Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero,
Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe Shi
Twitter, Inc.
{cledig,ltheis,fhuszar,jcaballero,aaitken,atejani,jtotz,zwang,wshi}@twitter.com

Abstract

Despite the breakthroughs in accuracy and speed of single image super-resolution using faster and deeper convolutional neural networks, one central problem remains largely unsolved: how do we recover the finer texture details when we super-resolve at large upscaling factors? During image downsampling information is lost, making super-resolution a highly ill-posed inverse problem with a large set of possible solutions. The behavior of optimization-based super-resolution methods is therefore principally driven by the choice of objective function. Recent work has largely focussed on minimizing the mean squared reconstruction error (MSE). The resulting estimates have high peak signal-to-noise-ratio (PSNR), but they are often overly smoothed, lack high-frequency detail, making them perceptually unsatisfying. In this paper, we present super-resolution generative adversarial network (SRGAN). To our knowledge, it is the first framework capable of recovering photo-realistic natural images from $4\times$ downsampling. To achieve this, we propose a perceptual loss function which consists of an adversarial loss and a content loss. The adversarial loss pushes our solution to the natural image manifold using a discriminator network that is trained to differentiate between the super-resolved images and original photo-realistic images. In addition, we use a content loss function motivated by perceptual similarity instead of similarity in pixel space. Trained on 350K images using the perceptual loss function, our deep residual network was able to recover photo-realistic textures from heavily downsampled images on public benchmarks.

1. Introduction

The highly challenging task of estimating a high-resolution (HR), ideally perceptually superior image from its low-resolution (LR) counterpart is referred to as super-resolution (SR). Despite the difficulty of the problem, research into SR received substantial attention from within the computer vision community. The wide range of applications [38] includes face recognition in surveillance videos [63], video streaming and medical applications.

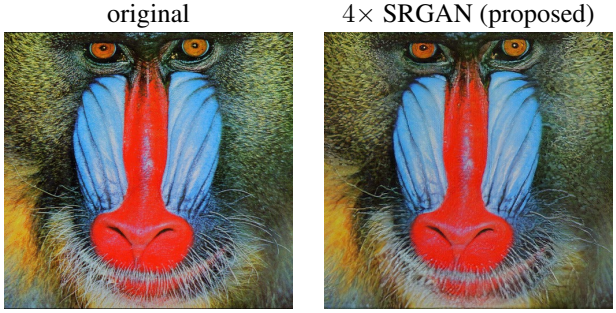

Figure 1: Super-resolved image (right) is almost indistinguishable from original (left) ($4\times$ downsampling).

One major difficulty when estimating the HR image is the ambiguity of solutions to the underdetermined SR problem. The ill-posed nature of the SR problem is particularly pronounced for high downsampling factors, for which texture detail in the reconstructed SR images is typically absent. Assumptions about the data have to be made to approximate the HR image, such as exploiting image redundancies or employing specifically trained feature models.

Over the last few decades substantial advances have been made in image SR [38, 55], with early methods based on interpolation, simple image features (e.g. edges) or statistical image priors. Later example-based methods very successfully detected and exploited patch correspondences within a training database or calculated optimized dictionaries allowing for high-detail data representation. While of good accuracy, the involved optimization procedures for

both patch detection and sparse coding are computationally intensive. More advanced methods formulate image-based SR as a regression problem that can be tackled for example with Random Forests [41]. The recent rise of convolutional neural networks (CNNs) also had a substantial impact on image SR [9], not only improving the state of the art with respect to accuracy but also computational speed, enabling real-time SR for 2D video frames [42].

The optimization target of supervised SR algorithms is usually the minimization of the mean squared error (MSE) between the recovered HR image and the ground truth. This is convenient as minimizing MSE also maximizes the peak signal to noise ratio (PSNR), which is a common measure used to evaluate and compare SR algorithms [55]. However, the ability of MSE (and PSNR) to capture perceptually relevant differences, such as high texture detail, is very limited as they are defined based on pixel-wise image differences [54, 51, 23]. This is illustrated in Figure 2, where highest PSNR does not necessarily reflect the perceptually better SR result. Figure 2 further illustrates that the ideal loss function depends on the application. Approaches that hallucinate finer detail might be perceptually convincing but less suited for medical applications or surveillance.

The perceptual difference between the super-resolved images and original images means that the super-resolved images are not photo-realistic as defined by Ferwerda [14]. Photo-realistic image super-resolution techniques including [46, 62, 59] have been focusing on minimizing the perceptual differences by using detail synthesis, a multi-scale dictionary or a structure aware loss function.

In this work we propose super-resolution generative adversarial network (SRGAN) for which we employ a deep residual network and diverge from MSE as the sole optimization target. Different from previous works, we define a novel perceptual loss using high-level feature maps of the VGG network [43, 28, 5] combined with a discriminator that encourages solutions perceptually hard to distinguish from the HR reference images. An example of a photo-realistic image that was super-resolved from a $4\times$ downsampling factor using SRGAN is shown in Figure 1.

1.1. Related work

1.1.1 Image Super-Resolution

There is a vast amount of literature and research that focuses on the problem of recovering high-resolution images from a low-resolution observation. Recent overview articles include Nasrollahi and Moeslund [38] or Yang et al. [55]. Here we will focus on single image super-resolution (SISR) and will not further discuss approaches that recover HR images from multiple images, such as object images acquired from varying view points or temporal sequences of image frames [4, 13].

Prediction-based methods are among the first and more straightforward methods to tackle SISR. While these filtering approaches, e.g. linear, bicubic or Lanczos [12] filtering, can be very fast, they oversimplify the SISR problem and usually yield overly smooth solutions failing to recover the high-frequency image information. Interpolation methods that put particular focus on edge-preservation have been proposed in for example Allebach and Wong [1] or Li et al. [34].

More powerful approaches aim to establish a complex mapping between low- and high-resolution image information and usually rely on training data.

Many methods that are based on example-pairs rely on LR training patches for which the corresponding HR counterparts are known. Early work was presented by Freeman et al. [16, 15]. Related approaches to the SR problem originate in compressed sensing and aim to estimate a sparse patch representation with respect to an over-complete dictionary [56, 10, 61]. In Glasner et al. [19] the authors exploit patch redundancies across scales within the image to drive the SR. This paradigm of self-similarity is also employed in Huang et al. [26], where insufficiently descriptive self dictionaries are extended by further allowing for small transformations and shape variations.

To reconstruct realistic texture detail while avoiding edge artifacts, Tai et al. [46] combine an edge-directed SR algorithm based on a gradient profile prior [44] with the benefits of learning-based detail synthesis. Zhang et al. [62] propose a multi-scale dictionary to capture redundancies of similar image patches at different scales with the goal to enhance visual details. To super-resolve landmark images, Yue et al. [59] retrieve correlating HR images with similar content from the web and propose a structure-aware matching criterion for alignment.

Neighborhood embedding approaches upsample a given LR image patch by finding similar LR training patches in a low dimensional manifold and combining their corresponding HR patches for reconstruction [47, 48]. In Kim and Kwon [30] the authors emphasize the tendency of neighborhood approaches to overfit and learn a more general map from low- to high-resolution images from example pairs using kernel ridge regression.

