Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network论文翻译——中英文对照

Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network论文翻译——中英文对照

文章作者：Tyan

博客：noahsnail.com | CSDN | 简书

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

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? The behavior of optimization-based super-resolution methods is principally driven by the choice of the objective function. Recent work has largely focused on minimizing the mean squared reconstruction error. The resulting estimates have high peak signal-to-noise ratios, but they are often lacking high-frequency details and are perceptually unsatisfying in the sense that they fail to match the fidelity expected at the higher resolution. In this paper, we present SRGAN, a generative adversarial network (GAN) for image superresolution (SR). To our knowledge, it is the first framework capable of inferring photo-realistic natural images for 4× upscaling factors. 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 motivated by perceptual similarity instead of similarity in pixel space. Our deep residual network is able to recover photo-realistic

textures from heavily downsampled images on public benchmarks. An extensive mean-opinion-score (MOS) test shows hugely significant gains in perceptual quality using SRGAN. The MOS scores obtained with SRGAN are closer to those of the original high-resolution images than to those obtained with any state-of-the-art method.

摘要

尽管使用更快更深的卷积神经网络在单图像超分辨率的准确性和速度方面取得了突破，但仍有一个主要问题尚未解决：当使用大的上采样系数进行超分辨率时，我们怎样来恢复更精细的纹理细节。基于优化的超分辨率方法的行为主要由目标函数的选择来决定。最近的工作主要专注于最小化均方重构误差。由此得出的评估结果具有很高的峰值信噪比，但它们通常缺乏高频细节，并且在感知上是不令人满意的，在某种意义上，它们在较高分辨率上没有满足期望的保真度。在本文中，我们提出了SRGAN，一种用于图像超分辨率(SR)的生成对抗网络(GAN)。据我们所知，这是第一个对于4倍上采样系数，能推断逼真自然图像的框架。为此，我们提出了一种感知损失函数，其由对抗损失和内容损失组成。对抗损失使用判别器网络将我们的解推向自然图像流形，判别器网络经过训练用以区分超分辨率图像和原始的逼真图像。此外，我们使用由感知相似性而不是像素空间相似性引起的内容损失。在公开的基准数据集上，我们的深度残差网络能从过度下采样图像中恢复出逼真的纹理。广泛的平均主观得分(MOS)测试显示，使用SRGAN可以显著提高感知质量。与任何最新方法获得的MOS得分相比，使用SRGAN获得的MOS得分更接近于原始高分辨率图像的MOS得分。

1. Introduction

The highly challenging task of estimating a highresolution (HR) image from its low-resolution (LR) counterpart is referred to as super-resolution (SR). SR received substantial attention from within the computer vision research community and has a wide range of applications [62, 70, 42].

1. 引言

从低分辨率(LR)图像来估计其对应高分辨率(HR)图像的高挑战性任务被称作超分辨率(SR)。SR在计算机视觉研究领域受到了广泛的关注并有大量应用[62, 70, 42]。

The ill-posed nature of the underdetermined SR problem is particularly pronounced for high upscaling factors, for which texture detail in the reconstructed SR images is typically absent. The optimization target of supervised SR algorithms is commonly 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 [60]. 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 [59, 57, 25]. This is illustrated in Figure 2, where highest PSNR does not necessarily reflect the perceptually better SR result. The perceptual difference

between the super-resolved and original image means that the recovered image is not photo-realistic as defined by Ferwerda [15].

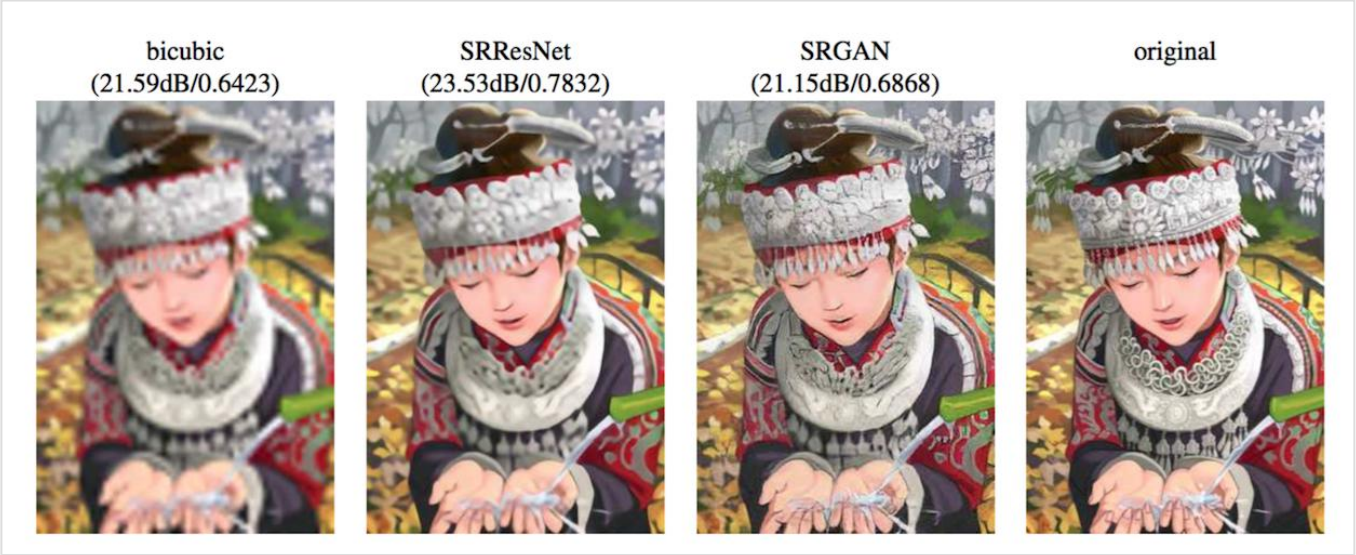

Figure 2: From left to right: bicubic interpolation, deep residual network optimized for MSE, deep residual generative adversarial network optimized for a loss more sensitive to human perception, original HR image. Corresponding PSNR and SSIM are shown in brackets. [4× upscaling]

欠定SR问题的不适定特性对于大的上采样系数尤其显著，重建的SR图像中通常缺少纹理细节。有监督SR算法的优化目标通常是最小化恢复的HR图像和真实图像之间的均方误差(MSE)。最小化MSE即最大化峰值信噪比(PSNR)是方便的，这是用来评估和比较SR算法的常用方法[60]。然而，MSE(和PSNR)捕获感知相对差异(例如高级纹理细节)的能力是非常有限的，因为它们是基于像素级图像差异[59, 57, 25]定义的。这在图2中进行了说明，其中最高的PSNR不一定能反映出感知上更好的SR结果。超分辨率图像和原始图像之间的感知差异意味着恢复图像不如Ferwerda[15]中定义的逼真。

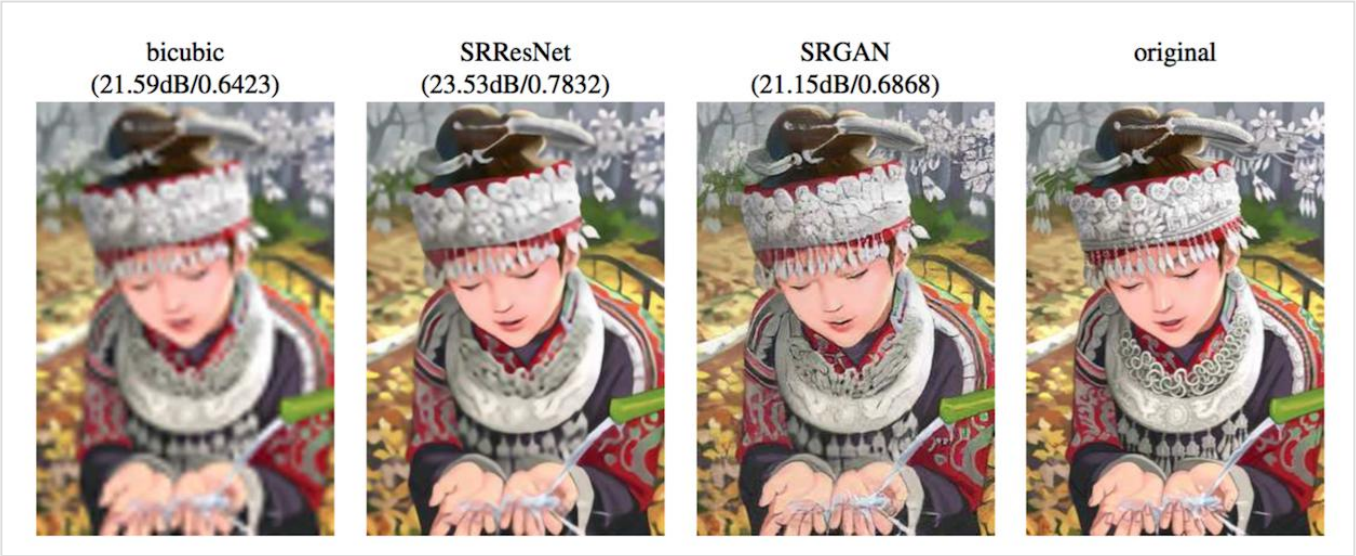

图2：从左到右：双三次插值，优化MSE的深度残差网络，优化人感知更敏感损失的深度残差生成对抗网络，原始HR图像。对应的PSNR和SSIM显示在括号中。[4倍上采样]

