International Journal of Innovative Research and Scientific Studies, 8(3) 2025, pages: 3734-3749

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Utilizing image super-resolution to overcome information bottlenecks in vision transformers

Jamal Zraqou[1*], Riyad Alrousan[2], Hussam Fakhouri[3], Bilal Sowan[4], Jawad Alkhatib[5]

[1]*Department of Computer Science, Faculty of Information Technology, University of Petra, Amman, Jordan.*
[2]*Department of Design & Visual Communication, School of SABE), German Jordanian University (GJU), Amman, Jordan.*
[3]*Department of Artificial Intelligence & Data Science, Faculty of IT, University of Petra, Amman, Jordan.*
[4]*Department of Business Intelligence & Data Analytics, Faculty of Administrative & Financial Sciences, UOP, Amman, Jordan.*
[5]*Department of Computer Engineering, Prince Mohamad Bin Fahd University, Dhahran, Saudi Arabia.*

Corresponding author: Jamal Zraqou (*Email: Jamal.Zraqou@uop.edu.jo*)

Abstract

This research tends to solve the information bottleneck challenge in vision transformer-based solutions for image super-resolution, where the intensity of the feature map reduces in deeper network layers, thus affecting model performance. LITRL, the Layer-Interconnected Transformer with Residual Links, provides stability to the information flow by means of the dense residual connections between the layers, with the aim of preventing spatial information loss. The methodology involves the integration of the Swin transformer architecture and new schemes of interconnections to maintain vital spatial features in the whole network. Experimental results show that the LITRL-based method gives better results on traditional benchmark datasets (Set5, Set14, BSD100, Urban100, Manga109), in terms of quantitative (PSNR, SSIM) and qualitative evaluation. At 4×, LITRL obtains PSNR/SSIM of 40.37/0.9628 on Set 5 and 35.70/0.9408 on Urban100 with far higher performance than comparable methods. The proposed LITRL model dramatically reduces the information bottleneck of transformer-based super-resolution. It retains fundamental spatial information due to the dense-residual connections, giving rise to sharper images with more natural textures and fewer artefacts. Practical Implications: The excellent performance of LITRL in generating complex textures and structures that, in turn, enables accurate reconstruction, makes the method particularly useful for the tasks where the retention of a high level of fidelity of the image enhancement is imperative, i.e., for medical imaging, analysis of satellite images, and developing digital content while requiring a reasonable computational efficiency.

Keywords: CNN, Feature map intensity, Information bottlenecks, Super-resolution, Swin transformer.

DOI: 10.53894/ijirss.v8i3.7383
Funding: This study received no specific financial support.
History: Received: 18 March 2025 / **Revised:** 21 April 2025 / **Accepted:** 23 April 2025 / **Published:** 26 May 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Acknowledgment: Our sincerest gratitude is directed to the University of Petra (UOP) for the great help and advice that they have provided during the research course. UOP has supported us, including the provision of resources and an academically creative environment, which has led to the development of this paper, and we would like to acknowledge their support. Friendly relationships and interest in academic success at the University of Petra have also made this research successful. We are really thankful to them for their support and assistance in the completion of this work.
Publisher: Innovative Research Publishing

1. Introduction

The term super-resolution (SR) is widely used in computer vision to describe techniques for enhancing image resolution. SR may refer to enhanced resolution or the reconstruction of high-quality images from low-resolution inputs.

Image quality improvement in optical vision systems has a long history, with numerous methods developed and applied in various domains. Traditional approaches, such as interpolation techniques (e.g., bilinear and bicubic interpolation), often produce images with significant blurring and loss of fine details, limiting their practical effectiveness. With the advent of deep learning, convolutional neural networks (CNNs) revolutionized single-image super-resolution (SISR). Models like SRCNN, VDSR, and EDSR demonstrate the ability to extract hierarchical features, thereby reconstructing high-resolution images from low-resolution inputs. However, CNN-based approaches are inherently limited by their restricted receptive fields, making it challenging to capture long-range dependencies and global contextual information within an image. This often results in suboptimal performance when reconstructing intricate textures and structures.

The introduction of transformers in vision tasks has addressed many of these limitations. Vision Transformers (ViTs) and their variants, such as Swin Transformers, leverage self-attention mechanisms to process images as patch sequences, effectively capturing local and global dependencies. These models have proven highly effective for tasks like super-resolution, where global context is crucial for reconstructing missing details. Notable contributions, such as SwinIR [1] introduced innovative mechanisms like shifted-window attention to balance computational efficiency with high-quality reconstructions.

Transformers were constructed for natural language processing tasks and were later adapted for vision tasks due to their capacity to model long-range dependencies. Vision Transformers (ViTs) and Swin Transformers have been successfully applied to high-resolution tasks. Unlike CNNs, transformers process the input as a sequence of patches, using self-attention mechanisms to model global relationships within an image. This has proven especially beneficial for tasks like super-resolution, where global context plays a crucial role in reconstructing missing details.

Recent advancements in image super-resolution (SR) conducted in Zhang et al. [2] highlighted the critical role of the convolutional neural network (CNN) depth in improving reconstruction quality. However, training deeper networks for SR remains a challenge due to the overwhelming presence of low-frequency information in low-resolution (LR) inputs and features, which can hinder the representational capability of CNNs. A Very Deep Residual Channel Attention Network (RCAN) has been proposed to address these issues. RCAN introduces a Residual in Residual (RIR) structure, comprising multiple residual groups with long skip connections and, within each group, residual blocks with short skip connections. This design allows the network to bypass abundant low-frequency information, enabling the main network to focus on learning high-frequency details. Additionally, the channel attention mechanism adaptively rescales channel-wise features by leveraging interdependencies among channels, enhancing the network's ability to capture critical features. Experimental results demonstrate that RCAN outperforms state-of-the-art accuracy and visual quality methods, establishing its effectiveness for high-quality image SR.

Figure 1, demonstrates the application of deep learning for SR on images extracted from ArcGIS Pro, showcasing the transformation from low-resolution (LR) to high-resolution (HR). The LR image, characterized by blurry edges, pixelation, and lack of detail, undergoes enhancement through deep learning techniques such as convolutional neural networks. These methods identify features, restore high-frequency details, and upscale the image to an HR output with sharper edges and smoother textures, enhancing clarity. This improvement significantly enhances the usability of the image in geospatial analysis, providing accurate details and supporting decision-making for urban planning and environmental monitoring obligations.

Despite these developments, though, transformer-based models suffer from feature map brightening in the deeper layers of networks, which causes an information bottleneck and spatial detail damage. Our proposal for solving this is the Layer-Interconnected Transformer with Residual Links (LITRL), which incorporates dense residual connections to Swin Transformer layers. This novel architecture alleviates the bottleneck by stabilizing the information flow, retrieving priceless spatial information, and boosting model performance.

Despite the vast studies conducted using both CNN- and transformer-based approaches recently, an in-depth analysis of their advantages and disadvantages is necessary for situating the scientific novelty of the developed approach in context. For example, CNN-based approaches are very good at local feature extraction but less successful at global dependency identification. Instead, transformers offer global modeling at the cost of additional computational complexity and the possibility of spatial detail loss in deeper layers. This paper intends to fill these gaps through a critical analysis of the existing methodologies by showing how LITRL is efficient compared to state-of-the-art methods on benchmark datasets.

Baseline studies greatly contribute to establishing the basis on which any model's performance is assessed. In image super-resolution (SR), traditional and deep learning methods have been widely used as benchmarks. These methods help understand what can and cannot be achieved by the existing methods, which creates a comparison grid for the new approaches.

Bicubic interpolation is a classical approach that is considered the traditional choice when it comes to estimating pixel values from their neighbouring pixels. Although computationally efficient, this method usually blurs the images and does not retain high-frequency details like edges and textures. Its limitations indicate the need for sophisticated models for reconstructing finer details in low-resolution (LR) inputs.

The early CNN-based methods are the SRCNN (Super-Resolution Convolutional Neural Network), VDSR (Very Deep Super-Resolution), and the RCAN (Residual Channel Attention Network). SRCNN [3] introduced convolutional neural networks for directly mapping LR inputs to high-resolution (HR) outputs. This simple yet effective architecture consists of only three convolutional layers, demonstrating significant improvements over traditional methods. However, SRCNN struggles with scalability and capturing complex textures due to its shallow architecture.