The regression problem can also be solved directly with Random Forests and thus the explicit training of a sparse dictionary is avoided [41]. In Dai et al. [6] the authors learn a multitude of patch-specific regressors during training and select the most appropriate regressors for a given LR patch during testing using kNN.

Recently CNN based SR algorithms have shown excellent performance. In Wang et al. [52] the authors encode a sparse representation prior into their feed-forward network architecture based on the learned iterative shrinkage and thresholding algorithm (LISTA) [21]. Dong et al. [8, 9]

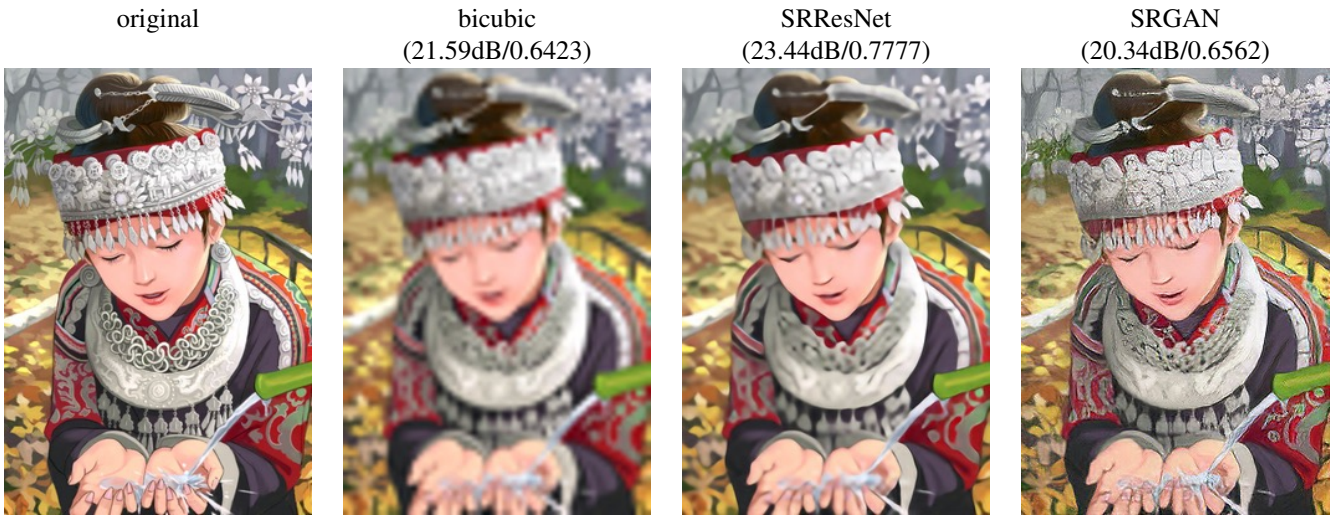

Figure 2: Illustration of performance of different SR approaches with downsampling factor: $4\times$. From left to right: original HR image, bicubic interpolation, deep residual network optimized for MSE, deep residual generative adversarial network optimized for a loss more sensitive to human perception. Corresponding PSNR and SSIM are shown in brackets.

used bicubic interpolation to upscale an input image and trained a three layer deep fully convolutional network end-to-end to achieve state of the art SR performance.

Subsequently, it was shown that enabling the network to learn the upscaling filters directly can further increase performance both in terms of accuracy and speed [42, 50]. In Shi et al. [42] the upscaling is only performed in the last layer of the network avoiding expensive computations in the high-resolution space of the SR image. With their deeply-recursive convolutional network (DRCN), Kim et al. [29] presented a highly performant architecture that allows for long-range pixel dependencies while keeping the number of model parameters small. Of particular relevance in the context of our paper are the works by Johnson et al. [28] and Bruna et al. [5], who rely on a loss function closer to perceptual similarity to recover visually more convincing HR images.

1.1.2 Design of Convolutional Neural Networks

The state of the art for many computer vision problems is meanwhile set by specifically designed CNN architectures following the success of the work by Krizhevsky et al. [32].

It was shown that deeper network architectures can be difficult to train but have the potential to substantially increase the network's accuracy as they allow modeling mappings of very high complexity [43, 45]. To efficiently train these deeper network architectures batch-normalization [27] is often used to counteract the internal co-variate shift. Deeper network architectures have also been shown to increase performance for SISR, e.g. Kim et al. [29] formulate a recursive CNN and present state of the art results. Another powerful design choice that eases the training of deep CNNs is the recently introduced concept of residual blocks [24] and skip-connections [25, 29]. Skip-connections relieve the network architecture of modeling the identity mapping that is trivial in nature, however, potentially non-trivial to represent with convolutional kernels.

In the context of SISR it was also shown that learning upscaling filters is beneficial both in terms of speed and accuracy [53, 42]. This is an improvement over Dong et al. [9] where data-independent, bicubic interpolation is employed to upscale the LR observation before feeding the image to the CNN. In addition, by extracting the feature maps in LR space as in [42, 28], the gain in speed can be used to employ a deep residual network (ResNet) to increase accuracy.

1.1.3 Loss Functions

Pixel-wise loss functions such as MSE struggle to handle the uncertainty inherent in recovering lost high-frequency details such as texture: minimizing MSE encourages finding pixel-wise averages of plausible solutions which are typically overly-smooth and thus have poor perceptual quality [37, 28, 11, 5]. Example reconstructions of varying perceptual quality are exemplified with corresponding PSNR in Figure 2. We illustrate the problem of minimizing pixel-wise MSE in Figure 3 where multiple potential solutions with high texture details are averaged to create a smooth reconstruction.

In Mathieu et al. [37] and Denton et al. [7] the authors tackled this problem by employing generative adversarial networks (GANs) [20] for the application of image genera-

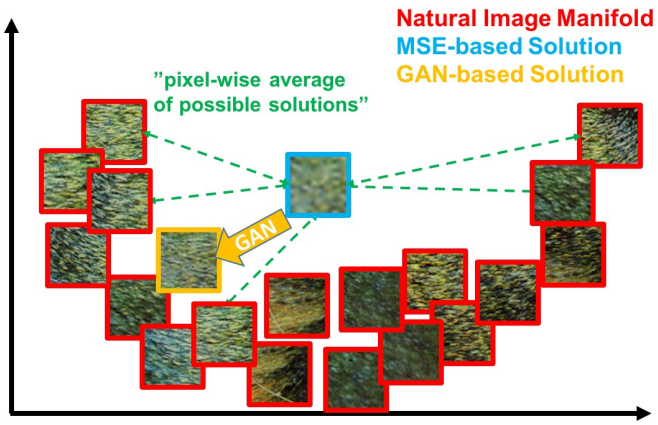

Figure 3: Illustration of patches from the natural image manifold (red) and super-resolved patches obtained with MSE (blue) and GAN (orange). The MSE-based solution appears overly smooth due to the pixel-wise average of possible solutions in the pixel space, while GAN drives the reconstruction towards the natural image manifold producing perceptually more convincing solutions.

tion. GANs were also used for unsupervised representation learning in Radford et al. [39]. The idea of using GAN to learn a mapping from one manifold to another is described by Li and Wand [33] for style transfer and Yeh et al. [57] for inpainting. Bruna et al. [5] minimize the squared error in the feature spaces of VGG19 and scattering networks.