In this work we propose a super-resolution generative adversarial network (SRGAN) for which we employ a deep residual network (ResNet) with skip-connection 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 [48, 32, 4] combined with a discriminator that encourages solutions perceptually hard to distinguish from the HR reference images. An example photo-realistic image that was super-resolved with a 4× upscaling factor is shown in Figure 1.

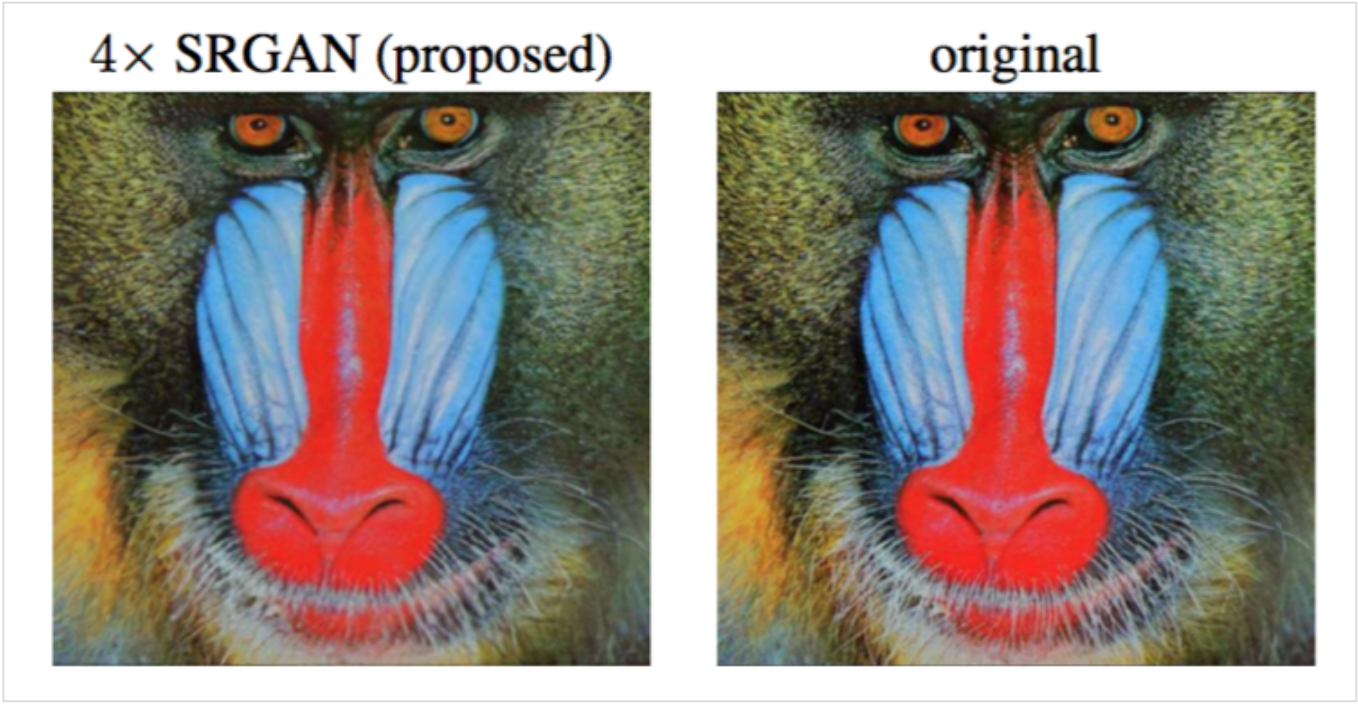

Figure 1: Super-resolved image (left) is almost indistinguishable from original (right). [4× upscaling]

在这项工作中我们提出了一种超分辨率生成对抗网络(SRGAN)，为此我们采用了具有跳跃连接的深度残差网络并舍弃了作为唯一优化目标的MSE。不同于以前的工作，我们定义了一种新的使用VGG网络[48, 32, 4]高级特征映射与判别器结合的感知损失，判别器会鼓励感知上更难与HR参考图像区分的解。图1中展示了一张示例逼真图像，其使用4倍上采样系数进行超分辨率。

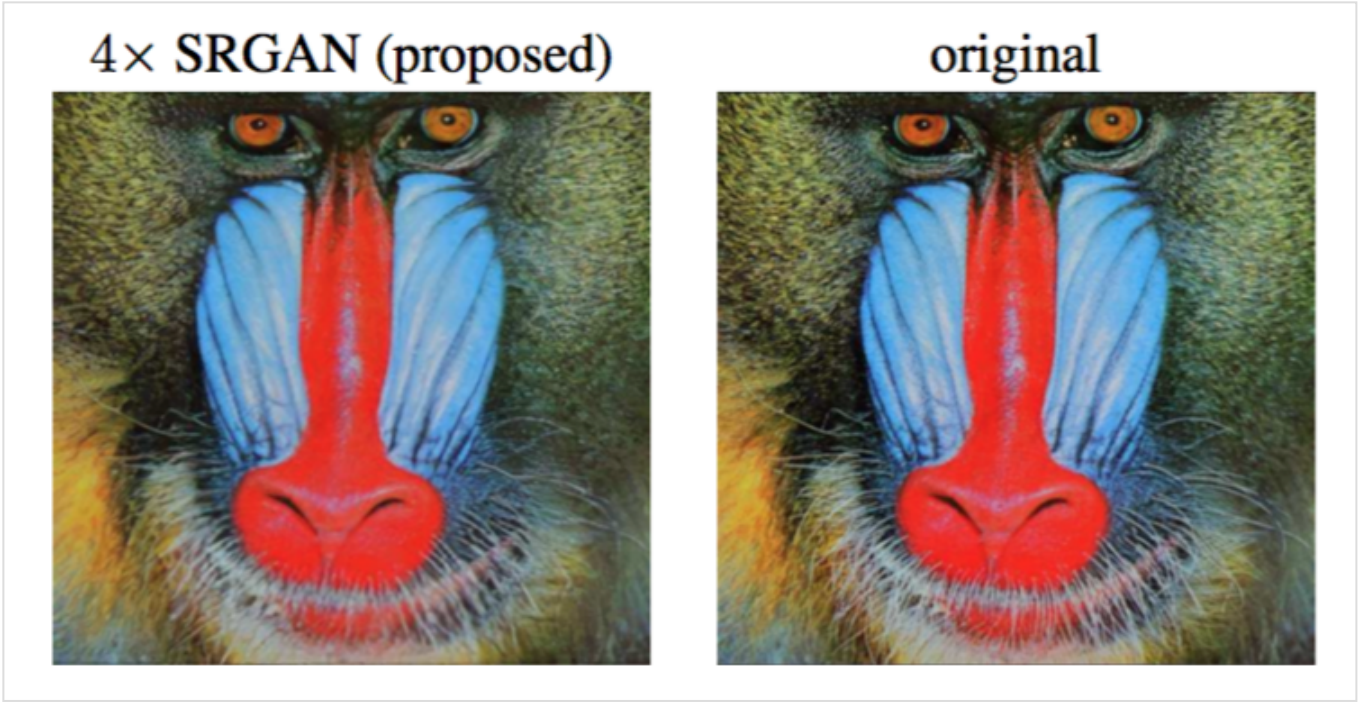

图1：超分辨率图像(左)是最难与原始图像(右)区分的. [4倍上采样]

1.1. Related work

1.1.1 Image super-resolution

Recent overview articles on image SR include Nasrollahi and Moeslund [42] or Yang et al. [60]. Here we will focus on single image super-resolution (SISR) and will not further discuss approaches that recover HR images from multiple images [3, 14].

