Robust Unpaired Single Image Super-Resolution of Faces

Saurabh Goswami, Rajagopalan A. N.
Indian Institute of Technology, Madras
Chennai- 600036, India
ee18s003@mail.iitm.ac.in, raju@ee.iitm.ac.in

Abstract
We propose an adversarial attack for facial class-specific Single Image Super-Resolution (SISR) methods. Existing attacks, such as the Fast Gradient Sign Method (FGSM) or the Projected Gradient Descent (PGD) method, are either fast but ineffective, or effective but prohibitively slow on these networks. By closely inspecting the surface that the MSE loss, used to train such networks, traces under varying degradations, we were able to identify its parameterizable property. We leverage this property to propose an adversarial attack that is able to locate the optimum degradation (effective) without needing multiple gradient-ascent steps (fast). Our experiments show that the proposed method is able to achieve a better speed vs effectiveness trade-off than the state-of-the-art adversarial attacks, such as FGSM and PGD, for the task of unpaired facial as well as class-specific SISR.

1 Introduction
Most real-world applications of SISR involve dealing with objects of a particular class (e.g. identifying faces in a speeding vehicle, spotting abnormalities in a medical image or monitoring the production line in a factory etc.). The capturing conditions are hardly ever ideal, and so the images often contain excessive blur, noise, illumination variation, and a host of different artefacts that can cause typical SISR networks employed in such situations to fail.

One way of making such networks robust to these degradations is to train them with examples that are accordingly degraded. However, compiling a dataset that covers such a wide range of diverse degradations is daunting to say the least. A much easier and more intuitive alternative is to train a conditional Generative Adversarial Network (GAN) that allows us to inflict a desired perturbation on an image as well as control its characteristics by tweaking a conditioning random vector. (Bulat, Yang, and Tzimiropoulos 2018; Goswami, Aakanksha, and Rajagopalan 2020; Maeda 2020) follow this methodology to synthetically degrade clean images before using them to train robust SISR networks. Given that (a) such networks are vulnerable to these degradations and (b) they do not produce stable outputs under varying degradations, this is akin to a black-box attack on the candidate SISR networks.

To turn this into a white-box adversarial attack, we need an algorithm that, given an image and an initial value of the conditioning random vector, would locate that value, in the neighbourhood of the initial value, which causes the loss function of the SISR network to attain its local maxima. State-of-the-art adversarial attacks such as FGSM (Goodfellow, Shlens, and Szegedy 2015) and PGD (Madry et al. 2018) fit this description. These methods ascend the gradient of a loss surface to find out what perturbation in the input would cause the maximum loss at the end of a network. To ensure preservation of semantic content, the perturbations are bounded and it is within these bounds that these attacks are optimal. However, learned degradation models are trained to always preserve semantic content irrespective of the value of conditioning random variables, making the norm unnecessary and rendering FGSM and PGD sub-optimal at best. Moreover, PGD, being an iterative procedure, takes a long time to converge to the desired maxima.

We observe that for Mean Square Error (MSE) loss function and normally distributed conditioning random vector, moving around on any two-dimensional (2D) plane in the space of this vector traces out, at the end of an unpaired facial SISR network, a surface that looks like the exponential of a second order polynomial and can be parameterized as such. Once we find these parameters, the maxima (if it exists) can be algebraically calculated without any gradient computation. This observation enables us to propose a novel white-box adversarial attack where, for every combination of a clean image and random normal vector: (a) we calculate the gradient of MSE loss at that random vector, (b) consider a triangle oriented in the direction of the computed gradient and sample loss values at multiple points on this triangle, (c) fit an exponential of a second order polynomial to this loss profile with weighted least squares and, (d) calculate the random vector corresponding to the worst degradation using the parameters of the estimated function. We do the above for every training iteration of the SISR network and feed the degraded image thus obtained to the network for training.

During inference, we simply feed Low Resolution (LR) images to the trained SISR module to get the super-resolved outputs.

