Machine Learning for Resolution: Using DF2K Dataset

Dhruv Lalchandani

Received December 21, 2025
Accepted June 29, 2025
Electronic access July 31, 2025

Unclear images pose a challenge when examining specific details, a problem that can be addressed using artificial intelligence. In particular, machine learning models for image super-resolution generate higher-resolution outputs without introducing the blurriness typical of traditional upscaling methods. Previous approaches have used architectures such as the Efficient Sub-Pixel Convolutional Neural Network (ESPCN) and the Super-Resolution Convolutional Neural Network (SRCNN). This project builds on that work with a customized convolutional neural network that first upsamples images beyond the target size, then strategically downsamples them to produce the end result. After experimenting with multiple architectures, the final model was selected based on its superior performance. It achieved a LPIPS loss of 0.0831 (lower values are better), indicating strong perceptual similarity between the generated and ground truth images. Training and evaluation were conducted on the DF2K dataset, a combination of DIV2K and Flickr2K—both high-quality datasets commonly used for super-resolution tasks. Future work will involve testing on more diverse and larger datasets to further improve generalization and accuracy. Super-resolution remains a critical area of research in deep learning due to its wide range of real-world applications and potential for continued advancement.

Keywords: Computer Science, Machine Learning, Deep Learning, Super Resolution, Image Augmentation

Introduction

Background and Context

Image resolution plays a critical role in how clear and readable an image is. By leveraging computer vision, we can enhance low-quality images to reveal finer details, significantly improving their usefulness in applications like surveillance, satellite imaging, and digital forensics. Blurry or unclear images often result from low-quality cameras, compression artifacts, or long-distance capture. A report from the University of Edinburgh explains that "many applications require zooming of a specific area of interest in the image wherein high resolution becomes essential, e.g. surveillance, forensic and satellite imaging applications"[1].

Two major models commonly used for image super-resolution are the Efficient Sub-Pixel Convolutional Neural Network (ESPCN) and the Super-Resolution Convolutional Neural Network (SRCNN). ESPCN maintains low-resolution images throughout most of the network and performs upscaling near the output layer[2], which may limit its ability to extract fine details early in the pipeline. The SRCNN, in contrast, employs only three convolutional layers.

This project introduces a different architecture, in which the following process is repeated: the image is upsampled and then a convolutional layer is applied. This design gives the network access to more spatial detail early on and goes beyond the target image size in order to allow the model to keep track of more data. It is hypothesized that this structure allows the model to better capture and enhance fine-grained features. With additional data and computational resources, this framework could be scaled further by future researchers, organizations, or governments to push the boundaries of what is possible in super-resolution. Ultimately, this could enable clearer analysis of visual data in domains such as surveillance, satellite imagery, and forensic investigation.

Related Studies and Projects

Several studies were found that also aimed to create a super resolution model utilizing deep learning techniques. Shi et al.[3] was the study that introduced the ESPCN in 2016 and they used both the BSDS 500 dataset and 50,000 random images from ImageNet. This study also used only the Y factor in the YUV color scale. They then compared the ESPCN to the SRCNN and found that it performed better on peak signal-to-noise ratio (PSNR). It received a PSNR of 33.92 – a higher PSNR is considered better and means the output images have a higher accuracy when compared to ground truth images.

In 2017, Lim et al.[4] proposed the Enhanced Deep Super-Resolution Network (EDSR) and Multi-Scale Deep Super Resolution System (MDSR) for single image super-resolution. They found that it performed better on PSNR and SSIM (Structural Similarity Index) than bicubic interpolation, the Super Resolution Residual Network (SRResNet), and Very Deep Super Resolution (VDSR). Bicubic interpolation refers to a non-machine-

learning-based algorithm used to resize images by estimating pixel values based on nearby pixels. The other two mentioned techniques are machine learning models.

In 2023, Y. Zhang et al.[5] presented the findings of the NTIRE 2023 challenge with single image super-resolution. The study uses a scale factor of 4, whereas this one uses a scale factor of 2. They are therefore reconstructing a high-resolution image from a 4x downsampled one. They found that transformer-based and hybrid models performed the best for 4x single-image super-resolution.

