Image Super Resolution using SR-ResNet

Shreyash Dwivedi
Indian Institute of Technology Kanpur

Kumar Aditya
Indian Institute of Technology Kanpur

Ganesh Srivathshava
Indian Institute of Technology Kanpur

Abstract

This project investigates the enhancement of Super-Resolution Residual Networks (SR-ResNet) for high-quality single image super-resolution, utilizing the CelebA-HQ dataset. We explore and compare three variants of the SR-ResNet architecture:

- *a vanilla SR-ResNet model trained using a composite loss function that combines Mean Squared Error (MSE), perceptual loss from a pre-trained VGG network, and Structural Similarity Index (SSIM) loss*
- *an attention-enhanced SR-ResNet, which introduces spatial focus mechanisms while maintaining the same loss function*
- *an attention-augmented SR-ResNet trained with an advanced hybrid loss function which integrates LPIPS perceptual similarity, SSIM, Charbonnier loss, and an edge-aware term based on Sobel gradients*

All models are implemented using PyTorch and trained on the CelebA-HQ dataset. Extensive evaluation using MSE, PSNR, SSIM, and qualitative inspection demonstrates that the combination of attention mechanisms and the hybrid loss leads to significantly improved visual quality and fidelity in the super-resolved outputs reaching a ***training loss of 0.106****,* ***validation loss of 0.1225****,* ***training PSNR of 28.19****,* ***validation PSNR of 27.85****,* ***training SSIM of 0.814****, and* ***validation SSIM of 0.8004*** *within just* ***20 epochs****.*

1. Introduction

Image Super-Resolution (ISR) refers to the task of reconstructing a high-resolution (HR) image from a low-resolution (LR) input, aiming to recover spatial details lost due to factors such as sensor limitations or compression. It is a fundamentally ill-posed problem, as multiple HR images can correspond to the same LR image [5]. Nevertheless, ISR plays a vital role in numerous real-world applications.

In medical imaging, super-resolution techniques can enhance low-quality scans, helping reveal anatomical structures crucial for accurate diagnosis [1]. In video surveillance, ISR improves the clarity of facial features or license plates captured under poor conditions [6]. Remote sensing relies heavily on ISR to sharpen satellite images, supporting tasks such as land cover classification, urban planning, and environmental monitoring [9].

Despite these applications, many ISR methods rely on optimizing pixel-wise losses like Mean Squared Error (MSE), which often produce overly smooth results and fail to capture fine textures or perceptual cues. This leads to high PSNR but low perceptual quality — a gap highlighted by Ledig et al. [15] with the introduction of perceptual-driven losses and GAN-based SR. The shift toward perceptual metrics (e.g., LPIPS [26], SSIM [22]) underscores the need for losses that better align with human visual perception.

Thus, the importance of ISR lies not only in achieving pixel-level accuracy but also in restoring semantically meaningful and perceptually convincing details — especially critical in domains such as face hallucination or identity-preserving SR, where small texture changes can significantly affect downstream recognition performance [25].

This project addresses these challenges by developing a novel SR framework that combines attention mechanisms with a perceptually enriched hybrid loss function. The goal is to enhance both the fidelity and realism of reconstructed images, particularly in face-centric applications.

1.1. Related Work

Early approaches to image super-resolution (SR) primarily relied on interpolation techniques such as bicubic or Lanczos interpolation, which are computationally efficient but often fail to reconstruct fine details and high-frequency

textures [12]. To address these limitations, learning-based methods introduced priors and sparse representations. Yang et al. [24] proposed sparse coding-based SR, where HR patches were reconstructed using overcomplete dictionaries learned from image pairs.

The advent of deep learning led to a paradigm shift in SR research. Dong et al. [4] introduced SRCNN, the first CNN-based method for SR, significantly outperforming previous handcrafted approaches. Deeper networks like VDSR [13] and EDSR [18] further improved performance by expanding receptive fields and refining architectural choices.

To improve perceptual realism, Ledig et al. [15] proposed SRGAN, which utilized a perceptual loss and adversarial learning to generate photo-realistic textures. ESRGAN [21] advanced this line by incorporating Residual-in-Residual Dense Blocks (RRDB) and a relativistic GAN loss, yielding state-of-the-art results.

Attention-based mechanisms have also shown promise. RCAN [27] introduced channel attention modules to adaptively rescale feature maps, while SAN [3] leveraged non-local attention to model long-range dependencies. More recently, SwinIR [17] combined hierarchical vision transformers with convolutional structures to enhance both fidelity and perceptual quality.

Loss functions have also evolved to better align with human visual perception. Perceptual losses using features from pretrained VGG networks were proposed by Johnson et al. [10], while LPIPS [26] provided a more accurate perceptual metric. SSIM-based losses [22] and edge-aware regularization [15] have also been used to preserve structural information.

In face super-resolution, identity preservation becomes crucial. Methods like FSRNet [2] and DFDNet [16] leverage facial priors such as landmarks and parsing maps to generate identity-consistent outputs even at high upscaling factors.