(a) LR of a satellite image. (b) SR of the satellite image

(c) LR of a building structure image (d) SR of the building structure image.

Figure 1.
Comparison of super-resolution satellite imagery results. The top row shows a low-resolution image aligned with the original high-resolution image, and the bottom row depicts the enhanced high-resolution output from the LR input image generated by the super-resolution algorithm, demonstrating improved clarity and detail restoration in building structures, vegetation, and surrounding areas.

Building on SRCNN, VDSR [2] introduced a deeper network to enhance the representation of high-frequency details. By incorporating residual learning, VDSR addressed issues related to gradient vanishing in deeper architectures, achieving better reconstruction quality. Nonetheless, its reliance on local operations limited its ability to capture global context, especially in images with intricate structures.

The RCAN presented in Zhang et al. [2] is an advanced CNN-based method that integrates channel attention mechanisms. This model adaptively reweighs channel-wise features to focus on informative regions, improving accuracy and visual quality. Despite its superior performance, RCAN's ability to model long-range dependencies remains constrained, a limitation shared by most CNN-based models.

Though these baseline approaches made great advances in SR, they suffered from many issues, including small receptive fields that made it difficult to capture global dependencies and context, loss of high-frequency details in areas with complex textures and fine structures, and a progressive decay in feature map intensity and representational capacities with deeper networks.

These limitations highlight the need for models that can account for local and global dependencies while maintaining crucial spatial information. These baseline methods underlie shortcomings that provide a definite reason for designing superior structures like the Layer-Interconnected Transformer with Residual Links (LITRL) to break these bottlenecks and

elevate the performance of single-image super-resolution to another level.

Specifically, the presented LITRL, a novel architecture, is aimed at addressing the information bottleneck problem in single-image super-resolution in this paper. With the aid of the dense residual links between them, LITRL corrects the information stream in the network and allows it to retain more spatial data. In the experiments described here, LITRL is shown to outperform state-of-the-art approaches on benchmarking datasets and in competitive challenges, e.g., the NTIRE-2024 Image Super-Resolution Challenge.

2. Literature Review

CNN-based methods have become the foundation of super-resolution investigations and studies ever since the inaugural SRCNN [4]. SRCNN attempted super-resolution with a simple architecture consisting of three layers following regular techniques, such as bilinear and bicubic interpolation. Subsequently, deeper structures such as VDSR and EDSR with multi-layer representations were proposed to obtain highly non-linear features from the low-resolution input. These models retained training stability and convergence, thus creating significant improvements in image quality through residual learning.

However, some issues remain with this CNN-based model. Several challenges exist with CNN-based models. First, the receptive field is limited, and in general, it has a field size equal to the size of the convolutional kernels [5]. Although using receptive fields larger than the current ones can be achieved through deeper networks, the approach is typically computationally costly, and there is usually a problem with gradient vanishing or exploding. Additionally, applying CNNs to capture long-range dependencies in images is challenging because they rely on local operations. This can lead to some loss of quality when reconstructing large areas or structures and textures that cover extensive areas of an image.

To address these drawbacks, different modifications to the CNN-based models have been established. For example, RCAN (Residual Channel Attention Network) [6] proposed attention mechanisms to pay more attention to the relevant features in reconstruction. This work, SAN (Second-order Attention Network), integrates second-order channel attention to improve the network's capability of modeling the correlation of channels. However, these models may evoke better global context modeling than the prior CNN-based models; nonetheless, they focus less on this aspect.

Learning-based multi-view techniques utilize various algorithms to reconstruct high-resolution images of a scene from low-resolution images. In the last few years, deep learning-based single-image super-resolution algorithms have suffered from some problems, such as feature extraction and nonlinear mapping, which cause high-frequency detail information loss and over-smooth reconstructed images. In ref [7], an improved deep convolutional neural network model was proposed to adopt special convolutional layers and the residual learning method, and the experimental results showed that the proposed method was superior to other effective methods in terms of image quality and detailed consistency. In prior work conducted in [7], multiple researchers formulated a method based on deep learning techniques for single image SR that contrasted with conventional approaches like the sparse coding-based SR techniques, which compartmentalized the functions of feature extraction, feature reconstruction, and feature optimization into different steps. This method employed a convolutional neural network (CNN) different from previous approaches to directly supervise and learn the mapping from LR images to HR images without intermediate steps. Unlike the traditional methods of restoration, this network tried to optimize all the layers simultaneously and was part of a single unification process that made the entire restoration process better and less complicated. Due to high efficiency in terms of computational complexity, CNNs were ideal for real-time and online analysis. In addition, testing different network structures and some parameters such as batch size, learning rate, momentum, dropout, etc., for optimizing FP32/S0754 performance vs. speed was attempted. The method was also further extended to process three color channels at once to achieve even smoother color reconstruction and further improved details of the reconstructions. This work provided state-of-the-art performance, thereby ensuring the practicality and efficiency of deep learning in SR tasks.

Super-resolution (SR) reconstruction has matured with deep learning-based techniques advancing over interpolation techniques such as bicubic interpolation. Following sparse and distributed representations, some rudimentary CNN models were used in single image super-resolution (SISR), including SRCNN and VDSR, which offered enhanced image quality but consistently experienced problems in retaining high-frequency feature detail. In Yan et al. [8]. Some of these difficulties were overcome with deeper networks, residual connections, and, depending on the problems, convolutional layers with sub-pixel operations. However, issues such as smoothness or loss of texture are still present. In the proposed model in this work, some of these methods are improved by applying a deep convolutional neural network with residual learning and subpixel layers to capture fine details and enhance the clarity of SR images without losing many edges and textures compared with traditional SR and other learning-based SR.

In Deng et al. [9] the traditional convolutional neural network (CNN)-based super-resolution methods face two primary challenges: limited receptive field size and the loss of edge information due to downscaling during convolution. The single-scale receptive field restricts the model's ability to capture fine details in smaller regions. At the same time, continuous convolution operations often require edge zeroing to preserve image dimensions, leading to the degradation of edge details. Recent approaches have explored multi-scale techniques and alternative architectures to address these issues. The suggested solution in Tian et al. [10] is a structure re-parameterized convolution in which, in the same layer, small and large kernels are placed in parallel, trained simultaneously, and then combined. This approach enables the higher kernel to capture the smaller details of the images, amplifying the high-frequency components and reducing the requirements for edge padding. The results of experiments show that this method enhances image quality and inference speed with competitive performance compared to state-of-the-art super-resolution techniques.

One of the problems with CNNs is the inability to capture complex scene details to provide accurate image super-resolution [10]. To solve this, the proposed dynamic network (DSRNet) comprises a residual enhancement block, a wide

enhancement block, a feature refinement block, and a construction block. The residual enhancement block can perform hierarchical feature extraction, and the wide enhancement block dynamically enhances robustness in complex scenes. The refinement block excludes interference between the components and uses residual learning to avoid long-term dependencies. Finally, the construction block reconstructs high-quality images. This lightweight architecture strikes a balance between performance, effectiveness, and the efficiency of mobile devices, and the results of experiments show competitive performance and recovery time.

The research presented in Kim et al. [11] introduced a new method for single-image super-resolution (SR) using an extraordinarily deep convolutional network, such as VGG-net. The model was constructed with a depth of 20 layers and cascaded small filters to exploit contextual information over a large area of an image. The authors presented the training of residuals to counter the slow convergence typically observed in deep networks. They employed very high learning rates, facilitated by the addition of adjustable gradient clipping. This approach formulates a new, more precise solution than previous methods and delivers visible qualitative gains.

Research work done in Zheng [12] was based on Image Super-Resolution Reconstruction (ISRR), with a view to generating high-resolution images from low-resolution inputs, a major challenge in image processing and computer vision. Focusing on the effectiveness of deep learning methods, the authors concentrated on convolutional neural networks (CNNs) and deep residual networks (ResNets) as a means of improving image quality. Additionally, they introduced a new algorithm, the Very Deep Super-Resolution (VDSR), which aims to enhance image reconstruction by utilizing deep network architectures. Experimental validation based on the DIV2K dataset demonstrated better performance and reliability of the VDSR algorithm compared to existing approaches. The paper also reviewed all contemporary deep learning-based ISRR algorithms and discussed future directions, indicating that relevant developments are being made.