The main contributions of our work are as follows:

• To the best of our knowledge, ours is the first work that considers the problem of adversarial robustness with degradations learned from real-world samples.
• We propose a novel white-box adversarial attack that is compatible with learned degradations and which leverages the parameterizable nature of the loss landscape of unpaired SISR networks.

• Even though our work mainly focuses on facial SISR, we perform extensive experiments to investigate the scope of our method as a useful adversarial attack even outside this specific scenario.

2 Previous Works

Real images are rife with degradations such as motion blur, defocus blur, random noise etc. Even though motion blur has been useful in detecting splicing forgeries in images (Rao, Rajagopalan, and Seetharaman 2014), recovering the latent motion (Purohit, Shah, and Rajagopalan 2019), and defocus blur has been used to infer depth from a single image (Paramanand and Rajagopalan 2011), in cases where the image is very small and contains only a single class, blur greatly reduces the recognizability of the image. Image deblurring, in itself, is a very challenging inverse problem and most previous works assume the blur to be uniform. (Purohit and Rajagopalan 2020) address it by using deformable convolution layers to adaptively change the size of receptive field to tackle non-uniform blur. However, it still leaves the task of SISR unattempted. Early works of Super-Resolution (Bhavsar and Rajagopalan 2012, 2010; Suresh and Rajagopalan 2007) used multiple shifted low-resolution images of the same scene to retrieve the latent high resolution image conditioned on the motion between the LR images. The performance of these algorithms, however, are highly dependent on the motion estimates. To address this, in (Rajagopalan and Kiran 2003), a motion free super-resolution was attempted by analytically deriving the relation for the reconstruction of the superresolved image from its blurred and downsampled versions. Recent works (Ledig et al. 2017; Wang et al. 2018; Vasu, Thekke Madam, and Rajagopalan 2018) though, use deep neural networks to directly learn the mapping from LR to HR domain.

2.1 Robust Facial SISR

It was first shown in (Yu and Porikli 2016) that facial images can be ultra-resolved using adversarial training. (Yu and Porikli 2017) extended this approach for noisy facial images by training an autoencoder SISR network that would have an additional decoder block at the beginning for denoising an input image. During training, they used a random noise model that does not cover blur, compression artifacts or even signal dependent noise. To train a facial SISR network suitable for real-world application, the noise model must adequately represent the degradations seen in real facial images. (Bulat, Yang, and Tzimiropoulos 2018) was the first work to take cognizance of this fact. They proposed a method whereby, in a two-step process, real degradation could be modeled with a Generative Adversarial Network (GAN) and an SISR network could be trained with LR images synthetically degraded with this model. During every epoch, differently degraded versions of the same images were shown to the SISR network by manipulating a random vector at the input of the degradation model. However, this method has two shortcomings: (a) it fails to maintain a steady output as the degradation in an image varies and, (b) it does not attempt to maximize the challenge for the SISR network by showing it degradations that it would find the most difficult to resolve. (Goswami, Aakanksha, and Rajagopalan 2020) addressed the first issue by fashioning their SISR network as an autoencoder the encodings of which were encouraged to stay consistent under varying degradations with the use of an explicit consistency constraint. This is an example of incorporating robustness through regularizing the SISR network to remain smooth (exhibit minimal variations) as the input transitions from clean to noisy. In our work, we address the second shortcoming by proposing an adversarial attack that is tailored for unpaired SISR network and which ensures that at every training iteration, our SISR network encounters only that degradation that it would find the most challenging to resolve.