1.1. 相关工作

1.1.1 图像超分辨率

最近的图像SR综述文章，包括Nasrollahi和Moeslund[42]或Yang等[60]。这里，我们将专注于单图像超分辨率(SISR)，不会进一步讨论从多张图像恢复HR图像的方法[3, 14]。

Prediction-based methods were among the first methods to tackle SISR. While these filtering approaches, e.g. linear, bicubic or Lanczos [13] filtering, can be very fast, they oversimplify the SISR problem and usually yield solutions with overly smooth textures. Methods that put particularly focus on edge-preservation have been proposed [1, 38].

基于预测的方法是解决SISR的首批方法之一。虽然这些滤波方法可能非常快，例如线性，双三次或Lanczos[13]滤波，但它们简化了SISR问题，通常会产生纹理过于平滑的解。特别关注边缘保留的方法已经被提出[1, 38]。

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. [17, 16]. Related approaches to the SR problem originate in compressed sensing [61, 11, 68]. In Glasner et al. [20] 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. [30], where self dictionaries are extended by further allowing for small transformations and shape variations. Gu et al. [24] proposed a convolutional sparse coding approach that improves consistency by processing the whole image rather than overlapping patches.

更强大的方法旨在在低分辨率图像和高分辨率图像之间建立一个复杂映射，并且通常依赖于训练数据。许多基于样本对的方法依赖于LR训练图像块，其对应的HR图像块是已知的。早期的工作由Freeman等[17, 16]提出。与SR相关的方法起源于压缩感知[61, 11, 68]。在Glasner等[20]中作者利用图像内跨尺度图像块冗余来推动SR。Huang等[30]也采用了这种自相似范式，通过进一步允许小的变换和形状变化扩展了自字典。Gu等[24]提出了一种卷积稀疏编码方法通过处理整张图像而不是重叠图像块提高了一致性。

To reconstruct realistic texture detail while avoiding edge artifacts, Tai et al. [51] combine an edge-directed SR algorithm based on a gradient profile prior [49] with the benefits of learning-based detail synthesis. Zhang et al. [69] propose a multi-scale dictionary to capture redundancies of similar image patches at different scales. To super-resolve landmark images, Yue et al. [66] retrieve correlating HR images with similar content from the web and propose a structure-aware matching criterion for alignment.

为了重建逼真的纹理细节同时避免边缘伪影，Tai等[51]将基于梯度轮廓先验[49]的边缘导向SR算法和基于学习的细节合成的优势相结合。张等[69]提出了一种多尺度字典来捕获不同尺度下相似图像块的冗余性。为了对地标图像进行超分辨率，Yue等[66]从网上采集了具有相似内容的相关HR图像，并提出了用于对齐的结构感知匹配标准。

Neighborhood embedding approaches upsample a LR image patch by finding similar LR training patches in a low dimensional manifold and combining their corresponding HR patches for reconstruction [53, 54]. In Kim and Kwon [34] the authors emphasize the tendency of neighborhood approaches to overfit and formulate a more general map of example pairs using kernel ridge regression. The regression problem can also be solved with Gaussian process regression [26], trees [45] or Random Forests [46]. In Dai et al. [5] a multitude of patch-specific regressors is learned and the most appropriate regressors selected during testing.

邻域嵌入方法通过在低维流形中查找相似的LR训练图像块并结合它们对应的用于重建的HR图像块对LR图像块进行上采样[53, 54]。在Kim和Kwon[34]中，作者强调了邻域方法过拟合的趋势，并使用核岭回归构建了样本对的更通用映射。回归问题也可以通过高斯过程回归[26]，树[45]或随机森林[46]来解决。戴等[5]学习了大量特定图像块的回归器，并在测试中选择最合适的回归器。

Recently convolutional neural network (CNN) based SR algorithms have shown excellent performance. In Wang et al. [58] the authors encode a sparse representation prior into their feed-forward network architecture based on the learned iterative shrinkage and thresholding algorithm (LISTA) [22]. Dong et al. [8, 9] 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 [10, 47, 56]. With their deeply-recursive convolutional network (DRCN), Kim et al. [33] presented a highly performant architecture that allows for long-range pixel dependencies while keeping the number of model parameters small. Of particular relevance for our paper are the works by Johnson et al. [32] and Bruna et al. [4], who rely on a loss function closer to perceptual similarity to recover visually more convincing HR images.

最近基于卷积神经网络(CNN)的SR算法已经展现出了出色的性能。在Wang等[58]中，作者基于学习的迭代收缩和阈值算法(LISTA)将稀疏表示先验编码到他们的前馈神经架构中[22]。Dong等[8, 9]使用双三次插值对输入图像进行上采样，并端到端地训练了一个三层的全卷积网络，取得了最佳的SR性能。之后的研究表明网络可以直接学习到上采样滤波器，并可以在准确性和速度方面进一步提高性能[10, 47, 56]。借助深度循环神经网络(DRCN)，Kim等[33]提出了一种高性能架构，在考虑长期像素依赖的同时保持了较少的模型参数数量。与本文特别相关的是约翰逊等[32]和Bruna等[4]的工作，其依赖于更接近于感知相似的损失函数来恢复视觉上更具说服力的HR图像。

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. [36].

1.1.2 卷积神经网络的设计

随着Krizhevsky等[36]工作取得成功的同时，专门设计的CNN架构设置了许多计算机视觉问题的最新技术。

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 [48, 50]. To efficiently train these deeper network architectures, batch-normalization [31] 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. [33] 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 [28] and skip-connections [29, 33]. 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.

研究表明，更深的网络架构更难训练，但具有大幅提高网络准确性的潜力，因为其允许建模非常复杂的映射[48, 50]。为了有效训练这些更深的网络架构，批归一化[31]通常用来抵消内部协变量转移。对于SISR，更深的网络架构已经表现出了性能提高，例如，Kim等[33]构建了一个循环CNN并介绍了最新的结果。缓解深度CNN训练的另一种强大设计选择是最近介绍的残差块[28]和跳跃连接[29, 33]概念。跳跃连接减轻了建模恒等映射的网络架构，本质上恒等映射是不重要的，然而对于卷积核表示而言，这可能是有意义的。

In the context of SISR it was also shown that learning upscaling filters is beneficial in terms of accuracy and speed [10, 47, 56]. This is an improvement over Dong et al. [9] where bicubic interpolation is employed to upscale the LR observation before feeding the image to the CNN.

SISR的背景下，研究表明学习上采样滤波器对于准确性和速度是有益的[10, 47, 56]。这是一种对Dong等[9]的改进，其中在将图片输入到CNN之前，采用双三次插值对LR观测进行上采样。

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 [41, 32, 12, 4]. Reconstructions of varying perceptual quality are exemplified with corresponding PSNR in Figure 2. We illustrate the problem of minimizing MSE in Figure 3 where multiple potential solutions with high texture details are averaged to create a smooth reconstruction.