Dosovitskiy and Brox [11] use loss functions based on Euclidean distances computed in the feature space of neural networks in combination with adversarial training. It is shown that the proposed loss allows visually superior image generation and can be used to solve the ill-posed inverse problem of decoding nonlinear feature representations. Similar to this work, Johnson et al. [28] and Bruna et al. [5] propose the use of features extracted from a pretrained VGG network instead of low-level pixel-wise error measures. Specifically the authors formulate a loss function based on the euclidean distance between feature maps extracted from the VGG19 [43] network. Perceptually more convincing results were obtained for both super-resolution and artistic style-transfer [18, 17]. Recently, Li and Wand [33] also investigated the effect of comparing and blending patches in pixel or VGG feature space.

1.2. Contribution

GANs provide a powerful framework for generating plausible-looking natural images with high perceptual quality. In the GAN framework, two neural networks, a generator and a discriminator, are trained simultaneously with competing goals. The discriminator network is trained to distinguish natural and synthetically generated images, while the generator learns to generate images that are indistinguishable from natural images by the best discriminator. In effect, the GAN procedure encourages the generated synthetic samples to move towards regions of the search space with high probability of containing photo-realistic images and thus closer to the natural image manifold as shown in Figure 3.

In this paper we describe the first very deep ResNet [24, 25] architecture using the concept of GANs to form a perceptual loss function close to the human perception for photo-realistic SISR. Our main contributions are:

- We set a new state of the art for image SR from a high downsampling factor ($4\times$) as measured by PSNR and structural similarity (SSIM) with our deep residual network (SRResNet) optimized for MSE. Specifically, we first employ the fast feature learning in LR space [42, 28] and batch-normalization [27] to robustly train a deep network of 15 residual blocks for better accuracy.

- We are able to recover photo-realistic SR images from high downsampling factors ($4\times$) by using a combination of content loss and adversarial loss as our new perceptual loss. The adversarial loss is driven by a discriminator network to encourage solutions from the natural image domain while the content loss ensures that the super-resolved images have the same content as their low-resolution counterparts. We further replace the MSE-based content loss function with the euclidean distance between the last convolutional feature maps of the VGG network [43], which are more invariant to changes in pixel space as illustrated in Figures 3 and 4 in Li and Wand [33].

We validate the proposed approaches using images from publicly available benchmark datasets and compare our performance against state-of-the-art works including SRCNN [8], SelfExSR [26] and DRCN [29]. We confirm our network's potential to reconstruct photo-realistic images under $4\times$ downsampling factors as compared to conventional methods.

In the following we describe the network architecture and the perceptual loss in Section 2. A quantitative evaluation on public benchmark datasets as well as visual illustrations are provided in Section 3. The paper concludes with a discussion of limitations in Section 4 and concluding remarks in Section 5. A random selection of images super-resolved by SRGAN are provided in the Appendix.

2. Method

In SISR the aim is to estimate a high-resolution, super-resolved image I^{SR} from a low-resolution input image

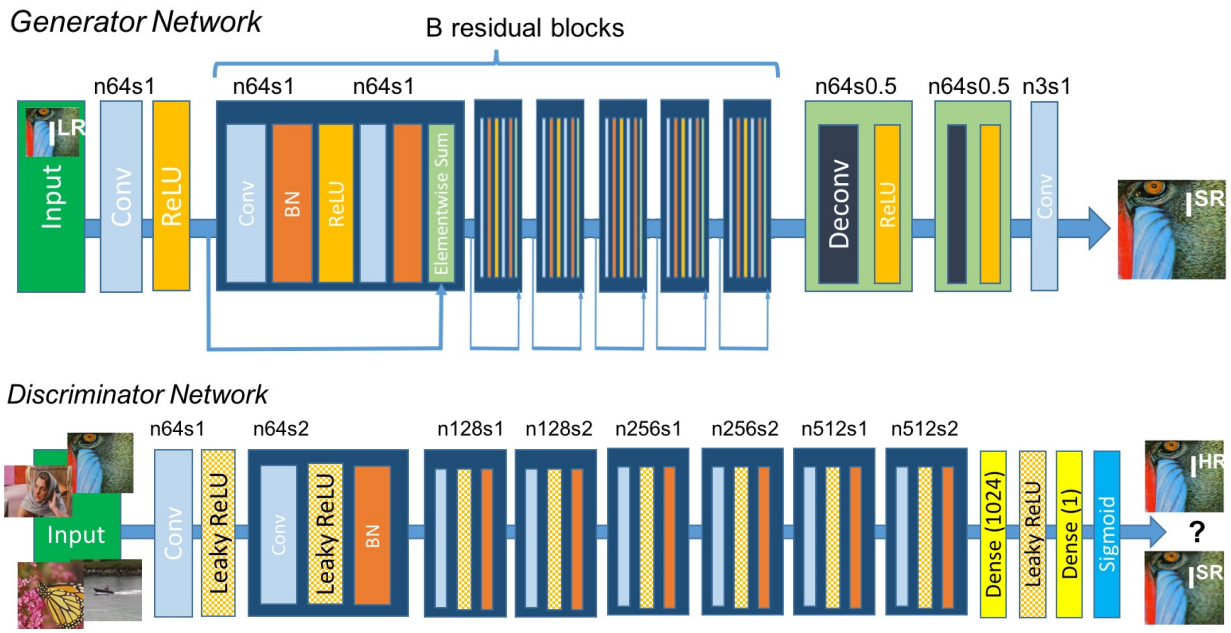

Figure 4: Architecture of Generator and Discriminator Network with corresponding number of feature maps (n) and stride (s) indicated for each convolutional layer.

I^{LR}. Here I^{LR} is the low-resolution version of its high-resolution counterpart I^{HR}. The high-resolution images are only available during training. In training, I^{LR} is obtained by applying a Gaussian-filter to I^{HR} followed by a downsampling operation with downsampling factor r. For an image with C color channels, we describe I^{LR} by a real-valued tensor of size $W \times H \times C$ and I^{HR}, I^{SR} by $rW \times rH \times C$ respectively.

Our ultimate goal is to train a generating function G that estimates for a given LR input image its corresponding HR counterpart. To achieve this, we train a generator network as a feed-forward CNN G_{θ_G} parametrized by θ_G. Here $\theta_G = \{W_{1:L}; b_{1:L}\}$ denotes the weights and biases of a L-layer deep network and is obtained by optimizing a SR-specific loss function l^{SR}. For given training images I_n^{HR}, $n = 1, \ldots, N$ with corresponding I_n^{LR}, $n = 1, \ldots, N$, we solve:

$$\hat{\theta}_G = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l^{SR}(G_{\theta_G}(I_n^{LR}), I_n^{HR}) \qquad (1)$$

In this work we will specifically design a perceptual loss l^{SR} as a weighted combination of several loss components that model distinct desirable characteristics of the recovered SR image. The individual loss functions are described in more detail in Section 2.2.