Despite these advances, a trade-off between perceptual quality and pixel-level accuracy persists. This motivates research into hybrid approaches that combine architectural innovations (e.g., attention and transformers) with perceptually motivated loss functions to better bridge the gap between visual quality and numerical precision.

1.2. Contribution

In this work, we aim to systematically evaluate and enhance the performance of Super-Resolution (SR) models by exploring the design space of loss functions and architectural refinements within the SR-ResNet framework [15]. Our key contributions are as follows:

- **Baseline Reproduction with Vanilla SR-ResNet:**
  We begin by implementing a Vanilla SR-ResNet model, trained using a composite perceptual loss function incorporating pixel-wise MSE, SSIM [22], and VGG-based perceptual similarity [10]. This provides a strong baseline on the CelebA-HQ dataset [11].

- **Attention-Augmented SR-ResNet:**
  We introduce spatial and/or channel-wise attention mechanisms [8, 23] into the SR-ResNet architecture to help the model dynamically focus on structurally and perceptually significant regions. This variant improves feature representation without increasing loss complexity.

- **Design of a Rich, Hybrid Loss Function:**
  We develop a novel hybrid loss function — combining LPIPS [26], SSIM [22], Charbonnier [14], and edge-aware losses [28] — to better preserve textures, perceptual sharpness, and structural fidelity. This loss function leads to significant improvements in perceptual quality as validated through PSNR, SSIM, and MSE metrics.

- **Empirical Evaluation Across Variants:**
  Through a thorough comparative study of the three SR-ResNet variants, we highlight the trade-offs between architectural complexity and loss sophistication. The study offers insights into how each component contributes to performance gains.

- **Open and Modular Codebase:**
  We provide a clean, modular PyTorch implementation with all three variants and training pipelines made openly available for further research and reproducibility.

In summary, our contributions target both architectural enhancements and loss function innovation within the SR-ResNet framework. By systematically analyzing each component's impact, we not only improve perceptual quality but also provide reproducible and extensible baselines for future research. A visual illustration of the output generated by the proposed models is shown in Figure 1.

2. Proposed Method

We propose a deep learning-based image super-resolution approach optimized by a novel ***Hybrid Loss Function incorporating Perceptual Similarity*** combined with an ***Attention-based SR-ResNet architechure***. We also compared the results with a ***Vanilla Loss Function*** which combines the MSE, SSIM and VGG Perceptual loss. These loss formulations aim to enhance perceptual fidelity, structural similarity, and edge-preserving properties of the super-resolved images.

Let $\mathbf{I}_{HR}$ be the high-resolution ground truth image and $\mathbf{I}_{SR} = \mathcal{F}_\theta(\mathbf{I}_{LR})$ be the network output from a low-resolution input $\mathbf{I}_{LR}$.

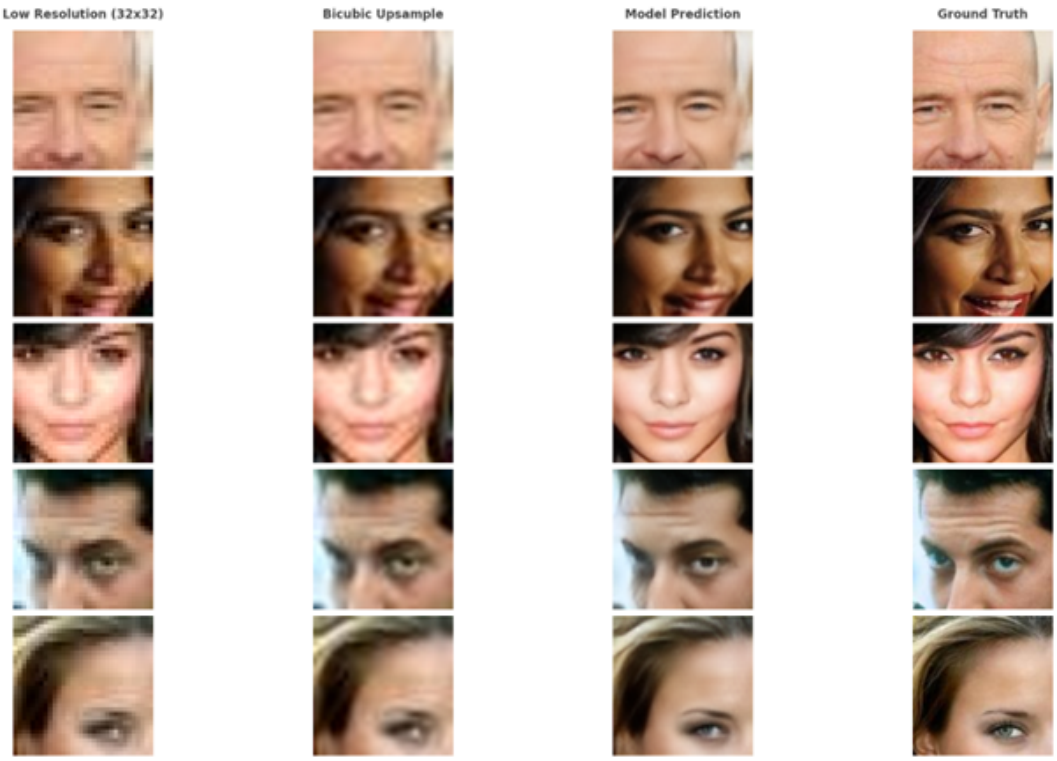


Figure 1. Sample outputs using our **attention-based model with hybrid perceptual-similarity based loss function** and compared with **bicubic sampling**