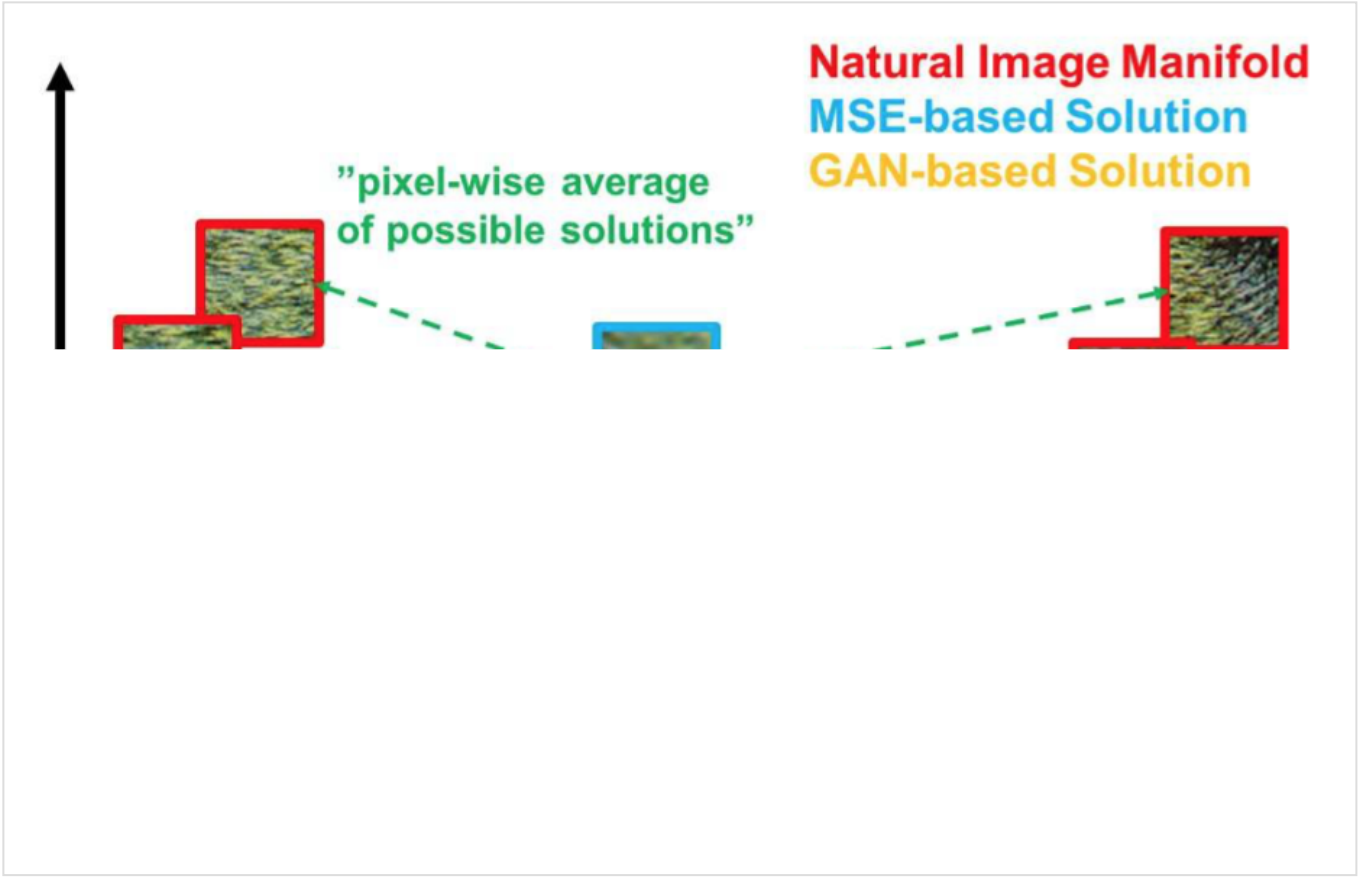

Figure 3: Illustration of patches from the natural image manifold (red) and super-resolved patches obtained with MSE (blue) and GAN (orange). The MSE-based solution appears overly smooth due to the pixel-wise average of possible solutions in the pixel space, while GAN drives the reconstruction towards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

1.1.3 损失函数

逐像素的损失函数(例如MSE)在努力处理恢复损失的高频细节(例如纹理)中的内在不确定性：最小化MSE鼓励寻找合理解的逐像素平均，这通常是过平滑的，因此会得到较差的感知质量[41, 32, 12, 4]。图2中以相应的PSNR为例说明了不同感知质量的重建。我们在图3中说明了最小化MSE的问题，其中对多个具有高级纹理细节的潜在解进行平均从而创建一个平滑的重建。

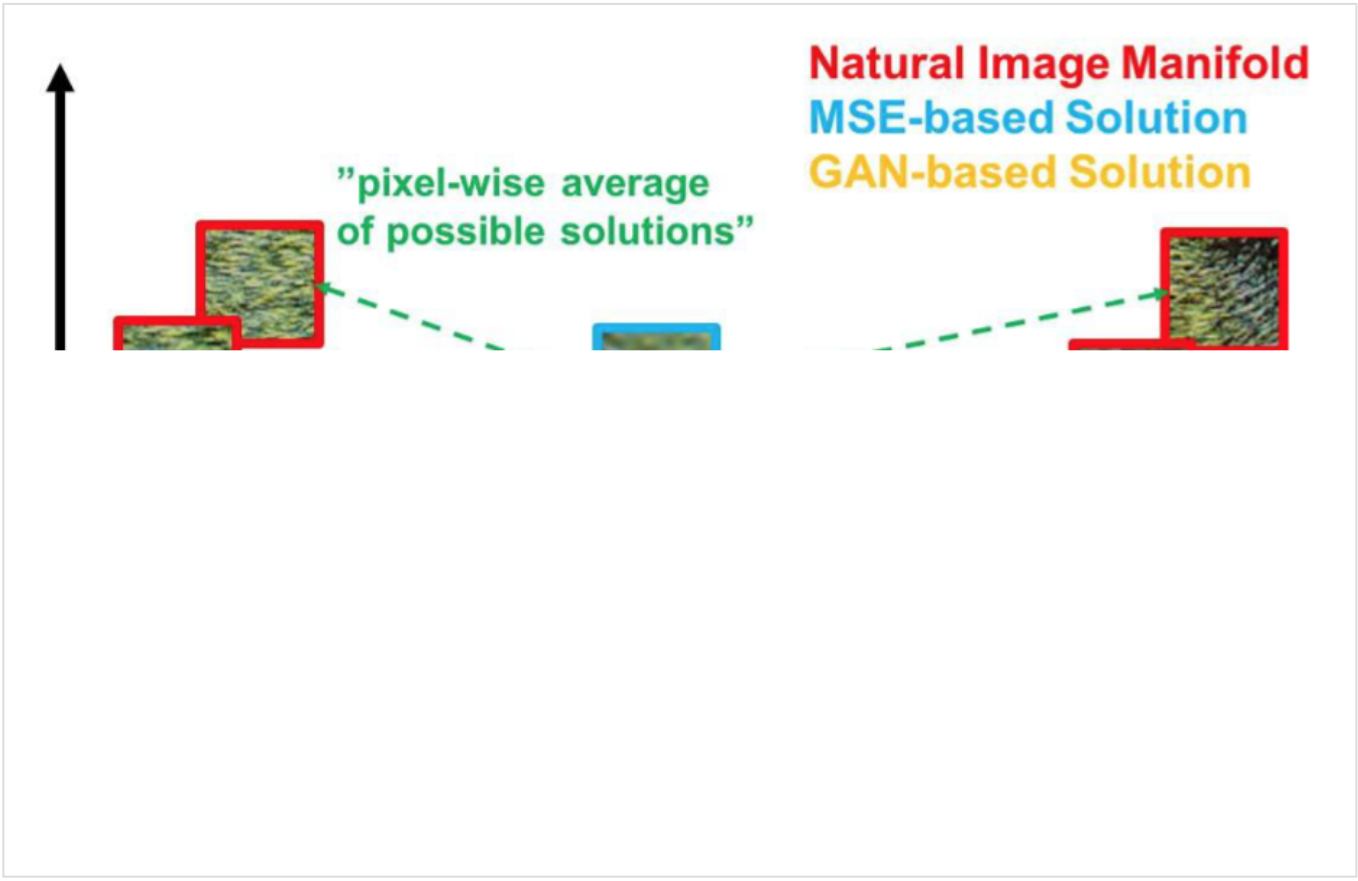

图3：自然图像流形图像块(红)，由MSE获得的超分辨率图像块(蓝)以及由GAN获得的超分辨率图像块(橙)。由于像素空间中可能解的逐像素平均，基于MSE的解似乎更平滑，而GAN将重建推向自然图像流形，产生了感知上更具说服力的解。

In Mathieu et al. [41] and Denton et al. [6] the authors tackled this problem by employing generative adversarial networks (GANs) [21] for the application of image generation. Yu and Porikli [65] augment pixel-wise MSE loss with a discriminator loss to train a network that super-resolves face images with large upscaling factors (8×). GANs were also used for unsupervised representation learning in Radford et al. [43]. The idea of using GANs to learn a mapping from one manifold to another is described by Li and Wand [37] for style transfer and Yeh et al.

[63] for inpainting. Bruna et al. [4] minimize the squared error in the feature spaces of VGG19 [48] and scattering networks.