2.2 Adversarial Robustness

The outputs of deep neural networks often change drastically with even a minimal perturbation in the input. Adversarial robustness immunizes networks against this by exposing them to their worst-case scenario. These attacks (i) inflict an additive perturbation on an input before feeding it to the network, (ii) climb the loss surface to converge to the perturbation that corresponds to a local maxima and, (iii) keep the norm of the perturbation constrained within a small value $\epsilon$ to preserve the semantic content of the input. In (Goodfellow, Shlens, and Szegedy 2015), the authors proposed an attack that linearizes the loss surface of the target network and takes a single step of fixed size along its gradient to calculate the adversarial perturbation. Requiring only a single forward and backward propagation, it is very fast. (Madry et al. 2018), showing that linearization is not effective for larger $\epsilon$, finds this worst degradation through multiple iterations, each comprising a forward and backward propagation, in an extremely slow process. (Zhang et al. 2019b) decomposes the robust error of a classification network into a natural classification error and a boundary error. They showed that a good trade-off between robustness and accuracy can be achieved by attacking the boundary error of a classification network. Performing multiple gradient ascent steps through the network per training iteration is frustratingly time-consuming, the inherently content-preserving nature of learned degradation models renders $\epsilon$ unnecessary, and SISR is not a classification task. These make all the above methods unsuitable for unpaired SISR. Observing and leveraging the parameterizable nature of the MSE-loss surface of such networks, we propose a simple, intuitive and effective attack for unpaired facial SISR that needs neither an $\epsilon$ nor multiple gradient ascent steps.

3 Proposed Method

3.1 Motivation

Choi et al. (2019) showed that SISR networks perform poorly on unseen degradations. In order to train a robust facial SISR network for real-world application, we would
1. For facial SISR, the loss landscape computed on most tri-

2. To validate our observation, we fit (learn a diverse degradation model) proposed by (Bulat, Yang, and Tzimiropoulos 2018) is to

3. The inner maximization term here tries to find the z that would cause the maximum loss. State-of-art attacks like PGD (Madry et al. 2018) and FGSM (Goodfellow, Shlens, and Szegedy 2015) are capable of performing this inner maximization. However, these are meant only for the following norm-bounded degradation model,

4. We needed a method that can handle general degradations \( f_{deg}(x, z) \) and is fast and accurate. We inspected the MLE loss landscapes of SISR networks under learnt degradation models and observed that these surfaces can be approximated with an exponential of second order polynomials. The worst degradation then becomes an algebraic function of the coefficients of these polynomials.

3.2 Observations

The \( f_{deg} \) modules, as used in (Bulat, Yang, and Tzimiropoulos 2018), (Goswami, Aakanksha, and Rajagopalan 2020), degrade an input image based on a random normal vector \( z \sim \mathcal{N}(0, I_n) \). We trained our \( f_{deg} \) with the specifications prescribed in these papers and observed the following about the MLE loss of our unpaired SISR network:

1. For facial SISR, the loss landscape computed on most triangles formed by 3 randomly sampled \( z \) had only a single peak (trough) and a gradual fall-off. Approximating them with an exponential of second order polynomials (with only 50 points) took us very close to the maxima (minima).

2. To validate our observation, we fit 800 component radial basis functions (RBFs) to the losses obtained for 8000 random \( z \) samples on a single image:

\[
l(z) = \sum_{i=1}^{800} a_i e^{-\frac{1}{2}(z-\mu_i)^T \Sigma_i^{-1}(z-\mu_i)}
\]

where \( l(z) \) is the loss at \( z \), and \( a_i, \mu_i, \Sigma_i \) are the coefficient, mean and standard deviation of the i-th RBF.

We diagonalized the estimated \( \Sigma \), to find out the standard deviations of the RBFs. The smallest standard deviations were found to be greater than 2.

3. Originating from \( \mathcal{N}(0, I_n) \), the \( L_2 \) distance \( d \) between two randomly sampled \( z \), according to (Thirey and Hickman 2015) has the following cumulative distribution function (CDF):

\[
F(d; n) = 1 - \frac{\Gamma\left(\frac{n}{2}, \frac{d^2}{4}\right)}{\Gamma\left(\frac{n}{2}\right)}
\]

where \( \Gamma \) is the upper incomplete gamma function. For a large \( n \), \( F(x; n) \approx 1 \) at \( x = 2 \) indicating that most pairs of \( z \) would be separated by distances less than 2. Since the peaks of loss surfaces are wider than gaussian function with standard deviation of 2, this verifies our observation that most triangles formed by 3 randomly sampled \( z \) would only contain one extrema.