2.1. Adversarial Network Architecture

Following Goodfellow et al. [20] we further define a discriminator network D_{θ_D} which we optimize in an alternating manner along with G_{θ_G} to solve the adversarial min-max problem:

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{I^{HR} \sim p_{\text{train}}(I^{HR})}[\log D_{\theta_D}(I^{HR})] + \mathbb{E}_{I^{LR} \sim p_G(I^{LR})}[\log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR})))] \qquad (2)$$

The general idea behind this formulation is that it allows one to train a generative model G with the goal of fooling a differentiable discriminator D that is trained to distinguish super-resolved images from real images. With this approach our generator can learn to create solutions that are highly similar to real images and thus difficult to classify by D. Eventually this encourages perceptually superior solutions residing in the subspace, the manifold, of natural images. This is in contrast to SR solutions obtained by minimizing pixel-wise error measurements, such as the MSE.

At the core of our very deep generator network G, which is illustrated in Figure 4 are B residual blocks with identical layout. Inspired by Johnson et al. [28] we employ the block layout proposed by Gross and Wilber [22]. Specifically, we use two convolutional layers with small 3×3 kernels and 64 feature maps followed by batch-normalization layers [27]

and ReLU as the activation function. We increase the resolution of the input image with two trained deconvolution layers (stride=0.5) as proposed by Shi et al. [42].

To discriminate real HR images from generated SR samples we train a discriminator network. The general architecture is illustrated in Figure 4. Here we follow the architectural guidelines summarized by Radford et al. [39] and use LeakyReLU activation and avoid max-pooling throughout the network. The discriminator network is trained to solve the maximization problem in Equation 2. It contains eight convolutional layers with an increasing number of filter kernels, increasing by a factor of 2 from 64 to 512 kernels as in the VGG network [43]. Strided convolutions are used to reduce the image resolution each time the number of features is doubled. The resulting 512 feature maps are followed by two dense layers and a final sigmoid activation function to obtain a probability for sample classification.

2.2. Perceptual Loss Function

The definition of our perceptual loss function l^{SR} is critical for the performance of our generator network and thus SR algorithm. While l^{SR} is commonly modeled based on the MSE [9, 42], we improve on Johnson et al. [28] and Bruna et al. [5] and design a loss function that can assess the quality of a solution with respect to perceptually relevant characteristics.

Given weighting parameters γ_i, $i = 1, \ldots, K$, we define $l^{SR} = \sum_{i=1}^{K} \gamma_i l_i^{SR}$ as weighted sum of individual loss functions. Specifically, our perceptual loss consists of a content loss, an adversarial loss and a regularization loss component that we specify in the following.

2.2.1 Content Loss

The pixel-wise **MSE loss** is calculated as:

$$l_{MSE}^{SR} = \frac{1}{r^2 WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} (I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y})^2 \quad (3)$$

This is the most widely used optimization target for image SR on which many state of the art approaches rely [9, 42]. However, while achieving particularly high PSNR, solutions of MSE optimization problems often lack high-frequency content which results in perceptually unsatisfying, overly smooth solutions (cf. Figure 2).

Instead of relying on pixel-wise losses we build on the ideas of Gatys et al. [17], Bruna et al. [5] and Johnson et al. [28] and use a loss function that is closer to perceptual similarity. We define the **VGG loss** based on the ReLU activation layers of the pre-trained 19 layer VGG network described in Simonyan and Zisserman [43].

With $\phi_{i,j}$ we indicate the feature map obtained by the j-th convolution before the i-th maxpooling layer within the VGG19 network, which we consider given. We then define the VGG loss as the euclidean distance between the feature representations of a reconstructed image $G_{\theta_G}(I^{LR})$ and the reference image I^{HR}:

$$l_{VGG/i.j}^{SR} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y})^2 \quad (4)$$

Here $W_{i,j}$ and $H_{i,j}$ describe the dimensions of the respective feature maps within the VGG network.

2.2.2 Adversarial Loss

In addition to the content losses described so far, we also add the generative component of our GAN to the perceptual loss. This encourages our network to favor solutions that reside on the manifold of natural images, by trying to fool the discriminator network. The generative loss l_{Gen}^{SR} is defined based on the probabilities of the discriminator $D_{\theta_D}(G_{\theta_G}(I^{LR}))$ over all training samples as:

$$l_{Gen}^{SR} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{LR})) \quad (5)$$

Here, $D_{\theta_D}(G_{\theta_G}(I^{LR}))$ is the estimated probability that the reconstructed image $G_{\theta_G}(I^{LR})$ is a natural HR image. Note that for better gradient behaviour we minimize $-\log D_{\theta_D}(G_{\theta_G}(I^{LR}))$ instead of $\log[1-D_{\theta_D}(G_{\theta_G}(I^{LR}))]$ [20].

2.2.3 Regularization Loss

We further employ a regularizer based on the total variation to encourage spatially coherent solutions [2, 28]. The regularization loss, l_{TV}, is calculated as:

$$l_{TV}^{SR} = \frac{1}{r^2 WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} ||\nabla G_{\theta_G}(I^{LR})_{x,y}|| \quad (6)$$

3. Experiments

3.1. Data and Similarity Measures

We perform experiments on the three widely used benchmark datasets **Set5** [3], **Set14** [61] and **BSD100** [36]. All experiments are performed with a downsampling factor of $4\times$. For fair quantitative comparison, all reported PSNR [dB] and SSIM [51] measures were calculated on the y-channel of center-cropped, removed 4 pixels on each border,

images using the daala package[1]. Super-resolved images for the reference methods bicubic, SRCNN [8] and SelfExSR [26] were obtained from Huang et al.[2] [26] and for DRCN from Kim et al.[3] [29]. Results obtained with **SRResNet** (for losses: l^{SR}_{MSE} and $l^{SR}_{VGG/2.2}$) and the **SRGAN** variants are available online for all three datasets[4].

Reader might also be interested in an independently developed GAN-based solution on GitHub[5]. However it only provides experimental results on a limited set of faces, which is a more constrained, easier task.

3.2. Training Details and Parameters

We trained all networks on a NVIDIA Tesla M40 GPU using a random sample from the **ImageNet** database [40]. These images are distinct from the **Set5**, **Set14** and **BSD100** testing images. We obtained the LR images by downsampling the HR images using bicubic kernel with downsampling factor $r = 4$. For each mini-batch we crop 16 random 96×96 sub images of distinct training images. Note that we can apply the generator model to images of arbitrary size as it is fully convolutional. For optimization we use Adam [31] with $\beta_1 = 0.9$. For the **SRResNet** network evaluated in Section 3.3 and the pre-training of the VGG2.2 network we used 50 thousand images of ImageNet and a learning rate of 10^{-4} with 10^6 update iterations. We pre-trained our generative model using $l^{SR} = l^{SR}_{VGG/2.2}$ for $\phi_{2,2}$ to provide an initialization when training the actual GAN to avoid undesired local optima. We found that the model pre-trained with VGG2.2 loss worked better than pre-trained with MSE. We used 350 thousand instead of 50 thousand images of ImageNet to train the adversarial **SRGAN** networks for about one day because we found them more difficult to train as compared to **SRResNet**. We alternate updates to the generator and discriminator network, which is equivalent to $k = 1$ as used in Goodfellow et al. [20]. Our generator network has 15 identical ($B = 15$) residual blocks. During test time we turn batch-normalization off to obtain a network output that deterministically depends only on the input [27].