在Mathieu等[41]和Denton等[6]中，作者通过采用图像生成应用生成对抗网络(GANs)来解决这个问题。Yu和Porikli[65]通过判别器损失增大了逐像素的MSE损失来训练网络，这个网络使用较大的上采样系数(8×)·对人脸图像进行超分辨率。在Radford等[43]中GAN也用来进行无监督表示学习。Li和Wand[37]的风格转换以及Yeh等[63]的图像修复都描述了使用GAN学习一个流形到另一个流形映射的想法。Bruna等[4]在VGG19[48]特征空间以及散射网络中都最小化了方差。

Dosovitskiy and Brox [12] 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. [32] and Bruna et al. [4] 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 [48] network. Perceptually more convincing results were obtained for both super-resolution and artistic style-transfer [18, 19]. Recently, Li and Wand [37] also investigated the effect of comparing and blending patches in pixel or VGG feature space.

Dosovitskiy和Brox使用基于神经网络特征空间中计算的欧式距离损失函数与对抗训练相结合。结果表明，提出的损失能够生成视觉上更好的图像并且可以用来解决解码非线性特征表示的不适定逆问题。与这个工作类似，Johnson等[32]和Bruna等[4]提出使用从预训练VGG网络中提取的特征来代替低级逐像素误差度量。具体来说，作者基于VGG19[48]网络提取的特征映射之间的欧式距离来构建损失函数。在超分辨率和艺术风格转换[18, 19]方面，都获得了感知上更具说服力的结果。最近，Li和Wand[37]还研究了在像素或VGG特征空间中对比和混合图像块的效果。

1.2. Contribution

GANs provide a powerful framework for generating plausible-looking natural images with high perceptual quality. The GAN procedure encourages the reconstructions 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.

1.2. 贡献

GAN提供了一种强大的框架，其可以生成看起来真实、具有高感知质量的自然图像。GAN过程鼓励重建朝向有很大可能包含逼真图像的搜索空间区域，因此更接近图3中所示的自然图像流形。

In this paper we describe the first very deep ResNet [28, 29] architecture using the concept of GANs to form a perceptual loss function for photo-realistic SISR. Our main contributions are:

• We set a new state of the art for image SR with high upscaling factors (4×) as measured by PSNR and structural similarity (SSIM) with our 16 blocks deep ResNet (SRResNet) optimized for MSE.

• We propose SRGAN which is a GAN-based network optimized for a new perceptual loss. Here we replace the MSE-based content loss with a loss calculated on feature maps of the VGG network [48], which are more invariant to changes in pixel space [37].

• We confirm with an extensive mean opinion score (MOS) test on images from three public benchmark datasets that SRGAN is the new state of the art, by a large margin, for the estimation of photo-realistic SR images with high upscaling factors (4×).

本文中我们描述了第一个很深的ResNet[28, 29]架构，使用GAN概念形成了逼真SISR的感知损失函数。我们的主要贡献如下：

• 我们在大的上采样系数下(4×)为图像SR设置了最新的技术水平，并用PSNR、结构相似性(SSIM)以及MSE进行了度量，使用了为MSE优化的16块深度ResNet(SRResNet)。

• 我们提出了SRGAN，一种为新感知损失优化的基于GAN的网络。这里我们将基于MSE的内容损失替换为在VGG网络特征映射上计算的损失，其对于像素空间[37]的变化更具有不变性。

• 我们通过在三个公开基准数据集的图像上进行大量的平均主观得分(MOS)测试，确认了SRGAN是最新的技术，在使用较大的上采样系数(4×)进行逼真SR图像评估上具有很大优势。

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 in Section 4 and concluding remarks in Section 5.

我们将在第二节中描述网络架构和感知损失。第三节中提供在公开基准数据集上的定量评估和视觉插图。本文在第4节中进行了讨论，并在第5节中作了总结。

2. Method

In SISR the aim is to estimate a high-resolution, super-resolved image I^{SR} from a low-resolution input image 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 × H × C and I^{HR}, I^{SR} by $rW \times rH \times C$ respectively.

2. 方法

SISR的目标是根据低分辨率输入图像I^{LR}来估计高分辨率、超分辨率图像I^{SR}。这里I^{HR}是高分辨率图像，I^{LR}是其对应的低分辨率版本。高分辨率图像仅在训练中可获得。训练中，I^{LR}可以通过对I^{HR}应用高斯滤波，然后执行下采样系数为r的下采样操作得到。对于有C个颜色通道的图像，我们分别用大小为$W \times H \times C$的实值张量描述I^{LR}，用大小为$rW \times rH \times C$的实值张量描述I^{HR}、I^{SR}。

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 training images $I_n^{HR}, n = 1, \ldots, N_n$ with corresponding I_n^{LR}, $n = 1, \ldots, N_n$, we solve:

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

我们的最终目标是训练一个生成函数G，用来估算给定LR输入图像的对应HR图像。为此，我们训练了一个生成网络，参数为θ_G的前馈CNNG_{θ_G}。其中$\theta_G = W_{1:L}; b_{1:L}$表示一个$L$层深度网络的权重和偏置，可以通过优化SR特定损失函数l^{SR}获得。对于训练图像I_n^{HR}，$n = 1, \ldots, N_n$，及其对应的I_n^{LR}，$n = 1, \ldots, N_n$，求解：

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

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.

在这项工作中，我们将专门设计一个感知损失l^{SR}作为几种损失分量的加权组合，这些损失分量对恢复的SR图像的不同要求特性进行建模。单个损失函数在2.2节中有更详细的描述。

2.1. Adversarial network architecture

Following Goodfellow et al. [21] 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:

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

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. 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.

2.1. 对抗网络架构

按照Goodfellow等[21]，我们进一步定义了一个判别器网络D_{θ_D}，我们对其与G_{θ_G}进行交替优化来解决对抗最小-最大问题：

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

这个公式的总体思想是，它允许训练生成模型G，生成模型目的是欺骗具有辨别能力的判别器D，判别器被训练用来区分超分辨图像与真实图像。通过这种方法，我们的生成器可以学习创建与真实图像高度相似的解，因此很难被D分类。这鼓励了位于自然图像子空间，流形中的感知上更优的解。这与通过最小化逐像素的误差测量(例如MSE)获得的SR解形成鲜明的对比。

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. [32] we employ the block layout proposed by Gross and Wilber [23]. Specifically, we use two convolutional layers with small 3×3 kernels and 64 feature maps followed by batch-normalization layers [31] and ParametricReLU [27] as the activation function. We increase the resolution of the input image with two trained sub-pixel convolution layers as proposed by Shi et al. [47].

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

如图4所示，我们的深度生成器网络G的中心是B个含有恒等设计的残差块。受Johnson等[32]启发，我们采用了Gross和Wilber[23]提出的块设计。具体来说，我们使用了两个卷积层，其核大小为3×3，具有64层特征映射，其后是批归一化层[31]，使用ParametricReLU[27]作为激活函数。如Shi等[47]的提议，我们使用两个训练好的子像素卷积层来增加输入图像的分辨率。

图4：生成器网络和判别器网络的架构，每个卷积层表明了对应的卷积核大小(k)，特征映射数量(n)和步长(s)。

To discriminate real HR images from generated SR samples we train a discriminator network. The architecture is shown in Figure 4. We follow the architectural guidelines summarized by Radford et al. [43] and use LeakyReLU activation (α = 0.2) 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 3 × 3 filter kernels, increasing by a factor of 2 from 64 to 512 kernels as in the VGG network [48]. 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.

为了从生成的SR样本中区分出真实的HR图像，我们训练了一个判别器网络。架构如图4所示。我们遵循Radford等[43]总结的架构指南，使用LeakyReLU激活(α=0.2)，在整个网络中避免使用最大池化。训练的判别器网络用来解决等式2中的最大化问题。它包含8个卷积层，其中3×3滤波器核的数量逐渐增加，与VGG网络一样[48]，从64个滤波器核增加到512个，增加了2倍。在每次特征数量加倍时，步长卷积用来降低图像分辨率。生成的512个特征映射之后是两个稠密层，最后的sigmoid激活用来获得样本分类的概率。

2.2. Perceptual loss function

The definition of our perceptual loss function l^{SR} is critical for the performance of our generator network. While l^{SR} is commonly modeled based on the MSE [9, 47], we improve on Johnson et al. [32] and Bruna et al. [4] and design a loss function that assesses a solution with respect to perceptually relevant characteristics. We formulate the perceptual loss as the weighted sum of a content loss l_X^{SR} and an adversarial loss component as:

$$l^{SR} = \underbrace{\underbrace{l_X^{SR}}_{content\ loss} + \underbrace{10^{-3} l_{Gen}^{SR}}_{adversarial\ loss}}_{perceptual\ loss (for\ VGG\ based\ content\ loss)} \tag{3}$$

In the following we describe possible choices for the content loss l_X^{SR} and the adversarial loss l_{Gen}^{SR}.

