SoK: Single Image Super-Resolution

Tao Li
root@tao.li

August, 2017

Outline

Outline

Interpolation-based Methods[1]

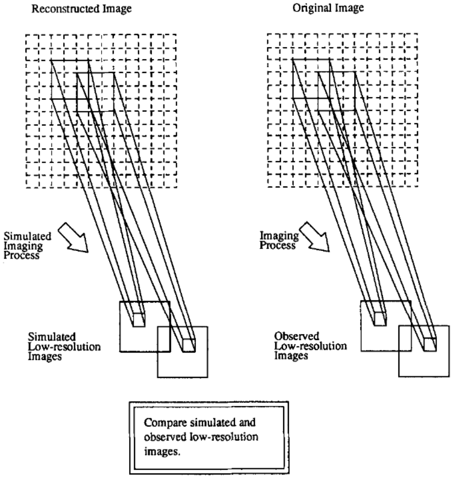


FIG. 1. Schematic diagram of the superresolution algorithm. A high-resolution reconstructed image (left) is sought, which gives simulated low-resolution images that are as close as possible to the observed low-resolution images.

[1] [Irani and Peleg, 1991]

Interpolation-based Methods

Interpolation-based methods (bilinear, bicubic, and Lanczos) generate HR pixel intensities by weighted averaging neighboring LR pixel values. Since interpolated intensities are locally similar to neighboring pixels, these algorithms generate good smooth regions but insufficient large gradients along edges and at high-frequency regions [Yang et al., 2014].

Outline

Edge-based Methods

Several SISR algorithms have been proposed to learn priors from edge features for reconstructing HR images [Yang et al., 2014]. [Fattal, 2007] proposed the depth and width feature of edges. [Sun et al., 2008] sugguested using gradient profiles.

Outline

Outline

Outline

Architecture of SR-GAN[2]

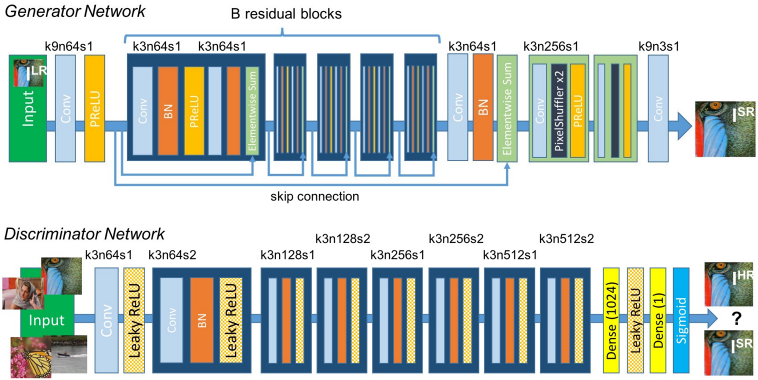


Figure 4: Architecture of Generator and Discriminator Network with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer.

[2] [Ledig et al., 2017]

Architecture of Super-FAN[3]

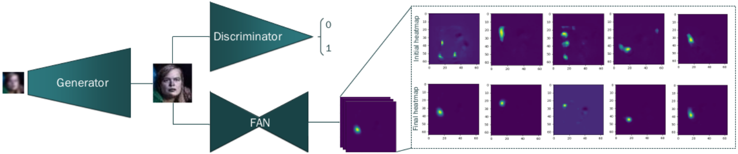


Figure 2: The proposed Super-FAN architecture comprises three connected networks: the first network is a newly proposed Super-resolution network (see sub-section 4.1). The second network is a WGAN-based discriminator used to distinguish between the super-resolved and the original HR image (see sub-section 4.2). The third network is FAN, a face alignment network for localizing the facial landmarks on the super-resolved facial image and improving super-resolution through a newly-introduced heatmap loss (see sub-section 4.3).

[3] [Bulat and Tzimiropoulos, 2017]

Outline

Pixel loss

Pixel Loss

Given a low resolution image I^{LR} and the corresponding high resolution image I^{HR}, pixel-wise MSE loss is used to minimize the distance between I^{LR} and I^{HR}.

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

where W and H denote and size of I^{LR} and r is the upsampling factor.

Perceptual Loss

The pixel-wise MSE loss achieves high PSNR values, however, it often results in blurry and unrealistic images. To address this issue, [Johnson et al., 2016, Ledig et al., 2017] proposed a perceptual loss where the super-resolved image and the original image must also be close in feature space.

