
Mehdi S. M. Sajjadi  Bernhard Schölkopf  Michael Hirsch
Max Planck Institute for Intelligent Systems, Tübingen, Germany
msajjadi.com

Overview
- Single-image super-resolution (SISR) is the task of inferring a high-resolution image from a single low-resolution input. It is a classic problem that has been well-studied in the past.
- Convolutional neural networks have achieved state-of-the-art results in recent years.
- Performance is traditionally measured using PSNR which has been shown to correlate poorly with the human perception of image quality. Thus, models trained for PSNR produce over-smoothed images that lack high-frequency textures.

Method
- We propose a novel application of automated texture synthesis in combination with a perceptual loss focusing on creating realistic textures rather than optimizing for a pixel-accurate reproduction of ground truth images during training.
- By using feed-forward fully convolutional neural networks in an adversarial training setting, we achieve a significant boost in image quality at high magnification ratios.
- Extensive experiments on a number of datasets show the effectiveness of our approach, yielding state-of-the-art results in both quantitative and qualitative benchmarks.

Illustration of the effect of different losses. The optimal solution under the Euclidean loss is the mean of all possible images, leading to blurry textures. Instead, our aim is to generate a pattern visually indistinguishable from the original, explicitly allowing large pixel-wise errors.

<table>
<thead>
<tr>
<th>Network</th>
<th>ENet-P</th>
<th>ENet-EA</th>
<th>ENet-PAT</th>
<th>ENet-PA + texture</th>
<th>ENet-P + adv.</th>
<th>ENet-EA + adv.</th>
<th>ENet-PA + adv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-5 error</td>
<td>0.427</td>
<td>0.438</td>
<td>0.456</td>
<td>0.471</td>
<td>0.484</td>
<td>0.497</td>
<td>0.506</td>
</tr>
<tr>
<td>Top-1 error</td>
<td>0.226</td>
<td>0.242</td>
<td>0.246</td>
<td>0.264</td>
<td>0.283</td>
<td>0.312</td>
<td>0.320</td>
</tr>
</tbody>
</table>

Problem formulation. A high resolution image \( I_{HR} \) is downsampled to a low resolution image \( I_{LR} \) with \( I_{LR} = \frac{1}{2} I_{HR} \). The task of SISR is to provide an approximate inverse \( f = \hat{I}_{HR} \): \( \hat{I}_{LR} = \hat{I}_{HR} \) that looks similar to the original image \( I_{HR} \). As it is non-injective, the task is ill-posed.

Approach. We use a feed-forward fully convolutional neural network as a function approximator for \( f \) which is trained to minimize several combinations of the following loss functions.

- (E)uclidean loss in space / MSE. The most commonly used loss function for generative models.

\[ L_{\text{MSE}} = \| I_{LR} - I_{HR} \|_{2} \]

- (P)erceptual loss in feature space, \( \text{D}_{\text{VGG}} \).

\[ L_{\text{VGG}} = \sum_{i} \left( \text{D}_{\text{VGG}}(I_{LR}) - \text{D}_{\text{VGG}}(I_{HR}) \right)^{2} \]

- (T)exture matching loss, \( \text{D}_{\text{L1}} \).

\[ L_{\text{L1}} = \sum_{i} | \text{D}_{\text{L1}}(I_{LR}) - \text{D}_{\text{L1}}(I_{HR}) | \]

- (A)dversarial training, \( \text{D}_{\text{GAN}} \).

\[ L_{\text{GAN}} = \text{D}(\hat{I}_{LR}) - \text{D}(I_{HR}) \]

Results
- ENet-EAT captures the most realistic looking image. This is the most realistic looking image.

Object recognition benchmark
- Proposed benchmark for perceptual image quality assessment.
- ImageNet images are downsampled and upsampled with methods to be evaluated.
- Performance of pre-trained ResNet model correlates with the perceived image quality of the results. Can be used as a no-reference metric for image quality assessment.

Quiz: Which image was produced by which model?