Research Paper Based Case Study & Implementation REPORT

Submitted By

Group No.	Roll No.	Name of the Student	Signature
13	17	Rutuja Balbudhe	Rutuja.
13	31	Aayush Zade	AZade
13	36	Arnab Chakraborty	
13	44	Jayushna Mahadule	

Referred Research Paper Title

"SwinIR: Image Restoration Using Swin Transformer"

Authors Name:

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, Radu Timofte

Journal Name:

"*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*"

DOI:

10.1109/ICCVW54120.2021.00162

Year:

2021

1. INTRODUCTION

The research paper explores the field of image super-resolution (SR), a branch of computer vision dedicated to enhance low-resolution images into their high-resolution versions. SR technology has applications in diverse fields like medical imaging, satellite imagery, and security, where high quality images are essential for accurate interpretation.

Traditional Super Resolution (SR) models, especially those which are based on convolutional neural networks (CNNs), have made substantial progress, but they were found to struggle in capturing long-range dependencies due to their limited focus on local features. Previous models like SRCNN and SRGAN brought advancements in deep learning for SR, but their performance is often limited in reproducing finer details and maintaining consistent textures across the entire image. Due to these limitations the SR approaches struggle to understand the overall structure of the image that could benefit from incorporating global context alongside local feature extraction.

To address these limitations, the authors of SwinIR developed a model build upon the Swin Transformer, a novel transformer architecture that introduces hierarchical self-attention within non-overlapping local windows. This design enables SwinIR to manage a balance between local detail extraction and global feature consistency, overcoming the CNN’s limited contextual understanding. The model thus achieves enhanced image quality while remaining computationally efficient, setting a new standard in SR performance.

SwinIR represents an advanced technique by introducing a transformer based model that can effectively handle image restoration tasks. This encourages further exploration of transformer architectures for Super Resolution tasks. It inspires researchers to adopt such methods that have the efficiency of CNNs with the context-awareness of transformers.

2. METHODOLOGY/APPROACH EXPLAINED IN RESEARCH PAPER

Network Architecture

The SwinIR framework processes input low-quality (LQ) images and outputs high-quality (HQ) images. The pipeline consists of three stages:

1. Shallow Feature Extraction

A **3×3 convolution layer** is used to extract initial features (F_0) from the input image (ILQ). This module preserves low-frequency components, which contribute to stable optimization and image fidelity during reconstruction.

$$F_0 = H_{SF}(ILQ)$$

2. Deep Feature Extraction

The core innovation of SwinIR lies in its **Deep Feature Extraction** module, which uses **Residual Swin Transformer Blocks (RSTBs)**. Each RSTB comprises several **Swin Transformer Layers (STLs)** for localized and global feature extraction. The features are enhanced using convolutional layers and aggregated through residual connections.

- **Swin Transformer Layers (STLs):**
  STLs divide the image into non-overlapping patches or windows and apply **local self-attention** within each window. To overcome the limitations of isolated windows, a **shifted window mechanism** alternates between regular and shifted partitions, enabling cross-window connections.

$$F_{i,j}=H_{STL(i,j}(F_{i,j-1}),\ i=1,2,...,K$$

- **RSTB Output:**
  Each RSTB outputs refined features that are fed to subsequent layers. A convolutional layer follows each block to refine the extracted features.

$$F_{DF}=H_{DF}(F_0)$$

High-Quality Image Reconstruction

The reconstruction module aggregates shallow (F_0) and deep (F_{DF}) features for generating the HQ output image (I_{HQ}). The module varies depending on the task:

- **For Super-Resolution (SR):**
  A sub-pixel convolution layer is used for upscaling.
- **For Denoising and JPEG Artifact Reduction:**
  A simple convolutional layer is applied for refinement.

$$I_{HQ}=H_{REC}(F_0+F_{DF})$$

Additionally, residual learning is used for tasks like denoising, where the model predicts the difference between the LQ and HQ images.

$$I_{HQ}=H_{SwinIR}(I_{LQ})+I_{LQ}$$

Residual Swin Transformer Block (RSTB)

An **RSTB**, combines Swin Transformer Layers with convolutional layers and residual connections. Each block ensures efficient feature extraction by modeling long-range dependencies and enhancing local details.

Key Components:

1. **Swin Transformer Layers (STLs):**
   - Uses **multi-head self-attention (MSA)** within local windows to capture context.
   - Applies a **shifted window partition** for cross-window interaction.

$$\text{Attention}(Q,K,V)=\text{SoftMax}(QK^{T}/\sqrt{d} + B)V$$

2. **Residual Connections:**
   Provides skip paths for better gradient flow and aggregation of multi-level features.

3. **Convolutional Layers:**
   Enhances translational invariance and refines features extracted from STLs.

Loss Functions

The model uses task-specific loss functions to optimize the parameters:

- **L1 Loss (for SR):**

  $\mathscr{L}=\| I_{HQ}-I_{GT}\|$ L=‖IHQ−IGT‖2+ϵ2
- **Charbonnier Loss (for denoising and artifact reduction):**

  $\mathscr{L} = \sqrt{\| \mathrm{IHQ} - \mathrm{IGT}|| 2 + \epsilon 2}$

3. CLAIMS BY THE AUTHORS:

The author claims that using a three part architecture consisting of: Shallow Feature Extraction which utilizes a convolutional layer for low-frequency information, Deep Feature Extraction which Employs residual Swin Transformer blocks (RSTBs) for local attention and cross-window interactions, and High Quality Image Reconstruction which fuses shallow and deep features for detailed and accurate image restoration.