3.3. Performance of MSE-based Network

In this paper, we choose a network architecture for the generator (cf. Figure 4) which combines the effectiveness of the efficient sub-pixel convolutional neural network (ESPCN) [42] and the high performance of the ResNet [24]. We first evaluate the performance of the generator network for $l^{SR} = l^{SR}_{MSE}$ without adversarial component. We refer to this MSE-based SR network as **SRResNet**.

Table 1: Comparison of methods: bicubic, SRCNN [8], SelfExSR [26], DRCN [29], ESPCN [42] on Set5, Set14, BSD100 benchmark data. Highest calculated measures (PSNR [dB], SSIM) in bold.

Set5	bicubic	SRCNN	SelfExSR	DRCN	ESPCN	SRResNet
PSNR	28.43	30.07	30.33	31.52	30.76	**31.92**
SSIM	0.8211	0.8627	0.872	0.8938	0.8784	**0.8998**
Set14						
PSNR	25.99	27.18	27.45	28.02	27.66	**28.39**
SSIM	0.7486	0.7861	0.7972	0.8074	0.8004	**0.8166**
BSD100						
PSNR	25.94	26.68	26.83	27.21	27.02	**27.52**
SSIM	0.6935	0.7291	0.7387	0.7493	0.7442	**0.7603**

We compare the performance of **SRResNet** to bicubic interpolation and four state of the art frameworks: the super-resolution CNN (SRCNN) described by Dong et al. [8], a method based on transformed self-exemplars (SelfExSR) [26], a deeply-recursive convolutional network (DRCN) described by Kim et al. [29] and the efficient sub-pixel CNN (ESPCNN) allowing real-time video SR [42]. Quantitative results are summarized in Table 1 and confirm that **SRResNet** sets a new state of the art on the three benchmark datasets. Please note that we used a publicly available framework for evaluation (cf. Section 3.1), reported values might thus slightly deviate from those reported in the original papers.

3.4. Investigation of Content Loss

We further investigate the effect of different content loss choices in the perceptual loss for the GAN-based networks to which we refer to as **SRGAN**. Specifically we investigate the following losses:

$$l^{SR} = \underbrace{\underbrace{l^{SR}_{\mathrm{X}}}_{\text{content loss}} + \underbrace{10^{-3} l^{SR}_{Gen}}_{\text{adversarial loss}}}_{\text{perceptual loss (for VGG based content losses)}} + \underbrace{2 \cdot 10^{-8} l^{SR}_{TV}}_{\text{regularization loss}} \tag{7}$$

Here we set l^{SR}_{X} to either:

- **SRGAN-MSE**: l^{SR}_{MSE}, to investigate the adversarial network with the standard MSE as content loss.
- **SRGAN-VGG22**: $l^{SR}_{VGG/2.2}$ with $\phi_{2,2}$, a loss defined on feature maps that represent lower-level features [60].
- **SRGAN-VGG54**: $l^{SR}_{VGG/5.4}$ with $\phi_{5,4}$, a loss defined on feature maps of higher level features from deeper network layers with more potential to focus on the content of the images [60, 58, 35]. We refer to this network as **SRGAN** in the following.

[1] https://github.com/xiph/daala (commit: 8d03668)
[2] https://github.com/jbhuang0604/SelfExSR
[3] http://cv.snu.ac.kr/research/DRCN/
[4] https://twitter.box.com/s/2divbopdbeaxru7kqs4pame56mupl02m
[5] https://github.com/david-gpu/srez

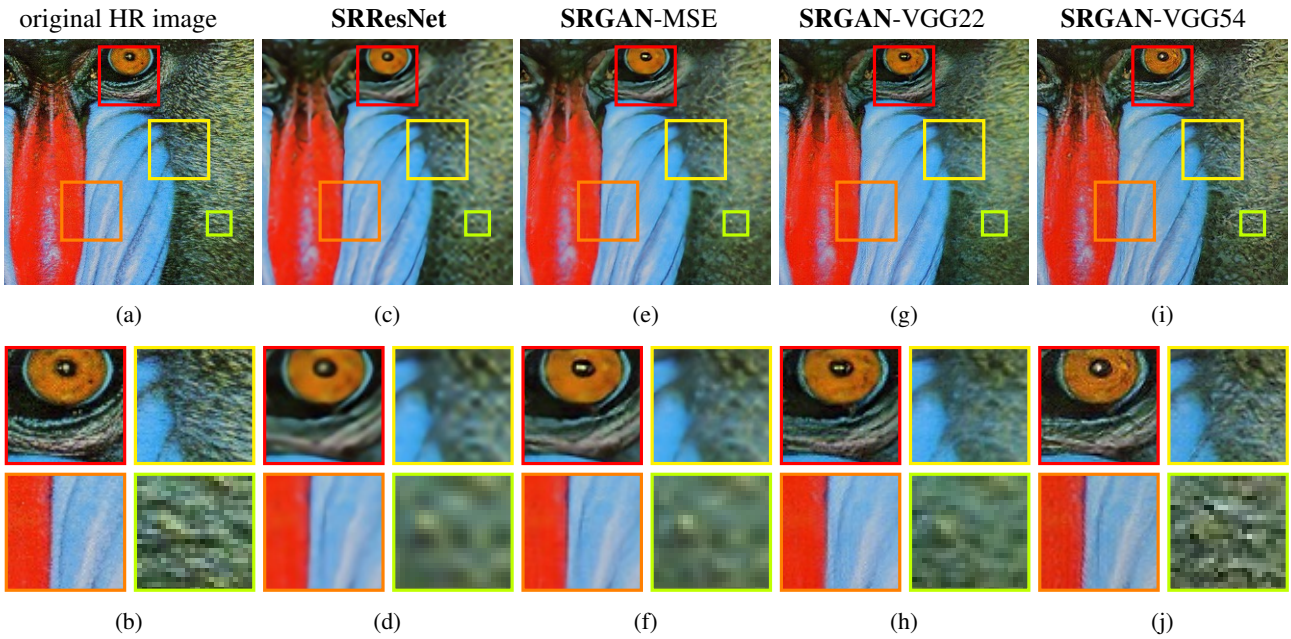

Figure 5: Reference HR image (left: a,b) with corresponding SRResNet (middle left: c,d), SRGAN-MSE (middle: e,f), SRGAN-VGG2.2 (middle right: g,h) and SRGAN-VGG54 (right: i,j) reconstruction results.

Table 2: Quantified performance (PSNR [dB], SSIM) of different loss functions within the adversarial network on Set5, Set14 and BSD100 benchmark data.

Set5	SRGAN-MSE	SRGAN-VGG22	SRGAN-VGG54
PSNR	30.36	29.88	28.74
SSIM	0.8727	0.8524	0.8435
Set14			
PSNR	27.02	26.48	25.75
SSIM	0.7817	0.7513	0.7376
BSD100			
PSNR	26.51	25.69	24.65
SSIM	0.7237	0.6882	0.6502

Quantitative results are summarized in Table 2 and visual examples provided in Figure 5. Even combined with the adversarial loss, MSE provides solutions with the highest PSNR values that are, however, perceptually rather smooth and less convincing than results achieved with a loss component more sensitive to visual perception. This is caused by competition between the MSE-based content loss and the adversarial loss. In general, the "further away" the content loss is from pixel space the perceptually better the result. Thus, we observed a better texture detail using the higher level VGG feature maps $\phi_{5,4}$ as compared to $\phi_{2,2}$. Further examples of perceptual improvements through **SRGAN-VGG54** over **SRResNet** are visualized in Figure 6 and provided in the Appendix.