Feature Reconstruction Loss

The loss over the ResNet features at a given level i is defined as

$$l_{feature/i} = \frac{1}{W_i H_i} \sum_{x=1}^{W_i} \sum_{y=1}^{H_i} (\phi_i(I^{HR})_{x,y} - \phi_i(G_{\theta G}(I^{LR}))_{x,y})^2 \tag{2}$$

where ϕ_i denotes the feature map obtained after the last convolutional layer of the i^{th} block, and W_i and H_i are its size.

Feature Reconstruction Loss[4]

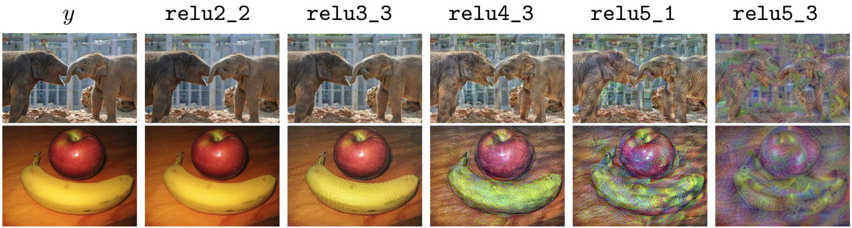


Fig. 3. Similar to [6], we use optimization to find an image $\hat{y}$ that minimizes the feature reconstruction loss $\ell_{feat}^{\phi,j}(\hat{y}, y)$ for several layers j from the pretrained VGG-16 loss network ϕ. As we reconstruct from higher layers, image content and overall spatial structure are preserved, but color, texture, and exact shape are not.

[4] [Johnson et al., 2016]

Style Reconstruction Loss[5]

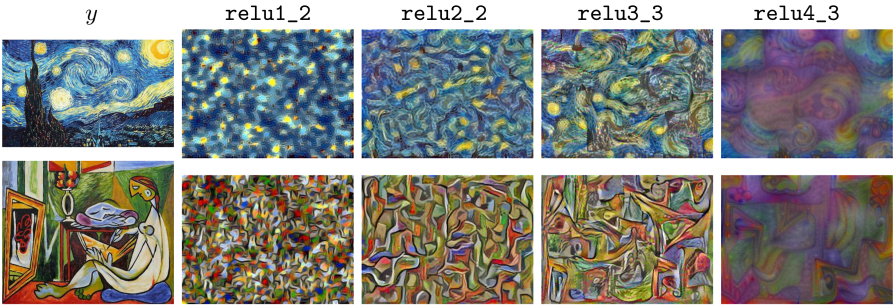


Fig. 4. Similar to [10], we use optimization to find an image $\hat{y}$ that minimizes the style reconstruction loss $\ell_{style}^{\phi,j}(\hat{y}, y)$ for several layers j from the pretrained VGG-16 loss network ϕ. The images $\hat{y}$ preserve stylistic features but not spatial structure.

[5][Johnson et al., 2016]

WGAN Loss

Wasserstein GAN Loss

$$l_{\mathrm{WGAN}} = \mathop{\mathbb{E}}_{\hat{I} \sim \mathbb{P}_g} [D(\hat{I})] - \mathop{\mathbb{E}}_{\hat{I} \sim \mathbb{P}_r} [D(I^{HR})] + \lambda \mathop{\mathbb{E}}_{\hat{I} \sim \mathbb{P}_{\hat{I}}} [(||\nabla_{\hat{I}} D(\hat{I})||_2 - 1)^2] \quad (3)$$

where $\mathbb{P}_r$ is the data distribution and $\mathbb{P}_g$ is the generator G distribution defined by $\hat{I} = G(I^{LR})$. $\mathbb{P}_{\hat{I}}$ is obtained by uniformly sampling along straight lines between pairs of samples from $\mathbb{P}_r$ and $\mathbb{P}_g$.

Heatmap Loss

[Bulat and Tzimiropoulos, 2017] proposed a heatmap loss to enforce structural consistency between the super-resolved and the corresponding HR facial image.

