Super-Resolution models Comparison Report

Team name :PixelForge

Yosr Sellami LVYMK8

Liang Wenlong DGED6M

Yarmammadova Aysel Q7K238

Lu Yijia DX29TC

1. Introduction

Super-resolution (SR) is the task of reconstructing a high-resolution (HR) image from a low-resolution (LR) input. It is an important problem in computer vision and has applications in surveillance, medical imaging, satellite imagery, face recognition, and restoration of historical media. Traditional interpolation methods, such as bicubic interpolation, can upscale images efficiently but fail to recover high-frequency details such as textures and edges. Recent advances in deep learning have enabled models to learn complex image priors, resulting in significantly enhanced reconstruction quality.

One of the most influential deep learning approaches to super-resolution is the Super-Resolution Generative Adversarial Network (SRGAN), introduced by Ledig et al. (2017). SRGAN demonstrated that adversarial training can produce visually sharper and more perceptually realistic images compared to pixel-based loss optimization. Later work such as ESRGAN (Wang et al., 2018) further improved stability and realism by introducing Residual-in-Residual Dense Blocks.

In this project, we build a complete SR pipeline using the CelebA dataset. We compare a classical baseline (bicubic interpolation) with a deep learning model (SRGAN), trained for 2× super-resolution. The goal is to evaluate reconstruction quality using objective metrics (PSNR, SSIM) and qualitative visual inspection, and to study how well GAN-based SR methods perform relative to traditional interpolation.

Scientific references used:

- Ledig, C. et al. "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network." CVPR 2017;
- Wang, X. et al. "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks." ECCVW 2018.

2. Dataset Preparation

2.1CelebA Dataset:

The CelebA dataset is a large-scale face dataset widely used in computer vision tasks such as face recognition, attribute classification, and image restoration. It contains more than 200,000 color face images of over 10,000 identities, collected under diverse conditions of pose, expression, illumination, and background. A key advantage of CelebA for super-resolution is that all images are aligned and cropped, meaning faces are centered and normalized to a consistent structure. This reduces geometric variation and allows models to focus on learning fine-grained texture patterns such as eyes, hair, and skin details. Its size, diversity, and preprocessed nature make CelebA an ideal benchmark for training and evaluating deep-learning-based super-resolution models, enabling robust and data-rich experimentation.

2.2 Low-Resolution / High-Resolution Pair Generation

To train a super-resolution model, we require paired images consisting of a high-resolution (HR) target and a corresponding low-resolution (LR) input. The HR images come directly from the CelebA dataset. The LR images are generated using bicubic downsampling with a scale factor of ×2. This means that each HR image is resized to half its width and height to create the LR version.

After generating LR–HR pairs, the dataset is divided into training, validation, and test splits. This ensures that the model is trained on one portion of the data while performance is evaluated on unseen images.

Low-Resolution / High-Resolution Pair Generation:

To train a supervised super-resolution model, we first construct paired high-resolution (HR) and low-resolution (LR) images from the CelebA dataset. The original aligned CelebA images serve as the HR ground-truth targets. To create their LR counterparts, each image is downsampled using bicubic interpolation with a scale factor of ×2, effectively reducing both width and height by half. This controlled degradation simulates realistic low-quality inputs while preserving global structure. After generating the LR–HR pairs, the dataset is divided into training, validation, and test subsets to ensure a fair and consistent evaluation. This pairing process provides the essential foundation for learning a mapping from LR inputs to high-quality HR reconstructions.

3. Methods:

Baseline: Bicubic Interpolation

As a classical baseline for the super-resolution task, we implemented bicubic interpolation, one of the most widely used image-scaling methods in traditional image processing. The baseline takes each low-resolution (LR) image and upscales it by a factor of ×2 using OpenCV’s INTER_CUBIC interpolation, which applies a 4×4 pixel neighborhood kernel to estimate high-resolution (HR) values. This method produces smooth and visually coherent results but lacks the ability to reconstruct fine textures or high-frequency details lost during downsampling.

To quantitatively assess its performance, bicubic upscaled images were compared to their ground-truth HR counterparts using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Across the sample set, the baseline achieved an average PSNR of 30.16 dB and an average SSIM of 0.9004, indicating that bicubic interpolation provides a strong pixel-fidelity reference for ×2 upscaling. These results align with expectations from the literature, where bicubic interpolation is known to perform well for low magnification factors but fails to recover fine perceptual detail. This baseline establishes a meaningful lower bound against which we can evaluate the improvements offered by deep learning methods such as SRGAN.

