IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 August 2024, accepted 22 August 2024, date of publication 26 August 2024, date of current version 11 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3450300

RESEARCH ARTICLE

DHTCUN: Deep Hybrid Transformer CNN U Network for Single-Image Super-Resolution

JAGRATI TALREJA[1], (Graduate Student Member, IEEE),
SUPAVADEE ARAMVITH[2], (Senior Member, IEEE),
AND TAKAO ONOYE[3], (Senior Member, IEEE)
[1]Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand
[2]Multimedia Data Analytics and Processing Unit, Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand
[3]Graduate School of Information Science and Technology, Osaka University, Suita 565-0871, Japan

Corresponding author: Supavadee Aramvith (supavadee.a@chula.ac.th)

This research is funded by the Second Century Fund (C2F), and the Ratchadapiseksompotch Fund, Chulalongkorn University Bangkok, 10330, Thailand. This research is also supported by the 90th Anniversary of Chulalongkorn University Scholarship under the Ratchadapisek Somphot Endowment Fund.

ABSTRACT Recent advances in image super-resolution have investigated various transformer and CNN techniques to improve quantitative and perceptual outcomes. Reconstructing high-resolution images from their low-resolution equivalents by combining the power of transformers and CNN has been a crucial task in recent times. We propose a novel U-shaped architecture that integrates transformers and convolutional neural networks (CNNs) to leverage the strengths of both approaches. The network incorporates a novel Parallel Hybrid Transformer CNN Block (PHTCB) on the backbone of the U-shaped design, ensuring computational efficiency and robust hierarchical feature representation. Our architecture incorporates triple-enhanced spatial-attention mechanisms and a Transformer CNN (TCN) Block in PHTCB. The TCN Block helps preserve sharp edges and intricate details often lost in traditional SISR methods and enhances the visual fidelity of the reconstructed high-resolution images. Additionally, we introduce the triple-enhanced spatial attention (TESA) approach that helps precisely localize of important features. Blurring can be reduced for crucial features by focusing on these critical areas because of the network's ability to control features at various scales. Experiments demonstrate that our proposed method yields better quantitative measurements, including visually appealing high-resolution image reconstructions, peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM).

INDEX TERMS CNN, enhanced spatial attention, single-image super-resolution, Transformer.

I. INTRODUCTION

In the rapidly evolving field of image processing and computer vision, the pursuit of high-fidelity image super-resolution (SISR) stands as a cornerstone challenge. The primary goal of SISR is to reconstruct high-resolution images from their low-resolution counterparts. This task that holds significant importance across a multitude of applications, including medical imaging [1], satellite imagery analysis [2], surveillance [3], and multimedia enhancement [4]. In recent years, there has been a surge in research focused on improving the quantitative metrics and perceptual quality of super-resolved images. These advancements are central to the synergistic integration of transformer models and convolutional neural networks (CNNs), leveraging each other's unique strengths to achieve superior results.

Due to the creation of numerous High-Resolution (HR) images that correspond to a single Low-Resolution (LR) image, SISR is an ill-posed problem. In recent years, single image super-resolution (SISR) has seen remarkable advancements, with convolutional neural networks (CNNs) emerging as the dominant approach. These CNN-based models

The associate editor coordinating the review of this manuscript and approving it for publication was Hengyong Yu.

have significantly improved the quality of generated high-resolution (HR) images, making them the mainstream method for SISR tasks. The advent of convolutional neural networks (CNNs) was pioneered in Super-Resolution Convolutional Neural Networks (SRCNN) [5], by Dong et al. Following the SRCNN approach, researchers have delved into various aspects of image super-resolution, including model frameworks, up-sampling methods, network design, and learning strategies in models like FSRCNN [6], VDSR [7], LapSRN [8], MemNet [9], EDSR [10], RCAN [11], NLSN [12] and DANS [13]. This exploration has yielded a various sophisticated techniques to enhance the performance and efficiency of SISR models. However, despite their success, CNN-based models encounter several limitations, including constrained receptive fields, the introduction of blurring, jagged patterns, or over-smoothing. These limitations hinder their efficiency and scalability, especially when dealing with large and complex images.

On the other hand, Vision Transformers (ViTs) [14] have demonstrated superior modeling capabilities and a larger receptive field, enabling them to capture long-range dependencies and global context more effectively than CNNs. ViTs leverage the self-attention mechanism to process the entire image as a sequence of patches, allowing them to model relationships between distant pixels and capture holistic image features. While powerful, the self-attention mechanism in Transformers is computationally expensive, especially as the resolution of the input image increases. This high resource demand can be prohibitive for practical applications, particularly in environments where computational efficiency and memory usage are critical considerations. The introduction of ViT led to several advancements and modifications to address the challenges associated with applying Transformers to SISR, such as Swin IR [15], Swin Transformer [16] SRFormer [17], ELAN [18], TCDFN [19], and HNCT [20]. However, the application of Transformers to SISR tasks comes with its own set of challenges. Notably, Transformers require extensive computing power and memory, which can be prohibitive for practical applications. The high resource demands of Transformers limit their widespread adoption in scenarios where computational efficiency and memory usage are critical considerations.

Given the trend towards fast processing devices, minimizing model size is crucial for achieving rapid and state-of-the-art (SOTA) comparable results for higher scale factors. Although CNN and transformers greatly enhance network performance, they still encounter certain limitations.

(1) Sometimes, these methods can introduce distortions into super-resolved images, such as jagged patterns, blurring or over-smoothing in extremely textured regions.

(2) The previously mentioned techniques require significantÂ processing power and time and are computationally expensive. Moreover, processing large datasets or high-resolution images further exacerbates the computational burden, potentially leading to longer processing times and reduced real-time performance.

(3) In some cases, attempts to enhance image resolution may inadvertently amplify noise or artifacts, resulting in degraded image quality and increased visual distortion. Consequently, achieving satisfactory super-resolution outcomes for extremely low-resolution and noisy input images remains a significant challenge in image processing.

In some cases, attempts to enhance image resolution using transformers may amplify noise [21], while CNNs may introduce artifacts [22]. Transformers often misinterpret noise patterns as image features, increasing noise amplification in the enhanced image [21]. CNNs, on the other hand, can create artifacts such as ringing or blurring, especially around edges or high-frequency regions, due to limitations in accurately reconstructing fine details [23]. Additionally, models trained on specific datasets may not generalize well, causing quality degradation on new data shown by [9], [24], [31]. This is particularly challenging with extremely low-resolution and noisy images, where algorithms may need help differentiating between true details and noise.

One way to efficiently address these challenges is to merge the complementary abilities of CNNs and Transformers and integrate their strengths. Transformers have been proven to capture long-range dependencies effectively, whereas CNNs have advanced in enhancing the quality of noisy images without compromising computational expense. Some of the Hybrid models have already been proposed, such as EHNet [25], HNCT [20], and TCDFN [19]. Utilizing comparable methodologies, we put forth a unique strategy for single-image super-resolution that combines the CNN and Transformer in a U-shaped architecture with skip connections. The U-shaped design with skip connections is the backbone of our proposed method. It helps the network to extract features differently from different layers without increasing the computational burden. The transformer helps to capture long-range dependencies and global context and reduce artifacts like jagged patterns or blurring. CNN helps enhance the quality of noisy images and prevent noise amplification while restoring noisy images. This synergetic combination allows the model to enhance the super-resolution performance while reducing its computational burden.

The following is a summary of the primary contributions made by our suggested model:

(i) Propose a Parallel Hybrid Transformer CNN Block (PHTCB) with a Transformer CNN (TCN) Block that combines the powers of CNNs and Transformer to capture long range dependencies, reduce artifacts, and simultaneously prevent the amplification of noise in super-resolution noisy images.

(ii) Put forth a Triple Enhanced Spatial Attention (TESA) block to improve the model's performance by focusing on relevant image regions while suppressing irrelevant or noisy areas.

(iii) Suggest a U-shaped backbone with a skip connection to extract features differently from different layers without increasing the computational burden.

The article's remaining sections are Section II, which examines relevant research on the suggested approach, and Section III, which outlines the network's methodology. Section IV presents the experimental results and a comparative analysis using cutting-edge techniques. Sections V and VI contain a discussion, conclusion and suggestions for further work.

II. RELATED WORK

Throughout the past ten years, image super-resolution (ISR) has made great strides, mostly due to the development of various deep learning algorithms. Convolutional neural networks (CNNs), generative adversarial networks (GANs), attention mechanisms, and transformer-based models are some of the major categories into which these developments can be divided.