4. This behavior persists across training epochs, making it a property that we can exploit.

3.3 Overview of the Method

To train all the modules from scratch, we need two datasets: (i) a dataset \( X \) with clean HR images, (ii) a dataset \( Y \) containing degraded LR images.

Learning a Degradation Model

- Our \( f_{deg} \) takes in an HR image \( x \in X \) and an n dimensional random normal vector \( z \sim \mathcal{N}(O, I_n) \) and produces a synthetically degraded LR image \( \hat{y} \) with similar semantical content as \( x \) and degradation similar to the samples in \( Y \).

\[
\hat{y} = f_{deg}(x, z)
\]

Solving the Inner Maximization Problem

- For each \( x \), we randomly sample \( z_1 \) and \( z_2 \), calculate the gradient of loss on \( z_1 \), and consider another point \( z_3 \) along the direction of this gradient. For simplicity, \( z_2 \) and \( z_3 \) are kept equidistant from \( z_1 \). This forms a triangle oriented in the direction where the loss starts increasing from \( z_1 \).

- We sample \( n \) number of points (\( z_i \)) inside these triangles as a convex combination of \( z_1, z_2 \) and \( z_3 \) with coefficients \( (\alpha_i, \beta_i) \) as follows

\[
z_i = \alpha_i z_1 + \beta_i z_2 + (1 - \alpha_i - \beta_i) z_3
\]

We combine each \( z_i \) with \( x \), compute the SR outputs \( f_{sr}(f_{deg}(x, z_i)) \) and calculate the corresponding MSE losses (\( l_i \)).

- We fit the following function to the curvated set of \( (\alpha_i, \beta_i, l_i) \),

\[
l_i = e^{-\alpha_i x^2+b_i z_i^2+c \alpha_i x+b \beta_i+d \alpha_i+e \beta_i+f}
\]
with weighted least square as 
\[(A^TW^TA)^{-1}A^TW^TWB\]
where:
\[
A = \begin{bmatrix}
\alpha_2^2 & \beta_2^2 & \alpha_1 \beta_1 & \alpha_1 & \beta_1 & 1 \\
\alpha_2 \beta_2 & \alpha_2 \beta_2 & \alpha_2 & \beta_2 & 1 \\
\alpha_3 \beta_3 & \alpha_3 \beta_3 & \alpha_3 & \beta_3 & 1 \\
\alpha_4 \beta_4 & \alpha_4 \beta_4 & \alpha_4 & \beta_4 & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\alpha_n^2 & \beta_n^2 & \alpha_n \beta_n & \alpha_n & \beta_n & 1 \\
\end{bmatrix},
\]
\[
B = \begin{bmatrix}
-\ln l_1 \\
-\ln l_2 \\
-\ln l_3 \\
-\ln l_4 \\
\vdots \\
-\ln l_n \\
\end{bmatrix}
\]  
and
\[
W = \text{diag}([w(l_1) w(l_2) w(l_3) w(l_4) \ldots w(l_n)])
\]
with
\[
w(l_i) = \frac{1}{\sqrt{2\pi}\sigma(l)}(\exp(\frac{l_i - \text{max}(l)}{\sigma(l)})+\exp(\frac{l_i - \text{min}(l)}{\sigma(l)}))
\]
where \(l\) is the vector containing all the \(l_i\)s and \(\sigma(l)\) is the standard deviation of \(l\). Since we are interested in accurately locating the local maxima (or minima), we put more weight on the extreme points with \(W\).

- Next, we find the \((\alpha_i, \beta_i)\) values corresponding to the extrema using the following relation.
\[
\begin{bmatrix}
\alpha_s \\
\beta_s \\
\end{bmatrix} = \begin{bmatrix}
2a & c \\
-c & 2b \\
\end{bmatrix}^{-1} \begin{bmatrix}
-d \\
-e \\
\end{bmatrix}
\]
The derivation for the same can be found in the supplementary material.

- We use Eq. 8 to calculate the corresponding loss \(l_s\), append it to the list of \((\alpha_i, \beta_i, l_i)\) and calculate the \(z_s\) corresponding to the maximum loss value in the list.

A batch implementation of this algorithm has been provided in the supplementary material.

