Hybrid Swin Transformer for High-Fidelity Image Super-Resolution

ANSHUMAN R PRASAD

ARENUKAP@depaul.edu

February 18, 2025

Abstract

Image Super-Resolution (ISR) is a critical task in computer vision that enhances the resolution of low-quality images, benefiting applications like medical imaging, satellite imaging, and autonomous driving. Traditional ISR methods, primarily based on CNNs, struggle with capturing long-range dependencies, leading to a loss of fine-grained details. Transformer-based models, such as SwinIR, improve global feature representation but often fail to preserve local textures effectively. This project proposes a Hybrid Swin Transformer ISR model that combines CNN-based feature extraction with Swin Transformer-based global attention to achieve high-fidelity super-resolution. The model will be trained on high-quality datasets like DIV2K, Flickr2K, and Urban100, using perceptual loss and MSE loss for improved detail retention. By leveraging both local and global features, the proposed approach aims to outperform existing ISR methods in terms of sharpness, structural similarity, and artifact reduction**.**

Introduction

Super-resolution (SR) is a fundamental problem in image processing where the goal is to reconstruct high-resolution (HR) images from low-resolution (LR) inputs. This problem is particularly significant in areas such as medical imaging, satellite surveillance, and autonomous driving, where high-resolution images provide crucial information. However, traditional deep learning approaches based on Convolutional Neural Networks (CNNs) often struggle with capturing long-range dependencies, leading to blurry or artifact-laden results. While transformer-based models, such as SwinIR, improve upon this by introducing global self-attention mechanisms, they still lack local texture refinement, often producing unnatural results.

This project aims to develop a Hybrid Swin Transformer ISR model that combines the local feature extraction power of CNNs with the global context-awareness of Swin Transformers. The key challenges include balancing local and global feature extraction, reducing computational overhead, and optimizing perceptual loss for high-fidelity reconstruction. The project will focus on training the model on diverse, high-resolution datasets and evaluating its effectiveness against state-of-the-art methods in terms of PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index Measure).

Related Work

Several deep learning-based image super-resolution (ISR) approaches have been developed over the years, each with unique strengths and limitations. CNN-based super-resolution models such as SRCNN were among the first deep learning approaches, utilizing a shallow convolutional network to enhance low-resolution images. Later models like EDSR improved performance by introducing deeper architectures, though they required significantly higher computational power. To address feature representation challenges, RCAN introduced residual channel attention mechanisms, improving texture preservation and detail enhancement. However, CNN-based models struggle with capturing long-range dependencies, limiting their ability to reconstruct fine details effectively.

On the other hand, Transformer-based super-resolution models have emerged as a promising alternative. SwinIR, for instance, leverages the Swin Transformer's self-attention mechanism to model global dependencies, achieving superior feature extraction. However, it often lacks the ability to recover fine local textures, leading to slightly unnatural reconstructions. More recent models like HAT have attempted to refine SwinIR's limitations, but they come at a high computational cost, making them difficult to deploy in real-world applications. Despite these advancements, hybrid models that effectively integrate CNN-based local feature refinement with Transformer-based global context learning remain an underexplored research area. The proposed Hybrid Swin Transformer ISR model aims to bridge this gap by combining CNN-based texture enhancement with Swin Transformer's self-attention, leading to high-fidelity super-resolution with improved detail preservation and reduced artifacts.

Methodology

The proposed model combines the strengths of CNNs and Swin Transformers to achieve high-fidelity image super-resolution. The CNN-based feature extractor captures fine local details and textures, ensuring accurate edge reconstruction. Meanwhile, the Swin Transformer encoder enhances global feature representation by modeling long-range dependencies. The final upsampling and reconstruction module utilizes PixelShuffle or transposed convolution to upscale low-resolution images to high-resolution outputs. The model is trained on datasets such as DIV2K, Flickr2K, and Urban100, ensuring diverse and high-quality image restoration. To optimize performance, MSE loss is used for pixel-wise accuracy, while perceptual loss improves visual fidelity. The AdamW optimizer with a learning rate scheduler ensures stable and efficient training. Data augmentation techniques such as flipping, cropping, and rotation enhance model generalization. The proposed approach aims to outperform traditional CNN-based and Transformer-based super-resolution models. By integrating both local and global feature extraction, the model ensures sharp, detailed, and artifact-free reconstructions.

Experimental Evaluation

The proposed model will be evaluated using well-established benchmark datasets to ensure its effectiveness in real-world scenarios. The primary datasets include DIV2K, which consists of 800 high-resolution images for training and 100 for validation, Flickr2K, which provides additional high-quality images for better generalization, and Urban100 & Set5, which contain diverse scenes and textures for robust evaluation. These datasets offer a mix of natural and urban environments, allowing the model to generalize well across different image types.

To assess the model's performance, multiple evaluation metrics will be employed. Peak Signal-to-Noise Ratio (PSNR) will measure pixel-wise reconstruction accuracy and will attempt to use Structural Similarity Index Measure (SSIM) to evaluate the perceptual similarity between the super-resolved and ground-truth images.

The expected results of this hybrid approach include enhanced sharpness and improved texture preservation compared to traditional CNN-based super-resolution models. The integration of local feature extraction from CNNs and global self-attention from Swin Transformers is anticipated to yield higher PSNR and SSIM scores than existing models such as SwinIR and EDSR.

Reference

[1] Transformer for Single Image Super-Resolution Zhisheng Lu1† , Juncheng Li2† , Hong Liu1⇤ , Chaoyan Huang3, Linlin Zhang1, Tieyong Zeng2 1Peking University Shenzhen Graduate School 2The Chinese University of Hong Kong 3Nanjing University of Posts and Telecommunications.

[2] A Hybrid Network of CNN and Transformer for Lightweight Image Super-Resolution Jinsheng Fang1 , Hanjiang Lin1 , Xinyu Chen1 , Kun Zeng2* 1Minnan Normal University, China 2Minjiang University, China

[3] SwinIR: Image Restoration Using Swin Transformer Jingyun Liang1 Jiezhang Cao1 Guolei Sun1 Kai Zhang1,* Luc Van Gool1,2 Radu Timofte1 1Computer Vision Lab, ETH Zurich, Switzerland 2KU Leuven, Belgium

[4] Image Super-Resolution Using Very Deep Residual Channel Attention Networks Yulun Zhang1 , Kunpeng Li1 , Kai Li1 , Lichen Wang1 , Bineng Zhong1 , and Yun Fu1,2 1Department of ECE, Northeastern University, Boston, USA 2College of Computer and Information Science, Northeastern University, Boston, USA

[5] Enhanced Deep Residual Networks for Single Image Super-Resolution Bee Lim Sanghyun Son Heewon Kim Seungjun Nah Kyoung Mu Lee Department of ECE, ASRI, Seoul National University, 08826, Seoul, Korea