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

이다경

Abstract

- 더 빠르고 깊은 CNN을 사용한 single image super-resolution의 정확도와 속도에도 불구하고, 한 가지 중요한 문제가 남아있다.
  : ***large upscaling에서 미세한 texture details은 어떻게 복구할 것인가?***
  * upscaling : ex) 4X upscaling -> 16X pixel

- 최근 연구들은 **meas squared reconstruction error(MSE)**를 minimizing함으로써 super resolution method를 optimization 했다.

- 그 결과, 높은 high peak signal-to-noise ratios(PSNR - super resolution을 평가하는 수치)를 가지지만, high-frequency details가 결핍되어 있고, perceptually 불만족스럽다.
  * 즉, super resolution을 평가하는 수치는 높아도, 실제 눈으로 확인 했을 땐, 해상도가 그리 높지 않다.

- ***본 논문에서, SRGAN(a generative adversarial network(GAN) for image super-resolution(SR))을 제안한다.***

- 4X upscaling이 가능한 최초의 framework이다.

Abstract

- ***본 논문에서는 adversarial loss와 content loss를 포함하는 a perceptual loss function을 제안한다.***
- adversarial loss는 super-resolved images와 original photo-realistic images를 구별하는 discriminator network를 train한다.
- **content loss는 pixel space에서의 similarity 대신, perceptual similarity를 train한다.**
- 우리의 deep residual network는 heavily downsampled된 이미지를 복구할 수 있다.
  * 즉, 저해상도 이미지를 고해상도로 복구할 수 있다.

1. Introduction

- low-resolution image(LS)를 high-resolution(HR)으로 추정하는 것을 super-resolution(SR)이라 한다.
- SR의 문제는 특히 high upscaling에서 나타나는데, texture detail이 부족하다.
- 일반적으로 SR algorithm의 optimization target은 회복된 HR image와 original photo-realistic image의 MSE를 minimization하는 것이다.
- 하지만, MSE와 PSNR는 pixel-wise image의 차이 기반으로 정의되었기 때문에, high texture detail과 같은 지각적으로 관련 있는 차이를 잡기에는 제한적이다.

