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

Christian Ledig, Lucas Theis, Ferenc Husz´ar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe Shi

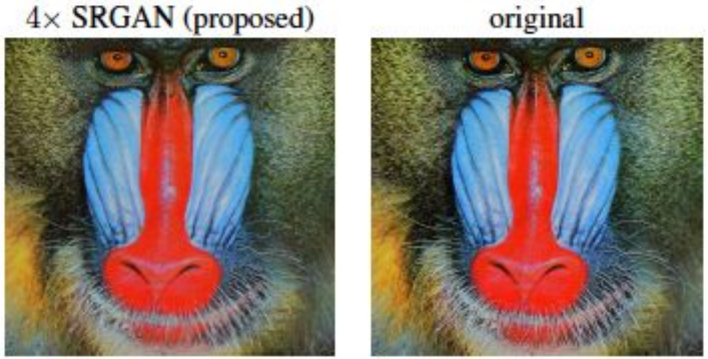
Bedirhan Uzun

Nazlıcan Gengeç

# Contents

# Problem statement

Super-resolution is to take a low resolution image and produce an estimate of a corresponding high-resolution image.



# Motivation

This task has numerous applications including in:

- Satellite imaging
- Media content
- Medical imaging
- Face recognition
- Survelliance

# Related work

Image super-resolution can be separated into 3 groups:

- Traditional filtering methods
- Training based methods
- Neural network approaches

# Traditional filtering methods


Jain, Anil K. *Fundamentals of digital image processing.* Englewood Cliffs, NJ: Prentice Hall. 1989.

R. Keys. *Cubic convolution interpolation for digital image processing.* IEEE Transactions on Acoustics, Speech, and Signal Processing. 29 (6): 1153–1160. 1981.

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

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

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


- Simple
- Very fast
- Overly smooth textures
- Not photo-realistic results

- ❑ Basic filtering techniques
- ❑ Particularly focused on edge-preservation

# Training based methods

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

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

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.

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.

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.

- ❏ Based on example-pairs rely on low-resolution (LR) training patches with high-resolution (HR) counterpart.
- ❏ Dictionary-based approach
- ❏ Multi-scale
- ❏ Whole image or overlapping patches
- ❏ Self-similarities

- Not photo-realistic results

# Neural network approaches

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

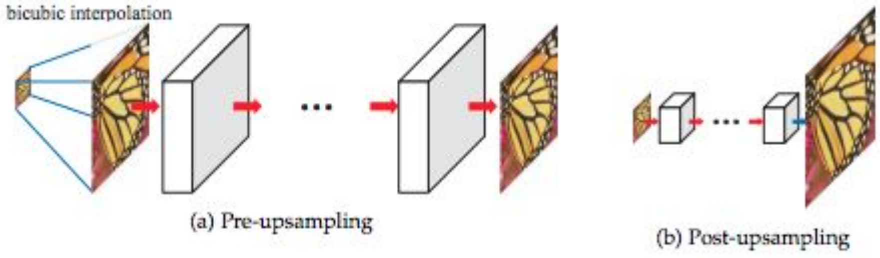
Kim, Jiwon, Jung Kwon Lee, and Kyoung Mu Lee. *Accurate image super-resolution using very deep convolutional networks.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. **[VDSR]**

Kim, Jiwon, Jung Kwon Lee, and Kyoung Mu Lee. *Deeply-recursive convolutional network for image super-resolution.* Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. **[DRCN]**

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.

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

- ❑ Using bicubic interpolation, to upscale LR input images to target spatial resolution before feed to deep neural network (SRCNN, VDSR, DRCN)
- ❑ Train with residual image (VDSR)
- ❑ Enable network to learn the upscaling filters directly
- ❑ Loss function closer to perceptual similarity



# Design of convolutional neural networks

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

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.

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.

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.

- Deeper network architecture
- Residual blocks and skip-connections
- Learning upscaling filters

# Loss functions

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

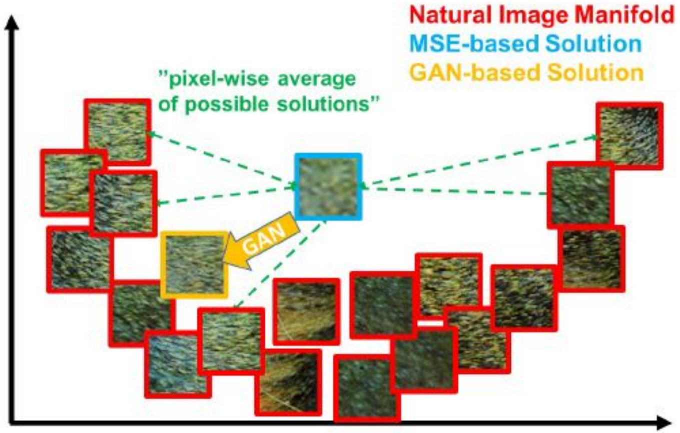
E. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In Advances in Neural Information Processing Systems (NIPS), pages 1486–1494, 2015.