In 2024, Conde et al.[6] presented the results of the NTIRE Challenge on RAW image super-resolution. They work with the RAW Bayer dataset which consists of images directly from cameras, with no camera processing. This introduces an aspect of difficulty as camera filters clean up noise. The Samsung team achieved the best results, using two stages in their model. The first stage was aimed at recovering image structure while the second focused on recovering details and hence further refining the reconstruction.

Problem Statement and Rationale

The aim of this project is to discover the optimal deep learning model architecture for image sub-pixel super-resolution, a technique focused on enhancing the resolution of low-quality images. By approaching the problem of image super-resolution in a novel way, this research advances the field of computer vision and provides valuable insights for future researchers. Specifically, the proposed network upscales the image before reverting it to its original size, in contrast to the traditional U-Net model that first downsamples the image. This method retains more data throughout the process, allowing for a more significant increase in resolution.

Objectives

The objective of this project is to compare the suggested model architecture to others commonly used in the field, namely the ESPCN and SRCNN. In doing so, this project hopes to develop a model for image super-resolution that is both accessible and beneficial to the community. By enhancing the quality of low-resolution images, the model aims to improve the overall effectiveness of image super-resolution techniques. Given the recent scarcity of studies in this field, this project seeks to fill a research gap, exploring innovative approaches and methodologies to advance image enhancement.

Scope and Limitations

Given that this is a machine learning project involving a complex neural network, computational power is a significant limitation. The A5000 GPU was used, but more computation power could have allowed for processing of a larger dataset. The dataset used had 3450 training images but with more computation power, a dataset with tens of thousands of images could have been used to bring in more diversity and allow the model to have more data to learn from.

Methods

Fundamental Processes

Deep-learning uses convolutional neural networks to interpret images. The network consists of multiple layers, starting with the input layer, followed by hidden layers, and finally an output layer. Each layer has neurons that have specific algorithms that they run before passing on to the next layer[7]. The specific neural network used is the convolutional neural network. It is a neural network whose layers contain filters, also known as kernels, that are 3 by 3 matrices. These matrices are applied to 3 by 3 sections of the image, moving in steps determined by a specified stride. This process is repeated throughout the image to produce the final output image[8].

The methodology for this project comprised four main steps: data gathering and filtering, data preprocessing, model training, and evaluation. First, preprocessing steps were applied to standardize the images for model input. The preprocessed images were subsequently used to train a neural network, allowing it to learn relevant features from the dataset. Finally, the trained model was tested on a separate set of images, and its outputs were compared with expected results to assess accuracy and performance.

Implementation

Algorithm 1: Inverted U-Net Forward Pass

1. Upsample x
2. Apply Conv1 followed by ReLU to x
3. Upsample x
4. Apply Conv2 followed by ReLU to x
5. Upsample x
6. Apply Conv3 followed by ReLU to x
7. Apply Conv4 followed by ReLU to x
8. Apply Conv5 followed by ReLU to x
9. Apply Conv6 followed by ReLU to x
10. 10: Return x

The Inverted U-Net architecture was implemented as a PyTorch class. It consists of three upsampling blocks followed by three convolutional downsampling layers. Each upsampling block performs nearest-neighbor interpolation, followed by a convolutional layer with ReLU activation. Therefore, the upsampling itself is not trainable, but the following convolutional layer is. This structure increases the spatial resolution of the input. After the final upsampling, a series of convolutions with strides reduce the dimensions and produce the final output. These layers are trainable. The model architecture definition is summarized in Algorithm 1. The Adam optimizer was used with a learning rate of 10^{-4}, and the model was trained using LPIPS [9] loss with pretrained features from the VGG (Visual Geometry Group) network [10].

The SRCNN is generally made to convert blurry images to high-resolution ones, but not upscale them, whereas the ESPCN and proposed architecture produce outputs double the size of inputs. Therefore, in implementation of the SRCNN, an upsample layer was added to the very beginning, using bicubic interpolation. This is untrainable, so it has the same effect as upscaling all the inputs using this method before training.