1. Introduction

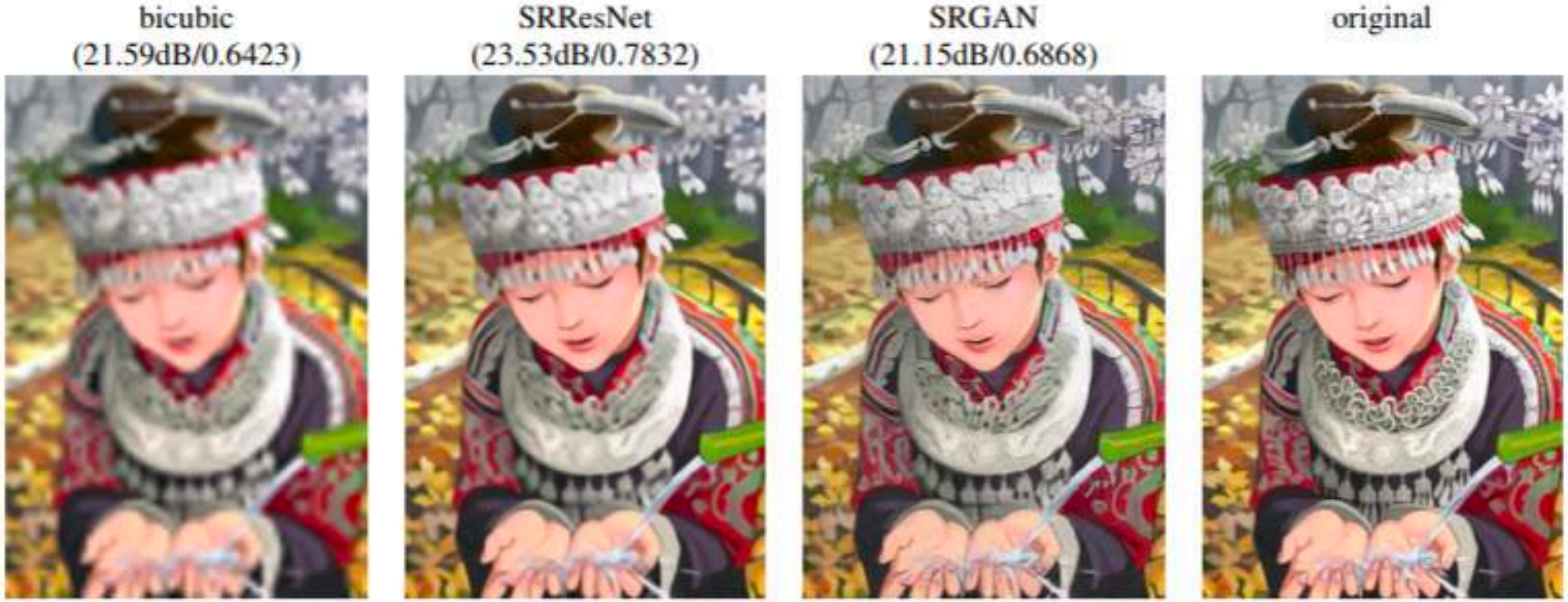


Figure 2: From left to right: bicubic interpolation, deep residual network optimized for MSE, deep residual generative adversarial network optimized for a loss more sensitive to human perception, original HR image. Corresponding PSNR and SSIM are shown in brackets. [4× upscaling]

* 즉, Figure 2와 같이 PSNR은 perceptual SR을 반영하지 못한다.

- 이전 연구와 다른 점은, 본 논문은 VGG network의 high-level feature maps와 discriminator를 결합한 새로운 perceptual loss를 제안한다.

1.1 Related work

1.1.3 Loss function

- MSE는 pixel-wise average loss이기 때문에, 과하게 smooth하고, 따라서 poor perceptual quality이다.

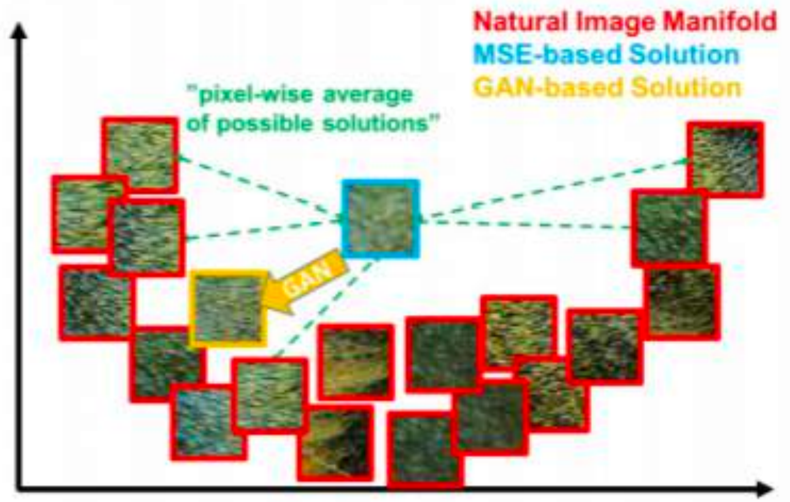


Figure 3: Illustration of patches from the natural image manifold (red) and super-resolved patches obtained with MSE (blue) and GAN (orange). The MSE-based solution appears overly smooth due to the pixel-wise average of possible solutions in the pixel space, while GAN drives the reconstruction towards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

- Figure 3과 같이 MSE는 평균을 내기 때문에 지나치게 smooth하지만, GAN은 natural image에서 reconstruction하기 때문에 더 설득력 있는 solution이다.

1.2 Contribution

- GAN은 reconstruction을 natural image를 포함할 가능성이 높은 영역으로 이동시킨다.
  * GAN이 분포를 추정하는 모델이라 그런 듯!

- 우리의 contribution은
  - We set a new state of the art for image SR with high upscaling factors(4X) 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, which are more invariant to changes in pixel space.
  - 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(4X).

✓ 즉, SRResNet과 비교해도 성능이 좋은, GAN base의 SR기술인 SRGAN을 제안하는데, 이는 MOS test에서도 좋은 성능을 보인다.

2. Method

- single image super-resolution(SISR)의 목표는 low-resolution input image I^{LR} 에서 high-resolution image(super-resolved image. I^{HR})를 추정하는 것이다.

- 우리의 최고의 목표는 주어진 LR input image를 그에 상응하는 HR image 짝을 생성하는 generating function G를 train하는 것이다.

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

I^{SR} : 이 논문에서 design한 perceptual loss

2. Method

2.1 Adversarial network architecture

- Goodfellow가 제안한 GAN loss :

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

- G에 의해 생성된 image가 D에 의해 진짜 image인지, 생성된 이미지인지 판별된다. 이것이 SR에서 MSE와 같은 pixel-wise error를 minimizing하는 것과 다른 점이다.