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

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

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.

- Pixel-wise loss
- Adversarial loss
- Feature-level loss



# Proposed method

- Deeper network architecture
- Residual blocks w/ skip connections
- Learning upscaling filters ( w/ sub-pixel convolutional layer )
- GAN based solution
- Perceptual loss ( features from 5th layer of VGG19 )

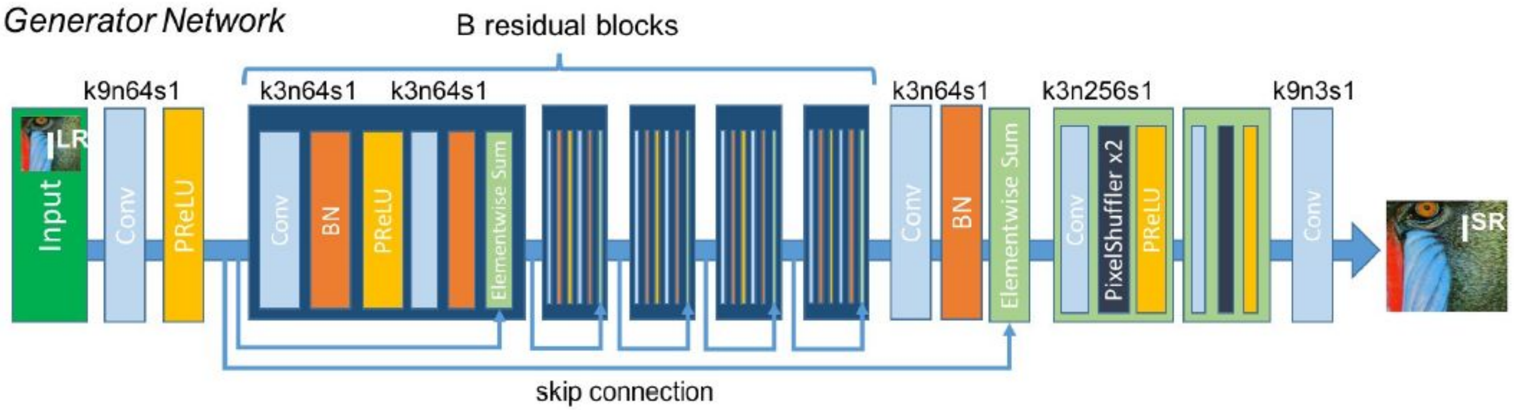
# Contribution

- 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.
- 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.
- With an extensive mean opinion score (MOS) test on images from three public benchmark datasets, 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).