The foundation of many advances in ISR has been Convolutional Neural Networks (CNNs). A major turning point in the discipline was when Dong et al. introduced SRCNN [5], showcasing deep learning's potential for super-resolution applications. Later efforts have concentrated on strengthening CNNs' architecture and training methodologies to improve performance. A significant contribution in this domain is the Fast Super-Resolution Convolutional Neural Network (FSRCNN) [6] by Dong et al. FSRCNN [6] was designed to be a faster and more efficient model than its predecessors. LapSRN [8], developed by Lai et al., uses a progressive reconstruction approach via a pyramid of images, improving training and testing efficiency. A persistent memory network, MemNet [9], was introduced by Tai et al. to handle the memory requirements of the network better. Li et al. proposed the Super-Resolution Feedback Network (SRFBN) [26], which uses feedback connections to refine feature representations iteratively, resulting in enhanced super-resolution quality. To get better results, Kim et al., for example, proposed the VDSR [7] model, which used extremely deeper CNN layers and residual learning to enhance super-resolution performance. Deeply Recursive Convolutional Network (DRCN) [27] and Deeply-Recursive Residual Network (DRRN) [28] by Tai et al. leverage recursive learning to improve depth and performance with fewer parameters. Li et al.'s Multi-Scale Residual Network MSRN [29] utilizes hierarchical information for image SR and adaptive feature extraction, and the Adaptive Weighted Learning Network AWSRN [30] for Lightweight Image Super-Resolution efficiently improves image resolution by utilizing adaptive weighted learning to balance performance and computational complexity. DBPN [31] Deep Back-Projection Network for image super-resolution enhances image quality through iterative up- and down-sampling processes, refining details and improving accuracy with each back-projection step. As deeper networks were not suitable ofr the currect cutting-edge devices, by deleting pointless modules from conventional CNNs, the EDSR [10] model by Lim et al. stretched the envelope even further and produced a more potent and effective network by improving optimization strategy and winning the New Trends in Image Restoration and Enhancement (NTIRE) 2017 challenge on single image Super-Resolution: Dataset and Study. A Multi-Scale Deep Cross Network for Image Super-Resolution (MDCN) [32] was introduced to handle multiple scale factors in a single model. RDN [33] Residual Dense Network for image super-resolution improves image quality by leveraging dense residual connections and feature fusion, allowing the network to learn and preserve fine details and textures effectively. Generative Adversarial Networks (GANs) introduced a new paradigm in ISR by focusing on generating more realistic and perceptually pleasing images to improve the visual quality further. Ledig et al. introduced SRGAN [34], which utilized GANs for ISR, producing high-resolution images with sharper details. Subsequent models, such as ESRGAN [35] by Wang et al., improved upon SRGAN by incorporating a deeper and more complex generator and discriminator architecture; furthermore EnhanceNet [36] was also introduced, which led to further enhancements in image quality.

Attention mechanisms have been pivotal in enhancing feature representation in ISR models. The integration of attention mechanisms into ISR models has significantly improved their ability to focus on important features of the image. The work by Zhang et al. on the Residual Channel Attention Network (RCAN) [11] demonstrated the effectiveness of channel attention mechanisms in enhancing feature representation. Similarly, Dai et al. proposed the Second-order Attention Network (SAN) [37], which leverages second-order channel attention to capture more complex feature interactions. Cross-Scale Non-Local (CSNL) [38] attention network by Mei et al. captures dependencies across different scales to enhance the representation of complex image structures. The Holistic Attention Network (HAN) [39] by Niu et al. integrates spatial and channel-wise attention mechanisms at multiple levels of the network to effectively capture fine-grained details. MFCC [40] leverages multi-frequency information through channel attention, improving the network's ability to distinguish between different textures and details in the image. Mei et al. developed the Non-Local Sparse Attention Network (NLSN) [12], which uses non-local operations to capture long-range dependencies and sparse attention mechanisms to reduce computational complexity while maintaining performance. These advancements have shown that attention mechanisms can significantly boost the performance of CNN-based ISR models by allowing them to adjust their focus based on the input image dynamically. DANS [13] refine feature maps at various stages of the network, significantly enhancing the quality of super-resolved images, and SENext [41] by Wazir et al. incorporates advanced channel-wise attention mechanisms to recalibrate feature responses dynamically,

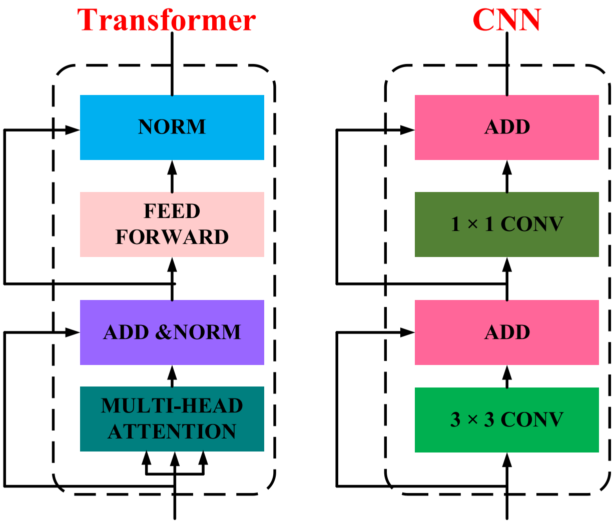

FIGURE 1. General architecture of transformer and CNN networks.

improving super-resolution performance while simultaneously reducing computational cost and prevents over-fitting. Howevere, careful consideration and investigation of these issues are still required.

Enhanced spatial attention mechanisms have been explored to improve the performance of ISR models. Woo et al. introduced the Convolutional Block Attention Module (CBAM) [42], which combines spatial and channel attention mechanisms to refine feature representations. This approach has effectively enhanced image quality by allowing the model to focus on relevant regions and suppress noise. Similarly, the dual attention mechanism proposed by Li et al. in the DANet [43] model demonstrated the benefits of incorporating both spatial and channel attention for image super-resolution tasks.

Recently, transformer-based models have gained popularity in ISR due to their capability to capture long-range dependencies and model global context more effectively than CNNs. Vision Transformers (ViTs) [14], introduced by Dosovitskiy et al., demonstrated the potential of transformers in various vision tasks. The Swin Transformer by Liu et al. [16]. proposed a hierarchical transformer model with shifted windows, significantly reducing computational complexity while maintaining high performance. Liang et al. introduced the SwinIR [15] model, which integrated the Swin Transformer into a residual network, showcasing its effectiveness in image restoration tasks. ELAN [18] architecture is designed to harness the power of long-range attention mechanisms while maintaining computational efficiency. SRFormer [17] introduces multi-head self-attention mechanisms to capture complex relationships between distant pixels, improving the network's ability to reconstruct fine details and textures in super-resolved images. SRFormer incorporates positional encoding, which helps the model understand the relative positions of pixels, enhancing its spatial awareness. Hybrid models combining the strengths of CNNs and transformers have been proposed to leverage local feature extraction and global context modeling. For instance, the TransCNN [44] model by Waseem et al. combined CNN and transformer blocks to achieve state-of-the-art results in ISR. The IPT [21] model by Chen et al. used a large-scale pre-trained transformer for various image processing tasks, including ISR, Hybrid Network of CNN, and Transformer for Lightweight Image Super-Resolution HNCT [20] and TCDFN [19] are also some of the hybrid approaches highlighting the versatility and power of transformer-based approaches.

The landscape of image super-resolution has been significantly enriched by advancements in CNNs, GANs, attention mechanisms, transformer-based models, hybrid approaches, and efficient model design. Each of these techniques has contributed to improving the quality and performance of ISR models, addressing various challenges such as computational complexity, memory requirements, and the need for real-time processing on edge devices. The ongoing research and development in these areas continue to push the boundaries of what is possible in image super-resolution, paving the way for more sophisticated and practical applications. In this work, a novel approach to U-shaped network architecture combining CNN and transformer is laid out. These parts are brought together to improve image super-resolution performance by utilizing the advantages of both transformers and CNNs.

To sum up, advances in image super-resolution have been made possible by deep learning-based methods, CNNs, transformers, and network designs. By combining these methods, reconstruction quality has significantly improved, allowing for the creation of more detailed high-resolution images. Research issues include adapting SR approaches to video super-resolution, generalizing across domains, and maximizing computational efficiency. These developments are needed to realize the full potential of image super-resolution in various applications, including digital content creation [45], medical imaging [1], surveillance [3], remote sensing [46], and facial image super-resolution [47].

III. PROPOSED METHODOLOGY

This section presents our proposed novel hybrid approach in single image super-resolution by fusion of Transformer and CNN in a Parallel Hybrid Transformer CNN Block (PHTCB) in a U-shaped design framework. Furthermore, we employed enhanced spatial attention in the network architecture to refine feature representations and allow the model to focus on relevant regions and suppress noise. Moreover, the information from PHTCB is transferred using skip connections to transmit low-frequency information at each stage of the network to reduce the parameters for the computation. Finally, we use pixel shuffle to reconstruct the high-resolution image.