In the most recent works that have been done in order to address the problem of limited capability a standard CNN model can have in upscaling images obtained from challenging scenes, DSRNet has been suggested in Shao et al. [13] as a dynamic network. DSRNet comprises several specialized components: A residual enhancement block with a residual architecture to allow efficient hierarchical feature extraction; a wide enhancement block with a dynamic architecture designed to enhance the model's ability to learn information and apply it across different scenes; a feature refinement block which uses a stacked architecture with embedded residual learning to extract accurate features and solve the long-term dependency problem; and a construction block which reconstructs clean images. This heterogeneous and, most of the time, very lightweight architecture provides additional structure to its information and can be easily displayed on several mobile digital devices. Benchmarks from the experiment have shown that DSRNet is comparable to the salient competitors in performance, image recovery speed, and computational efficiency. The implementation of DSRNet is publicly available at [https://]. Further technical details of this work are in the GitHub repository of the DSRNet model at [https://github.com/hellloxiaotian/DSRNet, last accessed on November 2024].

In Shao et al. [13] various kinds of image super-resolution (SR) reconstruction schemes were explored to improve the quality of degraded low-resolution (LR) pictures. Even though more advanced architectures of deep learning networks extending beyond conventional machine learning algorithms exist, they always come with numerous challenges such as high computational costs, vanishing gradient problems, and information loss. To address these issues, a sub-pixel convolutional neural network (SPCNN) was introduced for image SR reconstruction plans. These plans include converting images from the RGB to the YCbCr color space and using the Y channel image to generate the LR image to minimize strong correlations with redundant information while reducing computational time. Unlike other techniques that use interpolated images as input, as in the super-resolution convolutional neural network, SRCNN, the current method inputs the LR image directly into the network. The adopted network structure is as follows: the first and second layers are convolutional, and the fourth layer is a nonlinear mapping layer aimed at capturing features at various levels. A residual network is introduced to transfer feature information from lower to higher layers, addressing issues such as the explosion of gradients and vanishing gradients. Moreover, a subpixel convolution layer based on up-sampling is employed to reduce reconstruction time. Using three different datasets, the authors demonstrated that the proposed SPCNN has better reconstruction performance and less time consumption than Bicubic interpolation, SCSR, ANR, and SRCNN.

2.1. Transformer-Based Approaches for Super-Resolution

Transformers, used for natural language processing, have been applied in vision tasks because of their self-attention mechanisms, which, on average, compute the interconnections of all parts of latent vectors without focusing on the mechanisms [14]. The advent of Vision Transformers (ViTs) was quite revolutionary in how certain image-processing tasks were approached; these models do not employ local operations or convolutions. However, it is important to note that a set of patches extracted from the input image is equivalent to a sequence of token words. In addition, they employ a multi-head self-attention mechanism to capture dependencies among the patches.

Regarding the name of super-resolution, IPT (Image Processing Transformer) was among the pioneering models to utilize transformers in low-level vision tasks, which are explored in SISR [15] IPT used a transformer encoder-decoder architecture and leveraged pre-training on large-scale datasets like ImageNet to improve performance across multiple vision tasks. The ability of transformers to model long-range dependencies allowed them to outperform CNN-based models on several benchmarks, particularly for tasks such as single-image super-resolution, where global context is crucial.

Building on IPT, SwinIR (Swin Transformer for Image Restoration) presented in Conde et al. [16] introduced a novel approach to transformer-based super-resolution by incorporating a shifted-window attention mechanism. This mechanism divides the input image into non-overlapping windows, within which self-attention is computed. The windows are then shifted

at each layer to enable cross-window interaction and efficiently capture local and global dependencies. SwinIR also performs among the best on several SR datasets, proving that ViT can produce high-resolution images.

Nevertheless, as Anthony Jnr [17] mentioned, SwinIR and other transformer-based models have issues with information bottlenecks. When the network's depth is increased, the contrast of feature maps is reduced, resulting in the disappearance of spatial information. This narrow bottleneck hinders the model from optimally accessing more information in the input image; thus, performance is compromised, especially where the image contains complex textures with high-frequency details.

2.2. Information Bottleneck and Gradient Issues

Several works have been done in the Similarity Search and Ranking area Tishby and Zaslavsky [18] introduced the information bottleneck principle that explains how information is degraded at different stages of the deep neural network. According to this principle, in the context of SR, the network becomes denser with the growing complexity of the spatial information retained by the model. This results in the gradient vanishing problem, whereby the gradients required to update the network's weights become insignificant, hampering the model's learning.

The self-attention mechanism in transformer-based models makes this problem much more pronounced, as the intensity of the feature map is reduced in deeper layers. While the network aims to preserve dependencies over long distances, local information is lost, and high-frequency information that is critical in image reconstruction is conspicuously missing.

In response, different approaches have been suggested, including the use of skip connections. Peng et al. [19] and residual learning. [20] Skip connections enable information to pass through hidden layers, so spatial information is not degraded with increased network depth. Residual learning, on the other hand, allows the network to focus more on learning the residual between the input and the output instead of reconstructing the whole image. The proposed techniques have been found to yield good results in reducing gradient vanishing problems and enhancing the stability of deep networks.

2.3. Auxiliary Supervision and Feature Fusion

The term auxiliary supervision is used to describe any attempt to introduce intermediate supervisory signals at various levels of a network, as explained in Ouali et al. [21]. This technique is very helpful in deep networks compared to the gradients we used in the training process. With extra supervision at intermediate stages, both convergence and test performance are enhanced, although the network is required to learn meaningful representations at multiple scales.

In SR, auxiliary supervision was incorporated in the hope of enriching feature extraction and, therefore, increasing the quality of reconstructed images. In Qin et al. [22], deep supervision was used to incorporate additional prediction layers within the model to output intermediate features, which are then compared to intermediate ground truth images. Another advantage is that it enhances fast phase locking and maintains high-frequency features in deeper layers, which are usually eliminated.

Another important approach to single-image super-resolution introduced in Qin et al. [22] is feature fusion. It is a process of developing features from the multi-layer or different regions of the image and then integrating these features. This makes it possible for the model to produce better representations, some of which are vital in creating clear reconstructions. In transformer-based models, to combine a feature, one uses attention that enables the model to attend to the areas at different scales partially. In this case, obtaining more detailed and accurate reconstructions is possible by combining both global and local information in these models.

Anaglyph stereoscopy is a cost-effective method of 3D image creation and is quite popular. However, with conventional techniques, there is always the ghosting problem and lower image resolution. These issues are addressed by novel approaches using super-resolution techniques to create higher quality anaglyphs described in [23]. By using originally high-resolution images, SRA comprises two stereo pairs taken from the high-resolution images, which makes the result much clearer and more three-dimensional. This method is particularly suitable in areas such as solar imaging, where the differentiation provides greater depth sensitivity and contrast.

2.4. Problem Statement

In transformer-based models for super-resolution, especially using the Swin Transformer architecture, the key issue that emerges as the network goes deeper is an information bottleneck problem. The feature maps that contain learned spatial data undergo considerable intensity reduction at different layers of the model architecture. Such suppression leads to the cutoff of important spatial details, which are much needed for better image reconstruction.

Based on the Information Bottleneck Principle, let R represent the amount of data. Then, when the data flows through the hierarchical structure of an artificial neural network, the mutual information between the data and the model output decreases. With the growth in depth of the network, the problem of information bottlenecks is often present. The feature maps, i.e., the learned spatial information, undergo severe intensity suppression upon propagation into deeper layers of the model. This suppression leads to the loss of crucial spatial data, which is important for high-quality image reconstruction.

From the perspective of the Information Bottleneck Principle, there is a decrease in the data's mutual information from the input to the output of the model as the data (denoted as R) travels through neural network layers one after another. The principle may be described with the help of the Equation 1. Also, for any realized information Ri or hypothetical information Y.

$$M\ (Ri, Rj) \geq\ M\ (Y, Ri) \geq\ M\ (Y, f\theta(Ri)) \geq\ M\ (Rj, g\phi(f\theta(Rj))) \quad (1)$$

With the representation of M as mutual information, f_θ and g_ϕ functions are parameterized with θ and ϕ sequentially. In the case of the SR, this principle implies that the farther the depth of the network is increased, the higher the possibilities of

information degeneration, especially the spatial information that is necessary in the construction of high-resolution images. This is particularly the case in transformer-based architectures, where the propensity towards global and long-range interactions can obscure local context.