SwinIR can -

- Outperform state of the art techniques for a multitude of tasks related to image restoration using super resolution, image denoising, and artifact reduction.
- Accomplish a top performance in testing situations, offering up to a 0.45dB better result in PSNR(peak signal to noise ratio).
- Reduce the number of parameters required by 67%.
- Provide detailed and accurate results for image restoration using enhanced detail processing.
- Use a shifted window strategy to model interactions in order to avoid border artifacts.
- Provide strong, robust, and accurate results even when trained on a smaller dataset.

- Better than BSRGAN at restoring architectural images as BSRGAN may be better at face restoration, but worse at building restoration because it uses different datasets in training.

4. SUMMARY OF THE STUDIED RESEARCH PAPER

- Proposed SwinIR, a Transformer-based model for image restoration tasks like super-resolution, denoising, and artifact reduction.
- Utilized Residual Swin Transformer Blocks (RSTBs) with shifted window-based attention for effective long-range dependency modeling.
- Incorporated shallow feature extraction (3×3 convolution), deep feature extraction (RSTBs), and task-specific reconstruction modules.
- Achieved state-of-the-art performance across benchmarks with fewer parameters than CNN-based methods.
- Demonstrated versatility in real-world image restoration, excelling in PSNR and SSIM metrics.
- Emphasized parameter efficiency, better visual quality, and adaptability to diverse restoration tasks.
- Combined local attention within windows and cross-window interaction using the shifted window mechanism.
- Leveraged residual learning for stable training and efficient multi-level feature aggregation.
- Outperformed traditional methods in handling complex degradations while maintaining computational efficiency.

5. RELATED WORK

1. SwinIR: Image Restoration Using Swin-Transformer (2021): SwinIR, a model based on Swin Transformers, is presented in the study for image restoration tasks such as super-resolution, denoising, and reducing artifacts from JPEG compression. The components of SwinIR are high-quality image reconstruction, deep feature extraction using residual Swin Transformer blocks, and shallow feature extraction. SwinIR reduces the number of parameters by up to 67% and outperforms state-of-the-art techniques by up to 0.45 dB, according to extensive trials. Across a range of picture restoration tasks, such as classical, lightweight, and real-world image super-resolution, the model performs admirably. SwinIR will eventually be expanded to include other restoration tasks, such as image deraining and deblurring.

2. Hierarchical Conditional Flow: A Unified Framework for Image Super-Resolution and Image Rescaling (2021) : The study introduces HCFlow, a unified framework for image rescaling and super-resolution (SR) that can simulate distributions of both low-resolution (LR) and high-resolution (HR) images. In order to model high-frequency components based on the LR picture, HCFlow uses a hierarchical conditional technique to learn a bijective mapping between HR and LR images. Enhanced by perceptual and GAN losses, the model is trained with negative log-likelihood loss. HCFlow achieves state-of-the-art performance in picture rescaling, face image SR, and general image SR, according to extensive testing. In terms of both visual quality and quantitative metrics, this cohesive strategy performs better than current techniques.

3. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution(2021): The study presents a novel degradation model for deep blind image super-resolution (SISR) that incorporates noise, down-sampling, and random blur to more accurately represent actual picture circumstances. Multiple interpolation techniques and noise levels are used to vary down-sampling and noise, while isotropic and anisotropic Gaussian kernels are used to simulate the blur. Using this intricate degradation model, a deep ESRGAN model is trained, enhancing performance on both artificial and real images. The new model greatly improves blind super-resolution's performance in real-world situations, according to experimental results. A viable way to enhance deep SISR models for practical uses is provided by this method.

4. Pre-Trained Image Processing Transformer (2021): Image Processing Transformer (IPT), a pre-trained transformer model, is presented in the study for low-level computer vision tasks such as deraining, super-resolution, and denoising. Using multi-heads and multi-tails for task adaptability, the model is trained on extensive corrupted image pairs produced by ImageNet. IPT uses contrastive learning to enhance performance on a range of tasks. IPT surpasses current state-of-the-art techniques that rely solely on a single pre-trained model that can be adjusted for particular tasks, according to experimental results. With plans for further developments, the study emphasizes how transformers might improve image processing jobs.

5. MFAGAN: A Compression Framework for Memory-Efficient On-Device Super-Resolution GAN: A compression framework called MFAGAN is proposed in the paper for memory-efficient on-device super-resolution GANs. MFAGAN prevents memory explosion and lowers memory access costs by using a multi-scale feature aggregation network as the generator. Both the generator and the discriminator employ knowledge distillation, and PatchGAN is introduced as a discriminator. For mobile device optimization, hardware-aware neural architecture search (NAS) is used. With minimum compromise of visual quality, MFAGAN can save up to 8.3× of memory and reduce computation by 42.9×, allowing for faster inference on mobile devices such as the Qualcomm Snapdragon 865 chipset.