Figure 2 shows our proposed Deep Hybrid transformer U-Network for Single Image Super-Resolution (DHTCUN)

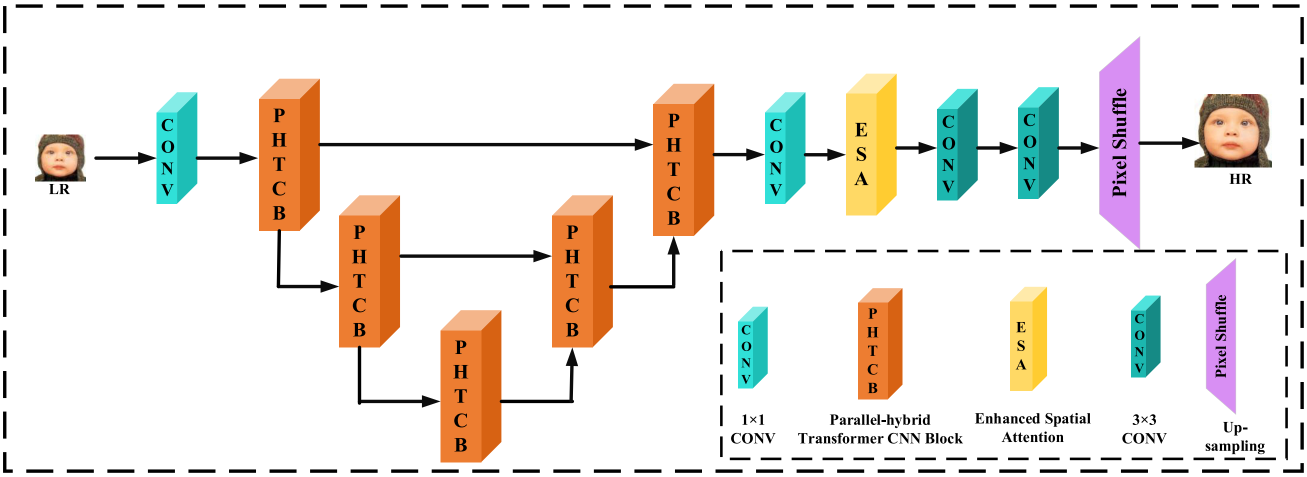

FIGURE 2. Deep hybrid transformer CNN U-Net for single image super resolution's (DHTCUN) proposed network architecture.

consisting of five Parallel Hybrid Transformer CNN Blocks (PHTCB) described in section IIIA. The initial low-resolution features are extracted using a normal convolution 1 × 1. Then, the features are transmitted to the U-shaped architecture designed using the PHTCB. The less complex features are directly transmitted from the first PHTCB to the last PHTCB using a skip connection, where the more complex features traverse through the second PHTCB to the second-last PHTCB. The higher the complexity of the feature, the more they traverse through the PHTCB to attain better refinement. After passing from the U-shaped framework, these features pass through the Enhanced spatial attention block followed by a 1 × 1 convolution layer to achieve better refinement by focusing on the relevant features and and noise suppression. Finally, the High-resolution image is reconstructed using a Pixel Shuffle folowed by two layers of the 3 × 3 convolution.

The initial feature extraction is indicated in Equation 1.

$$H_0 = F_{Conv1}(H_{LR}), \tag{1}$$

Here, H_{LR} is the input Low-Resolution (LR) image, $F_{Conv1}(.)$ represents 1 × 1 convolution operation for extracting initial features, and H_0 is the output of the convolution layer.

After passing through the initial feature extraction stage, the features are transmitted to the PHTCB in the U-shaped framework.

A. PARALLEL HYBRID TRANSFORMER CNN BLOCK (PHTCB)

Parallel Hybrid Transformer CNN Block (PHTCB) captures long range dependencies, reduce artifacts, and simultaneously prevents noise amplification in super-resolution of noisy images. The architecture of PHTCB is shown in Figure 3. It consists of Triple Enhanced Spatial Attention (TESA) described in section III A.1 Block and Transformer CNN (TCN) Blocks described in section III A.2. Since the hybrid combination of Transformer and CNN is parallelly connected in this block, that's why the Block is named as Parallel Hybrid Transformer CNN Block (PHTCB).

As seen in Figure 3, the PHTCB consists of the TESA and the TCN Blocks. The input to the PHTCB first passess through the TESA and then is parallelly distributed through both the Hybrid TCN Blocks. The TCN block is the Hybrid block cascading together the Swin Transformer Layer (STL) and the convolutional layer. Finally, the features pass through a convolutional layer and pass through a TESA followed by an arithmetic addition.

The mathematical expression of PHTCB is given in Equation 2, 3, 4, 5, and 6.

The input to the PHTCB is fed to the TESA block. The euation of the TESA Block is given by Equation 2.

$$H_{TESA} = F_{TESA}(H_I), \tag{2}$$

Here, H_I is the input to the PHTCB block, $F_{TESA}(.)$ is the TESA Block function, and H_{TESA} is the output of the TESA block.

Equation 3 shows the output of each TCN block, which is distributed parallel to the TESA block.

$$H_{TCN} = \left(F_{\text{STL}}(F_{\text{Conv3}}(H_{\text{TESA}}))\right) \tag{3}$$

Here, $F_{\text{Conv3}}(.)$ is the 3 × 3 convolution function, $F_{\text{STL}}(.)$ is the STL function, and H_{TCN} is the output of the TCN block.

In Equations 4 and 5, the outputs of both TCN blocks are added and passed through the convolution layer.

$$H_{Con_{i/p}} = H_{TCN1} + H_{TCN2}, \tag{4}$$

Here, $H_{Con_{i/p}}$ is the input to the 1 × 1 convolution layer, and H_{TCN1} and H_{TCN2} are the outputs from both the TCN blocks. $H_{TCN1} = H_{TCN2} = H_{TCN}$

$$H_{Con_{o/p}} = F_{Conv1}(H_{Con_{i/p}}), \tag{5}$$

Here, $H_{Con_{o/p}}$ is the output of the 1 × 1 convolution layer, and $F_{Conv1}(.)$ represents the 1 × 1 convolution operation.

Finally, all the features are passed through the TESA block.

$$H_{PHTCB} = F_{TESA}(H_{Con_{o/p}}), \tag{6}$$

Here, H_{PHTCB} denotes the output of the PHTCB block

1) TRIPLE ENHANCED SPATIAL ATTENTION (TESA) BLOCK

The Triple Enhanced Spatial Attention (TESA), as seen in Figure 4, consists of three Enhanced Spatial Attention (ESA) modules stacked together. This approach enhances image quality by focusing on relevant regions of the image. TESA employs an iterative approach where the output of one attention module is refined by subsequent modules, progressively enhancing the focus on critical features. Equation 7 shows the output of TESA.

$$H_{TESA} = F_{ESA}(F_{ESA}(F_{ESA}(H_{i/p}))), \tag{7}$$

Here $F_{ESA}(.)$ represents the Enhanced Spatial Attention (ESA) module function, and $(H_{i/p})$ is the input fed into the TESA block, and H_{TESA} is the output of the TESA Block.

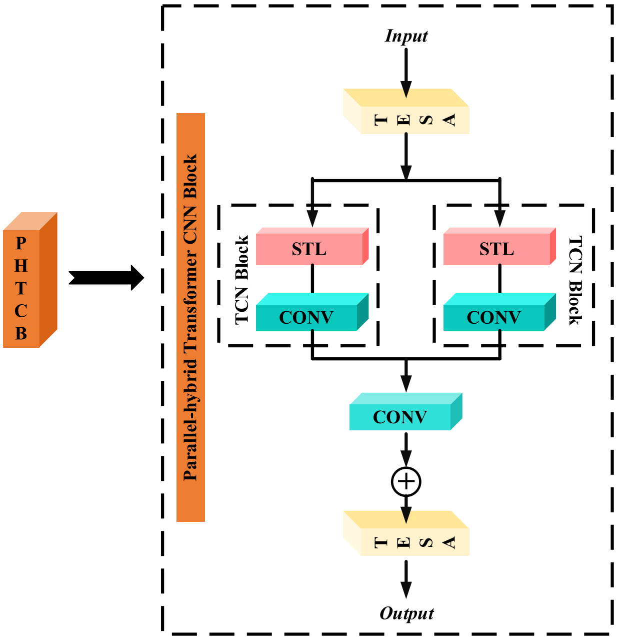

FIGURE 3. The structure of parallel hybrid transformer CNN block (PHTCB).

2) TRANSFORMER CNN (TCN) BLOCK

As already seen in Figure 2, the Transformer CNN (TCN) Block is a hybrid block formed by cascading the Swin transformer Layer and the CNN layer. It combines the powers of CNNS and Transformer to capture long-range dependencies, reduce artifacts, and simultaneously prevent the noise amplification in super-resolution noisy images. In our proposed architecture, we have used two TCNs connected in paraller in each PHTCH. The output of the TCN Block is given in Equation 8.

$$H_{TCN} = F_{Conv3}(F_{STL}(H_{In}), \tag{8}$$

Here, $F_{STL}(.)$ represents the STL function, (F_{Conv3}) is the 3 3 convolution operation, H_{In} is the input of the TCN block, and H_{TCN} is the output of the TCN block.

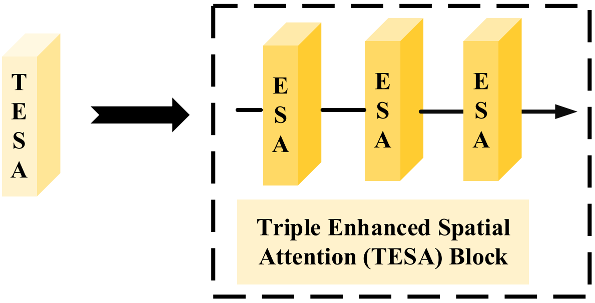

FIGURE 4. The structure of triple enhanced spatial attention (TESA) block.

B. ENHANCED SPATIAL ATTENTION (ESA)

Enhanced Spatial Attention (ESA) is a critical components in image super-resolution (ISR) models, designed to focus on significant features within an image selectively. This mechanisms aims to improve the network's ability to discern and enhance crucial details, such as edges, textures, and fine patterns, leading to higher-quality super-resolved images.