# Method

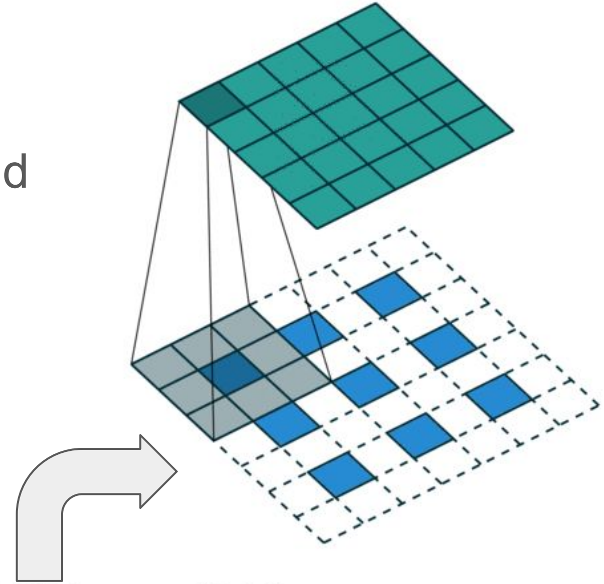
To start with SRResNet,

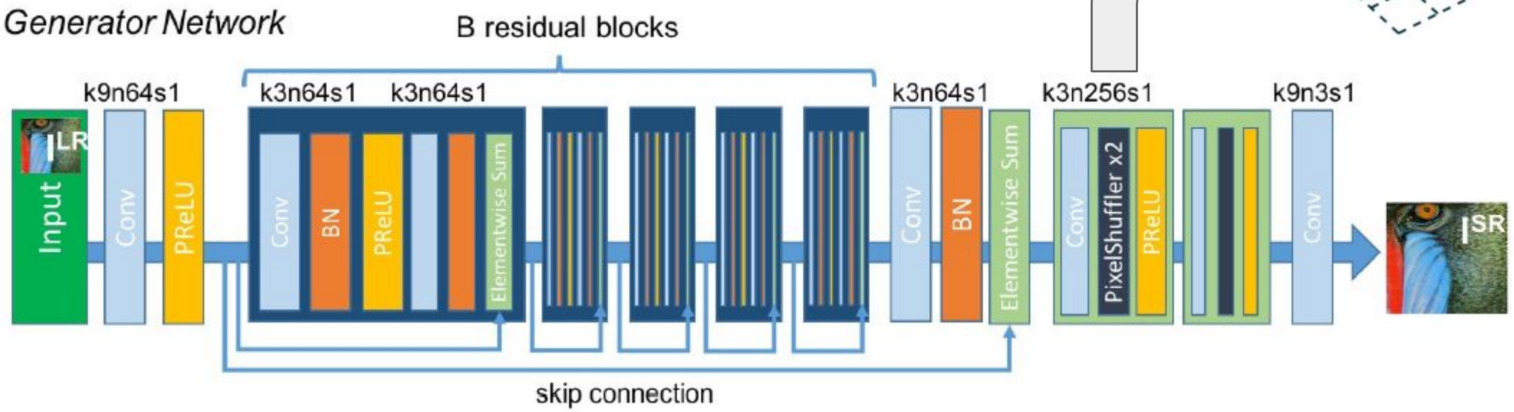
- It’s the same as **Generator** in SRGAN architecture.
- The base of the model architecture is the residual block. Each residual block has two convolutional layers, each followed by batch normalization (BN) layer with the parametric rectifying linear unit after the first one (PReLU).



# Method (Cont’d)

- Convolutional layers have 3 x 3 receptive field and each of them contains 64 filters.
- Image resolution is increased near the end of the model.

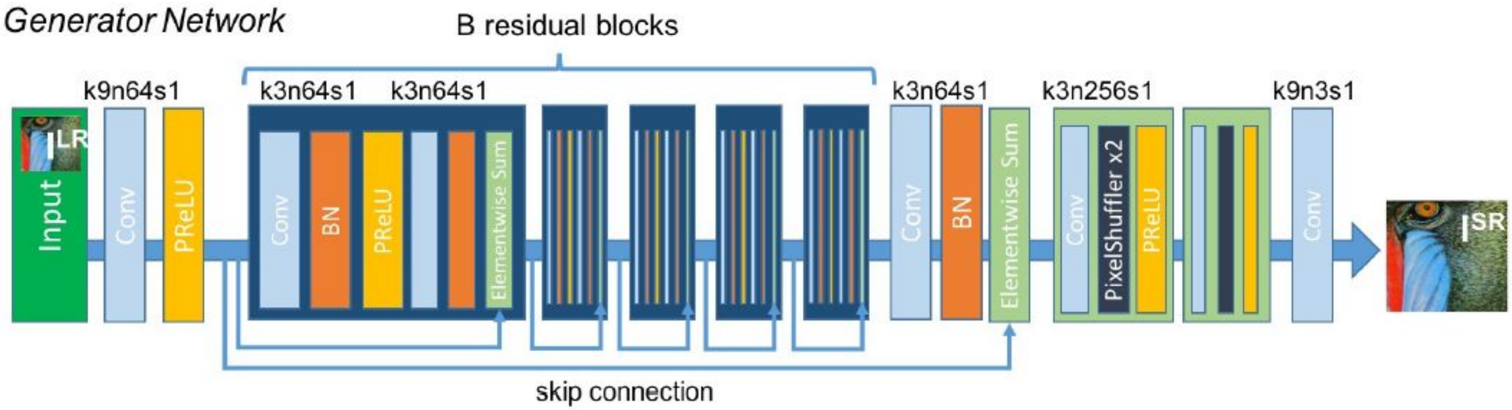




# Method (Cont'd)

The goal of generator network is optimizing loss function below.

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



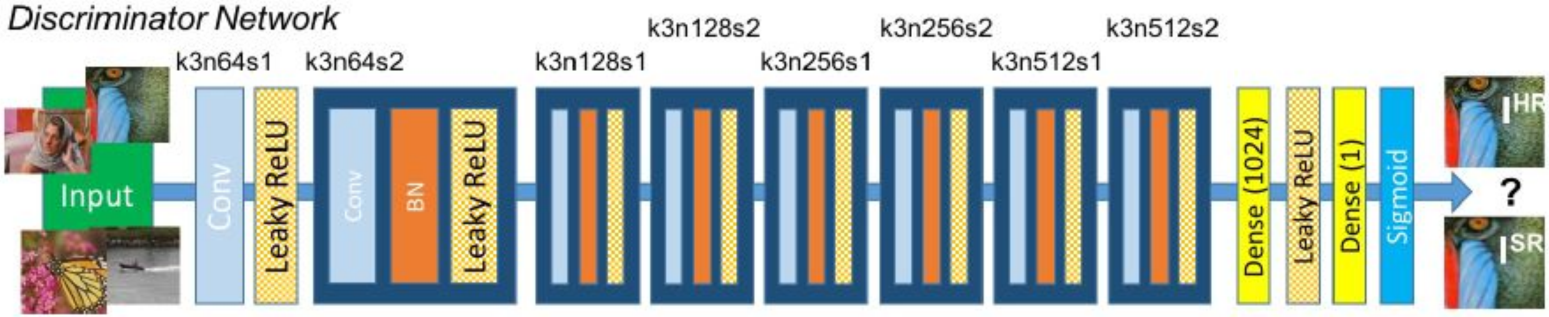
# Adversarial network architecture

The goal of generator is to fool discriminator D.