6. Meta-SR: A Magnification-Arbitrary Network for Super-Resolution(2019): Using a single model, the research presents Meta-SR, a model for super-resolution of arbitrary scale factors, including non-integer values. Based on the scale factor, it suggests the Meta-Upscale Module, which dynamically anticipates the upscale filters' weights. By doing away with the requirement for distinct models for various scale factors, this method offers computational efficiency. The technique produces high-resolution photos with higher performance when tested on common benchmark datasets. Meta-SR's versatility and efficiency stem from its ability to continuously zoom in on photos with any scale factor.

7. EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis (2017): EnhanceNet focuses on automated texture generation in conjunction with a perceptual loss to make a fresh approach to single image super-resolution. At high magnification ratios, this approach enhances image quality by emphasizing realistic texture production, in contrast to conventional techniques that maximize pixel-wise precision. Utilizing feed-forward fully convolutional neural networks in an adversarial training environment, the method produces cutting-edge quantitative and qualitative outcomes. Although the model produces images that appear more realistic, there are still certain drawbacks, like sporadic artifacts and pixel-by-pixel inconsistencies with the ground truth. Enhancing real-time performance and investigating video super-resolution are two areas of future research.

8. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation (2021): Swin-Unet is a new Transformer-based model with a U-shaped encoder-decoder architecture for medical picture segmentation. It uses the Swin Transformer to collect global and long-range semantic information through hierarchical, shifted windows, in contrast to CNN-based methods. For local-global feature learning and upsampling procedures to restore spatial resolution, the model makes use of skip connections. Swin-Unet exhibits superior performance than conventional convolutional or hybrid convolution-transformer algorithms in experiments on multi-organ and cardiac segmentation tasks. For tasks involving the segmentation of medical images, the model exhibits good performance and generalization capabilities.

9. Image Super-Resolution with Non-Local Sparse Attention(2021): The paper proposes a novel Non-Local Sparse Attention (NLSA) technique for Single Image Super-Resolution (SISR), combining the strengths of non-local operations and sparse representation. NLSA uses spherical locality-sensitive hashing (LSH) to partition the input space into related feature buckets, reducing computational complexity and enhancing efficiency. This sparse attention mechanism focuses on the most informative regions, improving both performance and robustness. The model, called Non-Local Sparse Network (NLSN), achieves state-of-the-art results on SISR tasks, outperforming traditional methods. Extensive experiments confirm its effectiveness in both quantitative and qualitative assessments.

10. Mutual Affine Network for Spatially Variant Kernel Estimation in Blind Image Super-Resolution(2021): The Mutual Affine Network (MANet) for blind image super-resolution (SR) is presented in this research. It tackles the problem of spatially variable blur kernels, which are prevalent in real images. To improve feature expressiveness while preserving computational efficiency, MANet makes use of a unique mutual affine convolution (MAConv) layer with a moderate receptive field. Affine transformations between channel splits are used by the model to take advantage of channel interdependence. In both spatially variable and invariant scenarios, MANet achieves state-of-the-art performance in blind SR, as demonstrated by extensive trials. Future research will examine GAN-based training and real-world degradations to enhance visual quality.

11. Residual Feature Aggregation Network for Image Super-Resolution(2020): The Residual Feature Aggregation (RFA) framework for image super-resolution (SISR) is presented in this research. It enhances feature extraction by aggregating hierarchical residual features through skip connections. To improve this framework's efficacy, the authors suggest focusing on key spatial features with an Enhanced Spatial Attention (ESA) block. RFA and ESA work together to create the RFANet, a network that performs better than the most advanced SISR techniques. RFANet's exceptional performance in both numeric and visual domains is demonstrated by experimental findings on conventional benchmarks. The suggested method makes good use of residual features to improve detail reconstruction.

12. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks (2018): With an emphasis on enhancing model performance through innovative structures and methodologies, the research articles examine developments in Single Image Super-Resolution (SISR). New standards in SISR are established by one work that presents the Non-Local Sparse Attention (NLSA) approach, which combines sparse attention and non-local operations for accuracy and efficiency. In a different publication, a Mutual Affine Network (MANet) is proposed to estimate spatially variable blur kernels, which reduces artifacts and improves image sharpness. ESRGAN, a third study that won first place in the PIRM-SR Challenge, uses a relativistic GAN with the Residual-in-Residual Dense Block (RRDB) to improve texture recovery and visual quality. Taken together, these methods help create image super-resolution that is more realistic, efficient, and perceptually correct.