2.2. 感知损失函数

感知损失函数l^{SR}的定义对于我们的生成器网络性能非常关键。虽然l^{SR}通常是基于MSE[9, 47]建模的，但我们在Johnson等[32]和Bruna等[4]的基础上进行了改进，设计了一个损失函数用来评估在感知相关特性方面的解。我们将感知损失构建为内容损失l_X^{SR}和对抗损失的加权和：

$$l^{SR} = \underbrace{\underbrace{l_X^{SR}}_{content\ loss} + \underbrace{10^{-3} l_{Gen}^{SR}}_{adversarial\ loss}}_{perceptual\ loss (for\ VGG\ based\ content\ loss)} \tag{3}$$

接下来我们描述内容损失l_X^{SR}和对抗损失l_{Gen}^{SR}的可能选择。

2.2.1 Content loss

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

$$l_{MSE}^{SR} = \frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}(I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y})^2 \tag{4}$$

This is the most widely used optimization target for image SR on which many state-of-the-art approaches rely [9, 47]. However, while achieving particularly high PSNR, solutions of MSE optimization problems often lack high-frequency content which results in perceptually unsatisfying solutions with overly smooth textures (c.f. Figure 2).

2.2.1 内容损失

逐像素的MSE损失计算如下：

$$l_{MSE}^{SR} = \frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}(I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y})^2 \tag{4}$$

对于图像SR，这是应用最广泛的优化目标，许多最新技术都依赖该目标[9, 47]。然而，虽然取得了特别高的PSNR，但MSE优化问题的解通常缺少高频内容，这会导致具有过于平滑纹理的解在感知上不令人满意（对比图2）。

Instead of relying on pixel-wise losses we build on the ideas of Gatys et al. [18], Bruna et al. [4] and Johnson et al. [32] 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 [48]. With $\phi_{i,j}$ we indicate the feature map obtained by the j-th convolution (after activation) 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}:

$$l_{VGG/i,j}^{SR} = \frac{1}{W_{i,j}H_{i,j}}\sum_{x=1}^{W_{i,j}}\sum_{y=1}^{H_{i,j}}(\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{HR}))_{x,y})^2 \tag{5}$$

在基于Gatys等[18]，Bruna等[4]和Johnson等[32]想法的基础上，我们构建并使用了更接近于感知相似性的损失函数，而不是依赖于逐像素损失。我们在Simonyan和Zisserman[48]中描述的预训练19层VGG网络的ReLU激活层的基础上定义了VGG损失。在给定的的VGG19网络中，我们用$\phi_{i,j}$指代在第i层池化层之前的第j层卷积(激活之后)获得的特征映射。我们使用重建图像$G_{\theta_G}(I^{LR})$的特征表示和参照图像I^{HR}之间的欧式距离来定义VGG损失：

$$l^{SR}_{VGG/i,j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{HR}))_{x,y})^2 \tag{5}$$

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

这里$W_{i,j}$和$H_{i,j}$描述了VGG网络中各个特征映射的维度。

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^{SR}_{Gen} is defined based on the probabilities of the discriminator $D_{\theta_D}(G_{\theta_G}(I^{LR}))$ over all training samples as:

$$l^{SR}_{Gen} = \sum_{n=1}^{N} -logD_{\theta_D}(G_{\theta_G}(I^{LR})) \tag{6}$$

2.2.2 对抗损失

除了目前为止描述的内容损失之外，我们也将GAN的生成组件添加到了感知损失中。通过设法欺骗判别器网络，这鼓励我们的网络支持位于自然图像流行上的解。基于判别器$D_{\theta_D}(G_{\theta_G}(I^{LR}))$在所有训练样本上的概率，生成损失$l^{SR}_{Gen}$定义为：

$$l^{SR}_{Gen} = \sum_{n=1}^{N} -logD_{\theta_D}(G_{\theta_G}(I^{LR})) \tag{6}$$

Here, $D_{\theta_D}(G_{\theta_G}(I^{LR}))$ is the probability that the reconstructed image $G_{\theta_G}(I^{LR})$ is a natural HR image. For better gradient behavior we minimize $-logD_{\theta_D}(G_{\theta_G}(I^{LR}))$ instead of $log[1 - logD_{\theta_D}(G_{\theta_G}(I^{LR}))]$ [21].

这里，$D_{\theta_D}(G_{\theta_G}(I^{LR}))$是重建图像$G_{\theta_G}(I^{LR})$为自然HR图像的概率。为了得到更好的梯度行为，我们对$-logD_{\theta_D}(G_{\theta_G}(I^{LR}))$进行最小化，而不是$log[1 - logD_{\theta_D}(G_{\theta_G}(I^{LR}))]$ [21]。

3. Experiments

3.1. Data and similarity measures

We perform experiments on three widely used benchmark datasets Set5 [2], Set14 [68] and BSD100, the testing set of BSD300 [40]. All experiments are performed with a scale factor of 4× between low- and high-resolution images. This corresponds to a 16× reduction in image pixels. For fair comparison, all reported PSNR [dB] and SSIM [57] measures were calculated on the y-channel of center-cropped, removal of a 4-pixel wide

strip from each border, images using the daala package. Super-resolved images for the reference methods, including nearest neighbor, bicubic, SRCNN [8] and SelfExSR [30], were obtained from online material supplementary to Huang et al. [30] and for DRCN from Kim et al. [33]. Results obtained with SRResNet (for losses: l^{SR}_{MSE} and $l^{SR}_{VGG/2.2}$) and the SRGAN variants are available online. Statistical tests were performed as paired two-sided Wilcoxon signed-rank tests and significance determined at $p < 0.05$.

3. 实验

3.1. 数据和相似性度量

我们在三个广泛使用的基准数据集Set5[2]，Set14[68]和BSD300的测试集BSD100[40]上进行实验。所有实验都在低分辨率和高分辨率图像之间以4倍的尺度因子执行。图像像素对应减少16倍。为了公平比较，所有报告的PSNR[dB]和SSIM[57]度量使用daala软件包，在中心裁剪的图像的y通道上进行计算，图像每个边界移除了4个像素宽的图像条。参考方法包括最近邻居，双三次，SRCNN[8]和SelfExSR[30]的超分辨图像是从Huang等[30]和Kim等的DRCN[33]的在线补充材料中获得的。SRResNet(损失：l^{SR}_{MSE}和$l^{SR}_{VGG/2.2}$)和SRGAN变体得到的结果可在线获得。统计测试以成对的双侧威尔科克森符号秩检验和显著性检验进行，显著性水平为$p < 0.05$。

The reader may also be interested in an independently developed GAN-based solution on GitHub. However it only provides experimental results on a limited set of faces, which is a more constrained and easier task.

读者可能还对GitHub上独立开发的基于GAN的解决方案感兴趣。然而，它只能提供一组有限人脸图像上的实验结果，这是一个更受限且更轻松的任务。

3.2. Training details and parameters

We trained all networks on a NVIDIA Tesla M40 GPU using a random sample of 350 thousand images from the **ImageNet** database [44]. These images are distinct from the testing images. We obtained the LR images by downsampling the HR images (BGR, $C = 3$) using bicubic kernel with downsampling factor $r = 4$. For each mini-batch we crop 16 random 96 × 96 HR 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 [35] with $\beta_1 = 0.9$. The SRResNet networks were trained with a learning rate of 10^{-4} and 10^6 update iterations. We employed the trained MSE-based SRResNet network as initialization for the generator when training the actual GAN to avoid undesired local optima. All SRGAN variants were trained with 10^5 update iterations at a learning rate of 10^{-4} and another 10^5 iterations at a lower rate of 10^{-5}. We alternate updates to the generator and discriminator network, which is equivalent to $k = 1$ as used in Goodfellow et al. [21]. Our generator network has 16 identical ($B = 16$) residual blocks. During test time we turn batch-normalization update off to obtain an output that deterministically depends only on the input [31]. Our implementation is based on Theano [52] and Lasagne [7].