The goal of discriminator is to determine super-resolved image as a fake.

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

Overall, this two neural networks supervise each other.



# Perceptual loss function

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

```
g_gan_loss = 1e-3 * tl.cost.sigmoid_cross_entropy(logits_fake, tf.ones_like(logits_fake), name='g')

mse_loss = tl.cost.mean_squared_error(net_g.outputs, t_target_image, is_mean=True)

vgg_loss = 2e-6 * tl.cost.mean_squared_error(vgg_predict_emb.outputs, vgg_target_emb.outputs, is_mean=True)
```

For SRResNet

```
g_loss = mse_loss
```

For Generator in SRGAN

```
g_content_loss = mse_loss + vgg_loss

g_loss = g_content_loss + g_gan_loss
```

# Content loss

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

```
g_gan_loss = 1e-3 * tl.cost.sigmoid_cross_entropy(logits_fake, tf.ones_like(logits_fake), name='g')

mse_loss = tl.cost.mean_squared_error(net_g.outputs, t_target_image, is_mean=True)

vgg_loss = 2e-6 * tl.cost.mean_squared_error(vgg_predict_emb.outputs, vgg_target_emb.outputs, is_mean=True)
```

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

For SRResNet

```
g_loss = mse_loss
```

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

For Generator in SRGAN

```
g_content_loss = mse_loss + vgg_loss

g_loss = g_content_loss + g_gan_loss
```

# Adversarial loss

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

```
g_gan_loss = 1e-3 * tl.cost.sigmoid_cross_entropy(logits_fake, tf.ones_like(logits_fake), name='g')

mse_loss = tl.cost.mean_squared_error(net_g.outputs, t_target_image, is_mean=True)

vgg_loss = 2e-6 * tl.cost.mean_squared_error(vgg_predict_emb.outputs, vgg_target_emb.outputs, is_mean=True)
```

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

## For SRResNet

```
g_loss = mse_loss
```

## For Generator in SRGAN

```
g_content_loss = mse_loss + vgg_loss

g_loss = g_content_loss + g_gan_loss
```

# Experiments

1. Data and similarity measures
2. Training details and parameters
3. Mean opinion score (MOS) testing
4. Investigation of content loss
5. Performance of the final networks

# Data and Similarity Measures

- Three benchmark datasets are used : **Set5**, **Set14** and **BSD100**. The testing set is obtained from **BSD300**.
- Experiments are performed with a scale factor of 4x between low-resolution and high-resolution images.
- All PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) measures were calculated on y-channel (luminance channel of YCbCr color space) of center-cropped. Mean of center-cropped is removing of 4-pixel wide strip from each border.

# Training Details and Parameters

- Training is done on a NVIDIA Tesla M40 GPU.
- All networks are trained using 350 thousand images from the **ImageNet**.
- LR images are obtained by downsampling (bicubic kernel with r=4) the HR images.
- LR input images are scaled in the range of [0,1] and HR image in [-1,1].
- Adam optimization with $\beta_1$ =0.9 is used (The method of stochastic optimization).
- The learning rate and iterations are $10^{-4}$ and $10^{6}$ in SRResnet networks.
- All SRGAN variants are trained with $10^{5}$ update iterations at a learning rate of $10^{-4}$ .

# Mean Opinion Score (MOS) Testing

- MOS is performed to quantify the ability of different approaches to reconstruct perceptually convincing images.
- 26 raters are asked and wanted to assign score from 1 to 5.
- The raters rated 12 versions of each image on Set5,Set14 and BSD100.
  - Nearest Neighbor (NN)
  - Bicubic
  - SRCNN
  - SelfExSR
  - DRCN
  - ESPCN
  - SRResNet-MSE
  - SRResNet-VGG22
  - SRGAN-MSE
  - SRGAN-VGG22
  - SRGAN-VGG54
  - Original HR Image