2.1. Vanilla SR-ResNet-Based Architecture

The baseline model in our pipeline is based on a modified SR-ResNet architecture, tailored for image super-resolution tasks. Given a low-resolution image $\mathbf{I}_{LR} \in \mathbb{R}^{C\times H\times W}$, the objective is to learn a mapping $\mathcal{F}_\theta$ such that the super-resolved output $\mathbf{I}_{SR} = \mathcal{F}_\theta(\mathbf{I}_{LR}) \in \mathbb{R}^{C\times sH\times sW}$ closely resembles the ground truth high-resolution image, where $s = 4$ is the *scale factor*. The model takes a **3x32x32** RGB image as input and return a **3x128x128** RGB as the output. We have used *PReLU* [7] as the activation function. The architecture consists of the following parts as described below and in Figure 2:

Initial Feature Extraction: The network begins with a convolutional layer with a large 9×9 kernel to capture low-level features over a wide receptive field:

$$F_1 = \text{PReLU}(\text{Conv}_{9\times 9}(\mathbf{I}_{LR}))$$

Residual Learning: We use $N = 16$ residual blocks, each without Batch Normalization, defined as:

$$F^{(l)} = F^{(l-1)} + \text{Conv}_{3\times 3}(\text{PReLU}(\text{Conv}_{3\times 3}(F^{(l-1)})))$$

These blocks encourage identity mapping and improve gradient flow. Let $\mathcal{R}(F_1)$ denote the output of the residual sequence:

$$F_{\text{res}} = \mathcal{R}(F_1)$$

Feature Aggregation: A 3×3 convolutional layer follows the residual blocks and is added to the initial features via a skip connection:

$$F_2 = \text{Conv}_{3\times 3}(F_{\text{res}}), \quad F_{\text{skip}} = F_2 + F_1$$

Upsampling: Two pixel-shuffle blocks [19], each doubling the resolution, are used to achieve $4\times$ upscaling:

$$F_{\text{up}}^{(i)} = \text{PReLU}(\text{PixelShuffle}(\text{Conv}_{3\times 3}(F^{(i-1)})))$$

where $i = 1, 2$ and $F_{\text{up}}^{(0)} = F_{\text{skip}}$.

Pixel Shuffling: Given an input tensor $\mathbf{X} \in \mathbb{R}^{B\times(C\cdot r^2)\times H\times W}$, **PixelShuffle** produces output $\mathbf{Y} \in \mathbb{R}^{B\times C\times(H\cdot r)\times(W\cdot r)}$ by:

$$\mathbf{Y} = \text{PixelShuffle}(\mathbf{X})$$

where r is the integer scale factor, and channels are reorganized into $r \times r$ spatial blocks.

Reconstruction Layer: Finally, a 9×9 convolutional layer reconstructs the high-resolution output:

$$\mathbf{I}_{SR} = \text{Conv}_{9\times9}(F_{\text{up}}^{(2)})$$

Layer Type	Output Channels	Kernel Size	Activation
Initial Conv	64	9×9	PReLU
Residual Blocks ×16	64	3×3	PReLU
Post-Residual Conv	64	3×3	None
Upsample ×2	64	3×3	PReLU
Final Conv	3 (RGB)	9×9	None

Table 1. Architecture summary of the baseline SR-ResNet model.

2.2. Attention-Enhanced SR-ResNet Architecture

To enhance the representation capability of the baseline SR-ResNet model, we integrate a lightweight attention mechanism using Squeeze-and-Excitation (SE) blocks [8]. These modules adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels. The enhanced network follows the same overall pipeline but replaces standard residual blocks with SE-augmented residual blocks and consists of the following parts as described below and in Figure 3:

Initial Feature Extraction: The network starts with a convolution layer to extract low-level features:

$$F_1 = \text{PReLU}(\text{Conv}_{9\times9}(\mathbf{I}_{LR}))$$

SE-Enhanced Residual Learning: Each residual block contains a Squeeze-and-Excitation module that applies global context recalibration. For a feature map $F \in \mathbb{R}^{C\times H\times W}$, the SE block operates as follows:

1. **Squeeze:** Global average pooling is applied across spatial dimensions:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_c(i, j) \tag{1}$$

2. **Excitation:** The squeezed vector is passed through two FC layers (implemented using 1×1 convolutions) with a reduction ratio r:

$$s = \sigma(W_2\delta(W_1 z_c))$$

where $\delta(\cdot)$ is ReLU, $\sigma(\cdot)$ is the Sigmoid function, and W_1, W_2 are learnable weights.