Training the Super-Resolution Network

- We degrade all the images in a minibatch using their corresponding \(z_s\) and train \(f_{sr}\) with a combination of MSE Loss and an Adversarial Loss to recover the clean HR image.
\[
\hat{x} = f_{sr}(f_{deg}(x, z_s))
\]

3.4 The Key Modules: \(f_{deg}\) and \(f_{sr}\)

In this subsection, we detail the architectures and loss functions used for our two key modules. Since we mainly focus on formulating the attack, we adopt modules used in the previous works with only a few minor modifications.

The Degradation Module \(f_{deg}\)

- **Architecture:** As shown in Eq. 5, our degradation module takes a clean HR image \(x\) with a random vector \(z\) and produces \(\hat{y}\) that appears to be degraded like the samples in \(\mathcal{Y}\) and is semantically similar to \(x\). As shown in Fig. 1h, the entire module consists of 12 residual blocks, 4 of them contain 2x average pooling layers for downscaling and 2 of them have pixelshuffle layers for upsampling. In Fig. 1k, Res\((n_i, n_o)\) and Conv\((n_i, n_o)\) are residual and convolutional layers with \(n_i\) input channels and \(n_o\) output channels. The residual blocks are similar to the ones used in \(\text{(Bulat, Yang, and Tzimiropoulos 2018)}\).

\[
\mathbb{L}_{MSE}^{deg} = \frac{1}{N} \left\| y - \hat{y} \right\|_2^2
\]

where \(N\) is the dimensionality of \(x\).

To ensure that the distribution of the outputs resembles \(\mathcal{Y}\), we impose hinge GAN loss \(\text{(Lim and Ye 2017)}\) (shown in Eq. 15 and 16) on \(\hat{y}\). We removed the sigmoid layer from the discriminator of SRGAN \(\text{(Ledig et al. 2017)}\) and imposed spectral normalization \(\text{(Miyato et al. 2018)}\) on each layer to obtain our degradation discriminator \(D_d\). The loss function of \(D_d\) is as follows:

\[
\begin{align*}
L_{D_d} &= \mathbb{E}_{y \in \mathcal{Y}} [\max(0, 1 - D_d(y))] + \\
&\mathbb{E}_{x \in \mathcal{X}} [\max(0, 1 + D_d(f_{deg}(x, z)))]
\end{align*}
\]

Likewise, the adversarial loss used for \(f_{deg}\) is:

\[
L_{Adv}^{deg} = -\mathbb{E}_{x \in \mathcal{X}} D_d(f_{deg}(x, z))
\]

Hence, the complete loss function for \(f_{deg}\) is

\[
L_{deg} = \lambda_1 L_{MSE}^{deg} + \lambda_2 L_{Adv}^{deg}
\]

The Super-Resolution Module \(f_{sr}\)

- **Architecture:** We use the same network architecture as \(\text{(Bulat, Yang, and Tzimiropoulos 2018)}\) for designing \(f_{sr}\). It has 3 residual units with 12, 3, 3 convolutional blocks respectively. The first 2 units have 2x upscaling blocks at the end. To avoid checkerboard artifacts, we have used Nearest Neighbour interpolation for upscaling instead of transposed convolution. As shown in Eq. 13, \(f_{sr}\) takes a synthetically degraded LR image supplied by the \(f_{deg}\) and super-resolves it into clean HR image. Fig. 1i shows a diagram of this network.

- **Losses:** We train this network with a combination of pixel-wise MSE loss \(L_{sr}^{MSE}\) and adversarial loss \(L_{sr}^{Adv}\) defined as follows:

\[
\begin{align*}
L_{sr}^{MSE} &= \|x - \hat{x}\|_2 \\
L_{sr}^{Adv} &= -\mathbb{E}_{x \in \mathcal{X}} D_s(f_{sr}(f_{deg}(x, z)))
\end{align*}
\]

where the discriminator \(D_s\) has the same architecture as \(D_d\) and is trained with the following loss function:

\[
\begin{align*}
L_{D_s} &= \mathbb{E}_{x \in \mathcal{X}} [\max(0, 1 - D_s(x))] + \\
&\mathbb{E}_{x \in \mathcal{X}} [\max(0, 1 + D_s(f_{sr}(f_{deg}(x, z))))]
\end{align*}
\]