2. Method

2.1 Adversarial network architecture

- Model architecture

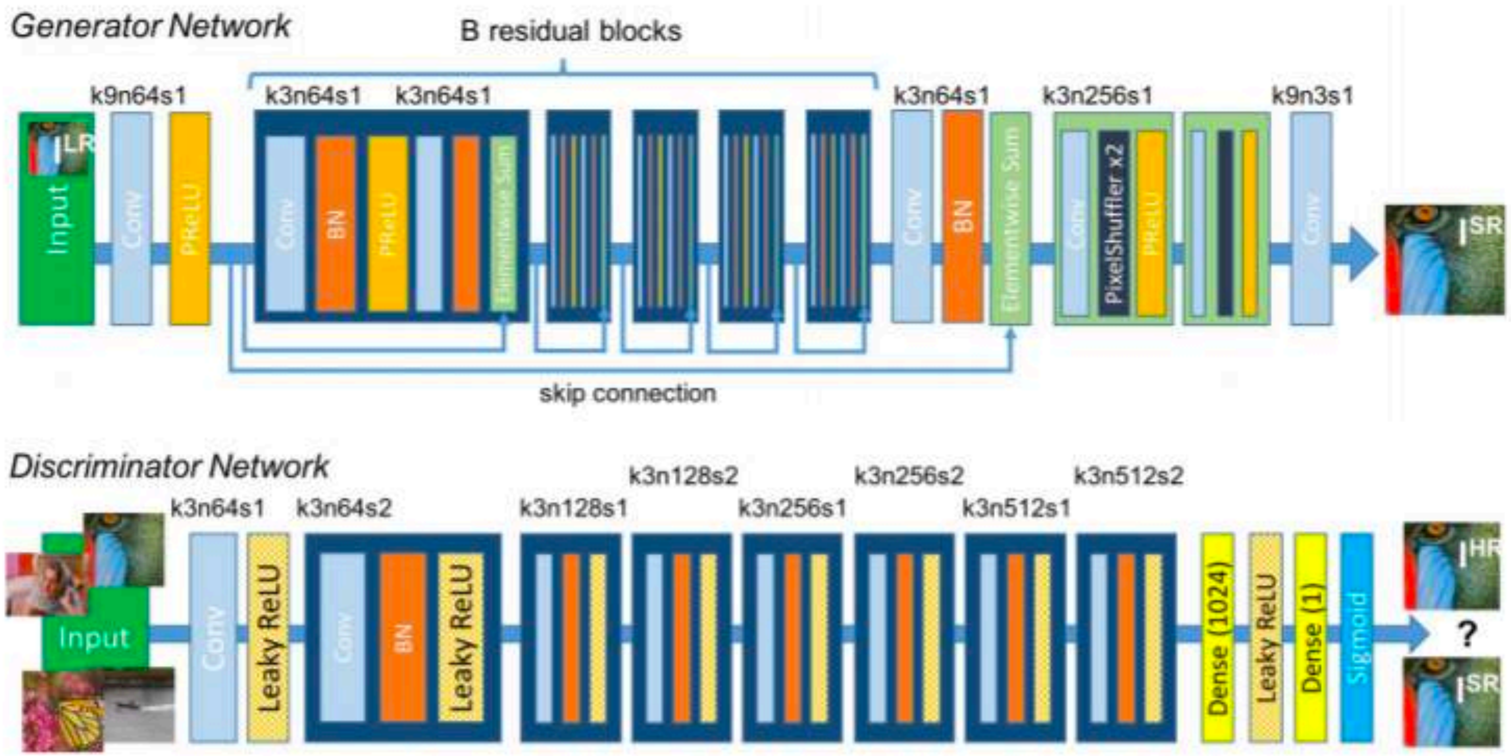


Figure 4: Architecture of Generator and Discriminator Network with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer.

2. Method

2.1 Adversarial network architecture

- We increase the resolution of the input image with two trained **sub-pixel convolution layers** as proposed by Shi et al.