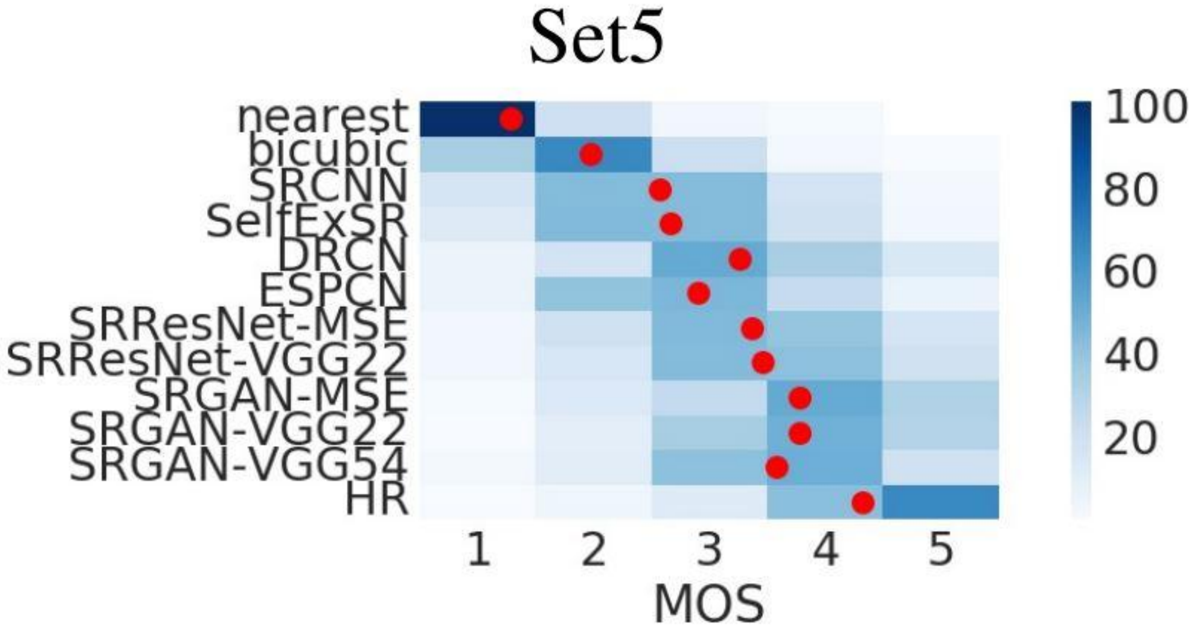



Figure shows MOS scores on Set5 dataset. Means of scores are shown as red marker for each method.

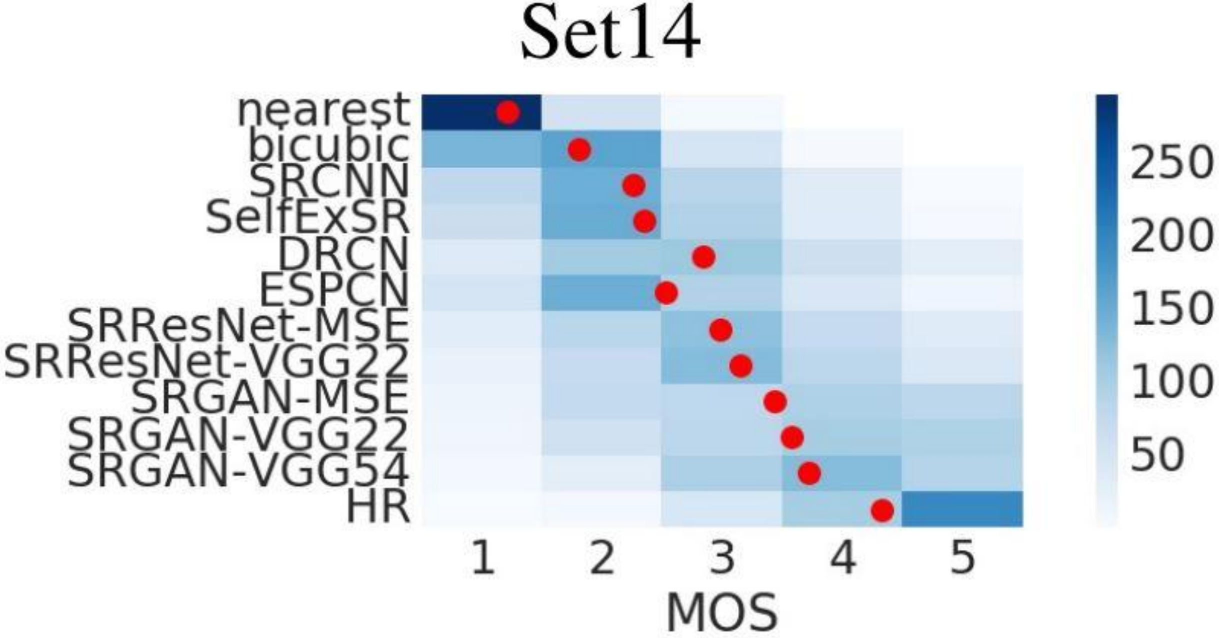



Figure shows MOS scores on Set14 dataset.