Heatmap Loss

$$l_{heatmap} = \frac{1}{N} \sum_{n=1}^{N} \sum_{i,j} (\tilde{M}_{i,j}^{n} - \bar{M}_{i,j}^{n})^2 \tag{4}$$

where $\tilde{M}_{i,j}^{n}$ is the heatmap corresponding to the n^{th} landmark at pixel (i, j) produced by running the FAN on the super-resolved image $\bar{I}^{HR}$, and $\bar{M}_{i,j}^{n}$ is obtained by running another FAN on the original image I^{HR}.

Outline

Peak Signal to Ratio (PSNR)[6]

PSNR

Given a reference image f and a test image g, both of size $M \times N$, the PSNR (in dB) between f and g is defined as

$$PSNR = 10 \cdot \log_{10}(\frac{MAX_I^2}{MSE}) \tag{5}$$

where

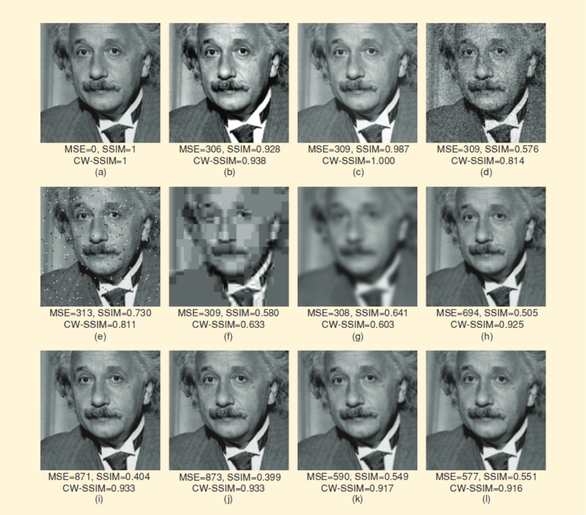
$$MSE(f, g) = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (f_{i,j} - g_{i,g})^2 \tag{6}$$

MAX_I is the maximum possible pixel value of the image. For example, when the pixels are represented using 8 bits per sample (grey-level), $MAX_I = 255$.

A small value of the PSNR implies high numerical differences between images (not necessarily to be of low quality!).

[6] [Hore and Ziou, 2010]

What's wrong with the MSE?[7]



[FIG2] Comparison of image fidelity measures for "Einstein" image altered with different types of distortions. (a) Reference image. (b) Mean contrast stretch. (c) Luminance shift. (d) Gaussian noise contamination. (e) Impulsive noise contamination. (f) JPEG compression. (g) Blurring. (h) Spatial scaling (zooming out). (i) Spatial shift (to the right). (j) Spatial shift (to the left). (k) Rotation (counter-clockwise). (l) Rotation (clockwise).

[7] [Wang and Bovik, 2009]

Structural Similarity Index Measure (SSIM) I

[Wang et al., 2004] separates the task of similarity measurement into three comparisons:

- Luminance
- Contrast
- Structure

Structural Similarity Index Measure (SSIM) II

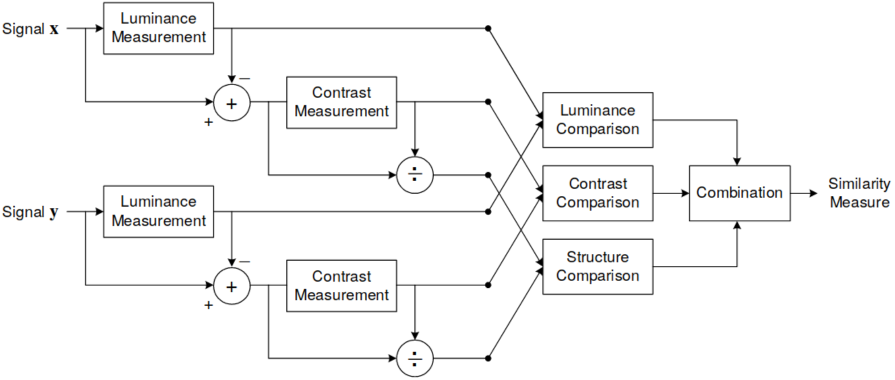


Fig. 3. Diagram of the structural similarity (SSIM) measurement system.

Structural Similarity Index Measure (SSIM) III

Luminance comparison function $l(x, y)$

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{7}$$

The luminance comparison function $l(x, y)$ is a function of μ_x and μ_y.