- super-resolution이라면 pixel의 수가 당연히 늘어날 텐데, 일반적으로 CNN filter를 거치면 image dimension은 줄거나 동일하다. 이 때, 여기서 pixel 수를 늘리는 즉, resolution을 increase하는 방법이 바로 저 **sub-pixel**인 것 같다.
- CVPR에 2016년 9월에 발간된 Super-Resolution 논문 (Real-Time Single Image and Video Super-Resolution Using and Efficient Sub-Pixel Convolution Neural Network - https://arxiv.org/abs/1609.05158)

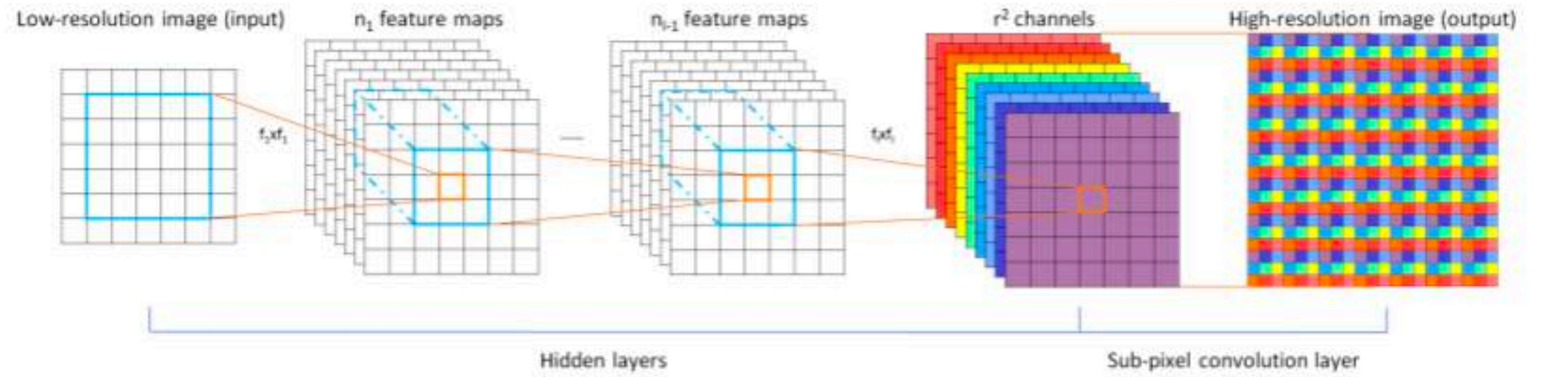


Figure 1. The proposed efficient sub-pixel convolutional neural network (ESPCN), with two convolution layers for feature maps extraction, and a sub-pixel convolution layer that aggregates the feature maps from LR space and builds the SR image in a single step.

input image의 feature map들을 이리저리 조합해서 pixel 수가 늘어나는 듯 하다.

2. Method

2.2 Perceptual loss function

- l^{SR} 은 다음과 같이 정의한다.

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

2. Method

2.2 Perceptual loss function

2.2.1 content loss

- pixel-wise MSE loss는 다음과 같다.

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

- 하지만 이는 high PSNR은 얻을 지라도, 너무 smooth되어 high-frequency content에서는 문제가 될 수 있다.

2. Method

2.2 Perceptual loss function

2.2.1 content loss

- 따라서 pixel-wise loss 대신, VGG loss(based on ReLU activation layers of pre-trained 19 layer VGG network)를 정의한다.

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

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

- ϕ_{ij} : feature map obtained by the j-th convolution(after activation) before the i-th maxpooling layer within the VGG19 network
- $G_{\theta G}(I^{LR})$ 과 I^{SR} 의 feature representation(VGG feature map)의 euclidean distance

2. Method

2.2 Perceptual loss function

2.2.2 Adversarial loss

- discriminator network를 속임으로써, natural image와 비슷하게 generating하도록 한다.
- l_{SR}^{Gen} iscriminator가 $G_{\theta G}(I^{LR})$ 를 natural HR image라고 구별할 확률 base로 정의된다.

$$l_{Gen}^{SR} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{LR})) \qquad (6)$$

3. Experiments

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]

	SRResNet-		SRGAN-		
Set5	MSE	VGG22	MSE	VGG22	VGG54
PSNR	32.05	30.51	30.64	29.84	29.40
SSIM	0.9019	0.8803	0.8701	0.8468	0.8472
MOS	3.37	3.46	3.77	3.78	3.58
Set14					
PSNR	28.49	27.19	26.92	26.44	26.02
SSIM	0.8184	0.7807	0.7611	0.7518	0.7397
MOS	2.98	3.15*	3.43	3.57	3.72*

- 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 [68].