3. **Recalibration:** The output is scaled channel-wise:

$$F' = F \odot s$$

The full residual block with SE becomes:

$$F^{(l)} = F^{(l-1)} + \text{SE}(\text{Conv}_{3\times3}(\text{PReLU}(\text{Conv}_{3\times3}(F^{(l-1)}))))$$

Feature Aggregation: Following the residual stack, another 3×3 convolution is used, and the original features are added via a skip connection:

$$F_{\text{agg}} = \text{Conv}_{3\times3}(F_{\text{res}}) + F_1$$

Upsampling: As with the vanilla model, two PixelShuffle-based blocks perform $4\times$ upsampling:

$$F_{\text{up}}^{(i)} = \text{PReLU}(\text{PixelShuffle}(\text{Conv}_{3\times3}(F_{\text{up}}^{(i-1)})))$$

with $F_{\text{up}}^{(0)} = F_{\text{agg}}$.

Reconstruction: Finally, a 9×9 convolution generates the super-resolved image:

$$\mathbf{I}_{SR} = \text{Conv}_{9\times9}(F_{\text{up}}^{(2)})$$

Layer Type	Output Channels	Kernel Size	Activation
Initial Conv	64	9×9	PReLU
SE Residual Blocks ×16	64	3×3	PReLU + SE
Post-Residual Conv	64	3×3	None
Upsample ×2	64	3×3	PReLU
Final Conv	3 (RGB)	9×9	None

Table 2. Architecture summary of the Attention-Enhanced SRResNet model.

Key Differences from Vanilla SR-ResNet

- SE attention modules introduce adaptive channel-wise feature recalibration.
- Each residual block performs global context modeling via pooling and excitation.
- Enhances representational power without significant computational overhead.

2.3. Vanilla Loss Function

The Vanilla Loss $\mathcal{L}_{\text{vanilla}}$ combines three loss terms:

$$\mathcal{L}_{\text{vanilla}} = \alpha_1 \cdot \mathcal{L}_{\text{MSE}} + \alpha_2 \cdot \mathcal{L}_{\text{SSIM}} + \alpha_3 \cdot \mathcal{L}_{\text{VGG}}, \tag{2}$$

where:

- $\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{I}_{SR}^{(i)} - \mathbf{I}_{HR}^{(i)})^2$ is the pixel-wise Mean Squared Error,
- $\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(\mathbf{I}_{SR}, \mathbf{I}_{HR})$ measures structural similarity,
- $\mathcal{L}_{\text{VGG}} = \|\phi(\mathbf{I}_{SR}) - \phi(\mathbf{I}_{HR})\|_2^2$ is the perceptual loss from VGG feature space $\phi(\cdot)$.

Weights α_1, α_2, and α_3 are empirically tuned as *1.0*, *0.1* and *0.1* respectively.