Contrast comparison $c(x, y)$

$$\sigma_x = \Big(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2\Big)^{1/2} \tag{8}$$

The contrast comparison $c(x, y)$ is the comparison of σ_x and σ_y.

Structural Similarity Index Measure (SSIM) IV

Structure comparison $s(x, y)$

$$\frac{x - \mu_x}{\sigma_x} \tag{9}$$

$$\frac{y - \mu_y}{\sigma_y} \tag{10}$$

The structure comparison $s(x, y)$ is conducted on these normalized signals.

Structural Similarity Index Measure (SSIM) V

SSIM

These three components are combined to generate an overall similarity measure

$$S(x,y) = f(l(x,y), c(x,y), s(x,y)) \tag{11}$$

The general form of the Structural SIMilarity (SSIM) index between signal x and y is defined as

$$SSIM(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma} \tag{12}$$

Specifically, when $\alpha = \beta = \gamma = 1$, the resulting SSIM index is

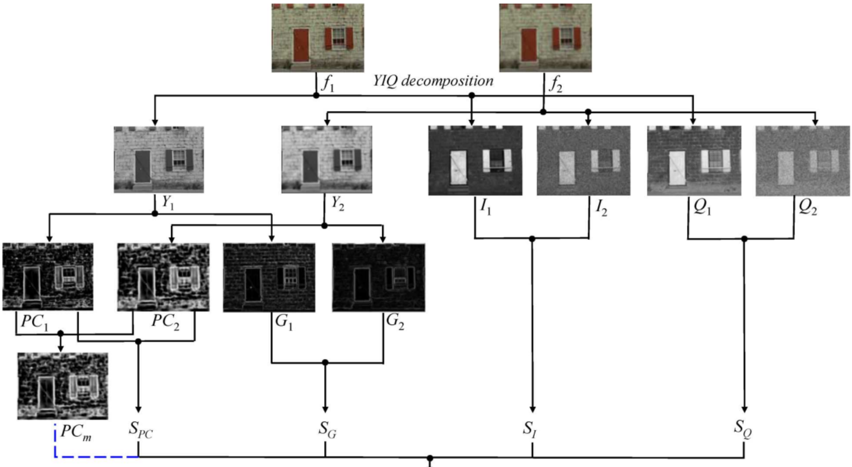
$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{13}$$

where C_1, C_2, and C_3 are small constants.

[7] [Wang et al., 2004, Wang and Bovik, 2009]

Fig. 2. Comparison of "Boat" images with different types of distortions, all with MSE = 210. (a) Original image (8bits/pixel; cropped from 512×512 to 256×256 for visibility); (b) Contrast stretched image, MSSIM = 0.9168; (c) Mean-shifted image, MSSIM = 0.9900; (d) JPEG compressed image, MSSIM = 0.6949; (e) Blurred image, MSSIM = 0.7052; (f) Salt-pepper impulsive noise contaminated image, MSSIM = 0.7748.

Feature Similarity Index Measure (FSIM)[8]



$$\mathrm{FSIM}_C = \frac{\sum_{\Omega} S_{PC}(\mathbf{x}) \cdot S_G(\mathbf{x}) \cdot \left[S_I(\mathbf{x}) \cdot S_Q(\mathbf{x}) \right]^{\lambda} \cdot PC_m(\mathbf{x})}{\sum_{\Omega} PC_m(\mathbf{x})}$$

Fig. 2. Illustration for the FSIM/FSIM$_C$ index computation. f_1 is the reference image, and f_2 is a distorted version of f_1.

[8] [Zhang et al., 2011]

PSNR, SSIM, FSIM versus human and deep nets (LPIPS)[9]

Figure 1: **Which patch (left or right) is "closer" to the middle patch in these examples?** In each case, the traditional metrics (L2/PSNR, SSIM, FSIM) disagree with human judgments. But deep networks, even across architectures (Squeezenet [20], AlexNet [27], VGG [51]) and supervision type (supervised [46], self-supervised [13, 40, 42, 63], and even unsupervised [26]), provide an *emergent embedding* which agrees surprisingly well with humans. We further calibrate existing deep embeddings on a large-scale database of perceptual judgments; models and data can be found at `https://www.github.com/richzhang/PerceptualSimilarity`.