- **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 [68, 65, 40]. We refer to this network as **SRGAN** in the following.

3. Experiments

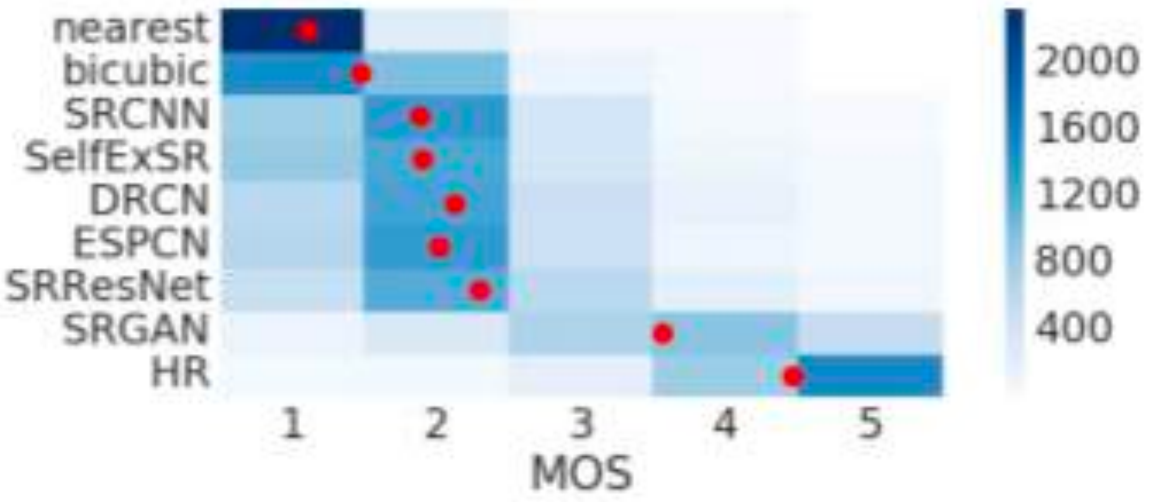


Figure 5: Color-coded distribution of MOS scores on **BSD100**. For each method 2600 samples (100 images × 26 raters) were assessed. Mean shown as red marker, where the bins are centered around value i. [4× upscaling]