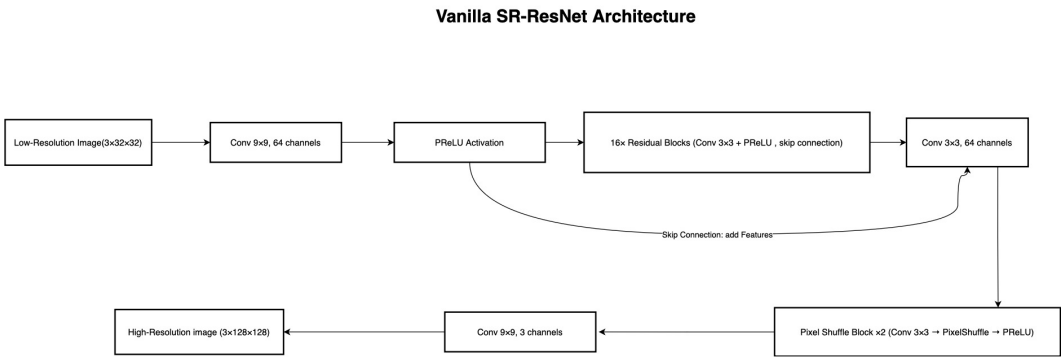


Figure 2. Vanilla SR-ResNet-Based Architecture

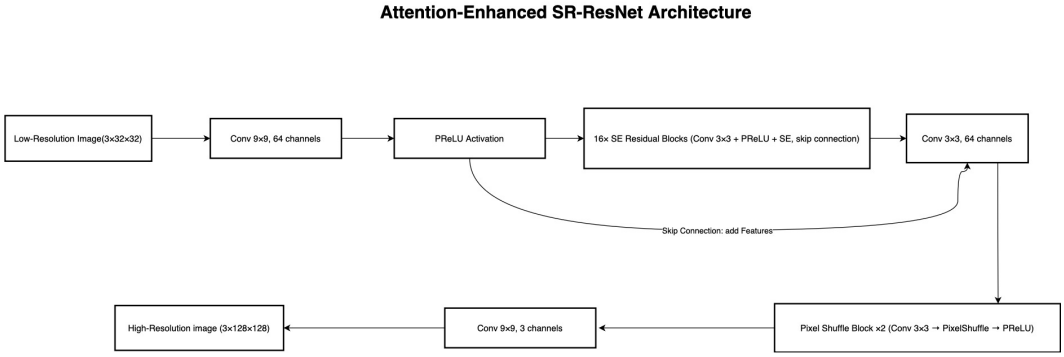


Figure 3. Attention-Enhanced SR-ResNet Architecture

2.4. Hybrid Loss based on Perceptual Similarity

To better preserve perceptual quality and structural edges, we define an enriched hybrid loss $\mathcal{L}_{\text{hybrid}}$:

$$\mathcal{L}_{\text{hybrid}} = \lambda_1 \cdot \mathcal{L}_{\text{LPIPS}} + \lambda_2 \cdot \mathcal{L}_{\text{SSIM}} + \lambda_3 \cdot \mathcal{L}_{\text{Charb}} + \lambda_4 \cdot \mathcal{L}_{\text{Edge}}, \quad (3)$$

with:

- $\mathcal{L}_{\text{LPIPS}}$: Learned Perceptual Image Patch Similarity using deep features,
- $\mathcal{L}_{\text{SSIM}}$: Same as above,
- $\mathcal{L}_{\text{Charb}}$: Charbonnier loss defined as

$$\mathcal{L}_{\text{Charb}} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{(\mathbf{I}_{SR}^{(i)} - \mathbf{I}_{HR}^{(i)})^2 + \epsilon^2}, \quad (4)$$

- $\mathcal{L}_{\text{Edge}}$: Edge-aware loss computed over Sobel-filtered versions of the images:

$$\mathcal{L}_{\text{Edge}} = \|\nabla \mathbf{I}_{SR} - \nabla \mathbf{I}_{HR}\|_1, \quad (5)$$

where ∇ denotes the gradient obtained using Sobel filtering.

The hybrid loss components are combined with fixed weights $\lambda_1 = 0.4$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$, and $\lambda_4 = 0.1$, chosen to balance perceptual and structural fidelity.

These loss formulations allow our network to generate outputs that are not only pixel-accurate but also visually pleasing and edge-consistent, which is crucial for high-quality face image super-resolution.

3. Experiments

In this section, we present the experimental setup, dataset used, and evaluation metrics applied to evaluate the performance of our proposed model.

3.1. Dataset

We conducted our experiments using the **CelebA-HQ** dataset, which contains high-quality images at a resolution of 256×256 pixels. For our experiments, we randomly selected a subset of $2{,}500$ images. This subset was then split into a training set ($2{,}000$ images) and a validation set (500 images), following an 80:20 ratio.

Each selected image was first randomly cropped to obtain high-resolution (HR) images of size 128×128. The corresponding low-resolution (LR) images were then generated by downsampling the HR images to 32×32 using

bicubic interpolation. This process ensured that each LR-HR pair was spatially aligned for supervised learning.

Finally, the dataset was converted into NumPy arrays for efficient loading during training. All images were processed in RGB format.

3.2. Evaluation Metrics

We employ three key metrics to assess the performance of our models:

- **Loss Function (MSE Loss):** We calculate the mean squared error between the predicted and ground truth images:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(I_i^{\text{HR}} - I_i^{\text{SR}}\right)^2 \tag{6}$$

  where I^{HR} and I^{SR} denote the high-resolution ground truth and super-resolved image pixels respectively, and N is the total number of pixels.

- **Structural Similarity Index (SSIM):** SSIM is used to evaluate the perceptual quality of the images, considering luminance, contrast, and structure. Higher SSIM values indicate better structural similarity between the predicted and ground truth images. It is defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{7}$$

  where μ_x, μ_y are the mean intensities, σ_x^2, σ_y^2 are the variances, and σ_{xy} is the covariance between x and y. C_1 and C_2 are small constants to stabilize the division.

- **Peak Signal-to-Noise Ratio (PSNR):** PSNR is another widely used metric to evaluate the quality of an image, where higher values indicate better reconstruction quality. It is defined in terms of MSE as:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{L^2}{\text{MSE}}\right) \tag{8}$$

  where L is the maximum possible pixel value (typically 255 for 8-bit images or 1 in case of *normalized pixel values*).