[9] [Zhang et al., 2018]

Outline

Comparison of bicubic, SRResNet, SRGAN[10]

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]

[10] [Ledig et al., 2017]

Comparison of SRGAN and Super-FAN[11]

Figure 1: A few examples of visual results produced by our system on real-world low resolution faces from WiderFace.

[11] [Bulat and Tzimiropoulos, 2017]

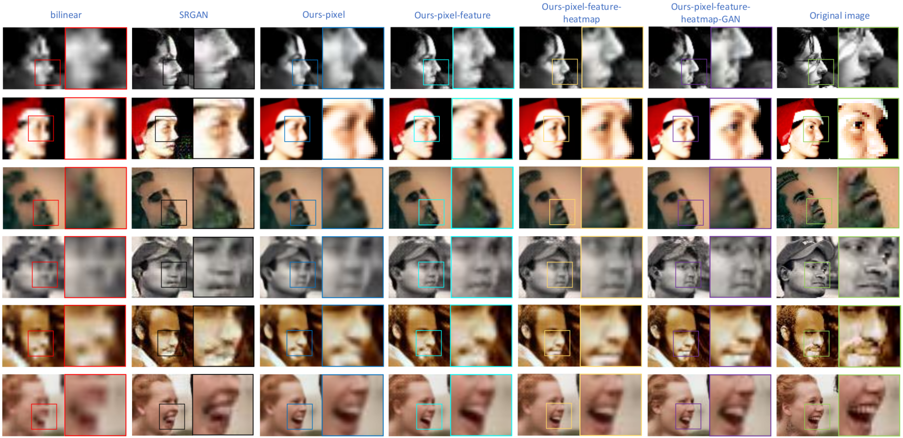


Figure 4: Visual results on LS3D-W. Notice that: (a) The proposed Ours-pixel-feature already provides better results than those of SR-GAN [20]. (b) By additionally adding the newly proposed heatmap loss (Ours-pixel-feature-heatmap) the generated faces are better structured and look far more realistic. Ours-pixel-feature-heatmap-GAN is Super-FAN which improves upon Ours-pixel-feature-heatmap by adding the GAN loss and by end-to-end training. Best viewed in electronic format.

Outline

Challenges: from toy data to real-world problems[13]

- Computation Efficiency
- Robustness
- Real-world Performance[12]
- And more . . .

[12] Higher MSE does not have to be visually more appealing! Bicubic interpolation usually achieves smaller MSE compared with those recovered by some example-based approaches [Yang et al., 2010].

[13] [Huang and Yang, 2010]

Summary

References I

Bulat, A. and Tzimiropoulos, G. (2017).
Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans.
arXiv preprint arXiv:1712.02765.

Fattal, R. (2007).
Image upsampling via imposed edge statistics.
In *ACM transactions on graphics (TOG)*, volume 26, page 95. ACM.

Hore, A. and Ziou, D. (2010).
Image quality metrics: Psnr vs. ssim.
In *Pattern recognition (icpr), 2010 20th international conference on*, pages 2366–2369. IEEE.

Huang, T. and Yang, J. (2010).
Image super-resolution: Historical overview and future challenges.
In *Super-resolution imaging*, pages 19–52. CRC Press.

Irani, M. and Peleg, S. (1991).
Improving resolution by image registration.
CVGIP: Graphical models and image processing, 53(3):231–239.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016).
Perceptual losses for real-time style transfer and super-resolution.
In *European Conference on Computer Vision*, pages 694–711. Springer.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A. P., Tejani, A., Totz, J., Wang, Z., et al. (2017).
Photo-realistic single image super-resolution using a generative adversarial network.
In *CVPR*, volume 2, page 4.

References II

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

Wang, Z. and Bovik, A. C. (2009).
Mean squared error: Love it or leave it? a new look at signal fidelity measures.
IEEE signal processing magazine, 26(1):98–117.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004).
Image quality assessment: from error visibility to structural similarity.
IEEE transactions on image processing, 13(4):600–612.

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

Yang, J., Wright, J., Huang, T. S., and Ma, Y. (2010).
Image super-resolution via sparse representation.
IEEE transactions on image processing, 19(11):2861–2873.

Zhang, L., Zhang, L., Mou, X., Zhang, D., et al. (2011).
Fsim: a feature similarity index for image quality assessment.
IEEE transactions on Image Processing, 20(8):2378–2386.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018).
The unreasonable effectiveness of deep features as a perceptual metric.
arXiv preprint.