3. Experiments

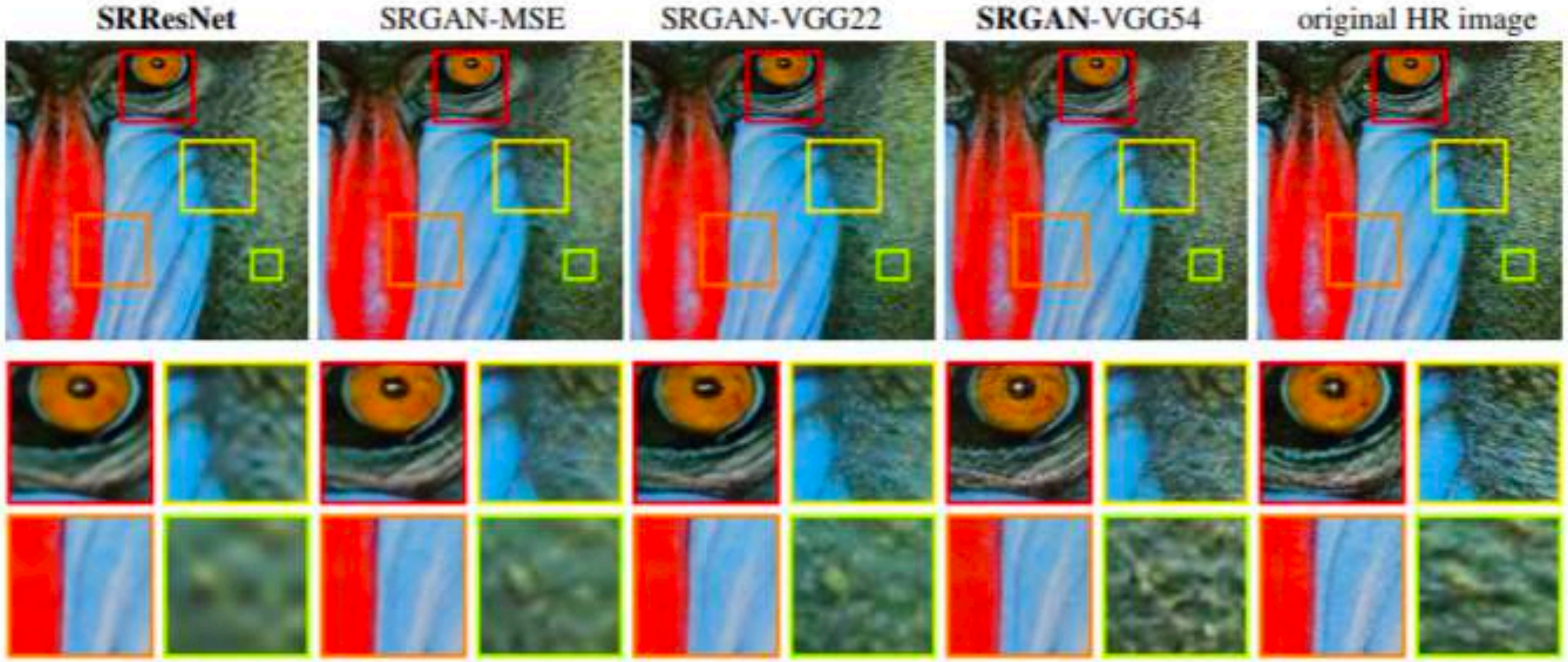


Figure 6: **SRResNet** (left: a,b), SRGAN-MSE (middle left: c,d), SRGAN-VGG2.2 (middle: e,f) and **SRGAN**-VGG54 (middle right: g,h) reconstruction results and corresponding reference HR image (right: i,j). [4× upscaling]