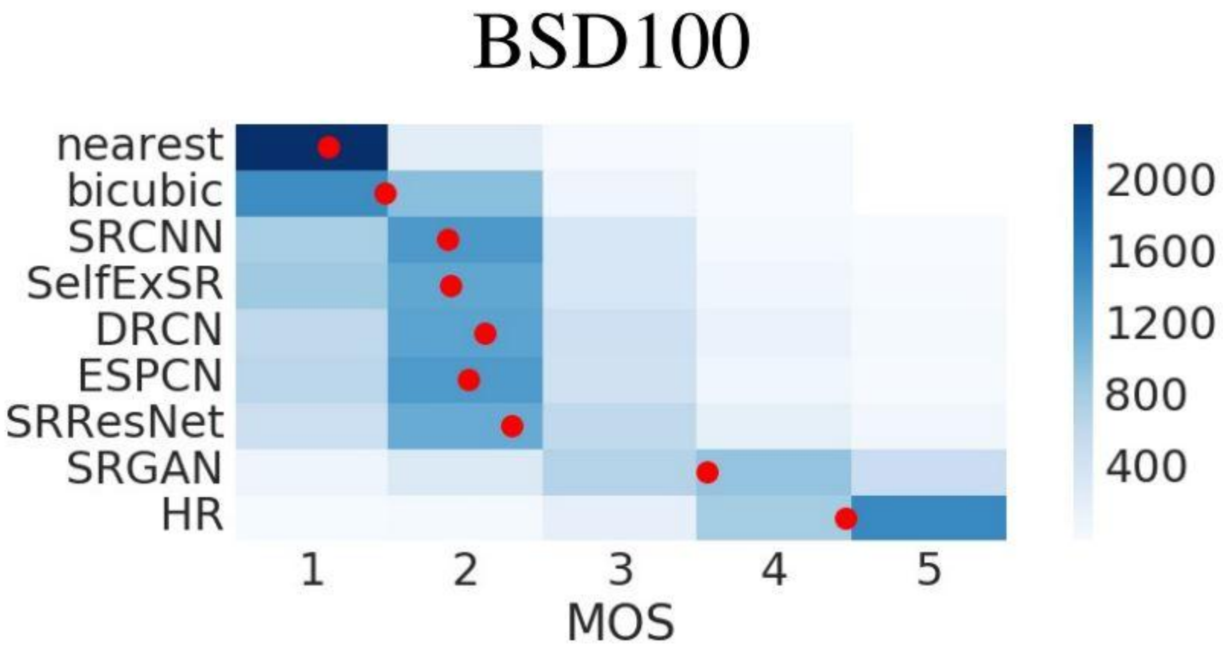



Figure shows MOS scores on BSD100 dataset.

# Investigation of Content Loss

The effect of different content loss choices is investigated in the perceptual loss.

- SRGAN-MSE : Adversarial network with standard MSE as content loss.
- SRGAN-VGG22 : A loss defined on feature maps representing lower-level features (with $\Phi_{2,2}$).
- SRGAN-VGG54 : A loss defined on feature maps representing higher-level features from deeper network layers (with $\Phi_{5,4}$).

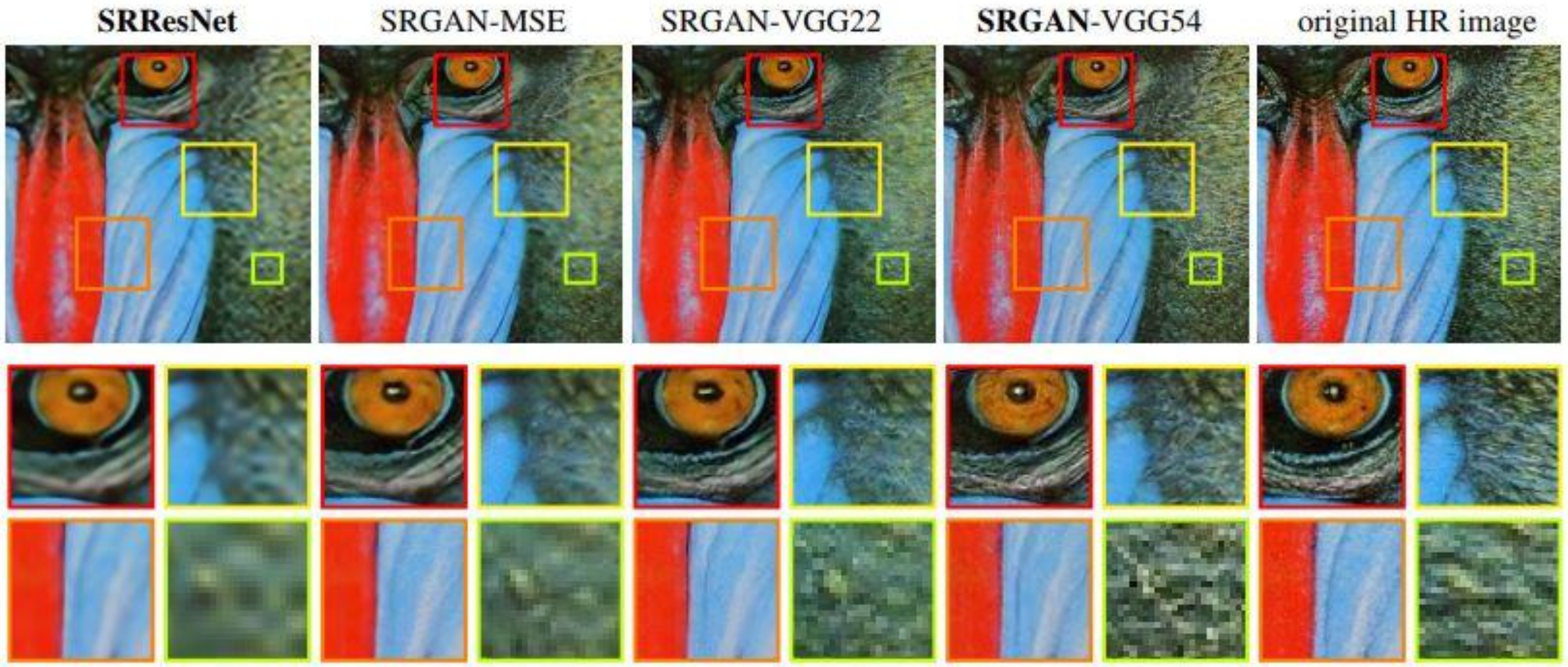
| Set5 | SRResNet-MSE | SRResNet-VGG22 | SRGAN-MSE | SRGAN-VGG22 | SRGAN-VGG54 |
|---|---|---|---|---|---|
| PSNR | 32.05 | 30.51 | 30.64 | 29.84 | 29.40 |
| SSIM | 0.9019 | 0.8803 | 0.8701 | 0.8468 | 0.8472 |
| MOS | 3.37 | 3.46 | 3.77 | 3.78 | 3.58 |
| **Set14** | | | | | |
| PSNR | 28.49 | 27.19 | 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* |