Enhanced Spatial Attention mechanisms refine traditional spatial attention by incorporating more sophisticated methods to identify and focus on critical image areas. This selective focus helps the network preserve and enhance fine details often lost in conventional methods. ESA dynamically adjusts the weights assigned to different regions based on the context and content of the image, allowing for adaptive attention that responds to varying image characteristics.

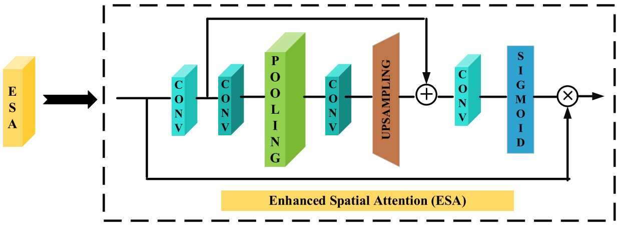

FIGURE 5. Enhanced spatial attention (ESA).

Figure 5 shows a diagram representing the ESA mechanism. Attention maps are generated by adding the input feature maps through convolutional, pooling, and up-sampling layers. These maps indicate the importance of each spatial location. The generated attention maps are then multiplied element-wise, followed by an activation function (often sigmoid) with the input feature maps, emphasizing the significant regions while suppressing less important areas.

The mathematical expression of ESA is described in Equations 9, 10, 11, and 12,

$$H_{E1} = F_{up}(F_{Conv3}(F_{pool}(F_{Conv3}(F_{Conv1}(H_{E1i/p}))))), \tag{9}$$

Here, $H_{E1i/p}$ is the input to the ESA, $F_{Conv1}(.)$ is the 1×1 convolution operation, $F_{Conv3}(.)$ is the 3×3 convolution operation function, $F_{up}(.)$ is the up-sampling function, and H_{E1} is one of the inputs to the addition operation in ESA.

$$H_{E2} = F_{Conv1}(H_{E1i/p}), \tag{10}$$

H_{E2} is another input for the addition operation in ESA, as seen in Figure 5.

$$H_{E3} = H_{E1} + H_{E2}, \tag{11}$$

H_{E3} is the added output of the addition operation in ESA.

$$H_{ESA} = H_{E1} \times \sigma \left(F_{Conv3} \left(H_{E3} \right) \right) \tag{12}$$

Here, × denotes the element-wise multiplication, σ denotes the sigmoid function, and H_{ESA} is the output of the Enhanced Spatial Attention (ESA) mechanism.

C. SWIN TRANSFORMER LAYER (STL)

The Swin Transformer layer is the proposed architecture's transformer component, designed for efficient and scalable vision tasks, including image super-resolution. Unlike the original STL, which uses the components twice, we have used the components inside STL just once to reduce the computation for self-attention across the input feature map. It uses a hierarchical approach to model long-range dependencies and global context efficiently. We apply global self-attention across the entire input feature map. After each self-attention operation, a two-layer MLP is applied to transform the features further. Layer Normalization preceds each SW-MSA and MLP block to stabilize and improve the training process. Residual connections around each SW-MSA and MLP block enhance gradient flow and mitigate the vanishing gradient problem.

Shifted Window Multi-Head Self-Attention is a technique used in the Swin Transformer [16] to balance computational efficiency and feature interaction. The input image is divided into non-overlapping windows where self-attention is performed independently, reducing computational complexity. In the next layer, the windows are shifted by a certain number of pixels, allowing cross-window interactions. This shift allows for cross-window interactions, ensuring that pixels from the boundaries of the initial windows now fall within the center of the new windows. This mechanism enables the model to capture long-range dependencies and global context without significantly increasing computational cost. By combining local and global feature interactions, the Swin Transformer Layer enhances performance on various vision tasks.

Figure 6 represents the STL used in our proposed method. Equations 13 and 14 give the mathematical expression for STL.

$$H_{S1} = H_{S1i/p} + (F_{LN}(F_{SWMSA}(H_{S1i/p}))), \tag{13}$$

Here, $H_{S1i/p}$ is the input to the STL, $F_{SWMSA}(.)$ is the Shifted Window Multi-Head Self Attention function, $F_{LN}(.)$ is the Layer Normalization function, and H_{S1} isthe output of the addition operation.

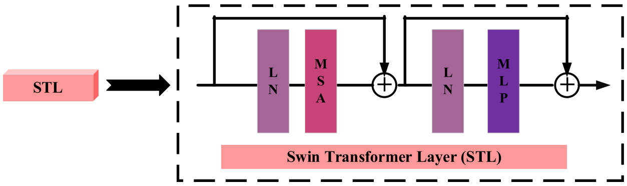

FIGURE 6. Swin transformer layer (STL).

$$H_{STL} = H_{S1} + (F_{LN}(F_{MLP}(H_{S1}))), \tag{14}$$

Here, $F_{MLP}(.)$ is a Multi-Layer Perceptron, and H_{STL} is the final output of the STL.

IV. EXPERIMENTAL RESULTS

In the experimental results section, we have demonstrated the results of various qualitative and quantitative experiments conducted using our proposed method. We have also shown the comparative analysis of our proposed DHTCUN for parameters, average PSNR and SSIM, execution time, and time complexiety of our proposed method. Furthermore an ablation study has also been conducted by checking different numbers of PHTCB blocks to be used in the model by changing the structure of the TCN block and by changing the configuration of the TCN block in PHTCH, PSNR versus Multi add analysis and finally, convergence analysis of the model.

A. EXPERIMENTAL SETUP

This section presents the evaluation metrics, training specifics, and datasets to test the efficacy of our suggested model on publicly available datasets. The testing and training sets are not the same.

1) TRAINING AND TESTING SETUP

Training our model involved randomly cropping low-resolution patches of 48 × 48. A window size of 24 × 24 has been selected.For ×2, ×3, ×4, and ×8, low-resolution images were created using MATLAB R2022b. A 24-GB NVIDIA GeForce GTX 2080ti GPU trains the proposed network. Python 3.6 and the PyTorch 1.1.0 platform were used to write the algorithm for the proposed model. Model training involves obtaining 800 high-quality samples from DIV2K [48] datasets. We choose an Adam optimizer with $\beta_2 = 0.99$ and $\beta_1 = 0.90$ for optimization. For every 200 epochs, the learning rate of the suggested model is halved from 10^{-4}.

Five common benchmark datasets, namely Set5 [49], Set14 [50], BSD100 [51], Urban100 [52], and Manga109 [53], were used to test our proposed model. Bicubic kernels are used to downsample the HR images to produce the LR image. Each batch size of eight training samples is divided up. Furthermore, data augmentation produces extra samples for the computation by flipping and rotating at random angles of 90, 180, and 270 degrees. The image intensity range [-1, 1] has been used to compute the Mean Squared Error (MSE). Standard evaluation metrics like PSNR and SSIM can be used to compare our model quantitatively with the most advanced techniques.

B. QUANTITATIVE EVALUATIONS IN COMPARISON TO STATE-OF-THE-ART METHODS

Five benchmark test datasets are tabulated and compared using standard metrics in Table 1. We quantitatively compare our proposed DHTCUN with seventeen SOTA methods: Bicubic, LapSRN [8], SENext [41], RCAN [11], MemNet [9], MFCC [40], EDSR [10], HAN [39], SwinIR [15], and NLSN [12], SRFormer [17], TCDFN [19], HNCT [20], MSRN [29], AWSRN [30], DBPN [31] and RDN [33]. Our suggested DHTCUN quantitative results have considerably surpassed the state-of-the-art techniques regarding PSNR and SSIM, as indicated in Table 1. For scale factors ×2, ×3, ×4, and ×8, our suggested DHTCUN model performs better across all test datasets. Additionally, our suggested approach produced a higher average PSNR/SSIM value on all image SR test datasets than other SOTA models.

C. COMPARATIVE STUDY USING THE QUANTITY OF MODEL PARAMETERS

The parameters and PSNR comparison for our suggested DHTCUN model are displayed in Figure 7. The effectiveness of our suggested model, DHTCUN, with a scale factor of ×2, is assessed using the Set5 [49] test dataset. Lowering the number of parameters indicates lower computational costs. The DHTCUN model more effectively reduces the model's size when compared to other state-of-the-art deep learning models. Around 92% fewer parameters are found in DHTCUN than in EDSR [10], 81% less in RCAN [11], 85% less in RDN [33], 42% less in NLSN [12], and 22% less in HAN [39]. Comparing our suggested method to five other cutting-edge approaches, Figure 7 demonstrates that our suggested method has fewer parameters. This indicates our model's efficiency in reducing the computational burden.

To further demonstrate the efficiency of the proposed model in reducing computational cost, the following points have been mentioned:

Unlike HNCT and SRFormer, which do not support ×8 super-resolution, DHTCUN excels at ×2, ×3, ×4, and ×8 magnifications. It provides superior results with similar computational costs, making it advantageous for high magnification applications at a lesser cost.

Although DHTCUN has more parameters than HNCT on ×2 Set 5 [49], it a shows a superior performance of 0.4 dB as compared to HNCT, which itself is a significant improvement. Additionally, as seen in Figure 11, the time required by a single epoch for the training our proposed DHTCUN is less than that of HNCT, which demonstrates the efficiency of our proposed DHTCUN in reducing the computational expense.