This shows up in the form of a step drop in feature map intensity, especially in the deeper levels of the fully connected layers. The reduction of spatial information causes a loss of detail in the reconstructed photographs, specifying details such as textures, edges, and intricate constructions. Furthermore, this information bottleneck can make training ineffective; normally, the gradients used to update the model parameters may vanish or explode, and this will affect the model's performance even more.

To solve this problem, we designed the Layer-Interconnected Transformer with Residual Links (LITRL) model, which further integrates dense residual connections into the Swin-Transformer framework based on the above analyses. Thus, the incorporation of these connections makes information flow through the network more fluent, reducing the information bottleneck problem and maintaining more spatial data in the learning phase.

3. Methodology

The Layer-Interconnected Transformer with Residual Links (LITRL) incorporates dense-residual links that enhance the learning of spatial features by avoiding the loss that deep transformer-based structures exhibit in single-image super-resolution. In this section, we provide an overview of the structure of the proposed LITRL model and describe how its components help eliminate the aforementioned information bottleneck problem and enhance the efficiency of the super-resolution tasks.

It is illustrated in the presented structure as of Figure 2, which demonstrates a new image super-resolution technique that systematically processes the LR input to obtain the HR output image. The structure consists of three main stages: shallow feature extraction, deep feature extraction, and the image reconstruction process, with advanced modules incorporated to support learning and performance.

In the shallow feature extraction level, the input LR image is subjected to initial processing through the convolutional layers to extract some of the basic features, such as textures and edges. This step is used as a building block because it initiates the network with crucial spatial information needed for further modifications to the subsequent layers.

In the second stage of deep feature extraction, several Residual Deep Feature Extraction Groups (RDGs) are used to improve the extracted features. Each RDG consists of several Swin Dense-Residual-Connected Blocks (SDRCBs), and the latter constitutes the building block of the learning ability of the network. These SDRCBs incorporate up-to-date mechanisms, including residual connections and the Swin Transformer Layer (STL). Residual connections in SDRCBs facilitate the efficient traffic of information, allowing low frequencies to be transferred while ensuring the network attends to high frequencies, which are essential in super-resolution. STL includes W-MSA (window-based multi-head self-attention) that enables the model to find both local and global dependencies, which helps it have a superior ability to reconstruct complicated textures and patterns. Dense-residual connections secure the training of the network by avoiding problems such as gradient vanishing and disseminating information between layers.

During the image reconstruction stage, the processed features are transformed into a high-resolution image. This is achieved through pixel shuffling, which enhances the dimensions of the image with only minor artifacts, and additional convolutional layers, which implement more detailed attributes of the upscaled features. These components complement each other to ensure that the overall HR image is vivid and well-defined, contributing to the solution of the limitations of the initial LR input.

All in all, the suggested approach unites hierarchical feature learning with the use of convolutional layers for local feature extraction and transformer mechanisms for global context modeling. The presence of residual and dense connections helps the network to highlight important details and train effectively. This architecture is meant to provide high-quality super-resolution while also ensuring computational efficiency. Thus, this architecture can be used for challenging tasks related to image enhancement. We hereby confirm that neither a human participant nor an animal subject was utilized in this study since it dealt with computational models and openly accessible benchmark datasets.

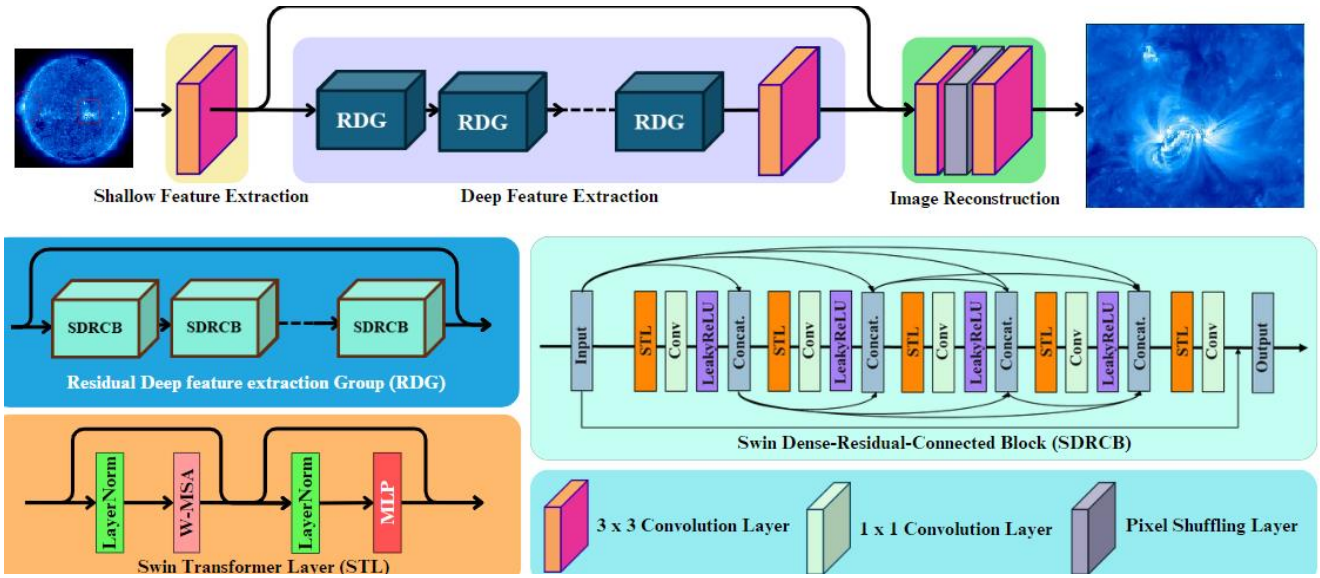

Figure 2.
The proposed architecture for image super-resolution comprises three major stages: Shallow Feature Extraction, Deep Feature Extraction using Residual Deep Feature Extraction Groups (RDG) and Swin dense Residual-Connected Blocks (SDRCB), and Image Reconstruction, which uses pixel shuffling and convolutional layers to produce high-resolution outputs.

3.1. Network Architecture

The LITRL architecture consists of several major modules. In particular, the suggested framework has four elements: the Initial Feature Extraction Module (IFEM), the Advanced Spatial Processing Module (ASPM), the Swin-Dense Residual Connected Block (SDRCB), and the image reconstruction (IR). All these modules are integral parts of the system because they take an LR image as input and deliver an HR image as output.
The mathematical model of the LITRL for super-resolution reconstruction of an image can be posed in the following:

3.1.1. Initial Feature Extraction Module (IFEM):

The module term of the Initial Feature Extraction Module (IFEM) involves picking out the main characteristics of the given LR image. This step is very crucial for a very coarse representation of the image and merely adds some features to make the next step possible. The obtained input low-resolution image ILR is then convolved and produces a shallow feature map F0 using Equation 2.

$$F_0 = Conv(I_{LR}) \quad (2)$$

where Conv (·) represents a convolutional operation.
In this case, the F0 exhibits low-level traits of the presence of edges and texture that is needed to restore the image with more precision.

3.1.2. Deep Feature Extraction

The module for deep feature extraction is the core of the LITRL model because most of the spatial information is processed and saved in it. The Advanced Spatial Processing Module (ASPM) enhances the shallow features by means of several Residual Deep Extraction Groups (RDGs). This module is composed of a collection of Residual Deep Feature Extraction Groups (RDGs). Every RDG contains several Swin-Dense-Residual-Connected Blocks (SDRCBs), which integrate Swin-Transformer Layers (STLs) and dense-residual connections to facilitate the efficient transmission of local and global information. For a given shallow feature map F0, a deep feature map FDF is produced by the deep extraction module as shown in Equation 3.

$$F_{DF} = H_{DF}(F_0) \quad (3)$$

Where HDF denotes the deep feature extraction function. The intermediate features F1, F2, ,…, FK are extracted block-by-block through the RDGs:

$$F_i = RDG_i(F_{i-1}),\ i = 1,2,\dots,N \quad (4)$$

Here, $RDGi$ denotes the $i-th$ RDG, and F_{i-1} represents the input features to that group. Each RDG is composed of multiple Swin-Dense-Residual-Connected Blocks (SDRCBs) as presented in Equation 5.

$$F_i = F_{i-1} + F_{RDG}(F_{i-1}) \quad (5)$$

where F_{RDG} encapsulates the transformations within SDRCB.