We report results for these metrics on both **training** and **validation** sets for all three variants of the model:

- Vanilla SR-ResNet
- SR-ResNet with Attention Mechanism
- SR-ResNet with Hybrid Loss

3.3. Experimental Setup

For all experiments, we used the following configuration:

- **Batch Size:** 8
- **Learning Rate:** 0.001
- **Optimizer:** Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$
- **Training Duration:** 50 epochs (*for 1st and 2nd variant model*) and 20 epochs (*for the 3rd model*)

3.4. Results

The following graphs in Figure 4 present the performance of the models based on the metrics discussed above. Further, the epoch-wise evaluation metrics for all three models are given in Tables 3, 4, and 5 below.

Table 3. Training and Validation Metrics for Vanilla SR-ResNet Model (Every 5 Epochs)

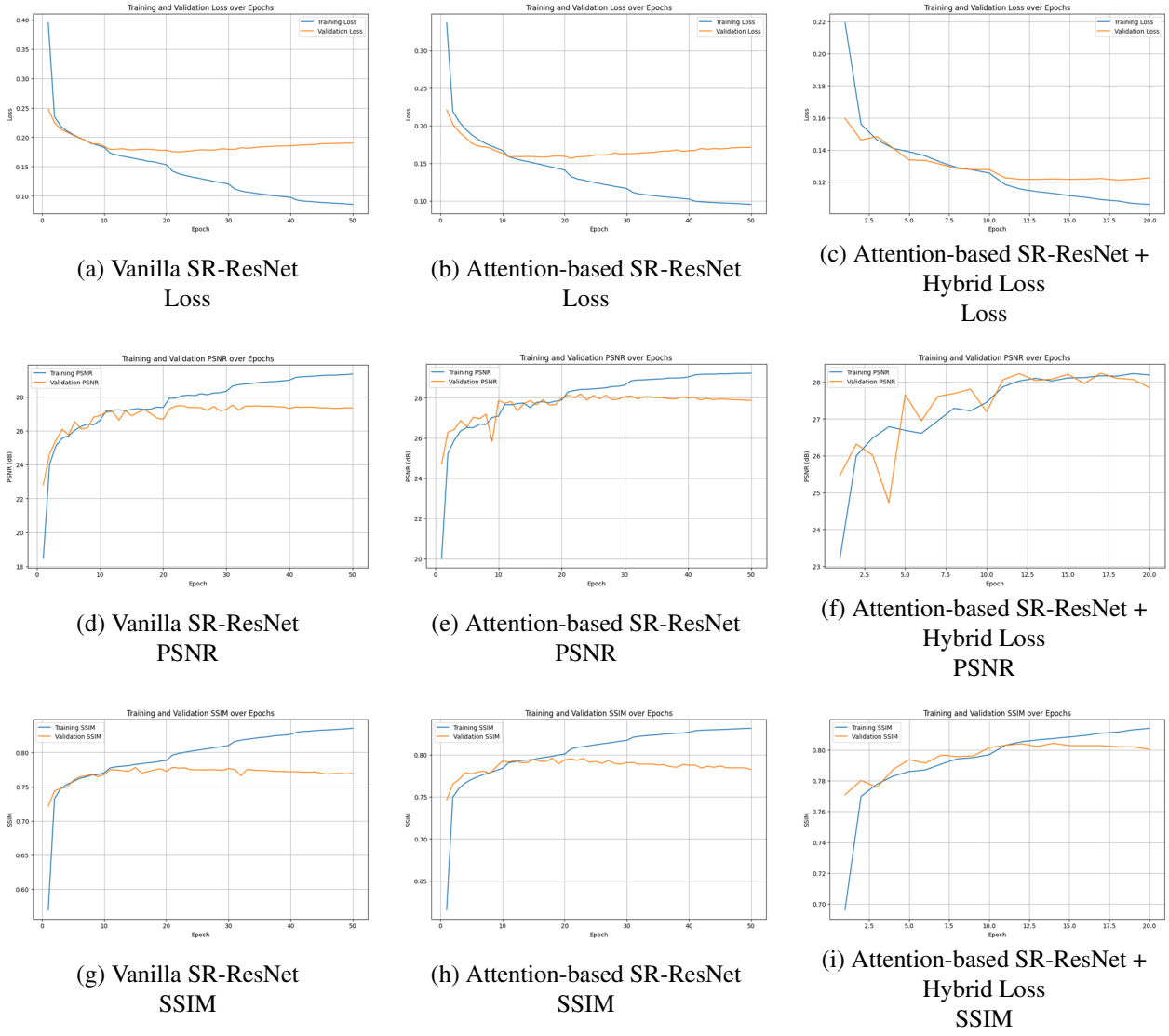
Epoch	Train Loss	Val Loss	Train PSNR	Val PSNR	Train SSIM	Val SSIM
5	0.2046	0.2036	25.70	25.76	0.7573	0.7590
10	0.1828	0.1843	26.63	26.91	0.7703	0.7677
15	0.1635	0.1785	27.26	26.89	0.7824	0.7780
20	0.1533	0.1780	27.38	26.68	0.7883	0.7724
25	0.1306	0.1783	28.08	27.38	0.8031	0.7747
30	0.1203	0.1793	28.33	27.26	0.8102	0.7765
35	0.1036	0.1833	28.84	27.47	0.8216	0.7736
40	0.0974	0.1854	28.99	27.32	0.8265	0.7715
45	0.0891	0.1889	29.26	27.35	0.8325	0.7696
50	0.0858	0.1901	29.35	27.35	0.8352	0.7696