As for the comparison with SRFormer, it should be noted that SRFormer is a Transformer-based method whereas our proposed DHTCUN combines Transformer and CNN techniques, leveraging the strengths of both. This dual architecture may result in a higher parameter count but offers superior performance across various scenarios.

While DHTCUN may have a higher parameter count due to its dual architecture, the trade-off with performance gains, especially in high magnification applications, can justify the increased parametric usage.

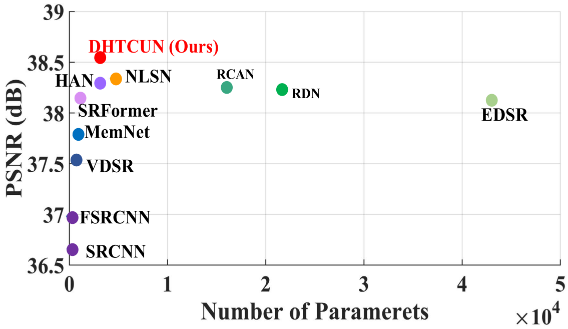

FIGURE 7. Analysis of model parameters about PSNR using the ×2 enlargement factor on the Set5 [49] image dataset.

D. COMPARISON OF THE MEAN PSNR AND SSIM OF THE IMAGE SR DATASETS FOR ×4 AND ×8 ENLARGEMENT FACTORS.

Using standard objective measures, Figures 8 and 9 compare the average PSNR and SSIM of various existing image SR methods on benchmark datasets (Set5 [49], Set14 [50], BSD100 [51], Urban100 [52], Manga109 [53]) for enlargement factors of ×4 and ×8. According to the quantitative results, our proposed DHTCUN outperforms HNCT [20], SRFormer [17], NLSN [12], RCAN [11], EDSR [10], and MemNet [9] when it comes to the enlargement factor of ×4, and SRCNN [5], LapSRN [8], MemNet [9], RCAN [11], HAN [39], RDN [33], EDSR [10], and AWSRN [30] when it comes to the enlargement factor of ×8. The average quantitative PSNR and SSIM values for Figure 8 and Figure 9 are given in Table 1.

E. PSNR VERSUS EXECUTION TIME: A QUANTITATIVE ANALYSIS

Execution time is required for an ISR model to process an image and produce a high-resolution output. As seen in Figure 10, this section displays the DHTCUN's performance regarding PSNR versus execution time. The state-of-the-art techniques were assessed using an NVIDIA GeForce GTX 2080ti GPU with 24GB of memory. Figure 10 illustrates the trade-off between PSNR and execution time on Set14 [50] scale factor ×4. Our suggested approach outperforms five state-of-the-art techniques (HNCT [20], SRFormer [17], NLSN [12], RNAN [53], and RDN [33]) with the highest PSNR of 29.06.

TABLE 1. Comparison of the most sophisticated SR methods for upscaling factors ×2, ×3, ×4, and ×8 with our suggested DHTCUN evaluated using standard metrics. The highest score is displayed in bold and is colored Red. Blue indicates and displays the score that comes in second place.

Method	Factor	#Param	Set5 [51]		Set14 [52]		BSD100 [53]		Urban100 [54]		Manga109 [55]		Average	
			PSNR↑	SSIM↑	PSNR↑	SSIM ↑	PSNR↑	SSIM ↑	PSNR↑	SSIM ↑	PSNR↑	SSIM ↑	PSNR↑	SSIM ↑
Bicubic	×2	-/-	33.68	0.9304	30.24	0.8691	29.56	0.8435	26.88	0.8405	31.05	0.9349	30.23	0.8832
MemNet [9]	×2	677K	37.78	0.9597	33.28	0.9142	32.08	0.8978	31.31	0.9195	37.72	0.9740	34.43	0.9330
LapSRN [8]	×2	812K	37.52	0.9591	32.99	0.9124	31.80	0.8949	30.41	0.9101	37.53	0.9740	33.87	0.9302
MSRN [29]	×2	6,226K	38.08	0.9605	33.74	0.9170	32.23	0.9013	32.22	0.9326	38.82	0.9868	35.02	0.9396
EDSR [10]	×2	43,000K	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773	35.28	0.9386
AWSRN [30]	×2	1,397K	38.11	0.9608	33.78	0.9189	32.26	0.9006	32.49	0.9316	38.87	0.9776	35.10	0.9379
DBPN [31]	×2	5,820K	38.08	0.9600	34.09	0.9210	32.31	0.9010	32.92	0.9350	39.28	0.9770	35.34	0.9388
MFCC [40]	×2	1,861K	38.16	0.9606	33.85	0.9195	32.28	0.9010	32.65	0.9331	39.11	0.9780	35.21	0.9384
RDN [33]	×2	21,900K	38.24	0.9614	34.01	0.9212	32.34	0.9017	32.89	0.9353	39.18	0.9780	35.33	0.9395
RCAN [11]	×2	16,000K	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786	35.52	0.9405
NLSN [12]	×2	4,475K	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789	35.57	0.9412
SENext [41]	×2	97K	38.04	0.9608	34.24	0.9181	32.21	0.8997	32.43	0.9287	38.79	0.9774	35.14	0.9369
HAN [39]	×2	3,230K	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785	35.53	0.9405
SwinIR [15]	×2	878K	38.38	0.9620	34.24	0.9233	32.47	0.9032	33.51	0.9401	**39.70**	**0.9794**	35.66	0.9416
ELAN [18]	×2	621K	38.36	0.9620	34.20	0.9228	32.45	0.9030	33.44	0.9391	39.62	0.9783	35.61	0.9412
SRFormer [17]	×2	853K	38.45	0.9622	34.21	**0.9236**	32.51	0.9038	33.86	**0.9426**	39.69	0.9786	35.74	**0.9422**
TCDFN [19]	×2	573K	38.16	0.9612	33.89	0.9206	32.26	0.9002	32.79	0.9341	39.12	0.9780	35.24	0.9388
HNCT [20]	×2	356K	38.08	0.9608	33.65	0.9182	32.22	0.9001	32.22	0.9294	38.87	0.9774	35.01	0.9372
DHTCUN (Ours)	×2	2,375K	**38.48**	**0.9624**	**34.25**	0.9234	**32.54**	**0.9040**	**33.88**	**0.9426**	**39.70**	0.9786	**35.77**	**0.9422**
Bicubic	×3	-/-	30.40	0.8686	27.54	0.7741	27.21	0.7389	24.46	0.7349	26.95	0.8566	27.31	0.7945
MemNet [9]	×3	677K	34.09	0.9248	30.00	0.8350	28.96	0.8001	27.56	0.8376	32.51	0.9369	30.62	0.8669
LapSRN [8]	×3	812K	33.82	0.9227	29.79	0.8320	28.82	0.7973	27.07	0.8271	32.21	0.9350	30.36	0.8631
MSRN [29]	×3	6,226K	34.38	0.9262	30.34	0.8395	29.08	0.8041	28.08	0.8554	33.44	0.9427	31.06	0.8736
EDSR [10]	×3	43,000K	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476	31.48	0.8792
AWSRN [30]	×3	1,476K	34.52	0.9281	30.38	0.8426	29.16	0.8069	28.42	0.8580	33.85	0.9463	31.26	0.8764
MFCC [40]	×3	2,230K	34.67	0.9294	30.51	0.8456	29.22	0.8080	28.64	0.8616	34.15	0.9478	31.43	0.8793
RDN [33]	×3	21,900K	34.71	0.9296	30.57	0.8468	29.26	0.8093	28.80	0.8653	34.13	0.9484	31.49	0.8798
RCAN [11]	×3	16,000K	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499	31.64	0.8818
NLSN [12]	×3	4,475K	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508	31.74	0.8824
SENext [41]	×3	54K	34.32	0.9255	**31.08**	0.8419	29.11	0.8047	28.60	0.8519	33.63	0.9451	31.35	0.8738
HAN [39]	×3	3,230K	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500	31.66	0.8819
SwinIR [15]	×3	886K	34.89	0.9312	30.77	0.8503	29.37	0.8124	29.29	0.8744	34.74	0.9518	31.81	0.8840
ELAN [18]	×3	629K	34.90	0.9313	30.80	0.8504	29.38	0.8124	29.32	0.8745	34.73	0.9517	31.82	0.8841
SRFormer [17]	×3	861K	34.94	0.9318	30.81	**0.8518**	29.41	0.8142	**29.52**	**0.8786**	34.78	0.9524	31.89	**0.8857**
TCDFN [19]	×3	582K	34.63	0.8287	30.56	0.8466	29.23	0.8082	28.71	0.8624	**34.98**	0.9477	31.62	0.8587
HNCT [20]	×3	363K	34.96	**0.9329**	30.88	0.8512	29.42	0.8132	29.31	0.8752	34.88	0.9519	31.89	0.8848
DHTCUN (Ours)	×3	2,386K	**34.98**	0.9320	30.89	0.8515	**29.44**	**0.8143**	29.34	0.8754	34.91	**0.9526**	**31.91**	0.8852
Bicubic	×4	-/-	28.43	0.8109	26.00	0.7023	25.96	0.6678	23.14	0.6574	25.15	0.7890	25.68	0.7250
MemNet [9]	×4	677K	31.74	0.8893	28.26	0.7723	27.40	0.7281	25.50	0.7630	29.42	0.8942	28.46	0.8094
LapSRN [8]	×4	812K	31.54	0.8866	28.09	0.7694	27.32	0.7264	25.21	0.7553	29.09	0.8900	28.27	0.8060
MSRN [29]	×4	6,226K	32.07	0.8903	28.60	0.7751	27.52	0.7273	26.04	0.7896	30.17	0.9034	28.88	0.8171
EDSR [10]	×4	43,000K	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148	29.32	0.8289
AWSRN [30]	×4	1,587K	32.27	0.8960	28.69	0.7843	27.64	0.7385	26.29	0.7930	30.72	0.9109	29.12	0.8245
DBPN [31]	×4	10,200K	32.65	0.8990	29.03	0.7910	27.82	0.7440	27.08	0.8140	31.74	0.9210	29.66	0.8338
MFCC [40]	×4	2,157K	32.42	0.8973	28.73	0.7849	27.67	0.7399	26.48	0.7977	30.98	0.9131	29.25	0.8265
RDN [33]	×4	21,900K	32.47	0.8990	28.81	0.7871	27.72	0.7419	26.61	0.8028	31.00	0.9151	29.32	0.8291
RCAN [11]	×4	16,000K	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173	29.46	0.8317
NLSN [12]	×4	4,475K	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184	29.49	0.8325
SENext [41]	×4	54K	31.50	0.8947	28.99	0.7812	**28.49**	0.7357	26.64	0.7839	30.48	0.9084	29.22	0.8208
HAN [39]	×4	3,230K	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177	29.52	0.8321
SwinIR [15]	×4	897K	32.72	0.9021	28.94	0.7914	27.83	0.7459	27.07	0.8164	31.67	0.9226	29.64	0.8356
ELAN [18]	×4	621K	32.75	0.9022	28.96	0.7914	27.83	0.7459	27.13	0.8167	31.68	0.9226	29.67	0.8357
SRFormer [17]	×4	873K	32.81	**0.9029**	**29.01**	0.7919	27.85	0.7472	27.20	0.8189	31.75	**0.9237**	29.72	0.8369
TCDFN [19]	×4	591K	32.44	0.8976	28.79	0.7861	27.71	0.7381	26.51	0.7981	30.90	0.9151	29.27	0.8270
HNCT [20]	×4	372K	32.78	0.9028	28.98	0.7928	27.97	0.7468	27.32	0.8189	31.74	0.9228	29.76	0.8368
DHTCUN (Ours)	×4	2,395K	**32.83**	**0.9029**	**29.01**	**0.7929**	27.98	**0.7474**	**27.34**	**0.8191**	**31.78**	0.9230	**29.79**	**0.8371**
Bicubic	×8	-/-	24.40	0.6580	23.10	0.5660	23.67	0.5480	20.74	0.5160	21.47	0.6500	22.68	0.5876
MemNet [9]	×8	677K	26.16	0.7414	24.38	0.6199	24.58	0.5842	21.89	0.5825	23.56	0.7387	24.11	0.6529
LapSRN [8]	×8	812K	26.15	0.7380	24.35	0.6200	24.54	0.5860	21.81	0.5810	23.39	0.7350	24.04	0.6520
MSRN [29]	×8	6,226K	26.59	0.7254	24.88	0.5961	24.70	0.5610	22.37	0.6077	24.30	0.7701	24.56	0.6520
EDSR [10]	×8	43,000K	26.96	0.7762	24.91	0.6420	24.81	0.5985	22.51	0.6221	24.69	0.7841	24.74	0.6824
AWSRN [30]	×8	2,348K	26.97	0.7747	24.96	0.6414	24.80	0.5967	22.45	0.6174	24.69	0.7842	24.77	0.6828
DBPN [31]	×8	23,100K	26.96	0.7762	24.91	0.6420	24.81	0.5985	22.51	0.6221	24.60	0.7732	24.75	0.6824
MFCC [40]	×8	2,453K	27.07	0.7762	25.01	0.6412	24.84	0.5980	22.54	0.6196	24.63	0.7791	24.81	0.6828
RDN [33]	×8	21,900K	27.21	0.7840	25.13	0.6480	24.88	0.6010	22.73	0.6312	25.14	0.7897	25.02	0.6907
RCAN [11]	×8	16,000K	27.31	0.7878	25.23	0.6511	24.98	0.6058	23.00	0.6452	25.24	0.8029	25.15	0.6985
SENext [41]	×8	97K	26.87	0.7415	**25.73**	0.6200	**26.79**	0.5847	21.90	0.5829	23.96	0.7389	25.05	0.6536
HAN [39]	×8	3,230K	27.33	0.7884	25.24	0.6510	24.98	0.6059	22.98	0.6437	25.20	0.8011	25.14	0.6980
DHTCUN (Ours)	×8	2,405K	**27.40**	**0.7888**	25.29	**0.6517**	25.08	**0.6064**	**23.08**	**0.6458**	**25.28**	**0.8033**	**25.22**	**0.6992**