So,

\[
L_{sr} = \lambda_3 L_{sr}^{MSE} + \lambda_4 L_{sr}^{Adv}
\]
4 Experiments

Since a differentiable degradation module is the key component to carry out an adversarial attack, (Bulat, Yang, and Tzimiropoulos 2018) and (Goswami, Aakanksha, and Rajagopalan 2020) are the only two methods the comparisons with which would make sense in terms of validating the effectiveness of an adversarial training for unpaired SISR. However, we also include ESRGAN (Wang et al. 2018) in our comparison to show how faithful the outputs of our network are to the corresponding ground-truth images.

Due to SISR being a regression task, we can only compare our adversarial attack with FGSM (Goodfellow, Shlens, and Szegedy 2015) and PGD (Madry et al. 2018), since the more recent alternatives, such as TRADES (Zhang et al. 2019b) and YOPO (Zhang et al. 2019a), are tailored for classification tasks.

4.1 Datasets

Since our goal is to devise an adversarial attack for unpaired facial SISR networks, we focus primarily on faces. We randomly sample 153446 images from the Widerface (Yang et al. 2016) dataset to compile the degraded LR dataset, namely LRFace dataset. 138446 of these images were used for training and 15000 for testing. To compile the clean dataset, i.e. HRFace dataset, we combined the entire AFLW (Martin Koestinger and Bischof 2011) dataset with 60000 images from CelebAMask-HQ (Lee et al. 2019) dataset and 100000 images from VGGFace2 (Cao et al. 2018) dataset.

To see whether our attack would enhance the robustness of SISR networks in the larger class-specific as well as class-agnostic general unpaired SISR setting, we perform identical experiments on datasets of handwritten digits and natural scenes respectively. For handwritten digits, we took the original MNIST (LeCun and Cortes 2010) dataset as clean LR dataset, 4 times upscaled (using ESRGAN) counterpart as clean HR dataset, and the n-MNIST (Basu et al. 2017) dataset as degraded LR dataset. For SISR of natural scenes, we used the entire RealSR (Cai et al. 2019) dataset. 80% of all datasets were used for training, and 20% for testing.

4.2 Implementation Details

For all our experiments that involved implementing our method, we used 50 points per sample to estimate the loss surface \( r = 10, s = 5 \). We have performed all the training with a learning rate of \( 10^{-4} \) with Adam Optimizer \( (\beta_1 = 0, \beta_2 = 0.9) \) and a hyperparameter setting of \( (\lambda_1 = 1, \lambda_2 = 0.05, \lambda_3 = 1, \lambda_4 = 0.05) \). For all experiments, we have kept the scale factor of SISR at 4. We keep the dimensionality of \( z \) at 64 and for every input image, we compute 50 points on the corresponding triangles of \( z \) to carry out the adversarial attack. For PGD, we use 20 iterations with a step size of 0.01. For both PGD and FGSM, \( \epsilon = 1 \) has been used.

4.3 Adversarial Attack

This experiment tests the effectiveness of our method as an adversarial attack for SISR networks with learned degradation modules. We sample 1900 clean HR images from HRFace and we paired them with 1900 different \( z \). We then locate their \( z^* \) using FGSM (Goodfellow, Shlens, and Szegedy 2015), PGD (Madry et al. 2018) and our method. We calculate the average of the PSNRs and SSIMs entailed by all \( z^* \)s at the end of an undefended SISR network. Table 1 shows the results as well as the average time taken to perform each attack.
Since adversarial attacks are meant to harm the performance of a target network, a lower PSNR/SSIM here indicates a more potent attack. On HRFace, FGSM proves to be the weakest as well as the quickest attack. Our method though, proves stronger and faster than PGD which is an iterative attack. This is because our method already knows how the parameters of a loss surface relates to its maxima, whereas PGD merely climbs the surface with only a limited number of steps and a fixed step-size. Using more and finer steps will make PGD even slower, and less steps may not be enough for reaching maxima. This shows that for unpaired facial SISR, our method is an optimal adversarial attack with a good speed vs effectiveness trade-off. Performing this experiment on MNIST (LeCun and Cortes 2010), we got the same results as in the case of HRFace, showing that our attack is optimal for a broader class-specific unpaired SISR setting. On RealSR (Cai et al. 2019), our method was outperformed by PGD, suggesting that our hypothesis of single peak on most random triangles may not hold true for SISR of real scenes, something which was verified in Section 7.