Table 4. Training and Validation Metrics for Attention-based SR-ResNet Model (Every 5 Epochs)

Epoch	Train Loss	Val Loss	Train PSNR	Val PSNR	Train SSIM	Val SSIM
5	0.1883	0.1767	26.52	26.55	0.7709	0.7776
10	0.1671	0.1633	27.09	27.86	0.7842	0.7926
15	0.1506	0.1589	27.51	27.86	0.7947	0.7947
20	0.1410	0.1593	27.88	27.98	0.8011	0.7941
25	0.1241	0.1613	28.44	28.12	0.8122	0.7923
30	0.1163	0.1629	28.63	28.06	0.8174	0.7907
35	0.1062	0.1655	28.93	28.01	0.8241	0.7880
40	0.1022	0.1667	29.03	27.98	0.8271	0.7878
45	0.0973	0.1691	29.19	27.94	0.8303	0.7869
50	0.0953	0.1714	29.22	27.87	0.8317	0.7828

Table 5. Training and Validation Metrics for Attention-based SR-ResNet Model with Hybrid Loss (Every 2 Epochs)

Epoch	Train Loss	Val Loss	Train PSNR	Val PSNR	Train SSIM	Val SSIM
2	0.1560	0.1461	26.00	26.32	0.7699	0.7802
4	0.1409	0.1410	26.79	24.73	0.7830	0.7873
6	0.1363	0.1334	26.61	26.95	0.7870	0.7914
8	0.1289	0.1283	27.29	27.69	0.7942	0.7957
10	0.1255	0.1277	27.45	27.20	0.7969	0.8015
12	0.1154	0.1216	28.03	28.23	0.8053	0.8040
14	0.1128	0.1219	28.03	28.08	0.8074	0.8043
16	0.1104	0.1218	28.12	27.96	0.8094	0.8028
18	0.1081	0.1212	28.16	28.10	0.8116	0.8021
20	0.1060	0.1225	28.19	27.85	0.8140	0.8004



(a) Vanilla SR-ResNet Loss

(b) Attention-based SR-ResNet Loss

(c) Attention-based SR-ResNet + Hybrid Loss Loss

(d) Vanilla SR-ResNet PSNR

(e) Attention-based SR-ResNet PSNR

(f) Attention-based SR-ResNet + Hybrid Loss PSNR

(g) Vanilla SR-ResNet SSIM

(h) Attention-based SR-ResNet SSIM

(i) Attention-based SR-ResNet + Hybrid Loss SSIM

Figure 4. Training and validation performance of the three model variants across different evaluation metrics. Each column corresponds to one model variant: Vanilla SR-ResNet (left), SR-ResNet with Attention Mechanism (middle), and Attention Based SR-ResNet with Hybrid Loss (Right). Rows show performance based on MSE Loss (Top), PSNR(Middle), and SSIM (Bottom) respectively.

4. Discussion and Scope of Future Work

From the results, we observe that the **SR-ResNet with Attention Mechanism** outperforms the **Vanilla SR-ResNet** model, especially on the validation set. The inclusion of the attention mechanism allows the model to focus more effectively on important regions of the image, leading to a better structural and perceptual reconstruction. The **Hybrid Loss** further improves performance, combining spectral consistency with structural similarity.

In particular, the **PSNR** scores indicate that the Hybrid Loss variant achieves the highest reconstruction quality, followed by the attention-based model. Meanwhile, the **SSIM** scores suggest that the attention-based model provides superior perceptual quality compared to the **Vanilla SR-ResNet** model.

4.1. Scope of Future Work

While our current study explores multiple improvements over the **Vanilla SR-ResNet** model, there remains substantial potential for further enhancement.

One promising direction is to integrate the strengths of both architectural and loss-based advancements. Specifically, using the **Attention-based SR-ResNet** variant in conjunction with the **Hybrid Loss Function** could serve as

a more expressive generator in a GAN-based framework. This hybrid setup may allow the model to learn both perceptual fidelity and fine-grained structural details more effectively through adversarial training [15, 21].

Another potential improvement is the incorporation of advanced attention mechanisms, such as the channel attention (CA) and spatial attention (SA) modules proposed in CBAM [23], or even transformer-based architectures for better global context modeling [17]. These can help the network selectively emphasize informative features and suppress irrelevant ones.

Moreover, the use of perceptual loss functions, which compare high-level features extracted from pretrained networks like VGG [10], could further improve the perceptual quality of the super-resolved images. This is especially useful when the goal is not just pixel accuracy but also better human visual perception.