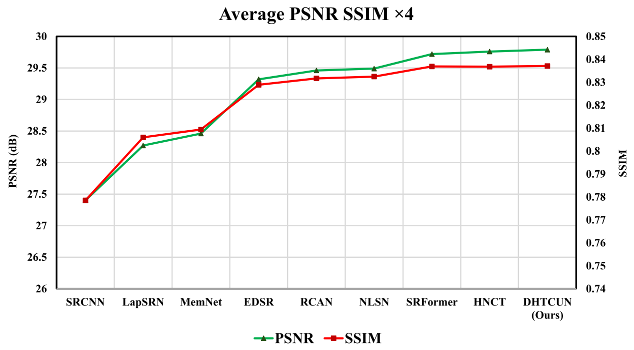

FIGURE 8. Comparison of image SR test data sets for PSNR and SSIM based on ×4 enlargement factor.

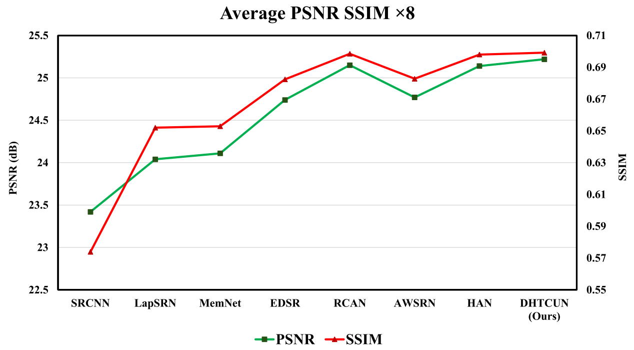

FIGURE 9. Comparison of image SR test data sets for PSNR and SSIM based on an enlargement factor of ×8.

Balancing the trade-off between execution time and performance in image super-resolution models involves a combination of approaches, architectural innovations, hardware acceleration, and careful evaluation of performance metrics. By applying these strategies, it is possible to achieve a model that performs well in terms of image quality while also being efficient in computation and suitable for deployment in resource-constrained environments.

To better understand the trade-off between execution time and performance, we have shown a tabular comparison of our proposed DHTCUN with four other state-of-the-art methods regarding performance, runtime, parameters and flops. Table 2 shows the comparison of our proposed DHTCUN with ELAN [18], SRFormer [17], TCDFN [19], and HNCT [20]. We have separated the transformer-based models and hybrid models with a horizontal line. As seen in Table 2, our proposed DHTCUN shows the best performance and lesser FLOPs compared to other state-of-the-art methods on Set 5 [49] ×4. SRFormer shows the second-best performance and best runtime but has a very high number of FLOPs. TCDFN [19] shows the second-best runtime, parameters count, and flops, but the performance is very low.

TABLE 2. Trade-off for Runtime, #Params, and #FLOPs with performance on Set5 [49] ×4.

Method	PSNR (dB)	Runtime (mS)	#Params (K)	#FLOPs (G)
ELAN [18]	32.75	298.11	8312	43.2
SRFormer [17]	32.81	**3.05**	873	125.6
TCDFN [19]	32.44	8.50	591	42.5
HNCT [20]	32.78	339.61	**372**	78.81
DHTCUN (Ours)	**32.83**	281.5	2395	**41.5**

Since ELAN [18] and SRFormer [17] are only transformer based models where as TCDFN [19], HNCT [20] and our proposed DHTCUN are hybrid methods. Our proposed DHTCUN shows better trade-off for performance, runtime, params and flops.

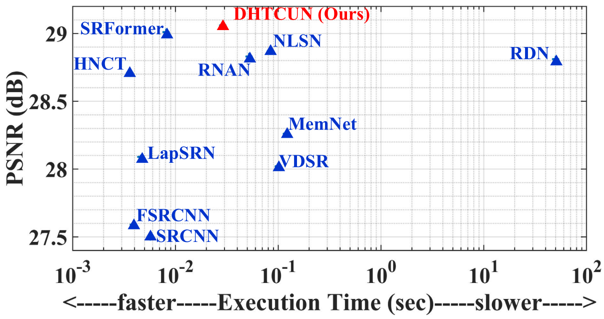

FIGURE 10. Execution time measurement against PSNR using a scale factor of ×4 on Set 14 [50].

F. EXAMINATION OF TIME COMPLEXITY

The time needed to finish each training epoch for a deep learning model illustrates how complex the model is in terms of time. The time complexity of these algorithms is a crucial factor, especially when aiming for real-time applications. The time complexity of image super-resolution is primarily influenced by the size of the input image ($n \times m$), the network d depth, and the convolutional kernel k.

