged versions of LR images from the first column. The third column (Ideal) shows what we expect the SRCNN to generate. For the clean LR image, we expect an SRCNN to generate the original HR image and for perturbed LR image, we expect that an SRCNN generates an HR image close to its adversarial HR image. And finally the last column (Super-resolved) shows the super-resolved HR images generated by the Deep Face SRCNN from LR images (shown in the first column).

	PSNR (avg (std)) ↑	SSIM (avg (std)) ↑	PerSim (avg (std)) ↓
Baseline	25.69 (2.104)	0.822 (0.051)	0.183 (0.0578)
Visual Similarity	18.76 (0.721)	0.558 (0.0478)	0.363 (0.054)
Noise Adjusted Visual Similarity	19.716 (0.638)	0.622 (0.053)	0.32 (0.058)

Table 7. For the scale of 8, we use the celebA face dataset that Deep Face SRCNN learned on. The similarity metrics (for both Visual Similarity & Noise Adjusted Visual Similarity) for super-resolved LR perturbed images is close to the baseline which shows UEP perturbation survives well through Deep Face SRCNN.

Face SRCNN ($\times 4$) for Facescrub dataset and Deep Face SRCNN ($\times 8$) for celebA dataset obtained similar values for similarity metrics. For larger scale (8) value, we expected these values to degrade but since this SRCNN learned on celebA dataset, it is able to reconstruct original images well and obtain high values for SSIM and PSNR and low values for PerSim.

Visual similarity: The second row shows the similarity between super-resolved LR UEP perturbed images and original HR images perturbed with UEP. As shown SSIM, PSNR and PerSim metric values (the second row in Tables 7) under perform the baseline similarity metrics (the first row in Tables 7) both due to the strength of UEP perturbation and the scaling of the perturbation by SRCNNs.

Noise adjusted visual similarity: The third row in Table 7 shows the similarity between super-resolved LR perturbed images and directly perturbed super-resolved LR clean images. Both these images have the SRCNN's noise and therefore have slightly better similarity metric values compared to the Visual Similarity in which we use the original adversarial image as reference image (See Figure 4). As shown in Table 7, the average PerSim metric is 0.32 which is even less than the average PerSim metric for Deep Face SRCNN ($\times 4$) and Facescrub dataset which is 0.36. Since Deep Face SRCNN ($\times 4$) trained on celebA dataset, so naturally it reconstructs the images from this dataset better.