Reference No.	Methodology Used	Dataset Used /Images Used	Performance Evaluation Parameters	Claims by Author(s)	Your Findings
[1]SwinIR: Image Restoration Using Swin Transformer (2021)	SwinIR (Swin Transformer-based Image Restoration),	Urban100 dataset	PSNR	Effective noise removal,High generalizability demonstrated on diverse restoration tasks	Promising results in preserving image textures and handling noise effectively
[2]Hierarchical Conditional Flow:A	Uses a hierarchical conditional flow-based architecture	DIV2K dataset	PSNR, SSIM	Successfully unifies image super-	HCFlow is efficient, Broaden

Unified Framework for Image Super-Resolution and Image Rescaling. (2021)	with invertible mapping and multi-scale modeling			resolution and image rescaling tasks	validation to include more datasets and diverse applications.
[3]Designing a Practical Degradation Model for Deep Blind Image Super-Resolution(2021)	A new degradation model with random combinations of blur, downsampling, and noise is used to train a deep blind super-resolution model	RealSRSet dataset for real-world degradation scenarios	PSNR LPIPS	improves real-world SISR applicability ,outperforming existing methods visually	effective visually but BSRGAN struggles with quantitative metrics, data diversity is needed
[4]Pre-Trained Image Processing Transformer (2021)	A pre-trained transformer model (IPT) with multi-heads, multi-tails, and contrastive learning	Synthesized ImageNet dataset	PSNR	The IPT model surpasses state-of-the-art methods across multiple low-level vision tasks	IPT demonstrates high adaptability and performance gains on large-scale data
[5]MFAGAN (2021)	A memory-efficient GAN-based framework (MFAGAN) using multi-scale feature aggregation, PatchGAN discriminator	Datasets include Set5, Set14, and B100, primarily for 4× super-resolution tasks	PSNR, LPIPS, memory usage, FLOPs, parameter size, and mobile latency	MFAGAN achieves up to 8.3× memory savings, 42.9× computation reduction	MFAGAN effectively balances compression, speed, and visual quality, making it a practical solution for deploying GAN

[6]Meta-SR: A Magnification-Arbitrary Network for Super-Resolution(2019)	Meta-Upscale Module,Dynamically predicts weights for upscale filters based on the input scale factor	DIV2K dataset	PSNR, SSIM	Achieves comparable or better results than traditional RDN(residual dense network)models trained independently for specific scales	Meta-SR's approach addresses the inefficiency of training multiple models for SR(super resolution) across different scales
[7]EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis (2017)	Combines texture synthesis, adversarial training, perceptual loss, and texture transfer loss for realistic SR	MSCOCO dataset, Urban100 dataset	PSNR, SSIM, IFC	ENet-PAT (perceptual quality) excels in perceptual quality; ENet-E (pixel accuracy) achieves state-of-the-art PSNR; both models are efficien	ENet-PAT generates realistic textures but may introduce artifacts; ENet-E provides high pixel accuracy
[8]Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation (2021)	Utilizes a Swin Transformer-based encoder-decoder with patch merging and patch expanding layers for hierarchical feature learning and resolution restoration	Trained and evaluated on Synapse multi-organ segmentation dataset and ACDC cardiac segmentation dataset	egmentation accuracy, robustness, generalization ability, and computational efficiency	Swin-Unet outperforms convolutional and hybrid approaches, demonstrates robustness and generalization across tasks.	The patch expanding layer and skip connections enhance segmentation performance, Transformer pre-training and 3D applications

					need further exploration.
[9]Image Super-Resolution with Non-Local Sparse Attention(2021)	Introduces Non-Local Sparse Attention (NLSA) to focus on the most informative regions using spherical locality-sensitive hashing (LSH)	Training dataset: DIV2K Testing dataset: Urban100, Manga109.	Evaluated using PSNR and SSIM on the Y channel in YCbCr space	Demonstrates robustness and efficiency with sparse attention mechanisms for SISR tasks	Adding NLSA blocks enhances global modeling by identifying informative regions without processing unrelated areas
[10]Mutual Affine Network for Spatially Variant Kernel Estimation in Blind Image Super-Resolution(2021)	Introduced MANet with moderate receptive field and MAConv layer for efficient spatially variant kernel estimation.	Synthetic datasets	Visual sharpness, artifact reduction, edge preservation, and kernel estimation accuracy	Achieves state-of-the-art blind SR performance with sharper, natural edges and fewer artifacts	MANet effectively handles spatially variant degradations, but further improvement is possible with GAN.
[11]Residual Feature Aggregation Network for Image Super-Resolution(2020)	Proposes a Residual Feature Aggregation (RFA) framework and Enhanced Spatial Attention (ESA) blocks for efficient and accurate super-resolution.	Urban100 dataset, standard SISR datasets	PSNR, parameter count,qualitative visualization of feature maps to assess accuracy, efficiency, and spatial detail recovery.	RFANet achieves state-of-the-art performance in super-resolution with fewer parameters	The RFA framework efficiently utilizes hierarchical residual features, while ESA enhances spatial attention, resulting in superior super-resolution quality with a

					lightweight architecture
[12]ESRGAN : Enhanced Super-Resolution Generative Adversarial Networks (2018)	ESRGAN, which introduces the Residual-in-Residual Dense Block (RRDB) without batch normalization, relativistic GAN for the discriminator, and an improved perceptual loss for texture recovery.	Pristine dataset	Perceptual index, PSNR, and visual quality	ESRGAN achieves superior visual quality with more realistic textures than SRGAN	ESRGAN improves SR performance by enhancing texture recovery and brightness consistency, outperforming previous methods

*PSNR- Peak Signal to Noise Ratio, CR- Compression Ratio, CT- Compression Time, DT- Decompression Time, SSIM- Self Similarity Index Measure, CM- Confusion Matrix, LPIPS - learned perceptual image patch similarity