Performance of different loss functions for SRResNet and the adversarial networks on Set5 and Set14 datasets.

# Investigation of Content Loss

- Best loss function for SRResNet or SRGAN with respect to MOS score on Set5 is not determined.
- But, SRGAN-VGG54 significantly outperforms other SRGAN and SRResNet variants on Set14 in terms of MOS.
- They observed that using the higher level VGG feature maps $\Phi_{5,4}$ yields better texture details when compare to $\Phi_{2,2}$.

# Investigation of Content Loss (Visual Examples)



# Performance of The Final Networks

- They compare the performance of SRResNet and SRGAN to NN, bicubic interpolation and four state-of-the-art methods.
- SRResNet sets a new state of the art on three benchmark datasets in terms of PSNR/SSIM.
- SRGAN outperforms all reference methods and sets a new state of the art for photo-realistic image SR (in terms of MOS).

# Performance of The Final Networks (Quantitative Results)

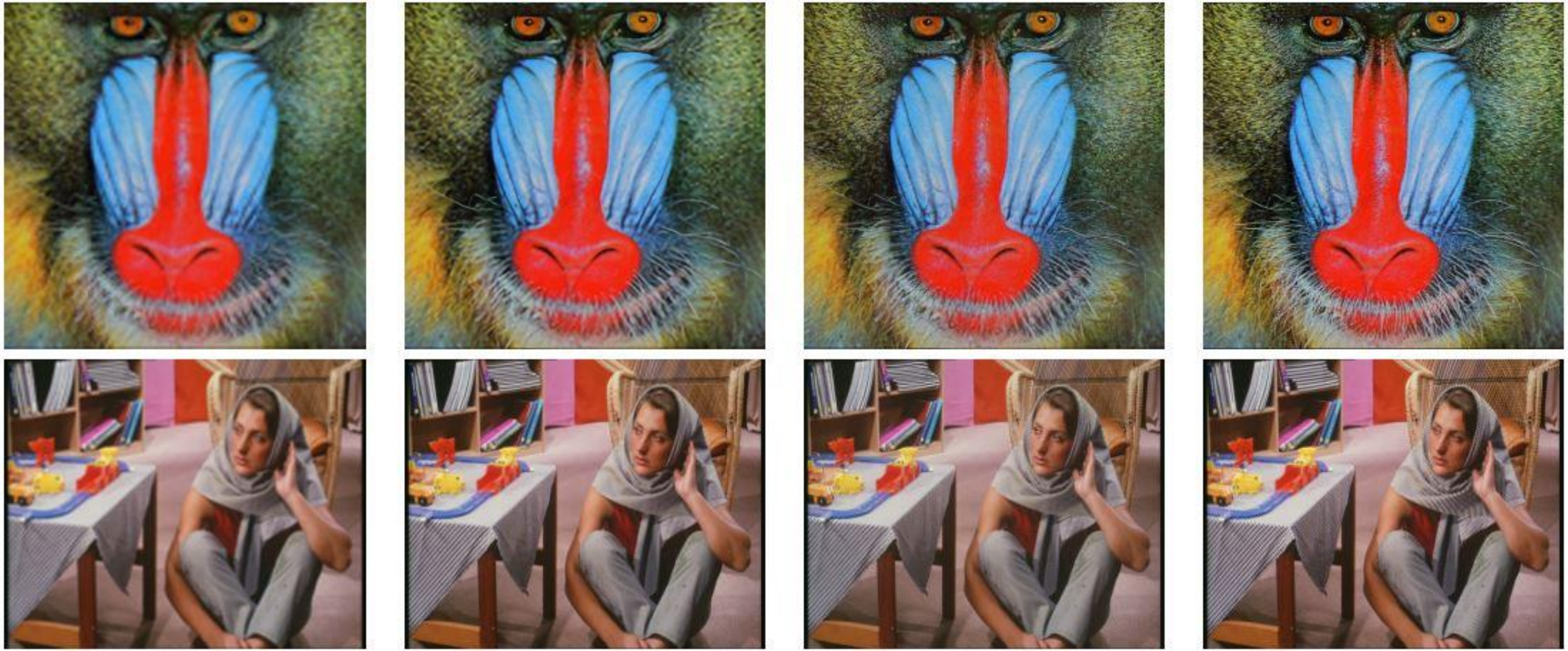
| **Set5** | nearest | bicubic | SRCNN | SelfExSR | DRCN | ESPCN | **SRResNet** | **SRGAN** | HR |
|---|---|---|---|---|---|---|---|---|---|
| PSNR | 26.26 | 28.43 | 30.07 | 30.33 | 31.52 | 30.76 | **32.05** | 29.40 | $\infty$ |
| SSIM | 0.7552 | 0.8211 | 0.8627 | 0.872 | 0.8938 | 0.8784 | **0.9019** | 0.8472 | 1 |
| MOS | 1.28 | 1.97 | 2.57 | 2.65 | 3.26 | 2.89 | 3.37 | **3.58** | 4.32 |
| **Set14** | | | | | | | | | |
| PSNR | 24.64 | 25.99 | 27.18 | 27.45 | 28.02 | 27.66 | **28.49** | 26.02 | $\infty$ |
| SSIM | 0.7100 | 0.7486 | 0.7861 | 0.7972 | 0.8074 | 0.8004 | **0.8184** | 0.7397 | 1 |
| MOS | 1.20 | 1.80 | 2.26 | 2.34 | 2.84 | 2.52 | 2.98 | **3.72** | 4.32 |
| **BSD100** | | | | | | | | | |