Figure 11 displays the time required per epoch in 100 training epochs of the five state-of-the-art methods TCDFN [19], ELAN [18], HNCT [20], NLSN [12], and our suggested DHTCUN. The curves show a noticeable difference, suggesting that our suggested DHTCUN requires less training time for every epoch. DHTCUN, therefore, exhibits lower time complexity than the five state-of-the-art methods.

G. SPACE COMPLEXITY EXAMINATION

We have discussed the memory footprint of each model, which is crucial for understanding the practical deployment of these models on hardware with limited resources. Balancing space complexity with performance is essential for practical deployment.

Figure 12 displays the memory space required by five state-of-the-art methods SRFormer [17], TransCNN [44], HNCT [20], TCDFN [19], and our suggested DHTCUN. The bar graph shows that our suggested DHTCUN requires less memory than other state-of-the-art methods. Therefore, DHTCUN exhibits lower space complexity and has a lesser

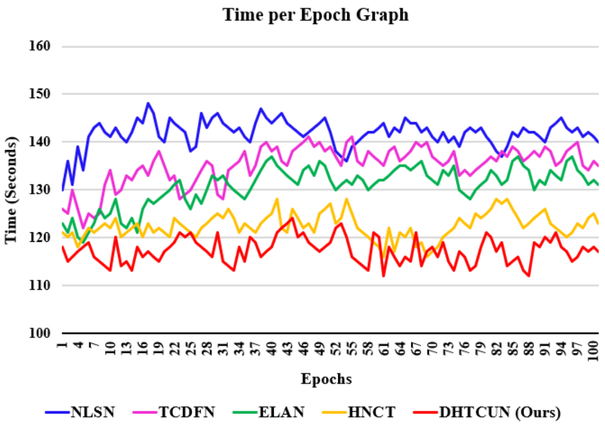

FIGURE 11. Estimation of time complexity on DIV2K [48] dataset for 100 epochs on scale factor ×4.

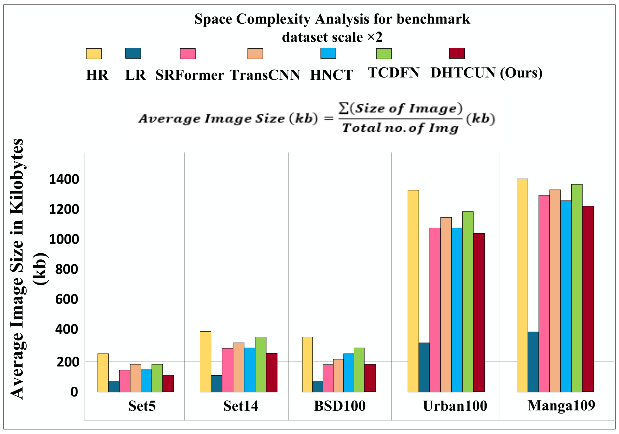

FIGURE 12. Convergence analysis of our model with the compared to another hybrid approach.

computational burden than the other four state-of-the-art methods.

H. A COMPARISON OF VISUAL QUALITY

For image SR test datasets, including Set5 [49], Set14 [50], BSD100 [51], Urban100 [52], and Manga109 [53], the visual quality of up-sampling factors ×4 and ×8 is displayed in Figures 13, 14, 15, 16, 17, 18, 19, and 20. Even so, improving an image for an enlargement factor of ×8 is challenging. Our proposed method shows finer and more detail-oriented results because of the hybrid transformer CNN approach used in the model.

We used the following images from respective datasets for scale factor ×4: Img_223061 from BSD100 [51], Img_098 from Urban100 [52], ARMS image from Manga109 [53], and zebra image from Set14 [50]. Similarly for the scale factor ×8, we used Img_119082 from BSD100 [51], Img_044 from Urban100 [52], the foreman form Set14 [50] dataset, and KuroidoGanka image from Manga109 [53] datasets. In comparison to other state-of-the-art methods like Bicubic, MSRN [29], EDSR [10], AWSRN [30], RCAN [11], NLSN [12], SwinIR [15], SRFormer [17], TCDFN [19], and HNCT [20] for ×4, better quantitative metrics (PSNR/SSIM) and aesthetically pleasing patches are displayed by our proposed DHTCUN. Likewise, for scale factor ×8 Bicubic, HAN [39], TCDFN [19], LapSRN [8], MSRN [29], RCAN [11], AWSRN [30], and DBPN [31] are comparable SOTA methods.

I. ABLATION EXAMINATION

Here, we analyze our proposed model through controlled experiments. Five Parallel Hybrid Transformer CNN Blocks (PHTCB) are included in the suggested model. The framework's pixel shuffle function was utilized for up-sampling. To make the model lightweight, we finally added a skip connection. The suggested model's ablation study was carried out using the following methods: (1) By calculating the PSNR versus Multi-Add, (2) By changing the number of PHTCBs in the network, (3) by analyzing the number of ESA for the Multi Enhanced Spatial Attention (MESA) inside the PHTCB, (4) through an examination of comparisons with conventional denoising methods, and (5) By calculating PSNR versus Epoch convergence. We conduct these tests to see how they affect the suggested model's performance.

1) ABLATION INVESTIGATION BY CALCULATING THE PSNR VERSUS MULTI-ADD

In deep learning models, particularly those involving convolutional neural networks (CNNs) and transformers, "multi adds" typically refer to the multiplication and addition operations required for matrix multiplications, which are fundamental to convolutional operations and transformer mechanisms. These operations are critical in determining the computational complexity and efficiency of the model. The total multi-adds for a deep learning model is the sum of the multi-adds for each layer. This sum provides an estimate of the computational complexity of the model, influencing the required computational power and execution time. Optimizing the number of multi-adds is crucial for making models efficient, particularly for deployment on resource-constrained devices like smart edge devices or embedded systems on chips (SoC). Hence, understanding and optimizing multi-adds is essential for developing efficient and effective deep-learning models.

From Figure 21, it is evident that our proposed method benchmarks a few representative SR methods like MSRN [29], RCAN [11], NLSN [12], SRFormer [17], HNCT [20], TCDFN [19] and EDSR [10] on the metrics of SR performance (PSNR), model size (number of parameters), and computation cost (number of Multi-Adds).

2) ABLATION STUDY BY CHANGING THE NUMBER OF PHTCBs IN THE NETWORK

Table 3 shows the model's performance and computation regarding PSNR, SSIM, and Multi-adds. **Red** represents the optimal value, while Blue represents the second-best value. We tried different numbers of Parallel Hybrid Transformer

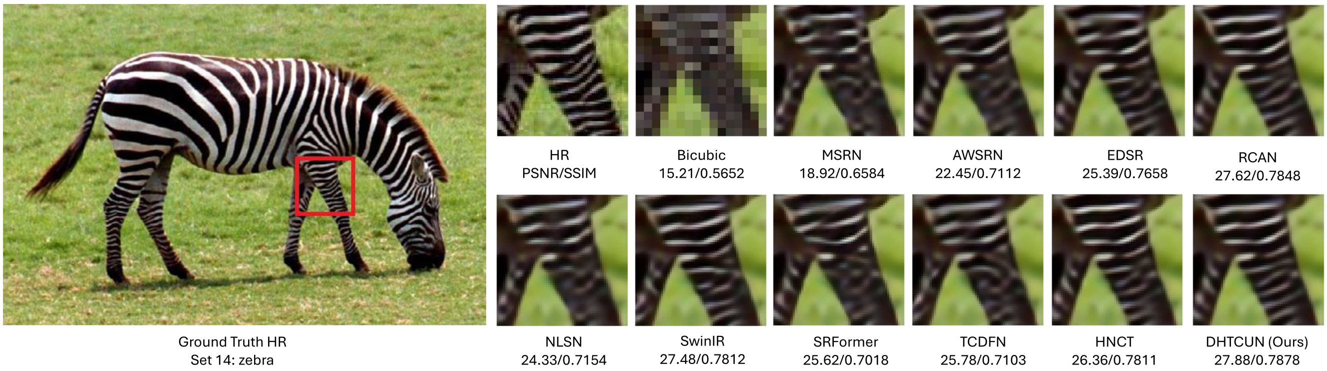

FIGURE 13. Zebra image quality improvement on a scale factor of ×4 from the Set14 [50] dataset.

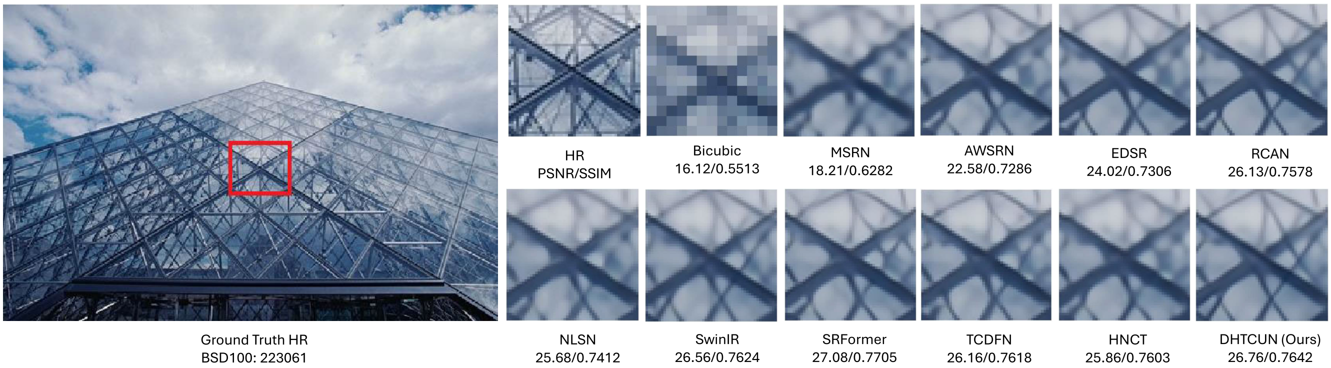

FIGURE 14. Img_223061 quality improvement on a scale factor of ×4 from the BSD100 [51] dataset.