Research Design

The research design for this project is primarily experimental, aimed at testing the hypothesis that the Inverted U-Net outperforms standard models–specifically the ESPCN and SRCNN–in image super-resolution tasks. This hypothesis is grounded in the theory that upsampling images beyond the required output resolution provides the model with more spatial information in a single plane, rather than relying solely on increasing the number of feature channels. Unlike traditional models that focus mainly on expanding channel depth, the Inverted U-Net leverages both increased spatial dimensions and feature channels to improve performance. Further, this project seeks to analyze how the Inverted U-Net performs on validation data over many epochs.

Dataset Description

The dataset used in this study was the DF2K dataset [11], a widely adopted benchmark for single image super-resolution tasks. DF2K is a composite dataset formed by merging the DIV2K [12] and Flickr2K [4] datasets. DIV2K provides high-quality, diverse images with 2K resolution or higher, while Flickr2K contributes a broader range of natural scenes sourced from Flickr. This combination increases both the volume and variability of training data, which is essential for improving the generalization and robustness of super-resolution models. The DF2K dataset was introduced as part of the NTIRE 2017 challenge and has since become a standard in the field. This dataset has 3450 training images and 100 validation images, as shown in Figure 1. All of these images are in high- and low-resolution forms. The validation images can be used for comparing the models because they have no impact on the training process - they are solely used for evaluation and tracking improvement over time.

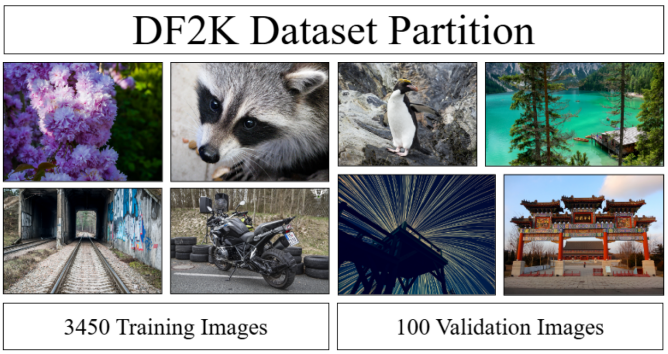

Fig. 1 Example images and partitioning of the DF2K dataset.

Procedure

The experiments were carried out on Jarvis Labs, an online GPU rental service. Specifically, the model was trained on the NVIDIA RTX A5000. The PyTorch framework facilitated the construction and training of a neural network model. The process began with data preparation, including filtering and preprocessing images for uniform size and normalization. After designing the model architecture with appropriate layers and activation functions, training was performed through iterative weight adjustments using backpropagation based on a defined loss function. The model's performance was then evaluated on a separate validation set, and adjustments were made as necessary before testing with new data to assess generalization capabilities. The PIL image module was used for bicubic interpolation and other preprocessing tasks.

Model Selection and Architecture

The optimal model was a very basic recreation of an inverted U-Net, as shown in Figure 2 This means that it expands first to create a lot of space for the model to work and then extract relevant information to output a high quality image. To do so, the model starts with an up-sampling layer to increase the size, and then a convolutional layer with 64 filters, repeated three times. Then the model begins to down-sample with the use of two convolutional layers, the first with 32 filters, and the second with 16 filters. Both have a stride of two, which causes the image size to shrink by a factor of 2. Finally, for output, there was one final convolutional layer with 1 filter to bring the image to the target size (double the size of the input). All convolutional layers use ReLU activation.