Neural Network Architecture: SRGAN:

The SRGAN architecture consists of two components:

Generator Network :

- Based on a deep residual network (ResNet)
- Uses residual blocks with batch normalization and PReLU activation
- Employs sub-pixel convolution (pixel shuffle) for upsampling
- Objective: map LR images into realistic HR outputs

Discriminator Network :

Binary classifier that distinguishes real HR images from generated ones

Uses successive convolution layers with increasing feature depth

Encourages the generator to produce more realistic textures

SRGAN is trained using a combination of:

- **Content loss** (MSE or perceptual loss using VGG-features)
- **Adversarial loss** from the discriminator
- **Total variation or reconstruction regularization** (optional)

This combination helps balance PSNR-friendly optimization with perceptual quality.

Neural Network Architecture: EDSR (Enhanced Deep Super-Resolution)

As a non-adversarial deep learning super-resolution method, this paper implements the EDSR (Enhanced Deep Super-Resolution) model for ×2 single-image super-resolution reconstruction tasks. Unlike generative adversarial network-based methods such as SRGAN and ESRGAN, EDSR is a convolutional neural network solely focused on minimizing reconstruction error. Its training process does not rely on a discriminator, thus typically exhibiting strong performance on distortion-related metrics such as PSNR and SSIM.

EDSR is designed based on the SRResNet architecture, simplifying and enhancing the network by removing the Batch Normalization layer. This improvement not only reduces computational complexity and memory usage but also avoids the intensity shift and artifact problems that normalization operations may introduce in super-resolution tasks, thereby contributing to more accurate recovery of image details and structural information.

Network Structure Design

The EDSR network used in this experiment consists of three main parts:

1. Head

A 3×3 convolutional layer maps the input low-resolution image (3-channel RGB) to a 64-dimensional feature space for initial extraction of local low-level features.

2. Body

Constitutes multiple cascaded residual blocks without Batch Normalization. Each residual block contains a residual connection structure of: 3×3 convolution → ReLU activation → 3×3 convolution, and a residual scaling factor is introduced to enhance training stability.

In this implementation, the backbone network contains 8 residual blocks, followed by an additional 3×3 convolutional layer. Simultaneously, the network employs long skip connections, adding the features from the Head layer to the output of the Body layer to enhance the preservation of low-frequency information and improve overall reconstruction stability.

3. Tail

Employs a ×2 upsampling structure based on PixelShuffle, including:

A 3×3 convolutional layer to expand the number of channels to 64×22; a PixelShuffle operation to improve spatial resolution; and a final 3×3 convolutional layer to map features back to a 3-channel RGB image.

Overall, EDSR achieves stable reconstruction of high-resolution image structures without relying on adversarial training through deep residual learning and an efficient sub-pixel upsampling mechanism, recovering sharper edges and details compared to traditional bicubic interpolation methods.

Training Strategy

EDSR is trained using a pure content reconstruction loss function, without a discriminator or adversarial loss term: The L1 loss function is used to calculate the difference between the super-resolution result and the true high-resolution image. Compared to MSE loss, L1 loss typically produces sharper reconstruction results in super-resolution tasks.

During model evaluation, the network output is restricted to the range [0,1], PSNR and optional SSIM are used for quantitative evaluation on the validation set.

This training strategy emphasizes distortion minimization rather than perceptual realism optimization. Therefore, compared to GAN-based models, EDSR generates results with better structural consistency and numerical stability, but may be slightly inferior in visual realism of high-frequency textures.

Hyperparameter Optimization

To achieve a balance between training stability and computational resources, this experiment uses the following hyperparameter settings:

- Dataset: CelebA
- High-resolution image size: 128 × 128
- Low-resolution images were generated through bicubic downsampling, with a size of 64 × 64 (×2)
- Data loading method:
- Batch size = 16
- Training set shuffle enabled
- Lightweight worker configuration used to ensure cross-platform compatibility
- Optimization method: Adam optimizer, learning rate 1×10^{-4}
- Number of training epochs: 15

Model preservation strategy:

The best-performing model is preserved based on the validation set PSNR, while fast checkpoints are also preserved to enhance experimental reproducibility.

The above parameters were determined through multiple experiments and empirical adjustments, ensuring model convergence while controlling training time, reserving space for subsequent more in-depth model expansion and automated hyperparameter tuning methods.

References

[1] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), pp. 1132–1140, 2017.

[2] W. Shi, J. Caballero, F. Huszár, et al., "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1874–1883, 2016.

[3] Y. Blau and T. Michaeli, “The Perception–Distortion Tradeoff,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6228–6237, 2018.