From an application perspective, future work could involve extending the current approach to multi-scale or progressive GAN frameworks [20], enabling better handling of extreme upscaling factors. Additionally, exploring temporal consistency for video super-resolution tasks and applying domain adaptation techniques to generalize the model across datasets (e.g., from *CelebA-HQ* to natural scenes or satellite imagery) are also promising avenues.

Ultimately, these extensions would contribute to building a more robust, generalizable, and visually coherent super-resolution pipeline.

5. Conclusion

In this work, we explored the task of single image super-resolution using three variants of the SR-ResNet architecture: the baseline Vanilla SR-ResNet, an enhanced SR-ResNet with an Attention Mechanism, and a further improved version incorporating a Hybrid Loss function. Our experiments were conducted on the CelebA-HQ dataset, with an 80:20 split of randomly selected 2500 sample images for training and validation.

Through a comprehensive evaluation using multiple metrics — MSE loss, Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR) — we demonstrated the effectiveness of our proposed enhancements. The **Attention Mechanism** allowed the network to focus on the most informative regions of the image, leading to noticeable improvements in both perceptual and quantitative results. The **Hybrid Loss**, by combining pixel-wise and perceptual objectives, further enhanced the reconstruction quality, especially on unseen validation samples.

Among all the models, the **SR-ResNet with Attention and Hybrid Loss** consistently achieved the best performance across all metrics for the *validation data*, showing better generalization and robustness to complex image textures. The qualitative outputs also aligned with these findings, producing sharper and more visually pleasing high-resolution reconstructions.

This study reaffirms the value of integrating attention-based modules and hybrid loss formulations in deep super-resolution networks and sets the stage for future improvements, such as integrating adversarial learning or frequency-domain priors for even finer detail preservation.

References

[1] J. Chen, L. Zhang, and Q. Shen. Single image super-resolution for mri using deep learning. In *IEEE EMBS*, pages 740–743, 2018. 1

[2] Yu Chen, Yu Tai, Xiaoyun Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2492–2501, 2018. 2

[3] Tao Dai, Jianrui Cai, Yongbing Zhang, Kui Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 2

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. 2

[5] W. Dong, L. Zhang, G. Shi, and X. Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011. 1

[6] M. A. Faraone, M. A. Sid-Ahmed, and H. H. Arslan. Super-resolution techniques for surveillance video enhancement. In *IEEE International Conference on Image Processing (ICIP)*, pages 1313–1317, 2016. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. 3

[8] Jie Hu et al. Squeeze-and-excitation networks. In *CVPR*, 2018. 2, 4

[9] S. Huang, X. Kang, and L. Zhang. A new pan-sharpening method with deep neural networks. *IEEE Geoscience and Remote Sensing Letters*, 12(5):1037–1041, 2015. 1

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 2, 8

[11] Tero Karras et al. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 2

[12] Robert Keys. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1153–1160, 1981. 2

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016. 2

[14] Wei-Sheng Lai et al. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017. 2

[15] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017. 1, 2, 8

[16] Yijun Li, Shuang Liu, Hao Lin, Ming-Hsuan Yang, and Jan Kautz. Dfdnet: Deep face dictionary network for pose-invariant face completion and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7693–7702, 2020. 2

[17] Jingyun Liang, Jie Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1833–1844, 2021. 2, 8

[18] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017. 2

[19] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1874–1883, 2016. 3

[20] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 606–615, 2018. 8

[21] Xintao Wang, Kelvin C K Yu, Chao Dong, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018. 2, 8

[22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1, 2

[23] Sanghyun Woo et al. Cbam: Convolutional block attention module. In *ECCV*, 2018. 2, 8

[24] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010. 2

[25] S. Yu and H. Shi. Face hallucination with identity priors for high-fidelity face recovery. In *ICCV*, pages 9395–9404, 2021. 1

[26] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 1, 2

[27] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 2

[28] Hang Zhao et al. Loss functions for image restoration with neural networks. In *IEEE Transactions on Computational Imaging*, 2017. 2

Appendix

Resources

- **Dataset:** CelebA-HQ dataset used in this study can be accessed from: https://www.kaggle.com/datasets/badasstechie/celebahq-resized-256x256
- **Code Repository:** The relevant PyTorch implementations are available at:
  - **GitHub Repository of the Project**: https://github.com/shreyash1110/Image-Super-Resolution-using-SRResNet
  - **Vanilla SR-ResNet Model**: https://github.com/shreyash1110/Image-Super-Resolution-using-SRResNet/blob/main/vanilla_model.py
  - **Attention-based SR-ResNet Model**: https://github.com/shreyash1110/Image-Super-Resolution-using-SRResNet/blob/main/model_v2.py
  - **Vanilla Loss Function**: https://github.com/shreyash1110/Image-Super-Resolution-using-SRResNet/blob/main/vanilla_loss.py
  - **Hybrid Loss Function incorporating Perceptual Similarity**: https://github.com/shreyash1110/Image-Super-Resolution-using-SRResNet/blob/main/loss_v3.py