### 4.4 Facial Super-Resolution

Here, we compare our adversarially robust facial SISR network with other facial SISR networks. Each network, in this experiment, is tested on two separate, unrelated datasets of 15,000 images each: (a) a dataset of clean LR images and, (b) a dataset of real-degraded LR images. For clean LR images, we report PSNR/SSIM between the outputs and the ground-truth HR images. A higher PSNR/SSIM indicates that the network is good at preserving semantic contents of an input. For degraded LR images, lacking the corresponding ground-truth, we report the Fréchet Inception Distance (FID) (Heusel et al. 2017) between the distribution of outputs and that of clean HR facial images. A lower FID indicates a more potent robustness. Since our work aims at enhancing the robustness of an unpaired SISR network, FID is our principal metric of interest. A PSNR/SSIM comparable with the previous methods only assures us that it does so without compromising the accuracy of reconstruction.

| Method   | PSNR/SSIM     | FID  |
|----------|---------------|------|
| ESRGAN   | 19.2533/0.8825 | 78.852 |
| Saurabh et al. | 18.5587/0.6012 | 25.572 |
| Bulat et al. | 16.9386/0.5038 | 23.158 |
| Ours-FGSM | 17.2348/0.4986 | 22.938 |
| Ours-PGD  | 17.1690/0.5127 | 15.583 |
| Ours*     | 17.0690/0.5038 | 9.454 |

Table 2: Performance comparison for facial SISR. PSNR/SSIM were computed on outputs of bicubically downsampled LR images. FID was computed on outputs of real degraded images.

In Tab. 2, we see that ESRGAN (Wang et al. 2018) achieves the highest PSNR/SSIM. This is expected since it was trained exclusively on clean (bicubically-downsampled) LR facial images. However, since it never...
saw real degradations during training, it gives the highest FID on degraded images. [Bulat, Yang, and Tzimiropoulos 2018] was trained exclusively on real-degraded images and ends up learning a slightly erroneous SR mapping which fails to preserve the identity, pose and expression of an input face. This leads to a lower FID but a poorer PSNR/SSIM performance on clean images. Our training batches are $1/3$ rd part clean LR image and $2$ rd part degraded LR image. So, the last 3 methods in Table 2 that refer to our network with different attacks, perform better than [Bulat, Yang, and Tzimiropoulos 2018] in terms of both PSNR/SSIM and FID. Ours*, the variant of our network that uses our proposed attack, gives the lowest FID among all methods without compromising too heavily in PSNR/SSIM. FGSM leads to the highest FID amongst all attacks and hence, is the least effective.

We perform similar experiments on both MNIST ([LeCun and Cortes 2010] and RealSR ([Cai et al. 2019]) datasets. They are included in the supplementary material.

5 Conclusion

We propose a fast and simple novel adversarial attack for class-specific SISR (4×) networks with learned degradation modules. By analyzing the MSE loss surface of these networks, we discovered their easily parameterizable nature and this paved the way for an adversarial attack that is, as we establish through our experiments, simple, fast and effective. Using this adversarial attack, we were able to train a facial SISR network that is more robust than the previous state-of-the-art networks. In our future works, we intend to study the loss surfaces of such networks in more depth and see whether there is a way to optimize our method even further.

References

Basu, S.; Karki, M.; Ganguly, S.; DiBiano, R.; Mukhopadhyay, S.; Gayaka, S.; Kannan, R.; and Nemani, R. 2017. Learning sparse feature representations using probabilistic quadtrees and deep belief nets. *Neural Processing Letters* 45(3): 855–867.