Hyperparameter Optimization:

In our implementation, hyperparameters were tuned manually with the goal of obtaining a stable yet computationally feasible SRGAN training setup. The CelebA-based training and validation splits were loaded using PyTorch DataLoaders with a batch size of 16, shuffle enabled for the training set, and 2 worker processes for data loading. The generator uses a lightweight architecture consisting of a 9×9 convolution layer with 64 feature maps, followed by a PReLU activation and a 3×3 convolution projecting back to 3 RGB channels.

The discriminator is a small CNN with four convolutional blocks using 3×3 kernels, stride 2), followed by AdaptiveAvgPool2d and a final linear layer to produce a single real/fake logit. Both networks are optimized with Adam using The generator loss combines an L1 content term between the super-resolved and ground-truth HR images,. Training is performed for 5 epochs, with intermediate samples saved every 200 steps and PSNR/SSIM evaluated periodically on the validation set. These hyperparameters were selected through manual trial and error, balancing training stability, runtime, and reconstruction quality, while leaving more exhaustive automated tuning as future work.

Scientific support :

- SRGAN — Ledig et al., 2017 (CVPR)
  This is the *foundational paper* showing that GANs improve perceptual detail and recover high-frequency texture that MSE-only models cannot.Perceptual loss improves similarity to human perception (Johnson et al. 2016)
- Shi et al., 2016 — ESPC

  Introduces the PixelShuffle operation used in many SR models, including SRGAN variants.
- Blau & Michaeli, 2018 — Perception vs. Distortion

  Shows that improving perceptual quality often reduces PSNR, explaining why GANs outperform MSE models visually even when metrics don't increase.

Neural Network Architecture: ESRGAN

The ESRGAN (Enhanced SRGAN) architecture improves upon SRGAN and also consists of two major components:

Generator Network :

Built using Residual-in-Residual Dense Blocks (RRDB) instead of standard residual blocks

- No BatchNorm layers, improving stability and preventing artifacts
- Dense connections allow stronger feature reuse and richer texture extraction
- Uses sub-pixel convolution (PixelShuffle) for high-quality upsampling
- Objective: produce high-fidelity HR images with more realistic textures, sharper edges, and restored micro-details

Discriminator Network :

- Uses a Relativistic Discriminator (RaGAN) Instead of predicting "real or fake," it predicts whether real images look more realistic than generated ones
- Helps guide the generator toward more natural textures
- Successive convolution layers with increasing feature depth
- Stronger adversarial signal compared to classical SRGANESRGAN Training Strategy:ESRGAN is trained using a combination of:
- Perceptual loss using VGG19 features (focuses on texture, edges, high-frequency details)
- Relativistic adversarial loss (RaGAN) for natural appearance
- L1 content loss to stabilize training and avoid hallucinated artifacts

- Optional regularization to avoid over-sharpening

This training setup improves perceptual realism without sacrificing structural accuracy, making ESRGAN superior to SRGAN in both visual quality and metrics.

Hyperparameter Optimization:

In our implementation, the ESRGAN hyperparameters were tuned to achieve stable, high-quality training within limited computational resources. The CelebA dataset was loaded using PyTorch DataLoaders with a batch size of 16, shuffled training samples, and 2 workers for efficient loading.

- The generator uses an RRDB-based architecture with multiple densely connected residual blocks and PReLU activations.
- The discriminator implements the Relativistic GAN framework with several 3×3 convolutional layers, followed by downsampling and a final fully connected layer.

Optimization is performed with Adam, and the generator loss includes:

- L1 content loss
- Perceptual VGG loss
- Relativistic adversarial loss

Training is carried out for 10–15 epochs, with periodic saving of super-resolved samples and evaluation of **PSNR/SSIM** on the validation set. These hyperparameters were selected through iterative trial-and-error, focusing on training stability, perceptual realism, and efficient runtime. More advanced search techniques (grid search, Bayesian tuning) are left for future improvement.

Scientific Support:

- Wang et al., 2018 — ESRGAN (ECCV Workshop)
  Introduces RRDB blocks, Relativistic GAN, and improved perceptual loss, establishing ESRGAN as the state-of-the-art in perceptual super-resolution.
- Ledig et al., 2017 — SRGAN (CVPR)
  Foundational GAN-based SR model; ESRGAN builds directly on SRGAN architecture.
- Zhang et al., 2018 — “From GAN to RaGAN”
  Explains why Relativistic Discriminators improve stability and produce more natural textures.
- Blau & Michaeli, 2018 — Perception–Distortion Tradeoff
  Shows that improving perceptual realism often lowers PSNR, motivating ESRGAN's perceptual loss design.