4. Discussion and Future Work

The presented experiments suggest superior perceptual performance of the proposed framework purely based on visual comparison. Standard quantitative measures such as PSNR and SSIM clearly fail to capture and accurately assess image quality with respect to the human visual system [49]. As PSNR is not sufficiently capable of quantifying the perceptual quality of the SR result, future work will include quantified user experiences by collecting subjective measures such as mean opinion scores (MOS). The focus of this work was the perceptual quality of super-resolved images rather than computational efficiency. The presented model is, in contrast to Shi et al. [42], not optimized for video SR in real-time. However, preliminary experiments on the network architecture suggest that shallower networks have the potential to provide very efficient alternatives at a small reduction of qualitative performance. A deeper investigation of the network architecture as well as the perceptual loss will be part of future work. In contrast to Dong et al. [9], we found deeper network architectures to be beneficial. We speculate that the ResNet design has a substantial impact on the performance of deeper networks and could potentially explain the different findings.

Of particular importance when aiming for photo-realistic solutions to the SR problem is the choice of the content loss as illustrated in Figure 5. In this work, we found $l^{SR}_{VGG/5.4}$ to yield the perceptually most convincing results, which we attribute to the potential of deeper network layers to

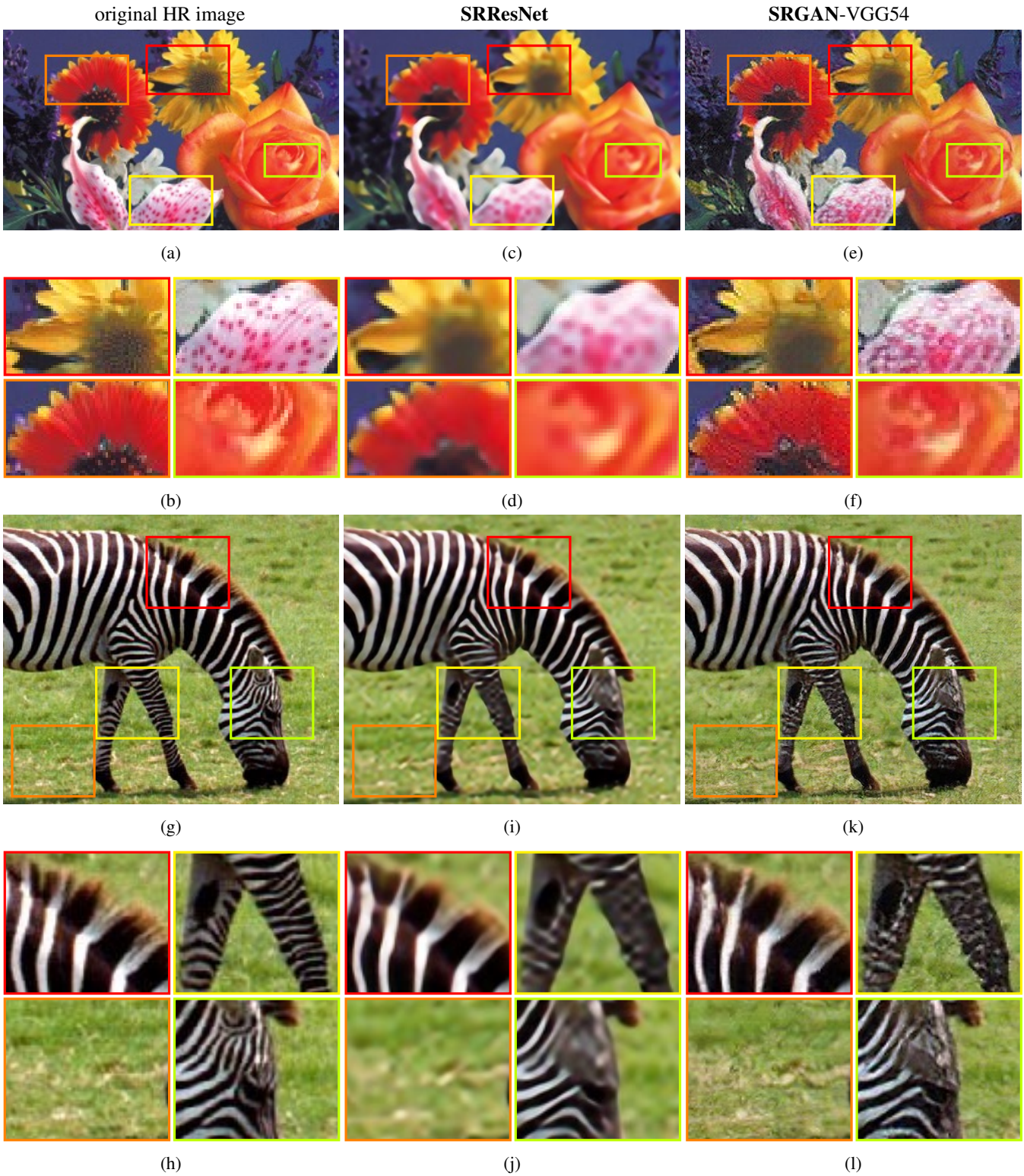

Figure 6: Reference HR images (left) with corresponding SRResNet (middle) and SRGAN-VGG54 (right) reconstruction results.

represent features of higher abstraction [60, 58, 35] away from pixel space. We speculate that feature maps of these deeper layers focus purely on the content while leaving the adversarial loss focusing on texture details which are the main difference between the super-resolved images without the adversarial loss and photo-realistic images. The development of content loss functions that describe image spatial content, but more invariant to changes in pixel space will further improve photo-realistic image SR results.

5. Conclusion

We have described a deep residual network **SRResNet** that sets a new state of the art on public benchmark datasets when evaluated with the widely used PSNR measure. We then further introduced **SRGAN** that yields photo-realistic super-resolved images from large downsampling factors ($4\times$) that are less distinguishable from their originals. We have highlighted some limitations of PSNR-focused image super-resolution and thus proposed a new perceptual loss for SR, which augments the content loss function with an adversarial loss by training a GAN.