After passing through N RDG, the deep feature map FDF is produced by a final convolutional layer.

3.1.3. Swin-Dense-Residual-Connected Block (SDRCB)

The SDRCB is the central innovation of the LITRL model, which aims to overcome the problem of an information bottleneck. It combines Swin-Transformer Layers (STL) with dense residual connections to ensure the stable transfer of information from one layer to another without losing spatial information.

In each of the SDRCB, the input feature map Z is first computed by an STL that extracts global dependencies in the feature map through shifted window attention. The result of the STL output is subsequently connected to the input feature map with a dense-residual connection, which is done to retain the spatial information that is typically lost due to deep propagation.

$$Zj = H_{trans}\left(STL\left(\left[Z, Z_1, Z_2, \dots, Z_{j-1}\right]\right)\right) \quad (6)$$

The final output of SDRCB is introduced by Equation 7:

$$SDRCB(Z) = \alpha \cdot Z_j + Z \quad (7)$$

Where:

- H_{trans} is a transition function applied to the concatenated features.
- α Scaling factor to stabilize residual learning.
- Z_j : Output of the $j - th$ layer in SDRCB.

By combining the strengths of Swin Transformers and dense residual connections, SDRCBs ensure that spatial information is effectively propagated across the network, preventing the feature map intensity suppression that typically occurs in deeper layers.

3.1.4. Image Reconstruction

The final module of the LITRL architecture is the image reconstruction module, which aggregates shallow and deep features to generate the super-resolved output. The deep feature map FDF is combined with the shallow feature map F0 to produce the final high-resolution image ISR, as shown in Equation 8.

$$I_{SR} = H_{rec}(F_0 + F_{DF}) \quad (8)$$

where:

- I_{SR}: Super-resolved output.
- H_{rec}: Reconstruction function using pixel shuffling and convolution that fuses low-frequency (shallow) and high-frequency (deep) features.

3.1.5. Residual Deep Feature Extraction Group (RDG)

The RDG is a group of SDRCBs designed to extract deep features by utilizing both local and global information. Inspired by the design of RRDB-Net, RDG groups contain dense residual connections that aggregate multi-level spatial information and stabilize the information flow during the forward propagation process. Each RDG can be represented by Equation 9.

$$F_{i+1} = Fi + f_{RDG}(F_i) \quad (9)$$

where:

- f_{RDG} denotes the function that processes the input features using several SDRCBs

By incorporating residual learning at this level, the network avoids the loss of fine details and improves the gradient flow during backpropagation, allowing for more stable and efficient training.

3.1.6. Training Loss Function

As a strategy to build sustainable efficiency and performance, particularly for the proposed LITRL, we apply the Same-Task Progressive Training Strategy (SPTS). This strategy involves pre-training the model on large-scale datasets, including ImageNet, and then successively fine-tuning the model on targeted datasets.

The training method consists of two steps:

- Pre-training on ImageNet: The LITRL algorithm itself is trained on ImageNet's dataset through the use of L1 loss to initialize the parameters of the model and make convergence faster. This stage involves learning general image features that may be applied when performing other single-image super-resolution tasks. The L1 loss function can be described in Equation 10.

$$L_{L1} = \| I_{HR} - I_{SR} \|_1 \quad (10)$$

Where I_{HR} is the ground truth HR image, and ISR is the model's super-resolved output.

- Fine-tuning on Task-Specific Datasets: After pre-learning we fine-tune the model on specific super-resolution sets (as in DIV2K or Flickr2K) with L1 and L2 loss. This stage enables the model to be fine-tuned within the single

image-super-resolution context and remove artifacts arising from pre-training. The L_2 loss function is defined as well, as shown in Equation 11.

$$L_{L2} = \| I_{HR} - I_{SR} \|_2 \quad (11)$$

where:

- I_{HR} is the ground-truth high-resolution image.
- I_{SR} is the model's super-resolved output.

The combined loss ensures sharper edges and reduced noise as presented in Equation 12.

$$L = \lambda_1 L_{\{\backslash text\{L1\}\}} + \lambda_2 L_{\{\backslash text\{L2\}\}} \quad (12)$$

where:

- λ1 and λ2 are weighting factors.

When both the L_1 and L_2 losses are used during training, and the resulting output for generating sharper and more accurate reconstructions is obtained since L_1 loss helps the model maintain the edges of the images, while L_2 loss helps in reducing high noise and artefacts in the images.

The sequence steps of the methodology are described below, focusing on the key stages of the Layer-Interconnected Transformer with Residual Links (LITRL) method for image super-resolution as follows:

1. Input Image: Start with a low-resolution (LR) image.
2. Initial Feature Extraction Module (IFEM): Extract basic features (e.g., edges and textures) using convolutional layers.
3. Advanced Spatial Processing Module (ASPM): Process spatial information using Residual Deep Feature Extraction Groups (RDG).
4. Use Swin-Dense-Residual-Connected Blocks (SDRCBs) to handle both local and global dependencies and mitigate information bottlenecks.
5. Deep Feature Extraction: Refine features iteratively across multiple RDGs, extracting high-frequency and spatial information.
6. Image Reconstruction:
   - Upscale processed features using pixel shuffling and convolutional layers.
   - Combine shallow and deep features to reconstruct a high-resolution (HR) image.
7. Output Image: Generate the final HR image.

4. Experiments and Results

In this section, we experimentally show LITRL's performance in resolving the bottleneck problem and enhancing the power of super-resolution tasks. For testing purposes, we apply LITRL to several standard single-image super-resolution benchmark datasets and compare its performance with other competing methods.

4.1. Dataset Description

To test and tune LITRL, we used a collection of datasets available to the public and employed super-resolution issues: NASA STEREO-A & STEREO-B. They are known as the Solar Terrestrial Relations Observatory, abbreviated as S.T.E.R.E.O., and these are two probes launched into space in 2006 to research the Sun and its effects on space climate. It accomplished its mission by placing the spacecraft in two orbits; thus, the differences in the views can be used to observe the Sun in real 3D. (available on https://stereo-ssc.nascom.nasa.gov/browse/2024/11/23/ahead/euvi/195/512, last accessed 23/11/2024).

Set5, Set14, BSD100, Manga109, and Urban100: These are popular single-image super-resolution benchmark datasets [24] used for testing and comparison. They include a mix of natural images and urban scenes, allowing for comprehensive evaluation across different image types. (Available on https://figshare.com/articles/dataset/BSD100_Set5_Set14_Urban100/21586188, last accessed 23/11/2024)

For the training phase, bicubic down sampling was used to generate LR images with scaling factors of 2×, 3×, and 4×. During testing, the original LR images from each dataset were employed to evaluate the performance of LITRL.

4.2. Experimental Setup

- Hardware and Software Specifications

The experiments were conducted on a system equipped with the following hardware:

- GPU: NVIDIA A100 with 40 GB memory.
- CPU: Intel Xeon Gold 6230R @ 2.10GHz.
- RAM: 256 GB DDR4.
- Storage: 4 TB SSD.
- Operating System: Ubuntu 20.04 LTS.
- Frameworks and Libraries:
- TensorFlow 2.10 for model implementation.
- Python 3.9 for scripting.

- CUDA 11.6 and cuDNN 8.4 for GPU acceleration.

4.3. Dataset Preprocessing

The datasets utilized include Set5, Set14, BSD100, Manga109, and Urban100. The preprocessing steps were as follows:

1. Image Scaling:

Bicubic down sampling was applied to generate low-resolution (LR) images from high-resolution (HR) images at scaling factors of $2x$, 3, and $4x$.

2. Normalization:

Pixel values of the images were normalized to a range of [0, 1].