6. WHAT DO YOU LEARN FROM STUDIED RESEARCH PAPER?

- The Swin Transformer architecture effectively combines the strengths of CNNs and Transformers, addressing challenges like long-range dependency modeling and content-based interactions.
- Shifted window attention is a powerful mechanism to balance computational efficiency and feature representation, enabling better local and global context modeling.
- The use of Residual Swin Transformer Blocks (RSTBs) demonstrates the importance of residual learning in stabilizing training and improving multi-level feature extraction.
- Task-specific reconstruction modules, such as sub-pixel convolution for super-resolution, can optimize performance across diverse image restoration tasks.
- Transformer-based models like SwinIR outperform CNN-based methods in image restoration tasks while requiring fewer parameters and less computational overhead.
- Loss functions tailored to specific tasks, such as L1 loss and Charbonnier loss, are crucial for optimizing model performance and achieving high visual fidelity.
- The research highlights the generalizability of the SwinIR architecture across various image restoration tasks, including super-resolution, denoising, and JPEG artifact reduction.
- Parameter efficiency and faster convergence make SwinIR suitable for real-world applications and resource-constrained environments.
- Real-world image degradation requires robust models like SwinIR that can adapt and perform well in uncontrolled conditions.

7. DEFINE THE IDENTIFIED PROBLEM RELEVANT TO THE STUDIED RESEARCH PAPER. ELABORATE THE PROBLEM DEFINITION.

The **SwinIR** paper addresses a key problem in super-resolution (SR): the limitations of traditional CNN-based methods, which struggle to capture long-range dependencies due to their local receptive fields. This constraint hinders SR performance, particularly in reconstructing detailed textures and maintaining large-scale structural coherence in high-resolution images.

Mathematical Formulation of the Problem: The SR task aims to learn a mapping function f_θ that transforms a low-resolution (LR) image I_{LR} to a high-resolution (HR) image I_{HR}, minimizing the loss $\mathscr{L}$between the generated $\hat{\square}_{HR}$ and ground truth I_{HR}:

$$\theta^* = \arg \boxed{f()} \min \boxed{f()}_{\theta} \mathscr{L}(f_\theta(I_{LR}), I_{HR})$$

where $\mathscr{L}$ is often **MSE** or **L1 loss**.

Challenges and SwinIR's Solution:

1. **Long-Range Dependencies**: CNNs overlook global dependencies, impacting SR quality.
2. **Model Complexity**: Larger receptive fields increase computational costs, limiting efficiency.

SwinIR Approach: SwinIR introduces **Swin Transformer**, using self-attention in non-overlapping windows to capture both local and global features efficiently, striking a balance in SR reconstruction quality and computational cost.

8. WHICH DATASET YOU HAVE USED FOR EXPERIMENTATION?

The paper utilized multiple datasets for different image restoration tasks, sourced from well-established

1. **Super-Resolution (SR):**
   - **DIV2K:** High-quality 2K resolution images used for training and evaluation.
   - **Flickr2K:** Additional training data for improved performance.
   - **Benchmark Datasets:**
     - **Set5, Set14, BSD100, Urban100, Manga109** – Standard datasets used to evaluate PSNR and SSIM for SR tasks.
2. **Image Denoising:**
   - **Set12:** A collection of grayscale images.
   - **BSD68:** Grayscale image dataset for benchmarking.

- **Urban100:** High-resolution urban scene dataset for testing noise reduction performance.
- **CBSD68:** Color version of BSD68, used for color denoising.
- **Kodak24 and McMaster:** High-quality color image datasets for evaluating denoising performance.

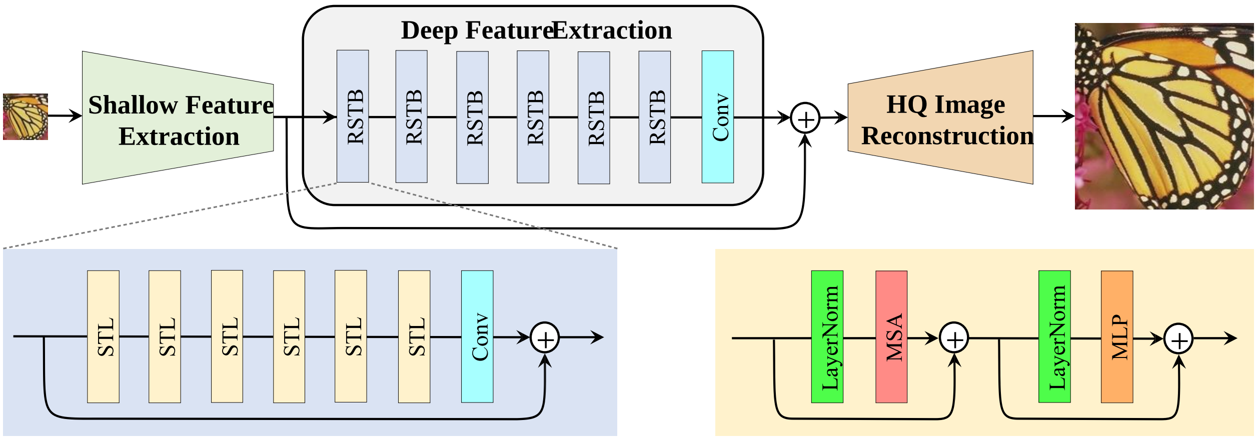
3. **JPEG Artifact Reduction:**
    - **Classic5:** A set of five standard images widely used for compression artifact testing.
    - **LIVE1:** A benchmark dataset for evaluating compression artifact removal.