3.2. 训练细节和参数

我们使用NVIDIA Tesla M40 GPU训练所有的网络，训练数据来自**ImageNet**数据集[44]中随机采样的35万张图像。这些图片不同于测试图片。我们使用双三次核对HR图像(BGR, $C = 3$)进行下采样得到LR图像，下采样系数为$r = 4$。对于每一份小批量数据，我们对不同的训练图像裁剪16个随机的96×96的HR子图像。注意我们可以对任意大小的图像应用生成器模型，因为它是全卷积的。我们使用Adam[35]，$\beta_1 = 0.9$来进行优化。SRResNet网络使用10^{-4}的学习率进行训练，更新迭代次数10^6。在训练实际的GAN时，为了避免不必要的局部最优值，我们采用预训练的基于MSE的SRResNet网络对生成器进行初始化。所有的SRGAN变种都以10^{-4}的学习率训练10^5次迭代，然后以10^{-5}的学习率再训练10^5次迭代。我们交替更新生成器和判别器网络，这等价于Goodfellow等[21]的$k = 1$。我们的生成器网络有16个恒等($B = 16$)残差块。测试期间，为了获得确定性地只依赖输入的输出，我们关闭了批归一化更新。我们的实现基于Theano[52]和Lasagne[7]。

3.3. Mean opinion score (MOS) testing

We have performed a MOS test to quantify the ability of different approaches to reconstruct perceptually convincing images. Specifically, we asked 26 raters to assign an integral score from 1 (bad quality) to 5 (excellent quality) to the super-resolved images. The raters rated 12 versions of each image on Set5, Set14 and BSD100: nearest neighbor (NN), bicubic, SRCNN [8], SelfExSR [30], DRCN [33], ESPCN [47], SRResNet-MSE, $SRResNet - VGG22^*$ (∗ not rated on BSD100), $SRGAN - MSE^*$, $SRGAN - VGG22^*$, SRGAN-VGG54 and the original HR image. Each rater thus rated 1128 instances (12 versions of 19 images plus 9 versions of 100 images) that were presented in a randomized fashion. The raters were calibrated on the NN (score 1) and HR (5) versions of 20 images from the BSD300 training set. In a pilot study we assessed the calibration procedure and the test-retest reliability of 26 raters on a subset of 10 images from BSD100 by adding a method's images twice to a larger test set. We found good reliability and no significant differences between the ratings of the identical images. Raters very consistently rated NN interpolated test images as 1 and the original HR images as 5 (c.f. Figure 5).

Figure 5: Color-coded distribution of MOS scores on **BSD100**. For each method 2600 samples (100 images × 26 raters) were assessed. Mean shown as red marker, where the bins are centered around value i. [4× upscaling]

3.3. 平均主观得分(MOS)测试

为了量化不同方法重建感知上令人信服的图像的能力，我们进行了MOS测试。具体来说，我们让26个评分员使用整数分1(质量差)到5(质量极好)对超分辨率图像进行打分。评分员对Set5，Set14和BSD100数据集上的每一张图像的12个版本进行了评分：最近邻(NN)，双三次，SRCNN[8]，SelfExSR[30]，DRCN[33]，ESPCN[47]，SRResNet-MSE，$SRResNet - VGG22^*$ (∗没有在BSD100上评分)，$SRGAN - MSE^*$，$SRGAN - VGG22^*$，SRGAN-VGG54和原始HR图像。因此每一个评分员对随机呈现的1128个实例（19张

图像的12个版本加上100张图像的9个版本）进行了评估。评分员对BSD300训练集的20张图像的NN（得分1）和HR（5）版本上进行了校准。在初步研究中，通过两次添加方法图像到更大的测试集中，我们评估了26个评分员在BSD100的10张图像子集上的校准程序和重测信度。我们发现了良好的可靠性，在相同图像的评分之间没有显著差异。评分员非常一致地将NN插值测试图像评分为1，原始HR图像评分为5（参加图5）。

□

图5：**BSD100**上MOS得分的颜色编码分布。每一种方法使用2600个样本(100张图片×26个评估者)评估。均值显示为红色标记，bin以值i为中心(4倍上采样)。

The experimental results of the conducted MOS tests are summarized in Table 1, Table 2 and Figure 5.

Table 1: Performance of different loss functions for SRResNet and the adversarial networks on Set5 and Set14 benchmark data. MOS score significantly higher ($p < 0.05$) than with other losses in that category. [4× upscaling]

□

Table 2: Comparison of NN, bicubic, SRCNN [8], SelfExSR [30], DRCN [33], ESPCN [47], SRResNet, SRGAN-VGG54 and the original HR on benchmark data. Highest measures (PSNR [dB], SSIM, MOS) in bold. [4× upscaling]

□

进行的MOS测试的实验结果总结在表1，表2和图5中。

表1：SRResNet不同损失函数的性能和对抗网络在Set5和Set14上的基准数据。MOS得分明显比其它损失在对应类别上更高($p < 0.05$)。[4×上采样]

□

表2：NN，双三次，SRCNN[8]，SelfExSR[30]，DRCN[33]，ESPCN[47]，SRResNet，SRGAN-VGG54和原始HR在基准数据上的比较. 最高的度量(PSNR[dB]，SSIM，MOS)以粗体显示。[4×上采样]

□

3.4. Investigation of content loss

We investigated the effect of different content loss choices in the perceptual loss for the GAN-based networks. Specifically we investigate $l^{SR} = l^{SR}_X + 10^{-3} l^{SR}_{Gen}$ for the following content losses l^{SR}_X:

• 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 representing lower-level features [67].

• 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 [67, 64, 39]. We refer to this network as SRGAN in the following.

3.4. 内容损失研究

对于基于GAN的网络，我们研究了感知损失中不同内容损失选择的影响。具体来说，对于下面的内容损失l^{SR}_X，我们研究了$l^{SR} = l^{SR}_X + 10^{-3} l^{SR}_{Gen}$：

• SRGAN-MSE：l^{SR}_{MSE}，以标准MSE作为内容损失来研究对抗网络。

• SRGAN-VGG22：具有$\phi_{2,2}$的$l^{SR}_{VGG/2.2}$，表示更底层特征[67]的特征映射上定义的损失。

• SRGAN-VGG54：具有$\phi_{5,4}$的$l^{SR}_{VGG/5.4}$，来自较深网络层的更高层特征的特征映射上定义的损失，更可能集中在图像内容上[67, 64, 39]。在下文中，我们将此网络称为SRGAN。

We also evaluate the performance of the generator network without adversarial component for the two losses l^{SR}_{MSE} (SRResNet-MSE) and $l^{SR}_{VGG/2.2}$ (SRResNet-VGG22). We refer to SRResNet-MSE as SRResNet. Quantitative results are summarized in Table 1 and visual examples provided in Figure 6. 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. We further attribute minor reconstruction artifacts, which we observed in a minority of SRGAN-MSE-based reconstructions, to those competing objectives. We could not determine a significantly best loss function for SRResNet or SRGAN with respect to MOS score on Set5. However, SRGAN-VGG54 significantly outperformed other SRGAN and SRResNet variants on Set14 in terms of MOS. We observed a trend that using the higher level VGG feature maps $\phi_{5,4}$ yields better texture detail when compared to $\phi_{2,2}$ (c.f. Figure 6).

Figure 6: SRResNet (left: a,b), SRGAN-MSE (middle left: c,d), SRGAN-VGG2.2 (middle: e,f) and SRGAN-VGG54 (middle right: g,h) reconstruction results and corresponding reference HR image (right: i,j). [4× upscaling]

对于两个损失l^{SR}_{MSE}(SRResNet-MSE)和$l^{SR}_{VGG/2.2}$(SRResNet-VGG22)，我们也对没有对抗组件的生成器网络性能进行了评估。我们将SRResNet-MSE称为SRResNet。在表1中总结了定量结果，图6中提供了直观的示例。即使结合对抗损失，MSE仍然提供了具有最高PSNR值的解，与视觉感知更敏感的损失组件取得的结果相比，其在感知上更平滑，更不令人信服。这是由基于MSE的内容损失和对抗损失之间的竞争引起的。我们进一步将

少量基于SRGAN-MSE的重构中观测到的那些较小的重构结果，归因于那些相互竞争的目标。关于Set5上的MOS得分，我们不能确定一个对于SRResNet或SRGAN明显最好的损失函数。但是，考虑到Set14上的MOS得分，SRGAN-VGG54显著优于其它SRGAN和SRResNet变种。我们观察到一种趋势，与$\phi_{2,2}$相比，使用更高层的VGG特征映射$\phi_{5,4}$得到了更好的纹理细节，参见图6。

图6：SRResNet（左：a，b），SRGAN-MSE（左中：c，d），SRGAN-VGG2.2（中：e，f）和SRGAN-VGG54（右中：g，h）的重建结果以及相应的参考HR图像（右：i，j）。 [4倍上采样]

3.5. Performance of the final networks

We compare the performance of SRResNet and SRGAN to NN, bicubic interpolation, and four state-of-the-art methods. Quantitative results are summarized in Table 2 and confirm that SRResNet (in terms of PSNR/SSIM) sets a new state of the art on three benchmark datasets. Please note that we used a publicly available framework for evaluation (c.f. Section 3.1), reported values might thus slightly deviate from those reported in the original papers.