Bhavsar, A. V.; and Rajagopalan, A. 2010. Resolution enhancement in multi-image stereo. *IEEE transactions on pattern analysis and machine intelligence* 32(9): 1721–1728.

Bhavsar, A. V.; and Rajagopalan, A. N. 2012. Range map superresolution-inpainting, and reconstruction from sparse data. *Computer Vision and Image Understanding* 116(4): 572–591.

Bulat, A.; Yang, J.; and Tzimiropoulos, G. 2018. To Learn Image Super-Resolution, Use a GAN to Learn How to Do Image Degradation First. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision – ECCV 2018*, 187–202. Cham: Springer International Publishing. ISBN 978-3-030-01231-1.

Cai, J.; Zeng, H.; Yong, H.; Cao, Z.; and Zhang, L. 2019. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE International Conference on Computer Vision*.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

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

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. URL http://arxiv.org/abs/1412.6572.

Goswami, S.; Aakanksha; and Rajagopalan, A. N. 2020. Robust Super-Resolution of Real Faces Using Smooth Features. In Bartoli, A.; and Fusilero, A., eds., *Computer Vision – ECCV 2020 Workshops*, 169–185. Cham: Springer International Publishing. ISBN 978-3-030-66415-2.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, 6629–6640. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

LeCun, Y.; and Cortes, C. 2010. MNIST handwritten digit database URL http://yann.lecun.com/exdb/mnist/.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2019. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. *arXiv preprint arXiv:1907.11922*.

Lim, J. H.; and Ye, J. C. 2017. Geometric GAN.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. URL https://openreview.net/forum?id=rJzIBfZA6.

Maeda, S. 2020. Unpaired image super-resolution using pseudo-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 291–300.

Martin Koestinger, Paul Wohlhart, P. M. R.; and Bischof, H. 2011. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*.
Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957.

Paramanand, C.; and Rajagopalan, A. N. 2011. Depth from motion and optical blur with an unscented Kalman filter. IEEE Transactions on Image Processing 21(5): 2798–2811.

Purohit, K.; and Rajagopalan, A. 2020. Region-adaptive dense network for efficient motion deblurring. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 11882–11889.

Purohit, K.; Shah, A.; and Rajagopalan, A. 2019. Bringing alive blurred moments. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6830–6839.

Rajagopalan, A. N.; and Kiran, V. P. 2003. Motion-free superresolution and the role of relative blur. JOSA A 20(11): 2022–2032.

Rao, M. P.; Rajagopalan, A.; and Seetharaman, G. 2014. Harnessing motion blur to unveil splicing. IEEE transactions on information forensics and security 9(4): 583–595.

Suresh, K. V.; and Rajagopalan, A. N. 2007. Robust and computationally efficient superresolution algorithm. JOSA A 24(4): 984–992.

Thirey, B.; and Hickman, R. 2015. Distribution of Euclidean distances between randomly distributed Gaussian points in n-space. arXiv preprint arXiv:1508.02238.

Vasu, S.; Thekke Madam, N.; and Rajagopalan, A. 2018. Analyzing perception-distortion tradeoff using enhanced perceptual super-resolution network. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 0–0.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Loy, C. C.; Qiao, Y.; and Tang, X. 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks.

Yang, S.; Luo, P.; Loy, C. C.; and Tang, X. 2016. WIDER FACE: A Face Detection Benchmark. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Yu, X.; and Porikli, F. 2016. Ultra-resolving face images by discriminative generative networks. In European conference on computer vision, 318–333. Springer.

Yu, X.; and Porikli, F. 2017. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3760–3768.

Zhang, D.; Zhang, T.; Lu, Y.; Zhu, Z.; and Dong, B. 2019a. You only propagate once: Accelerating adversarial training via maximal principle. arXiv preprint arXiv:1905.00877.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019b. Theoretically principled trade-off between robustness and accuracy. In International Conference on Machine Learning, 7472–7482. PMLR.