References

[1] J. Allebach and P. W. Wong. Edge-directed interpolation. In *Proceedings of International Conference on Image Processing*, volume 3, pages 707–710, 1996. 2

[2] H. A. Aly and E. Dubois. Image up-sampling using total-variation regularization with a new observation model. *IEEE Transactions on Image Processing*, 14(10):1647–1659, 2005. 6

[3] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. *BMVC*, 2012. 6

[4] S. Borman and R. L. Stevenson. Super-Resolution from Image Sequences - A Review. *Midwest Symposium on Circuits and Systems*, pages 374–378, 1998. 2

[5] J. Bruna, P. Sprechmann, and Y. LeCun. Super-resolution with deep convolutional sufficient statistics. In *International Conference on Learning Representations*, 2016. 2, 3, 4, 6

[6] D. Dai, R. Timofte, and L. Van Gool. Jointly optimized regressors for image super-resolution. In *Computer Graphics Forum*, volume 34, pages 95–104, 2015. 2

[7] E. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1486–1494, 2015. 3

[8] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 184–199. Springer, 2014. 2, 4, 6, 7

[9] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. 2, 3, 6, 8

[10] W. Dong, L. Zhang, G. Shi, and X. Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011. 2

[11] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. *arXiv preprint arXiv:1602.02644*, 2016. 3, 4

[12] C. E. Duchon. Lanczos Filtering in One and Two Dimensions. In *Journal of Applied Meteorology*, volume 18, pages 1016–1022. 1979. 2

[13] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar. Fast and robust multiframe super resolution. *IEEE Transactions on Image Processing*, 13(10):1327–1344, 2004. 2

[14] J. A. Ferwerda. Three varieties of realism in computer graphics. In *Electronic Imaging*, pages 290–297. International Society for Optics and Photonics, 2003. 2

[15] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65, 2002. 2

[16] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1):25–47, 2000. 2

[17] L. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 262–270, 2015. 4, 6

[18] L. A. Gatys, A. S. Ecker, and M. Bethge. A Neural Algorithm of Artistic Style. *arXiv preprint arXiv:1508.06576*, 2015. 4

[19] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *IEEE International Conference on Computer Vision (ICCV)*, pages 349–356, 2009. 2

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 3, 5, 6, 7

[21] K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 399–406, 2010. 2

[22] S. Gross and M. Wilber. Training and investigating residual nets, online at http://torch.ch/blog/2016/02/04/resnets.html. 2016. 5

[23] P. Gupta, P. Srivastava, S. Bhardwaj, and V. Bhateja. A modified psnr metric based on hvs for quality assessment of color images. In *International Conference on Communication and Industrial Application (ICCIA)*, pages 1–4. IEEE, 2011. 2

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 3, 4, 7

[25] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027*, 2016. 3, 4

[26] J. B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015. 2, 4, 7

[27] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:1502.03167*, 2015. 3, 4, 5, 7

[28] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super- resolution. *arXiv preprint arXiv:1603.08155*, 2016. 2, 3, 4, 5, 6

[29] J. Kim, J. K. Lee, and K. M. Lee. Deeply-recursive convolutional network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 4, 7

[30] K. I. Kim and Y. Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1127–1133, 2010. 2

[31] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012. 3

[33] C. Li and M. Wand. Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis. *arXiv preprint arXiv:1601.04589*, 2016. 4

[34] X. Li and M. T. Orchard. New edge-directed interpolation. *IEEE Transactions on Image Processing*, 10(10):1521–1527, 2001. 2

[35] A. Mahendran and A. Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, pages 1–23, 2016. 7, 10

[36] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 416–423, 2001. 6

[37] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 3

[38] K. Nasrollahi and T. B. Moeslund. Super-resolution: A comprehensive survey. In *Machine Vision and Applications*, volume 25, pages 1423–1468. 2014. 1, 2

[39] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 4, 6

[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 1–42, 2014. 7

[41] S. Schulter, C. Leistner, and H. Bischof. Fast and accurate image upscaling with super-resolution forests. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3791–3799, 2015. 2

[42] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016. 2, 3, 4, 5, 6, 7, 8

[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 3, 4, 6

[44] J. Sun, J. Sun, Z. Xu, and H.-Y. Shum. Image super-resolution using gradient profile prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 2

[45] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 3

[46] Y.-W. Tai, S. Liu, M. S. Brown, and S. Lin. Super Resolution using Edge Prior and Single Image Detail Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2400–2407, 2010. 2

[47] R. Timofte, V. De, and L. Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1920–1927, 2013. 2

[48] R. Timofte, V. De Smet, and L. Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision (ACCV)*, pages 111–126. Springer, 2014. 2

[49] G. Toderici, D. Vincent, N. Johnston, S. J. Hwang, D. Minnen, J. Shor, and M. Covell. Full Resolution Image Compression with Recurrent Neural Networks. *arXiv preprint arXiv:1608.05148*, 2016. 8

[50] Y. Wang, L. Wang, H. Wang, and P. Li. End-to-End Image Super-Resolution via Deep and Shallow Convolutional Networks. *arXiv preprint arXiv:1607.07680*, 2016. 3

[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 2, 6

[52] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang. Deep networks for image super-resolution with sparse prior. In *IEEE International Conference on Computer Vision (ICCV)*, pages 370–378, 2015. 2

[53] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang. Deeply improved sparse coding for image super-resolution. *arXiv preprint arXiv:1507.08905*, 2015. 3

[54] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multi-scale structural similarity for image quality assessment. In *IEEE Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 9–13, 2003. 2

[55] C.-Y. Yang, C. Ma, and M.-H. Yang. Single-image super-resolution: A benchmark. In *European Conference on Computer Vision (ECCV)*, pages 372–386. Springer, 2014. 1, 2

[56] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution as sparse representation of raw image patches. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 2

[57] R. Yeh, C. Chen, T. Y. Lim, M. Hasegawa-Johnson, and M. N. Do. Semantic Image Inpainting with Perceptual and Contextual Losses. *arXiv preprint arXiv:1607.07539*, 2016. 4

[58] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding Neural Networks Through Deep Visualization. In *International Conference on Machine Learning - Deep Learning Workshop 2015*, page 12, 2015. 7, 10

[59] H. Yue, X. Sun, J. Yang, and F. Wu. Landmark image super-resolution by retrieving web images. *IEEE Transactions on Image Processing*, 22(12):4865–4878, 2013. 2

[60] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014. 7, 10

[61] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, pages 711–730. Springer, 2012. 2, 6

[62] K. Zhang, X. Gao, D. Tao, and X. Li. Multi-scale dictionary for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1114–1121. IEEE, 2012. 2

[63] W. Zou and P. C. Yuen. Very Low Resolution Face Recognition in Parallel Environment . *IEEE Transactions on Image Processing*, 21:327–340, 2012. 1

Appendix (The following images are best viewed and compared zoomed in.)