9. HOW YOU HAVE CARRIED OUT IMPLEMENTATION?

9.1 Methodology Used for Implementation

The approach used in the paper to restore images utilizes the SwinIR (Swin Transformer for Image Restoration) architecture, which is designed for efficiently reconstructing high-quality images from low quality input image (degraded input).

The model works in three stages:



(a) Residual Swin Transformer Block (RSTB)

(b) Swin Transformer Layer (STL)

1. Shallow Feature Extraction

A single 3×3 convolutional layer is applied to extract shallow features from the low-quality input image. This layer preserves the low-frequency information and maps the image space into a higher-dimensional feature space.

2. Deep Feature Extraction

The deep feature extraction module consists of multiple Residual Swin Transformer Blocks(RSTBs). Each RSTB includes multiple Swin Transformer layers (STLs) for local attention and cross-window interactions. A residual connection is added to compile features across blocks for efficient training and better feature representation. The Swin Transformer layers use a shifted window mechanism to capture both local and long-range dependencies while avoiding border artifacts that arise during patch-based processing. At the end of the block, a 3×3 convolutional layer is used for further feature enhancement.

3. High-Quality Image Reconstruction

Both shallow and deep features are combined through a long skip connection to aggregate low- and high-frequency information. For super-resolution, the reconstruction module uses a sub-pixel convolution layer to upsample features to the desired resolution. For denoising and artifact reduction, a single convolutional layer is employed to directly reconstruct the high-quality output. The model uses residual learning to predict the residual between the degraded and high-quality images, which helps focus on the lost details.

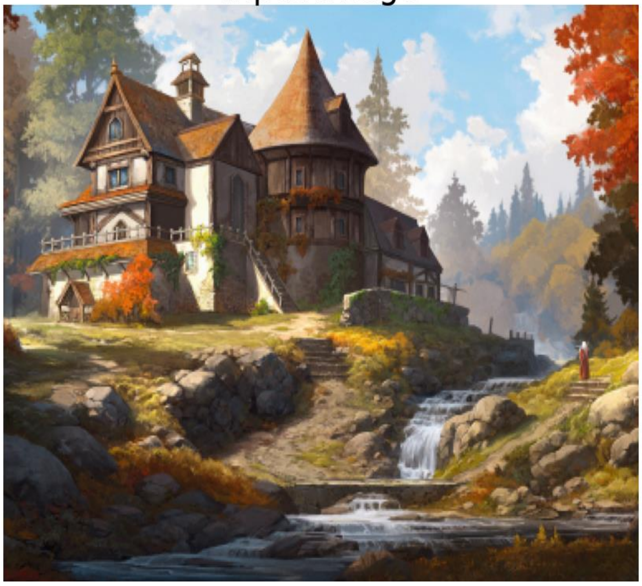
There are different loss functions used to optimize the restoration task:

- The **L1 pixel loss** used to optimize classical super resolution.
- **Pixel loss, GAN loss, and Perception loss** for real world super resolution.
- **Charbonnier loss** for denoising and artifact reduction.

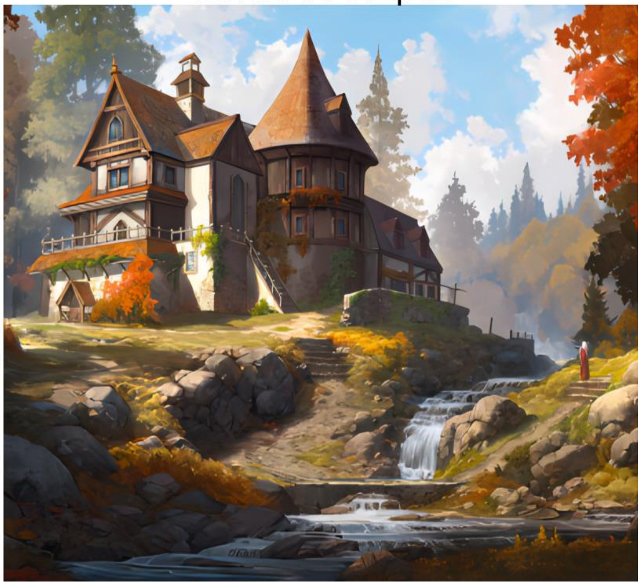
9.2 Obtained Experimental Results

In the obtained outputs where a degraded i.e. a low quality, low fidelity image was given as an input, the resultant SwinIR image was compared with the original as well as its predecessor, the BSRGAN model which introduced a Practical Degradation Model for Deep Blind Image Super-Resolution. The original SwinIR output is also compared with SwinIR-Large which is trained on a larger image dataset. In the obtained results it can be clearly seen that BSRGAN focuses mainly on improving the psychovisual quality of an image. As a result, the loss of pixel detail in BSRGAN is quite apparent. Since BSRGAN works on deprecated fundamentals, it may also introduce sharp edges and artifacts due to oversharpening in some areas. The output obtained from the SwinIR model, on the other hand, focuses on improving perception without losing fine detail striking a balance between perceptual quality and pixel accuracy. The SwinIR-Large is by and far the best model for image restoration tasks. It is evident from the obtained results that SwinIR-L manages to recover the most amount of detail from a degraded image whilst enhancing pixel accuracy and color representation. Since SwinIR-L is computationally expensive, it is recommended to use SwinIR-Base for lightweight tasks and SwinIR-Large should only be used for restoring heavily degraded images.