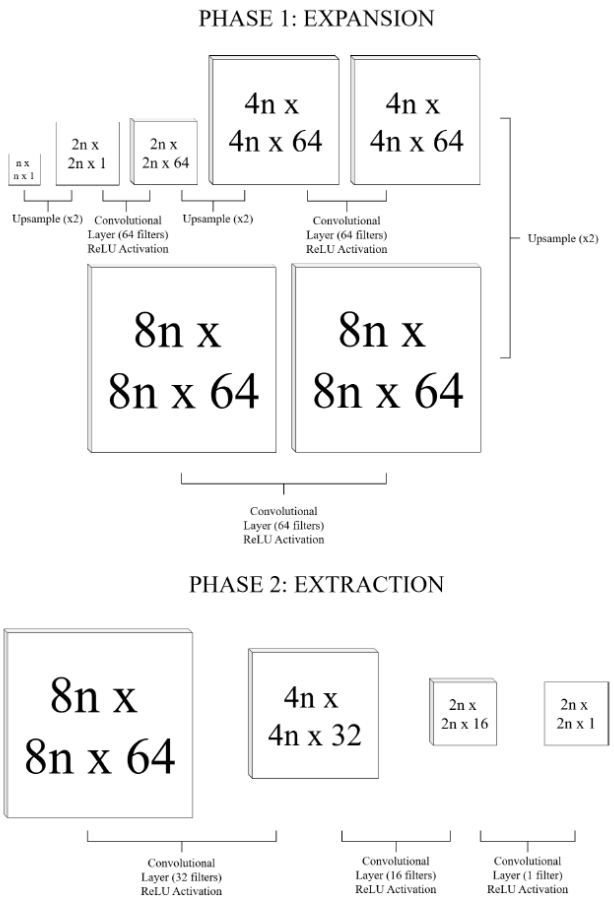

Fig. 2 Architecture Overview of the Inverted U-Net Model

Model Training Process

The training was with the Adam optimizer, with a learning rate of 0.0001, and a batch size of 8. Number of epochs run for each model was based on the cost to train for one epoch so that the models could be fairly compared. For example, the ESPCN costed about 14¢ per epoch while the Inverted U-Net costed about 52¢ per epoch. Therefore, the Inverted U-Net was only run for 10 epochs while the ESPCN was run for 40. Nonetheless, the Inverted U-Net performed better on the LPIPS metric (most representative of human-eye accuracy).

Data Preprocessing and Augmentation

A few preprocessing steps were applied to the raw data to prepare it for the machine learning algorithm. First, all images were converted from the RGB (Red-Green-Blue) to YCbCr color space. YCbCr color scale has Y, for luminance which is similar to grayscale, Cb (Chrominance Blue) which represents its blue signal, and Cr (Chrominance Red) which represents its red signal[13]. Only the Y channel was retained, as it carries the most information relevant to human visual perception. To ensure consistency and efficient training, all images were resized to 512×512. To maintain aspect ratio, images were scaled to fit within the target dimensions, then padded (either top/bottom or left/right) to match the required size exactly. Next, low-resolution images were generated by downscaling the 512x512 images by a factor of 2 to achieve 256x256, and these served as model inputs. This means that the model will operate on the generated 256x256 images (the inputs) to achieve as much visual similarity as possible to the 512x512 images (the outputs). An example of preprocessing is shown in Figure 3.

Fig. 3 Example of preprocessing applied to image from dataset.

Results

Table 1 Average LPIPS loss and MS-SSIM score achieved by each model on the DF2K Validation Set.
Note: Lower LPIPS and higher MS-SSIM indicate better perceptual quality.

Architecture	Epochs	LPIPS	MS-SSIM
Inverted U-Net	10	0.0831	0.9847
SRCNN	50	0.0928	0.9844
ESPCN	40	0.0994	0.9856

Data Analysis: Performance Evaluation

The most important metric in this evaluation is the Learned Perceptual Image Patch Similarity (LPIPS), introduced by Zhang et al. (2018). LPIPS is designed to reflect perceptual similarity as judged by humans. Unlike metrics that compare pixel-level differences, LPIPS employs a deep neural network to evaluate differences in high-level image features. The authors validate the effectiveness of this approach using a dataset of human similarity judgments, finding that LPIPS outperforms other perceptual similarity metrics in aligning with human assessments[9]. Another commonly used metric is the Multi-Scale Structural Similarity Index Measure (MS-SSIM), introduced by Wang et