4. Supplementary Material

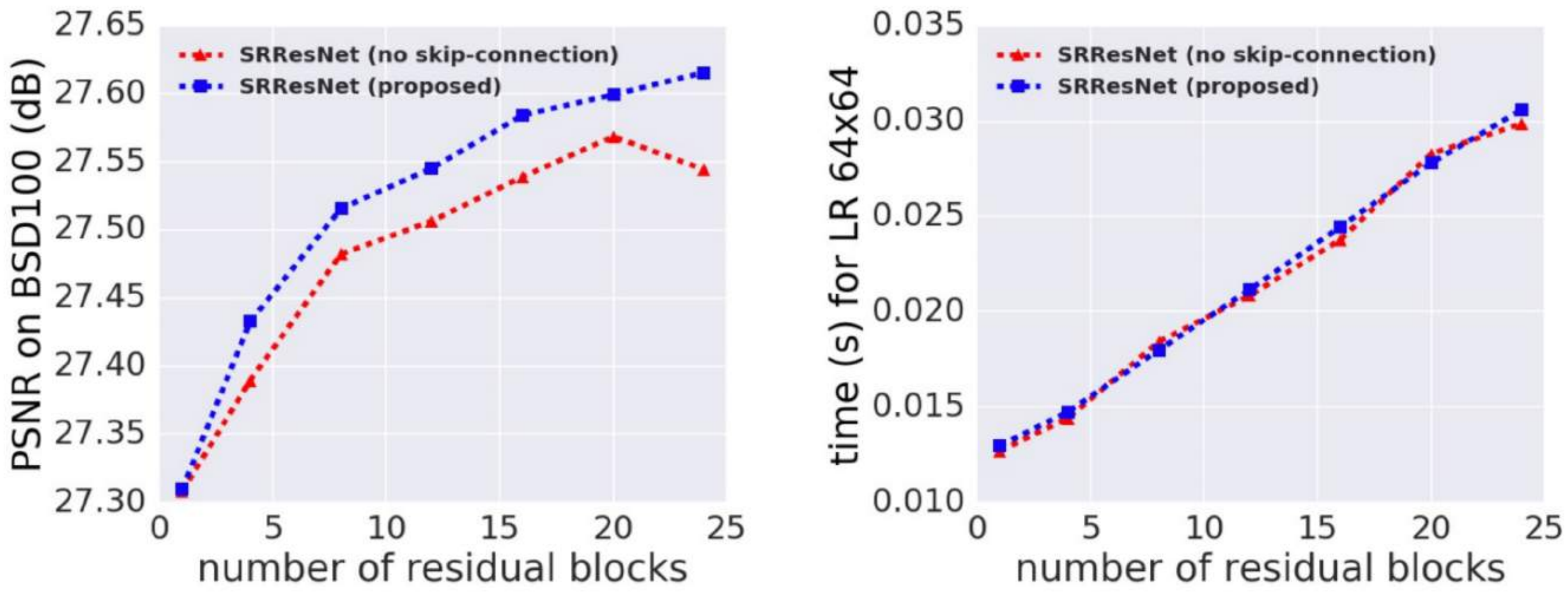
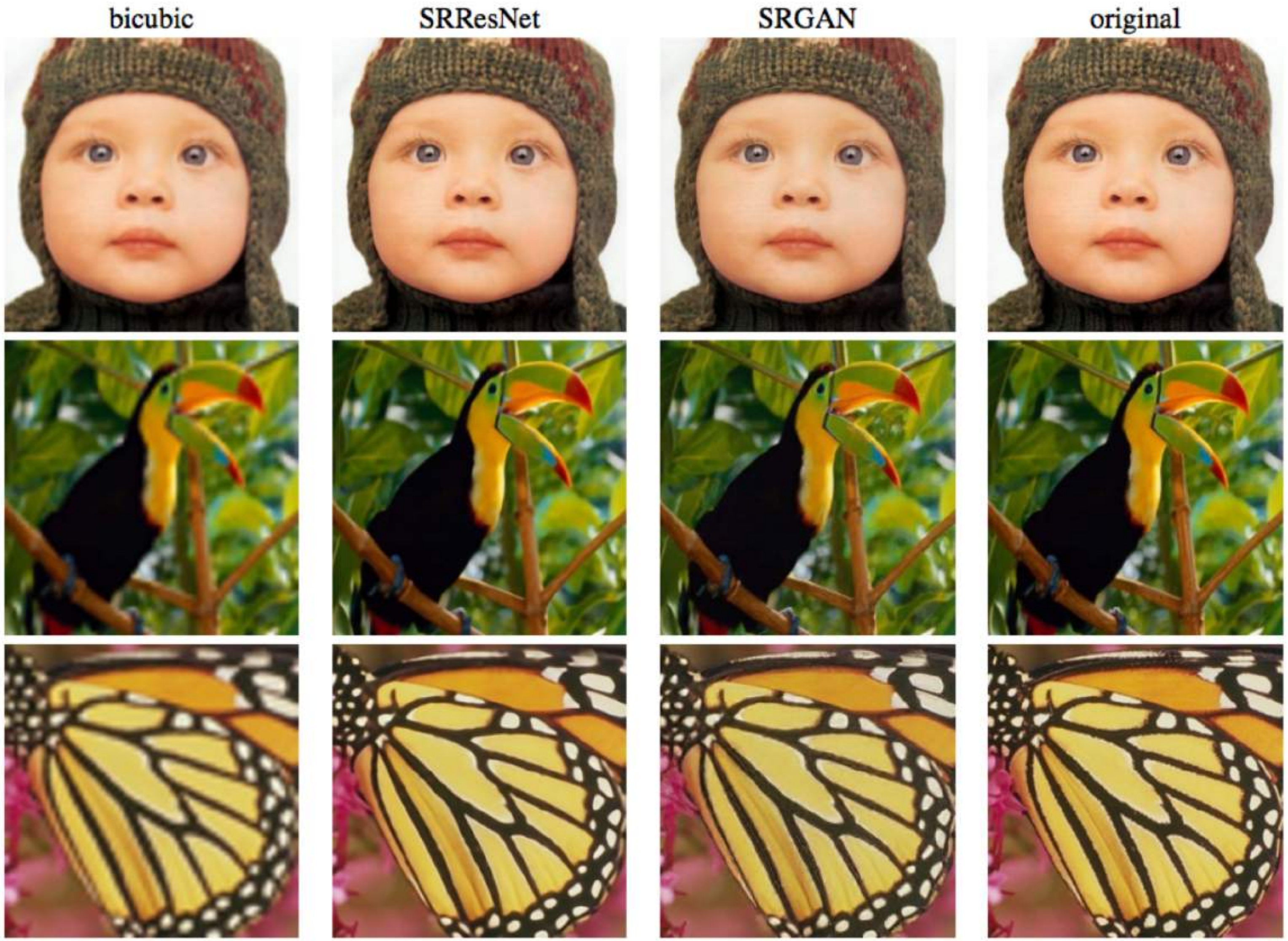


Figure 7: Dependence of network performance (PSNR, time) on network depth. PSNR (left) calculated on BSD100. Time (right) averaged over 100 reconstructions of a random LR image with resolution 64×64.

4. Supplementary Material



4. Supplementary Material

4. Supplementary Material