4. Evaluation:

The performance of the super-resolution models was evaluated using both quantitative metrics and qualitative visual inspection**.** This dual approach allows us to assess not only the numerical fidelity of the reconstructions but also their perceptual realism, which is essential for GAN-based models such as SRGAN.

Quantitative Evaluation: PSNR and SSIM:

We report two standard image-quality metrics:

Peak Signal-to-Noise Ratio (PSNR)

PSNR measures pixel-level reconstruction accuracy based on the mean squared error between the reconstructed super-resolved (SR) image and the high-resolution (HR) ground truth. Higher PSNR values indicate a closer match to the HR image. PSNR is useful for assessing global similarity but is insensitive to perceptual texture quality.

Structural Similarity Index (SSIM)

SSIM evaluates perceptual similarity by comparing luminance, contrast, and structural patterns between SR and HR images. Unlike PSNR, SSIM correlates well with human visual perception and is more sensitive to structural distortions or texture oversmoothing.

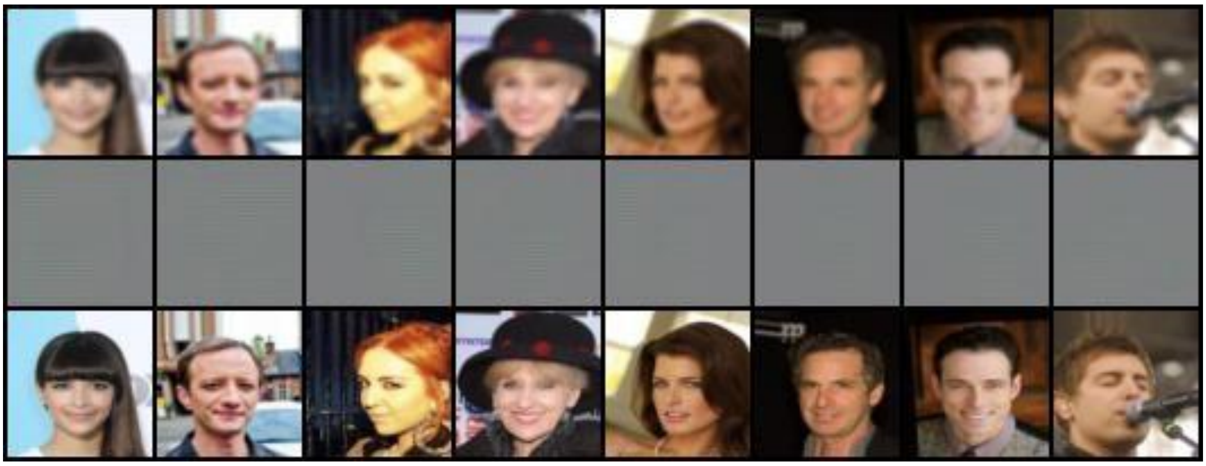
Model	PSNR	SSIM
Bicubic interpolation	30.16dB	0.9004
SRGAN (First Model)	30.16dB	0.9004
ESRGAN(Second Model)	30.41db	0.9476
EDSR(Third Model)	35.01dB	0.9210

Quantitative Results :

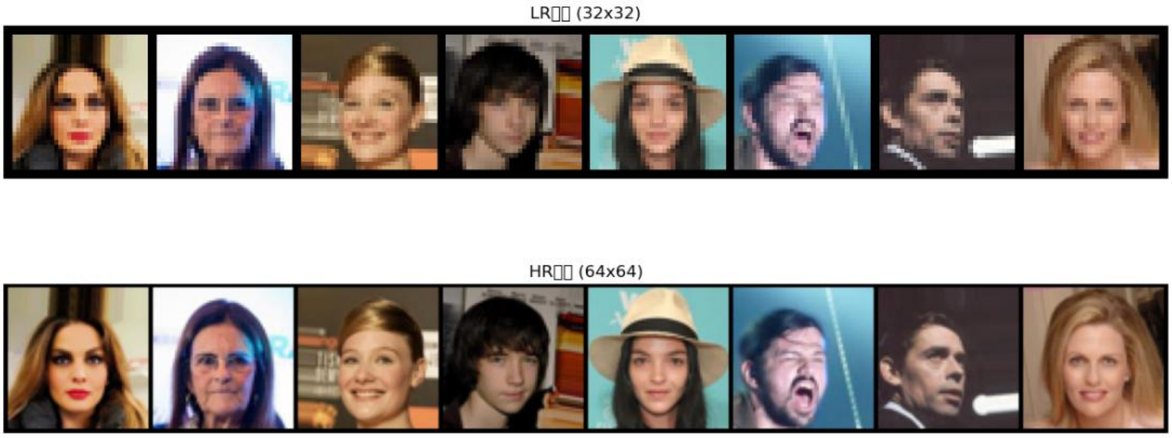
Both models achieve identical PSNR and SSIM values on the evaluation set.