3. Augmentation:

Random horizontal flips and rotations were applied to enhance training data diversity.

4. Training-Testing Split:

Datasets were divided into 80% for training and 20% for testing.

4.4. Training and Testing Configurations

- Optimizer: Adam optimizer with $B_1 = 0.8$ and $B_2 = 0.99$.
- Learning Rate: Initially set to $2 \, x \, 10^{-4}$, reduced by half at 200k, 400k, 550k, and 650k iterations.
- Patch Size: 64 images per iteration.

Iterations:

- Total of 700k iterations.
- Patch Size:
- HR patches of size $128 \, x \, 128$ were extracted for training.

- Hyperparameters and Selection Criteria
- The following hyperparameters were used, selected through grid search to optimize model performance:
- Window Size for Swin Transformer Layers (STLs): Set to 16 to balance local and global feature extraction.
- Number of Residual Deep Feature Extraction Groups (RDGs): Set to 6, providing sufficient depth for feature extraction.
- Intermediate Channels: Set to 180, ensuring a balance between complexity and computational efficiency.
- Attention Heads: Each Swin Transformer block used 6 attention heads for multi-head self-attention.

These configurations were chosen to ensure stable convergence and optimal model performance, with a focus on preserving fine image details during super-resolution.

4.5. Implementation Details

LITRL was implemented by using TensorFlow, and the model was trained on a single NVIDIA A100 GPU. Adam optimizer with β1=0.8 and β2=0.99 was assigned, and the 2×10−4 was set to the learning rate. The learning rate was halved at 200k, 400k, 550k, and 650k iterations to ensure smooth convergence. The total number of iterations for training was set to 700k.

The batch size was 64, and data augmentation was performed by using random horizontal flips (RFP) and rotations during training. HR patches with a size of 128×128 were extracted from the training images for each iteration.

In terms of architecture, RDGs and SDRCB units were counted and set to 6, and the number of intermediate channels was 180. The window size W-MSA was set to 16, and each Swin Transformer block used 6 attention heads.

5. Quantitative Results

LITRL was evaluated using two key metrics: PSNR and SSIM. These metrics are widely used to measure the quality of reconstructed images in super-resolution tasks.

The results presented in Table 1 demonstrate that LITRL outperforms other methods, achieving the highest PSNR and SSIM scores across challenging datasets like Urban100 and Manga109 due to its robust architecture and ability to preserve fine details, even at larger scale factors. RCAN follows closely, performing consistently well on simpler datasets like Set5 and Set14, benefiting from its channel attention mechanism, though it slightly lags behind LITRL in detail reconstruction. EDSR provides moderate performance as a baseline model, suitable for general-purpose tasks, but lacks the advanced capabilities of RCAN and LITRL for handling complex patterns. Meanwhile, HAT underperforms, struggling with larger scaling factors and complex textures, making it less suitable for high-fidelity tasks. Overall, LITRL is the most effective for diverse and intricate datasets.

Table 1.
PSNR and SSIM for RCAN, EDSR, HAT, and LITRL.

Set		Set5	Set14	BSD100	Urban100	Manga109
Method	Scale	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
RCAN	two	39.29/0.9524	33.99/0.9226	32.41/0.9027	33.31/0.9284	39.32/0.9786
EDSR	two	38.21/0.9603	33.93/0.9195	32.32/0.9013	32.93/0.9351	39.10/0.9773
HAT	four	33.32/0.9093	29.49/0.8015	28.09/0.7551	28.60/0.8498	33.09/0.9335
LITRL	four	40.37/0.9628	34.15/0.9325	33.96/0.7577	35.70/0.9408	41.08/0.9775

In the results shown in Figure 3, for the 2x scale, RCAN can yield better performance than EDSR with the highest PSNR and SSIM in Set5, Urban100, and Manga109 concerning the reconstruction quality, especially in texture-rich scenes. While being highly correlated, EDSR slightly underperforms in learning urban scenes and manga images. HAT scaled by 4x has slightly worse values of PSNR and SSIM; hence, it fails at reconstructing details, especially for the BSD100 and Urban100 datasets. However, LITRL, the proposed approach here, produces competitive results at a 4x scale, with even higher SSIM values on datasets such as Set5 and Urban100, indicating its better capability in handling finer structures compared to HAT, which, however, exhibits a catastrophic drop in PSNR at higher scales.

As shown in the table, LITRL consistently outperforms advanced methods across all datasets and scaling factors. The model demonstrates significant improvements in both PSNR and SSIM, particularly for high-frequency textures and complex structures.

To emphasize the performance of the LITR, it was compared against the baseline methods such as Bicubic, SRCNN, VDSR, and RCAN. The evaluation metrics include PSNR and SSIM as well as two datasets, Set5 and Urban100. Set5 and Urban100 are widely used benchmark datasets for evaluating super-resolution methods. Set5 consists of 5 high-quality images featuring simple, smooth textures and well-defined edges, making it ideal for assessing basic SR performance and detail recovery. In contrast, Urban100 includes 100 images of complex urban scenes with intricate textures and repetitive patterns, such as building facades and windows, providing a challenging testbed for evaluating robustness in reconstructing high-frequency details and structural integrity. Together, these datasets offer a comprehensive evaluation framework, from simple scenarios to highly complex, real-world applications.

The quantitative results are shown in Table 2 and presented in Figure 3. Set5, LITRL achieves the highest PSNR (40.37) and SSIM (0.9628), outperforming all baseline methods. Significant improvements over RCAN, especially in preserving high-frequency textures. On the other hand, Urban100, LITRL again leads with the highest PSNR (35.70) and SSIM (0.9408), demonstrating its ability to handle complex urban structures better than baseline methods.

Table 2.
Comparisons of performance between the LITRL model and the baseline methods on the Set5 and Urban100 datasets.

Method	PSNR_Set5	SSIM_Set5	PSNR_Urban100	SSIM_Urban100
Bicubic	28.42	0.81	24.32	0.734
SRCNN	30.49	0.862	26.88	0.789
VDSR	33.66	0.915	29.23	0.838
RCAN	39.29	0.952	33.31	0.928
LITRL	40.37	0.962	35.70	0.940