| PSNR | 25.02 | 25.94 | 26.68 | 26.83 | 27.21 | 27.02 | **27.58** | 25.16 | $\infty$ |
| SSIM | 0.6606 | 0.6935 | 0.7291 | 0.7387 | 0.7493 | 0.7442 | **0.7620** | 0.6688 | 1 |
| MOS | 1.11 | 1.47 | 1.87 | 1.89 | 2.12 | 2.01 | 2.29 | **3.56** | 4.46 |

# Visual Results (Set5)

Bicubic | SRResNet | SRGAN | Original

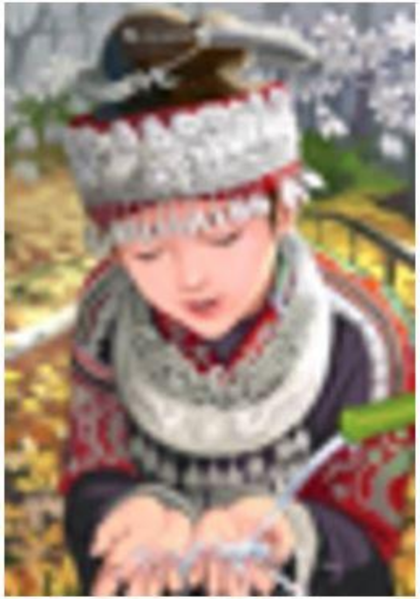
# Visual Results (Set14)

Bicubic | SRResNet | SRGAN | Original



# Visual Results (Set14)

Bicubic



SRResNet

SRGAN

Original

# Visual Results (BSD100)

Bicubic | SRResNet | SRGAN | Original

# Discussion and Future Work

- The superior perceptual performance of SRGAN is confirmed using MOS testing.
- Standard quantitative measures such as PSNR and SSIM fail to capture and accurately assess image quality with respect to the human visual system is shown.
- Preliminary experiments suggests that shallower networks provide very efficient alternatives at a small reduction of qualitative performance.
- But, they found deeper networks to be beneficial in contrast to Dong et al[1] .

1 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

# Discussion and Future Work

- ResNet design has a substantial impact on the performance of deeper networks.
- 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 perceptually convincing reconstruction of text or structured scenes is future work.

# Conclusion

- SRResNet and SRGAN have been described on public benchmark datasets.
- SRResNet gives good results in terms of PSNR/SSIM, but PSNR has some limitations.
- SRGAN which augments the content loss function with an adversarial loss by training a GAN have been introduced.
- As a result, SRGAN gives more photo-realistic results than state-of-the-art reference methods.

***Thank you for listening to us.***