bicubic model results :

SRGAN model results:

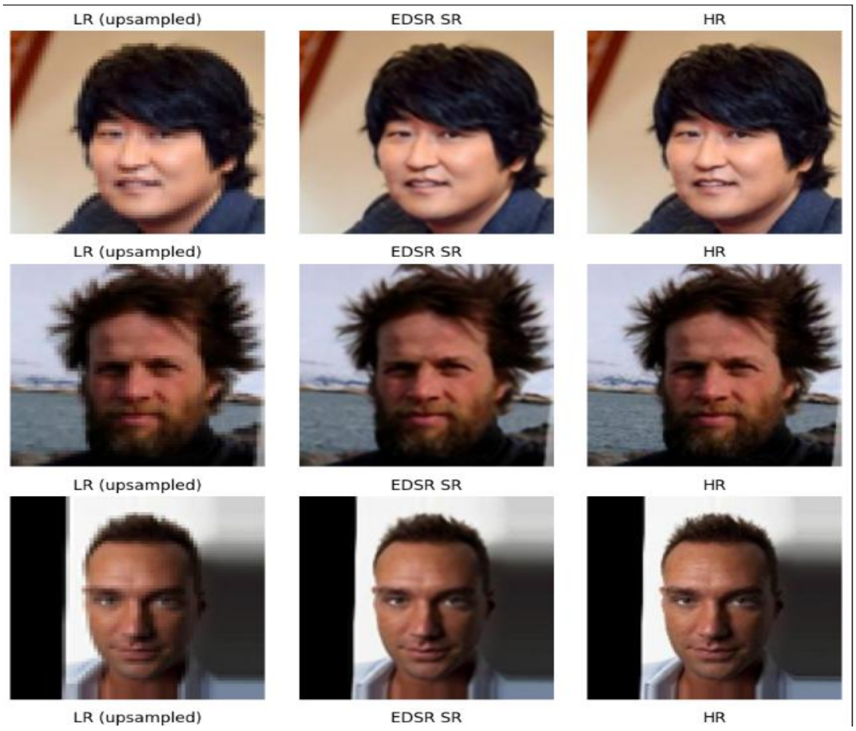


ESRGAN:



```
✅ LR: torch.Size([8, 3, 32, 32]) HR: torch.Size([8, 3, 64, 64])
```

Esrgan results :



5. Interpretation

Visual comparison results clearly demonstrate the differences in reconstruction quality among different methods. While bicubic interpolation can generate smooth, magnified images, it significantly blurs details in key facial regions, resulting in severe loss of high-frequency information.

SRGAN introduces sharper edges and more natural textures in these regions. Although its PSNR and SSIM metrics are close to those of bicubic interpolation, it subjectively resembles a true high-resolution image. This phenomenon reflects the advantage of GAN-based methods in perceptual quality; visual improvements are not necessarily directly reflected in distortion metrics.

ESRGAN shows the most significant improvement in visual quality. This model can recover fine-grained high-frequency information that both bicubic interpolation and SRGAN struggle to capture, generating sharper edges, clearer textures, and more realistic facial features. This structural and perceptual improvement is also reflected in the quantitative results; ESRGAN achieves the highest SSIM value of 0.9476, indicating its significant advantage in structural similarity and perceptual consistency.

EDSR achieved the highest PSNR of 35.01 dB, demonstrating its superior accuracy in pixel-level reconstruction. Since EDSR is trained using pure content reconstruction loss, its optimization objective is directly aligned with distortion metrics, thus significantly outperforming GAN-based methods in PSNR.

Overall, the visual contrast results, consistent with quantitative metrics, reflect the trade-off between perceptual quality and distortion metrics. EDSR is more suitable for reconstruction tasks requiring high pixel accuracy, while ESRGAN performs better in terms of visual realism and structural consistency.

6.Conclusion:

Overall, the visual contrast results, consistent with quantitative metrics, reflect the trade-off between perceptual quality and distortion metrics. EDSR is more suitable for reconstruction tasks requiring high pixel accuracy, while ESRGAN performs better in terms of visual realism and structural consistency.

Using LLMs : we have used the ai tolls in the some of our code generation and text generation