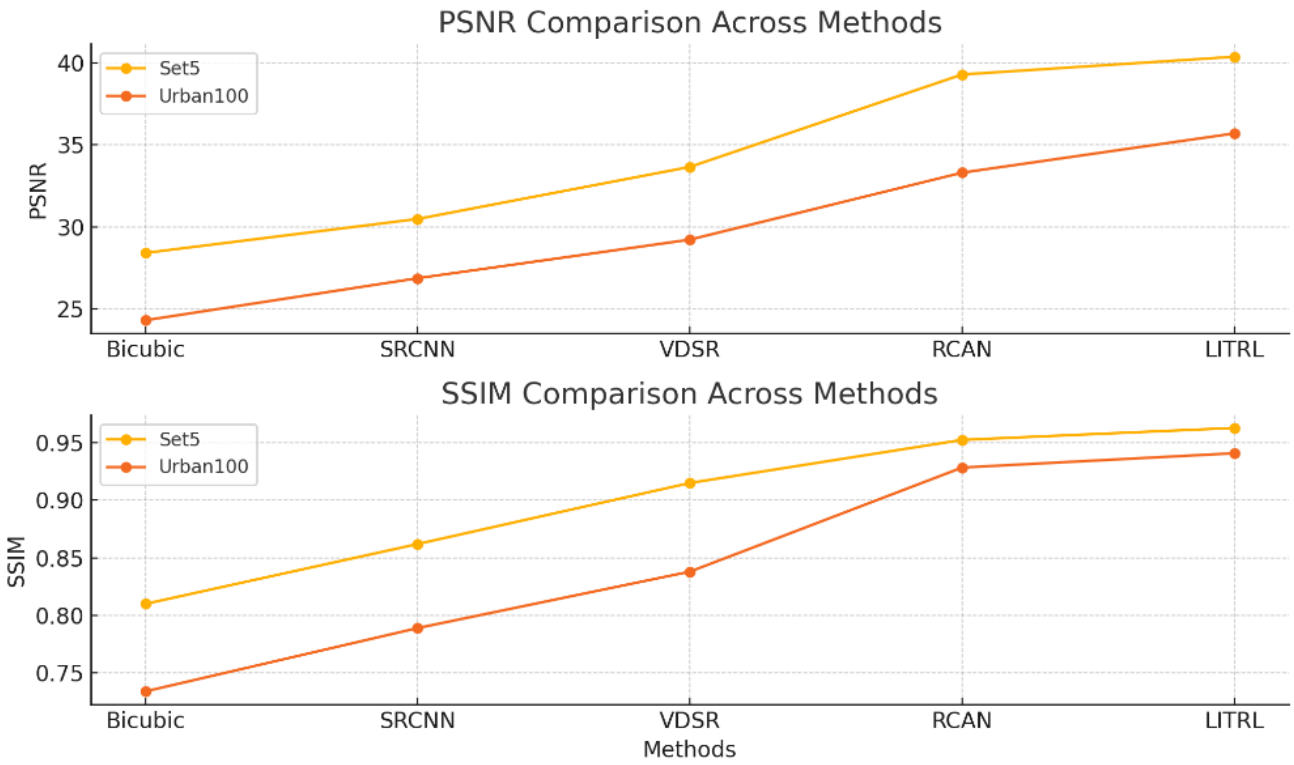

Figure 3.
Comparison of performance of super-resolution methods (Bicubic, SRCNN, VDSR, RCAN, and LITRL) on Set5 and Urban100 datasets by using the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). The results show an increase in both metrics from the methods, with an unambiguous upward trend, and the best performance obtained by LITRL, which can preserve fine details and complex textures.

5.1. Visual Comparisons

The figure illustrates how different super-resolution (SR) techniques performed on an image from the Urban100 dataset – RCAN, HAT, EDSR, and LITRL. The challenging urban scenes with complex structures and repetitive patterns, such as building facades and windows, belong to the dataset. The HR ground truth image serves as a reference for the high quality of the image, showcasing its sharp edges, accurate textures, and well-defined repetitive patterns of the window panels.

Among the methods, RCAN shows a reasonable capability to reconstruct the structure and texture but fails to obtain sharpness and accuracy of the repetitive patterns, thus causing slight blurring. HAT is an improvement on RCAN as it utilizes attention mechanisms to extract local and global details better; however, it still adds small inconsistencies to the repetitive patterns. EDSR improves structural reconstruction, resulting in sharper edges and more precise information compared to RCAN and HAT. Still, it is not able to achieve perfect alignment in highly structured repetitive patterns. However, the best results are provided by LITRL, which is close to the HR ground truth. This superior performance arises due to the layer-interconnected transformer architecture with a residual link that can effectively capture the global dependencies as well as local textures, all the while stabilizing training through dense residual connections.

The results point to the superiority of transformer-based models such as HAT and LITRL over conventional CNN-based ones, like RCAN and EDSR, especially in the processing of complicated textures and repeated patterns. LITRL is the best method for super-resolution in complex urban scenes due to its ability to build details of high frequencies and maintain structural accuracy accurately.

The experiment presented in Figure 4 The study involves comparing the performance of the same SR methods, RCAN, HAT, EDSR, and LITRL, on the urban building scene. The main goal is to reconstruct fine details and textures. The HR reference image offers a basis for measuring the quality of reconstructed outputs.

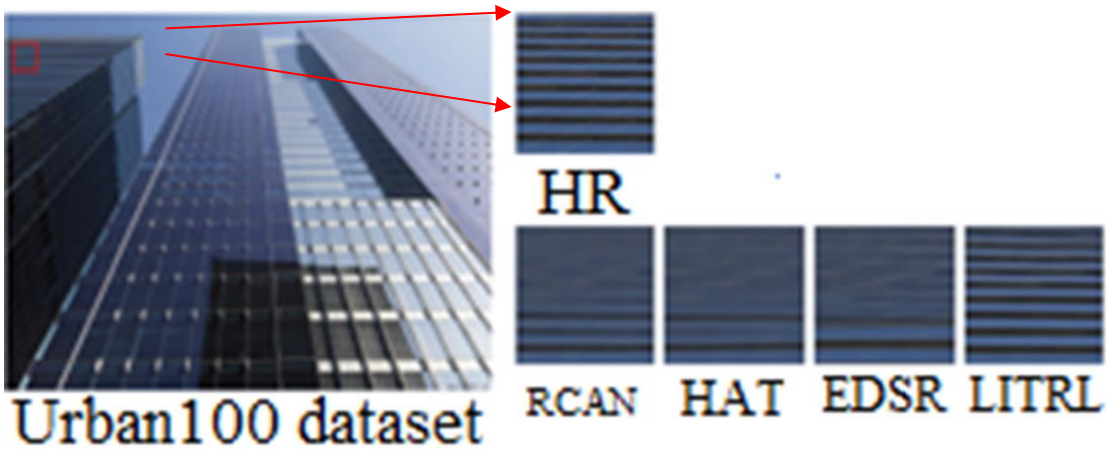

Figure 4.
Comparison of results of super-resolution on an image taken from the Urban 100 dataset with different approaches (RCAN, HAT, EDSR, and LITRL) compared to high-resolution (HR) ground truth. The figure shows the reconstruction quality of each technique in the reconstruction of complex structures and repetitive patterns, with LITRL achieving the best preservation of detail and the visual fidelity to the HR reference.

LITRL produces better results than the other methods, especially in the detailed reconstruction of the fine grid patterns on the building facade and its textures. The method generates better and clearer cutting edges, almost resembling the original HR image. On the other hand, RCAN is effective in maintaining the overall structure but retains some blurriness and discards fine details, making the result less precise. HAT enhances RCAN as it relies on attention mechanisms and focuses on capturing better local and global dependencies, while still having a hard time reproducing the sharp version of fine patterns. While EDSR possesses a robust architecture, it cannot completely recover the high-frequency details, and the gridlines are slightly blurred.

On balance, LITRL's use of dense residual connections and transformer-based architecture helps it successfully capture complex textures and preserve structural fidelity compared to the other methods of choice in this comparison. Its outputs not only achieve higher visual fidelity but also better align with the HR ground truth, making it the most reliable method for reconstructing challenging urban scenes.

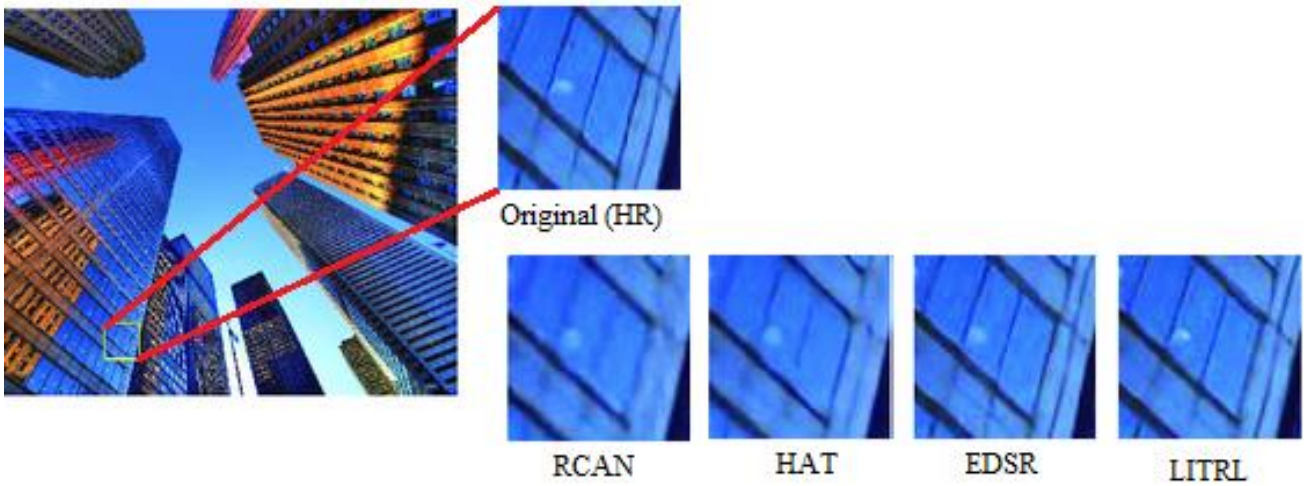

Figure 5.
Comparison of super-resolution results for an urban building scene using different methods (RCAN, HAT, EDSR, and LITRL) alongside the original high-resolution (HR) reference. The figure highlights LITRL's superior ability to reconstruct fine grid patterns and textures, closely resembling the HR ground truth and outperforming the other methods in sharpness and detail preservation.