3.5. 最终网络的性能

我们比较了SRResNet、SRGAN、NN、双三次插值和四种最新方法的性能。定量结果总结在表2中，证实了SRResNet(考虑PSNR/SSIM)在三个基准数据集上确立了最新的技术水平。请注意，我们使用了一个公开可获得的框架进行评估，（参加3.1节），因此报告的值可能会与原始论文中报告的值略有不同。

We further obtained MOS ratings for SRGAN and all reference methods on BSD100. The results shown in Table 2 confirm that SRGAN outperforms all reference methods by a large margin and sets a new state of the art for photo-realistic image SR. All differences in MOS (c.f. Table 2) are highly significant on BSD100, except SRCNN vs. SelfExSR. The distribution of all collected MOS ratings is summarized in Figure 5.

我们进一步获得了BSD100数据集上SRGAN和所有其他方法的MOS评分。表2中展示的结果证实了SRGAN大幅度优于所有的参考方法，并为逼真图像SR确立了最新的技术水平。除了SRCNN和SelfExSR之外，BSD100上的MOS得分差异（参加表2）是非常显著的。所有收集的MOS得分分布总结在图5中。

4. Discussion and future work

We confirmed the superior perceptual performance of SRGAN using MOS testing. We have further shown that standard quantitative measures such as PSNR and SSIM fail to capture and accurately assess image quality with respect to the human visual system [55]. 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. [47], 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. 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. We found that even deeper networks (B > 16) can further increase the performance of SRResNet, however, come at the cost of longer training and testing times. We found SRGAN variants of deeper networks are increasingly difficult to train due to the appearance of high-frequency artifacts.

4. 讨论和未来工作

我们使用MOS测试证实了SRGAN优秀的感知性能。我们进一步表明，对于人类视觉系统[55]，标准的定量度量，例如PSNR和SSIM，不能捕获并准确评估的图像质量。这项工作的重点是超分辨率的感知质量而不是计算效率。与Shi等[47]相反，提出的模型未针对实时视频SR进行优化。然而，网络架构的初步试验表明，更窄的网络有可能在质量性能降低的情况下提供非常有效的替代方案。与Dong等[9]相反，我们发现更深的网络架构是有益的。我们推测ResNet设计对更深网络的性能有实质性影响。我们发现更深的网络(B>16)可以进一步提升SRResNet的性能，但是以更长的训练和测试时间为代价。我们发现由于高频伪影的出现，更深网络的SRGAN变种越来越难训练。

Of particular importance when aiming for photo-realistic solutions to the SR problem is the choice of the content loss as illustrated in Figure 6. 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 represent features of higher abstraction [67, 64, 39] 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. We also note that the ideal loss function depends on the application. For example, approaches that hallucinate finer detail might be less suited for medical applications or surveillance. The perceptually convincing reconstruction of text or structured scenes [30] is challenging and part of future work. 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.

当针对SR问题的逼真解决方案时，内容损失的选择是非常重要的，如图6所示。在这项工作中，我们发现$l^{SR}_{VGG/5.4}$取得了感知上最令人信服的结果，这归因于更深的网络层可能表示远离像素空间的更加抽象[67, 64, 39]特征。我们推测这些深层的特征映射单纯的注重内容而剩下的对抗损失注重纹理细节，这是没有对抗损失的超分辨率图像和逼真图像之间的主要差异。我们也注意到理想的损失函数取决于应用。例如，虚幻的更精细的细节可能不适合医疗引用或监控。感知上令人信服的文本或结构化场景[30]重建是具有挑战性的，是未来工作的一部分。内容损失函数的开发描述了图像空间内容，但对像素空间变化的不变性将进一步改善逼真的图像SR结果。

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 have highlighted some limitations of this PSNR-focused image super-resolution and introduced SRGAN, which augments the content loss function with an adversarial loss by training a GAN. Using extensive MOS testing, we have confirmed that SRGAN reconstructions for large upscaling factors (4×) are, by a considerable margin, more photo-realistic than reconstructions obtained with state-of-the-art reference methods.

5. 结论

我们描述了一个深度残差网络SRResNet，当广泛使用PSNR度量进行评估时，其在公共基准数据集上树立了最新的技术水平。我们强调了以PSNR为中心的超分辨率的一些限制，引入了SRGAN，其通过训练GAN增加了具有对抗损失的内容损失函数。使用广泛的MOS测试，我们证实了对于大的上采样系数(4×)，SRGAN重构比最新的参考方法得到的重构更逼真。

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

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

[4] J. Bruna, P. Sprechmann, and Y. LeCun. Super-resolution with deep convolutional sufficient statistics. In International Conference on Learning Representations (ICLR), 2016.

[5] 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.

[6] 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.

[7] S. Dieleman, J. Schluter, C. Raffel, E. Olson, S. K. Snderby, ¨D. Nouri, D. Maturana, M. Thoma, E. Battenberg, J. Kelly, J. D. Fauw, M. Heilman, diogo149, B. McFee, H. Weideman, takacsg84, peterderivaz, Jon, instagibbs, D. K. Rasul, CongLiu, Britefury, and J. Degrave. Lasagne: First release., 2015.

[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.

[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.

[10] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. In European Conference on Computer Vision (ECCV), pages 391–407. Springer, 2016.

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

[12] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In Advances in Neural Information Processing Systems (NIPS), pages 658–666, 2016.

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

[14] 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.

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

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

[17] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning lowlevel vision. International Journal of Computer Vision, 40(1):25–47, 2000.

[18] L. A. 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.

[19] L. A. Gatys, A. S. Ecker, and M. Bethge. Image Style Transfer Using Convolutional Neural Networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2414–2423, 2016.

[20] 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.

[21] 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 (NIPS), pages 2672–2680, 2014.

[22] 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.

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

[24] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang. Convolutional sparse coding for image super-resolution. In IEEE International Conference on Computer Vision (ICCV), pages 1823–1831. 2015.

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

[26] H. He and W.-C. Siu. Single image super-resolution using gaussian process regression. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 449–456, 2011.

[27] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In IEEE International Conference on Computer Vision (ICCV), pages 1026–1034, 2015.

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

[29] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In European Conference on Computer Vision (ECCV), pages 630–645. Springer, 2016.

[30] 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.

[31] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of The 32nd International Conference on Machine Learning (ICML), pages 448–456, 2015.

[32] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super-resolution. In European Conference on Computer Vision (ECCV), pages 694–711. Springer, 2016.

[33] 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.

[34] 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.

[35] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), 2015.

[36] 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.

[37] C. Li and M. Wand. Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2479–2486, 2016.

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

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

[40] 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.

[41] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In International Conference on Learning Representations (ICLR), 2016.

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

[43] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In International Conference on Learning Representations (ICLR), 2016.

[44] 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.

[45] J. Salvador and E. Perez-Pellitero. Naive bayes super-resolution ´forest. In IEEE International Conference on Computer Vision (ICCV), pages 325–333. 2015.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

[52] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. arXiv preprint arXiv:1605.02688, 2016.

[53] 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.

[54] 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.

[55] 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.

[56] Y. Wang, L. Wang, H. Wang, and P. Li. End-to-End Image SuperResolution via Deep and Shallow Convolutional Networks. arXiv preprint arXiv:1607.07680, 2016.

[57] 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.

[58] 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.

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

[60] 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.

[61] 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.

[62] Q. Yang, R. Yang, J. Davis, and D. Nister. Spatial-depth super resolution for range images. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–8, 2007.

[63] 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.

[64] 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.

[65] X. Yu and F. Porikli. Ultra-resolving face images by discriminative generative networks. In European Conference on Computer Vision (ECCV), pages 318–333. 2016.

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

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

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

[69] 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, 2012.

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

Deep Learning

‹ SinGAN - Learning a Generative Model from a Single Natural Image论文翻译——中文版

Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network论文翻译——中文版 ›