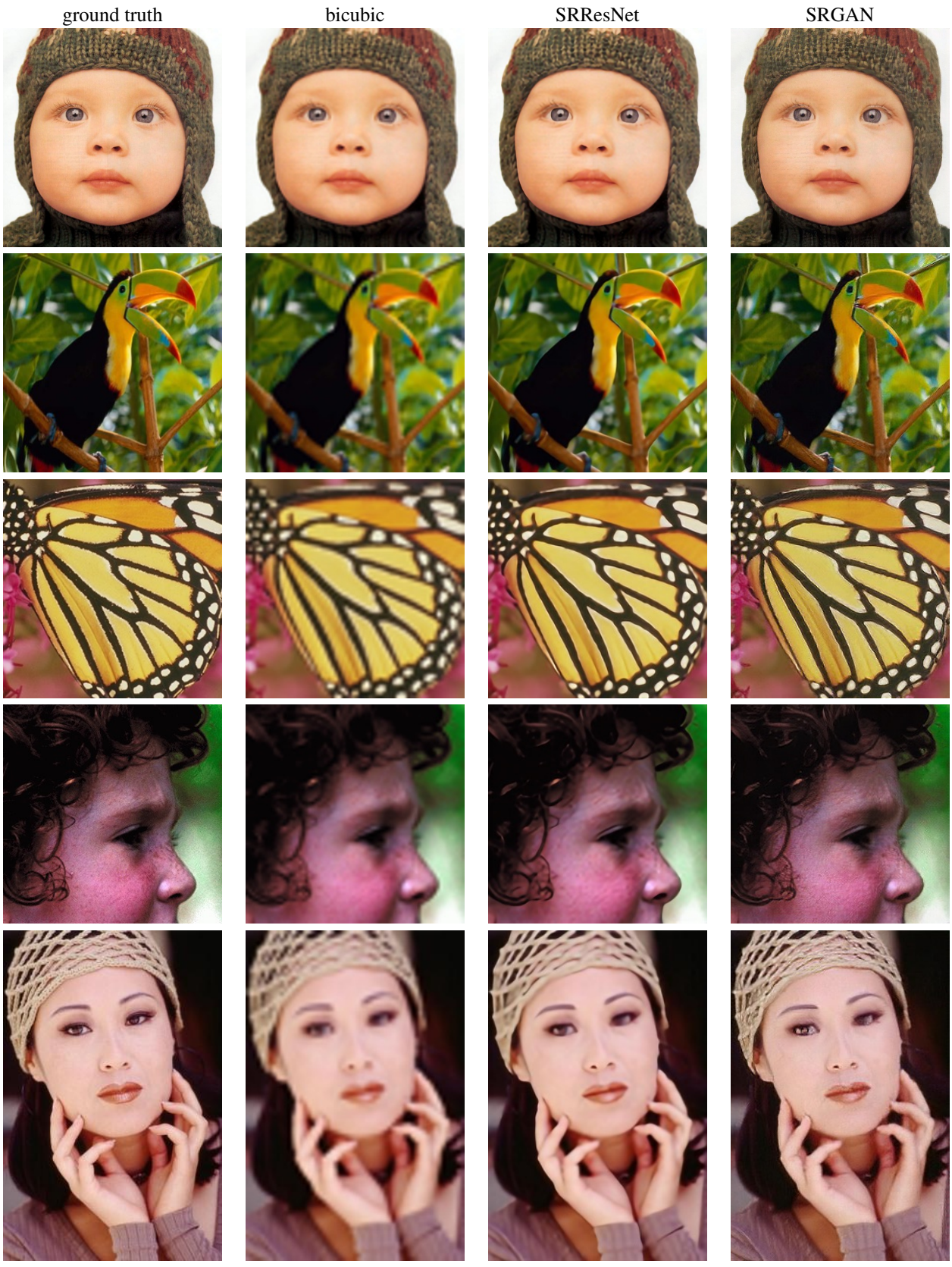

Figure 7: Results for **Set5** using bicubic interpolation (middle left), SRResNet (middle right) and SRGAN (right).

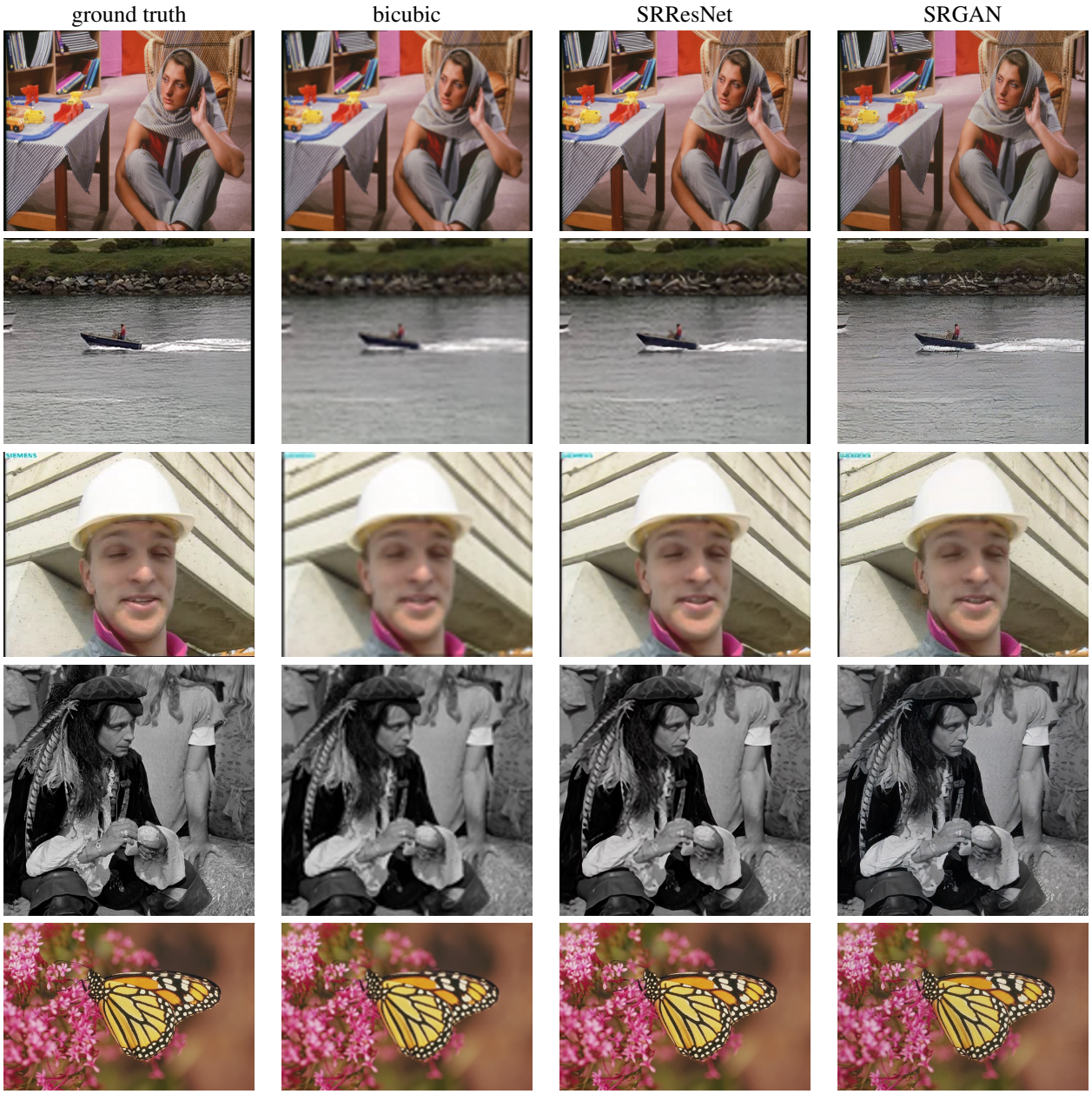

Figure 8: Results for five random samples (not shown in the paper) for **Set14** using bicubic interpolation (middle left), SRResNet (middle right) and SRGAN (right).

Figure 9: Results for five random samples of **BSD100** using bicubic interpolation (middle left), SRResNet (middle right) and SRGAN (right).