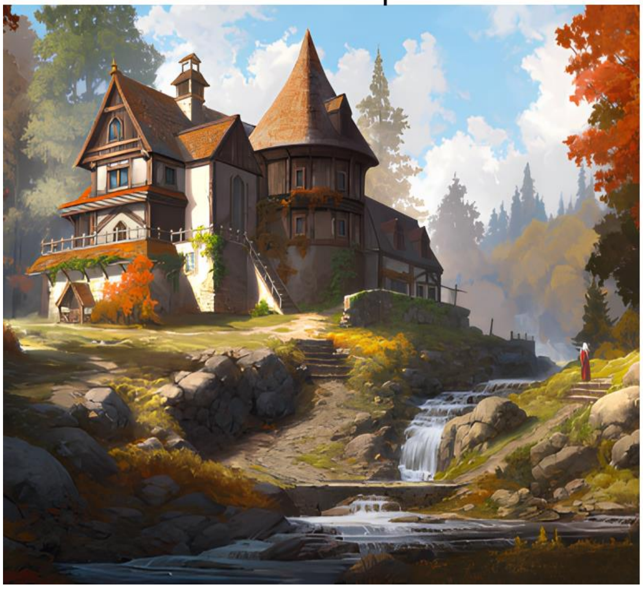
Input image



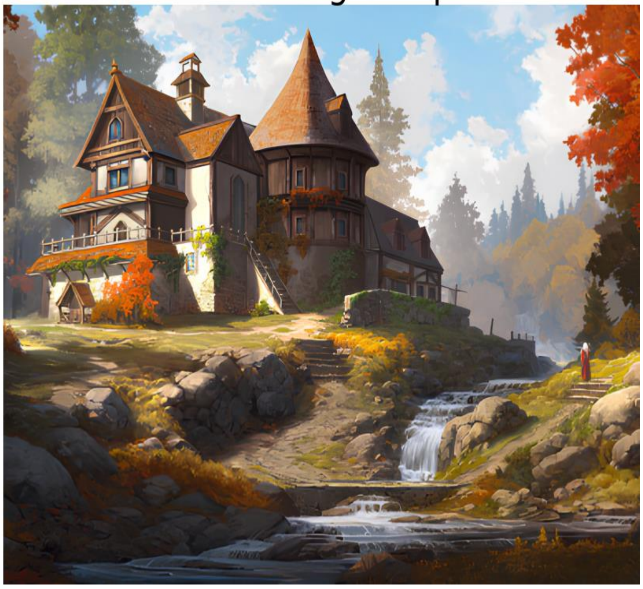
BSRGAN output



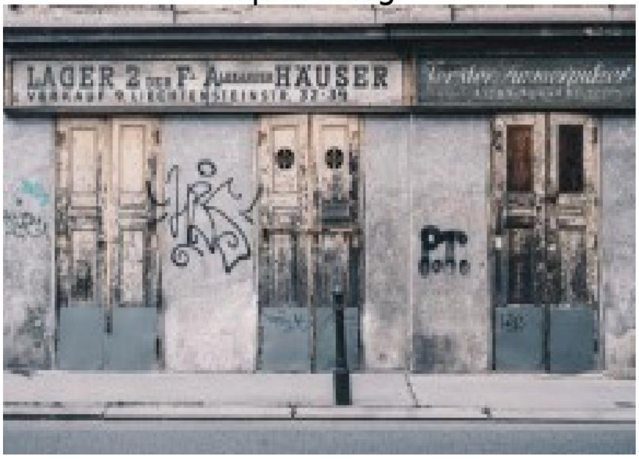
SwinIR output



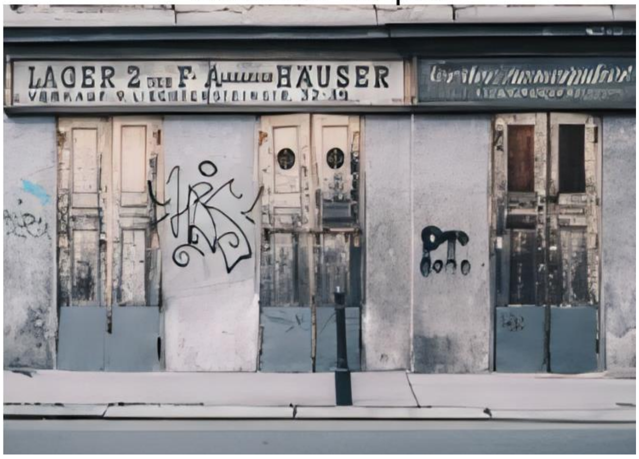
SwinIR-Large output



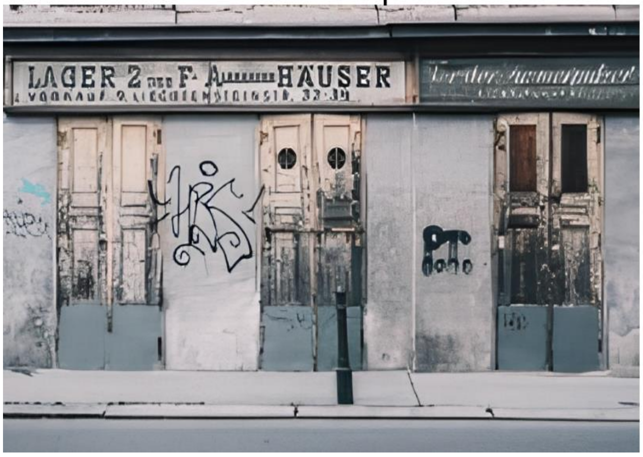
Input image



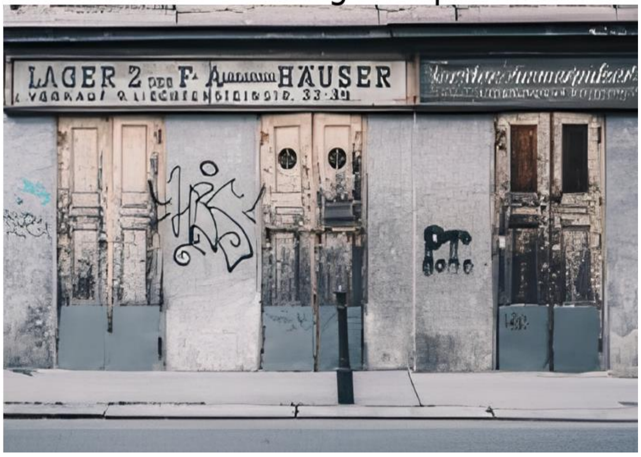
BSRGAN output



SwinIR output



SwinIR-Large output



9.3 Conclusion Based on Obtained Results & Future Scope

Conclusion:

- SwinIR is a state of the art model that can be used for image restoration tasks including, but not limited to, super-resolution, denoising, and artifact reduction.
- SwinIR can capture both the local and the global dependencies in images.
- SwinIR achieves better results than traditional image restoration techniques whilst being computationally efficient.
- The base model can handle lightweight image enhancement tasks with ease and maintains an unmatched level of accuracy.
- The enhanced model SwinIR large can handle complex and heavy image restoration tasks making it suitable for high-quality applications.

Future Scope:

- SwinIR could be used to improve real-world applications using image enhancement.
- It can be modified to do other tasks related to image enhancement like deblurring, band reduction, false contour reduction, dehazing, etc. making it the ultimate model for image restoration.
- SwinIR can be used with generative frameworks to generate brand new image representations based on user prompts to obtain artistic content with photo realistic detail.
- Images or manuscripts of historical importance can be studied better by enhancing the engravings and the relevant text with SwinIR making it an indispensable tool to archaeologists and historians as well as literature scientists improving advancements in the field of linguistics.

These advancements could make SwinIR one of the most important models in the field of image restoration in computer vision and digital image processing.

10. REFERENCES