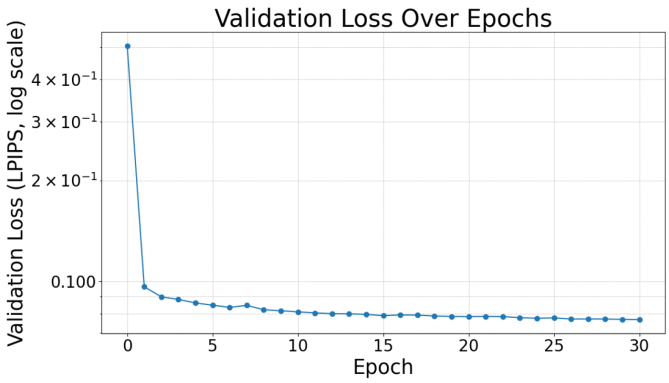

Fig. 4 Separate training of the Inverted UNet: Validation Loss through 30 Epochs

al. as an extension of the Structural Similarity Index Measure (SSIM). MS-SSIM improves upon SSIM by incorporating image details at multiple scales, which makes it more representative of perceptual similarity [14].

Data Analysis: Discussion of Findings

Given similar computational constraints, the Inverted U-Net demonstrates superior performance in terms of LPIPS, indicating that it produces images that are perceptually closer to the ground truth. Although MS-SSIM scores are nearly identical across all models, the ESPCN performs marginally better on this metric than the Inverted U-Net. However, the difference is minimal and may not be perceptible in practical scenarios. Overall, the Inverted U-Net provides the best perceptual quality based on LPIPS, which is considered the more human-aligned metric. Figure 5 demonstrates the result of the three different models on the same image, and bicubic interpolation, showing the superiority of the Inverted U-Net, which has sharpness and minimal noise. Figure 4 demonstrates the LPIPS loss of the Inverted U-Net over 30 epochs, showing quick learning at the beginning which slowed as the training progressed. Towards the end, the model continues to improve but in very small increments.

Discussion

Summary of Findings

The goal of this study was to evaluate the performance of the proposed Inverted U-Net model compared to the ESPCN and SRCNN, established super-resolution architectures. The results indicate that the Inverted U-Net achieved the lowest LPIPS score (0.0831), suggesting superior perceptual similarity to ground truth images. While MS-SSIM scores were comparable across models, the Inverted U-Net outperformed the others in LPIPS

Fig. 5 "Visual comparison of upscaling methods on the first image from the DF2K validation set. This image was selected purely based on its index in the dataset—before any model was evaluated—to avoid selection bias. Despite not being chosen to support a specific outcome, it clearly highlights the superior perceptual quality of the Inverted U-Net relative to both ESPCN and SRCNN, as well as a basic resized baseline (bicubic interpolation)."

while also appearing the best in Figure 5. These findings demonstrate the effectiveness of the Inverted U-Net's architecture in capturing perceptual details, validating its potential as a strong alternative to traditional approaches in image super-resolution tasks.

Implications and Significance

There are vast benefits of super resolution, as explained by D. C. Lepcha, including benefits in surveillance, medical, and media fields. In the case of surveillance, for example, with security cameras, super resolution can take a blurry image, enhance it, and allow it to be used in a facial recognition system. Next, it can be used to take blurry MRI imaging and develop high resolution images, for better analysis by doctors. Finally, in the media, it allows images to be sent at a lower resolution and upscaled afterwards [15]. This project advances the field by developing a high-quality technique for image upscaling and designing a model that future work can refine and deploy. The model's novel architecture unlocks perceptual quality improvements beyond

existing methods, enabling more effective image enhancement across a range of real-world applications.

Ethical Considerations

Initially, the BSDS 500 dataset [16] was used for both training and testing. However, it became evident that there was an issue in the loading process: the model was accidentally tested on the same images on which it had been trained. This approach is problematic as it compromises the accuracy of the test results; a reliable model must generalize well to new, unseen data rather than simply performing well on known images. To address this and to use a higher-quality dataset with more images, the model was retrained and retested using the DF2K Dataset, ensuring it was appropriately partitioned to avoid data leakage and provide a more valid assessment of model performance. The validation set used for evaluation has no impact on the training process whatsoever.