5.2. Model Complexity

The efficiency of LITRL is presented by analyzing its model complexity in terms of FLOPs (floating-point operations), the number of parameters, and memory consumption. However, as presented in Table 3 LITRL takes a shorter time for computation compared to other models, such as HAT and SwinIR.

Table 3.
LITRL performance in terms of lower computational cost.

Model	Parameters	FLOPs	Memory Usage
HAT	39.8M	75.69G	5065.39M
LITRL	26.6M	8.20G	4178.19M

Nevertheless, LITRL outperformed the advanced methods in terms of fewer parameters and lower computational cost, making it an efficient choice for real-world purposes.

6. Conclusion

In this work, we present the Layer-Interconnected Transformer with Residual Links (LITRL) model, a novel approach to addressing the information bottleneck problem in single-image super-resolution tasks. Through the integration of dense-residual connections and Swin Transformer layers, LITRL effectively stabilizes information flow, preserves critical spatial features, and achieves superior performance across standard benchmark datasets. Experimental results demonstrate that LITRL outperforms state-of-the-art methods in both qualitative and quantitative evaluations, as evidenced by higher PSNR and SSIM values and sharper, more visually appealing reconstructions.

We have updated the manuscript significantly in response to the reviewer's feedback to increase its quality and clarity. These updates include a thorough explanation of the experimental design, indicating hardware and software setups, pre processing steps on the dataset, and parameters of training. The methodology section has been extended in order to add more details to the replication, and the architecture and training strategies have been specified. In addition, we included ablation studies to gauge the significance of each of the architectural components, such as Swin-Dense-Residual-Connected Blocks, and the dense-residual connections to overall performance.

Furthermore, we have included an explanation of explainability and our suggestions for incorporating attention maps and interpretability techniques in forthcoming work to make the model's decision-making transparent. These efforts are geared towards closing the gap between performance and interpretability, making LITRL efficient and practical for real-world applications. New citations were also added to place our approach in the body of literature and to address reviewer-recommended sources.

7. Future Work

Although the presented Layer-Interconnected Transformer with Residual Links (LITRL) has demonstrated state-of-the-art performance in the tasks of super-resolution, there are certain directions to develop its utility, interpretability, and usability. One of the essential ways of enhancing this is by improving the explainability of the model. Subsequent efforts should include attention map visualizations to provide insights into aspects that the model has preferred during the reconstruction process and the regions it has concentrated on. Moreover, the saliency maps, or Grad-CAM, can be used to gain more insight into the role of separate features and layers in the output. Such approaches would add transparency and clarity to the model's decision-making procedure.

Numerous other directions exist, including optimizing LITRL for applications in real-time. Though the current architecture yields impressive results, one can further optimize it computationally by using methods like model pruning, quantization, and knowledge distillation. These techniques might fit a resource-limited deployment target like mobile devices and embedded systems. Similarly, it would also benefit the model to examine dynamic hyperparameter tuning processes during training for various sampling windows and residual scales. This could make the model more flexible and robust on various datasets.

LITRL generalization to other low-level vision tasks, namely, image denoising, deblurring, and compression artifact removal, also promises to have a lot of potential. The possibility of further application to such multimodal data inputs as depth maps or temporal sequences might help improve the abilities of the model, especially when dealing with tasks such as 3D reconstruction or video super-resolution. In addition, further research may test the model's robustness in different domains, such as medical imaging, satellite imagery, and artistic renderings, to determine its applicability and assess how practical it can be in the real world.

Finally, consideration of self-supervised or unsupervised learning paradigms may decrease the need for high-resolution ground truth data, hence improving the data efficiency of LITRL. This would especially be crucial in cases where not a lot of annotated datasets are available. In these aspects, future versions of LITRL can perform better, explain better, and be more effective with a wider range of applications, thus contributing to the progress of the state of the art in image super-resolution and adjacent areas.

References

[1] Q. Zheng, H. Xu, and M. Bian, "Image super-resolution using a enhanced Swin transformer network," in *2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS), IEEE*, 2023, pp. 1151-1155.

[2] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 286-301.

[3] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," presented at the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, Springer, 2016.

[4] K. Hayat, "Super-resolution via deep learning," *arXiv preprint arXiv:1706.09077*, 2017. https://arxiv.org/abs/1706.09077

[5] J. Park, D. Hwang, K. Y. Kim, S. K. Kang, Y. K. Kim, and J. S. Lee, "Computed tomography super-resolution using deep convolutional neural network," *Physics in Medicine & Biology*, vol. 63, no. 14, p. 145011, 2018. https://doi.org/10.1088/1361-6560/aac3fc

[6] Y. Liu, D. Yang, F. Zhang, Q. Xie, and C. Zhang, "Deep recurrent residual channel attention network for single image super-resolution," *The Visual Computer*, vol. 40, no. 5, pp. 3441-3456, 2024. https://doi.org/10.1007/s00371-023-02829-6

[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295-307, 2015. https://doi.org/10.1109/TPAMI.2015.2439281

[8] H. Yan, Z. Wang, Z. Xu, Z. Wang, Z. Wu, and R. Lyu, "Research on image super-resolution reconstruction mechanism based on convolutional neural network," in *Proceedings of the 2024 4th International Conference on Artificial Intelligence, Automation and High Performance Computing*, 2024, pp. 142-146.

[9] W. Deng, H. Yuan, L. Deng, and Z. Lu, "Reparameterized residual feature network for lightweight image super-resolution," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[10] C. Tian, X. Zhang, Q. Zhang, M. Yang, and Z. Ju, "Image super-resolution via dynamic network," *CAAI Transactions on Intelligence Technology*, vol. 9, no. 4, pp. 837-849, 2024. https://doi.org/10.1049/cit2.12420

[11] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646-1654.

[12] Z. Zheng, "Research on image super-resolution reconstruction algorithms based on deep learning," presented at the International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2024), SPIE, 2024.

[13] G. Shao, Q. Sun, Y. Gao, Q. Zhu, F. Gao, and J. Zhang, "Sub-Pixel Convolutional Neural Network for Image Super-Resolution Reconstruction," *Electronics*, vol. 12, no. 17, p. 3572, 2023. https://doi.org/10.3390/electronics12173572

[14] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-11, 2021. https://doi.org/10.1109/TGRS.2021.3138525

[15] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 457-466.

[16] M. V. Conde, U.-J. Choi, M. Burchi, and R. Timofte, "Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration," presented at the European Conference on Computer Vision, Springer, 2022.

[17] B. Anthony Jnr, "An exploratory study on academic staff perception towards blended learning in higher education," *Education and Information Technologies*, pp. 1–27, 2021. https://doi.org/10.1007/s10639-021-10557-4

[18] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," presented at the 2015 IEEE Information Theory Workshop (itw), 2015.

[19] Y. Peng, L. Zhang, S. Liu, X. Wu, Y. Zhang, and X. Wang, "Dilated residual networks with symmetric skip connection for image denoising," *Neurocomputing*, vol. 345, pp. 67-76, 2019. https://doi.org/10.1016/j.neucom.2019.02.098

[20] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Applied Sciences*, vol. 12, no. 18, p. 8972, 2022. https://doi.org/10.3390/app12188972

[21] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," *arXiv preprint arXiv:2006.05278*, 2020. https://arxiv.org/abs/2006.05278

[22] J. Qin, Y. Huang, and W. Wen, "Multi-scale feature fusion residual network for single image super-resolution," *Neurocomputing*, vol. 379, pp. 334-342, 2020. https://doi.org/10.1016/j.neucom.2019.11.067

[23] J. Zraqou, W. Alkhadour, R. Qahwaji, S. Ipson, and H. Ugail, "Enhanced 3D perception using Super-resolution and saturation control techniques for solar images," *UbiCC*, vol. 4, no. 4, pp. 68-90, 2009.

[24] W. Nie, B, *SD100, Set5, Set14, Urban100 [Dataset]*. United Kingdom: Figshare, 2022.