[1]Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In IEEE Conference on International Conference on Computer Vision Workshops, 2021

[2]Jingyun Liang, Andreas Lugmayr, Kai Zhang, Martin Danelljan, Luc Van Gool, and Radu Timofte. Hierarchical conditional flow: A unified framework for image superresolution and image rescaling. In IEEE Conference on International Conference on Computer Vision, 2021

[3]Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In IEEE Conference on International Conference on Computer Vision, 2021

[4]Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In IEEE Conference on Computer Vision and Pattern Recognition, pages 12299–12310, 2021

[5]Wenlong Cheng, Mingbo Zhao, Zhiling Ye, and Shuhang Gu. Mfagan: A compression framework for memory efficient on-device super-resolution gan. arXiv preprint arXiv:2107.12679, 2021

[6]Mehdi S. M. Sajjadi Bernhard Scholkopf Michael Hirsch ¨ Max Planck Institute for Intelligent Systems Spemanstr. 34, 72076 Tubingen, Germany.EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis 2017

[7]Xuecai Hu∗ 1,2 , Haoyuan Mu∗ 4 , Xiangyu Zhang3 , Zilei Wang1 , Tieniu Tan1,2 , Jian Sun3 1 University of Science and Technology of China 2 Center for Research on Intelligent Perception and Computing, NLPR, CASIA 3 Megvii Inc (Face++) 4 Tsinghua UniversityMeta-SR: A Magnification-Arbitrary Network for Super-Resolution 2019

[8]Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537, 2021

[9]Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image superresolution with non-local sparse attention. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3517–3526, 2021

[10]Jingyun Liang, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Mutual affine network for spatially variant kernel estimation in blind image super-resolution. In IEEE Conference on International Conference on Computer Vision, 2021

[11]Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2359–2368, 2020

[12] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In European Conference on Computer Vision Workshops, pages 701–710, 2018