Limitations and Future Work

One limitation of this study was the computational constraint that prevented training a model with four upsampling layers followed by three downsampling layers. Future research could explore whether such an architecture would lead to improved accuracy or diminishing returns. Another limitation was the relatively small training set of under 4,000 images. Scaling up the dataset could lead to better generalization and robustness of the model.

Future work could also focus on applying super-resolution to entire videos in a more integrated manner. Instead of processing each frame independently, a model could leverage information across frames and possibly incorporate audio cues to enhance reconstruction quality. Efficiency could further be improved by generating intermediate frames from learned features rather than processing every single frame, which may significantly reduce computational overhead while maintaining visual fidelity. However, this task would require massive computation ability, as each model pass would process significantly more data.

Conclusion

The goal of this project was to improve upon existing work in the field of super-resolution, finding a new model that could enhance images effectively and produce high-quality results, such that a human would consider it clear. Consequently, this project proposed and evaluated a novel Inverted U-Net architecture for image super-resolution, comparing its performance to two established models: ESPCN and SRCNN. In order to do so, models were designed, trained, and tested on the validation data, with costs made near equal by adjusting the number of epochs trained. They were trained on the Y (Luminance) channel of the YCbCr color space as the human-eye is the most sensitive to it. Experimental results show that the Inverted U-Net achieves superior LPIPS scores, indicating better perceptual quality, while MS-SSIM scores remain comparable across all models. These results suggest that the Inverted U-Net is an effective alternative for super-resolution tasks. Its unique approach of upscaling beyond the target resolution allows the model to leverage additional spatial information, enhancing its learning capacity. Improving super-resolution techniques remains a critical area of research, with applications ranging from surveillance to medical imaging. The success of this model underscores the potential for further innovation in image enhancement technologies and encourages continued exploration in this domain. Future studies should attempt to do the same task with a larger dataset, or on entire videos, with the goal of retaining data from one frame to the next.

Acknowledgments

The author received help from Dr. Robail Yasrab, who provided valuable guidance and feedback throughout the project.

References

1 P. Gaidhani, *Super-resolution*.

2 J. Nay, *Single image super resolution using espcn–with ssim loss*.

3 *Real-time single image and video super-resolution using an efficient subpixel convolutional neural network*.

4 B. Lim, S. Son, H. Kim, S. Nah and K. Lee, *Enhanced deep residual networks for single image super-resolution*.

5 Y. Zhang and K. Zhang, *Ntire 2023 challenge on image super-resolution (×4): Methods and results*.

6 M. Conde, F.-A. Vasluianu and R. Timofte, *Deep raw image super-resolution. A NTIRE 2024 challenge survey*.

7 W. Samek, A. Binder, G. Montavon, S. Lapuschkin and K.-R. Müller, *Evaluating the visualization of what a deep neural network has learned*.

8 J. Wu, *Introduction to convolutional neural networks*.

9 R. Zhang, P. Isola, A. Efros, E. Shechtman and O. Wang, *The unreasonable effectiveness of deep features as a perceptual metric*.

10 K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556,.

11 R. Timofte, S. Gu, J. Wu and L. Gool, *Ntire 2018 challenge on single image super-resolution: Methods and results*.

12 E. Agustsson and R. Timofte, *Ntire 2017 challenge on single image super-resolution: Dataset and study*.

13 S. Gopinathan and S. Gayathri, *A study on image enhancement techniques using YCbCr color space methods*.

14 Z. Wang, E. Simoncelli and A. Bovik, *Multiscale structural similarity for image quality assessment*.

15 D. Lepcha, B. Goyal, A. Dogra and V. Goyal, *Image super-resolution: A comprehensive review, recent trends, challenges and applications*.

16 D. Martin, C. Fowlkes, D. Tal and J. Malik, *A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics*.