CNN Blocks (PHTCBs) for our network. We trained the model for 300 epochs and compared the PSNR, SSIM, and Multi-adds. The number of PHTCBs showing the highest PSNR and SSIM value with the lowest Multi-Adds count is chosen for the final training of the model.

As such, it depends on the model's requirements. The model with four PHTCBs is the best option if we need to compute less; however, if we need more performance, the model with five PHTCBs is the best choice. It displays the second-lowest Multi-Adds count along with the highest PSNR and SSIM.

It is to be noted that the Parallel Hybrid Transformer CNN Blocks are connected in a U-shaped architecture (both series and parallel connection simultaneously) using skip connections to separate the high-frequency details (contextual details) and the low-frequency details to reduce the computational burden. Figure 22 shows better understanding of the arrangments of PHTCBs in the network as it is clear from Table 3 that the arrangement with 5 PHTCBs has the best trade-off for performance and computational expense.

Note: Figure 22 shows a different arrangement of PHTCBs in the proposed network. The rest of the layers (Convolutional, ESA, and Pixel shuffle) remain the same for all the arrangements as that of the final model shown in Figure 2.

TABLE 3. Different number of PHTCBs in network.

Number of PHTCB	Average PSNR	Average SSIM	Multi-Adds
4	37.73	0.9604	**345G**
5	**37.74**	**0.9606**	345.5G
6	37.68	0.9600	348.5G
7	37.72	0.9601	352G
8	37.66	0.9598	353.5G
9	37.71	0.9602	368G
10	37.72	0.9603	372.5G

3) ANALYSING THE NUMBER OF ESA FOR THE MULTI-ENHANCED SPATIAL ATTENTION (MESA) INSIDE THE PHTCB

We further experimented to check the most suitable number of ESA in a PHTCB. The module Triple Enhanced Spatial Attention (TESA) used in PHTCB comes from Multi-Enhanced Spatial Attention (MESA). MESA employs an iterative approach in which multiple ESA are cascaded together, and the output of one attention module is refined by subsequent modules, progressively enhancing the focus on critical features. The experiments included Dual Enhanced Spatial Attention (DESA), containing two ESA modules cascaded together; Triple Enhanced Spatial Attention (TESA), consisting of three ESA modules cascaded together; and Quadruple Enhanced Spatial Attention (QESA), consisting of

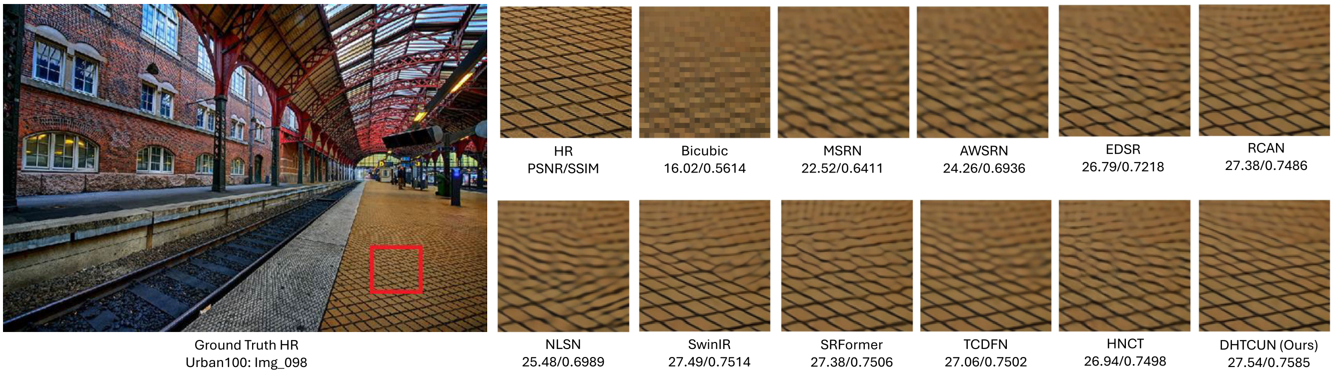

FIGURE 15. Img_098 quality improvement on a scale factor of ×4 from the Urban100 [52] dataset.

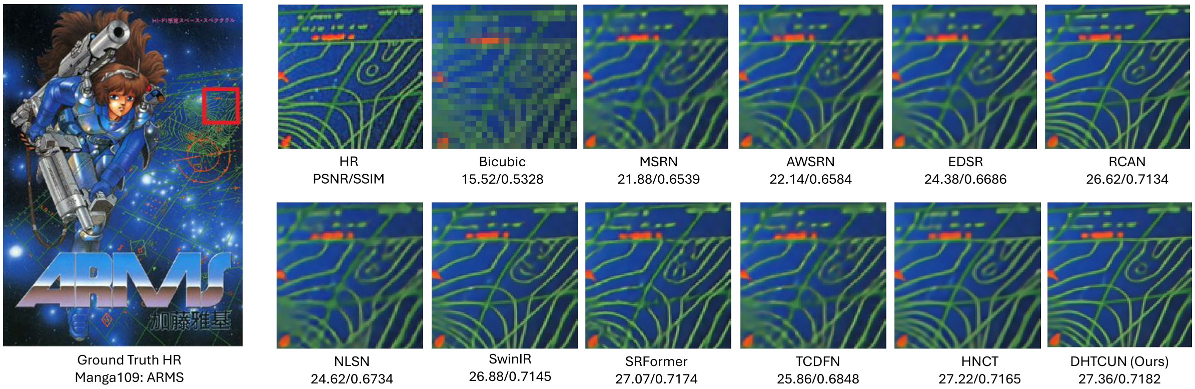

FIGURE 16. ARMS image quality improvement on a scale factor of ×4 from the Manga109 [53] dataset.

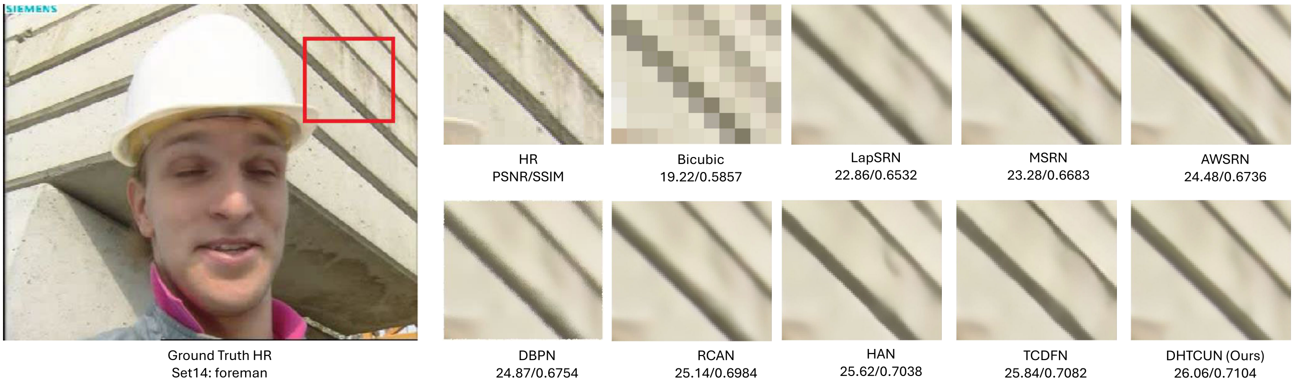

FIGURE 17. Foreman image quality improvement on a scale factor of ×8 from the Set14 [50] dataset.

four ESA modules cascaded together. Table 3 demonstrated that TESA can be used as MESA inside PHTCB as it gives better PSNR, SSIM, and Multi-Adds count than DESA or QESA on the Set5 dataset.

TABLE 4. Evaluation of MESA in the PHTCB for Set5 [49] dataset. Bold text with the color Red indicates the best quantitative value. The blue color and underline indicate the second best quantitative value.

Number of ESA in MESA	Multi-Adds	Set5 [50]	
		PSNR	SSIM
Dual Enhanced Spatial Attention (DESA)	**321G**	32.84	0.7808
Triple Enhanced Spatial Attention (TESA)	321.5G	**34.58**	**0.7848**
Quadruple Enhanced Spatial Attention (QESA)	326G	34.42	0.7826

4) ANALYTICAL COMPARISON WITH CONVENTIONAL DENOISING METHODS

Here, we present a comparative analysis of our proposed DHTCUN model on the Set5 [49] Dataset for a scale factore of ×2 with other classical denoising techniques, including Color image denoising via sparse 3D collaborative filtering

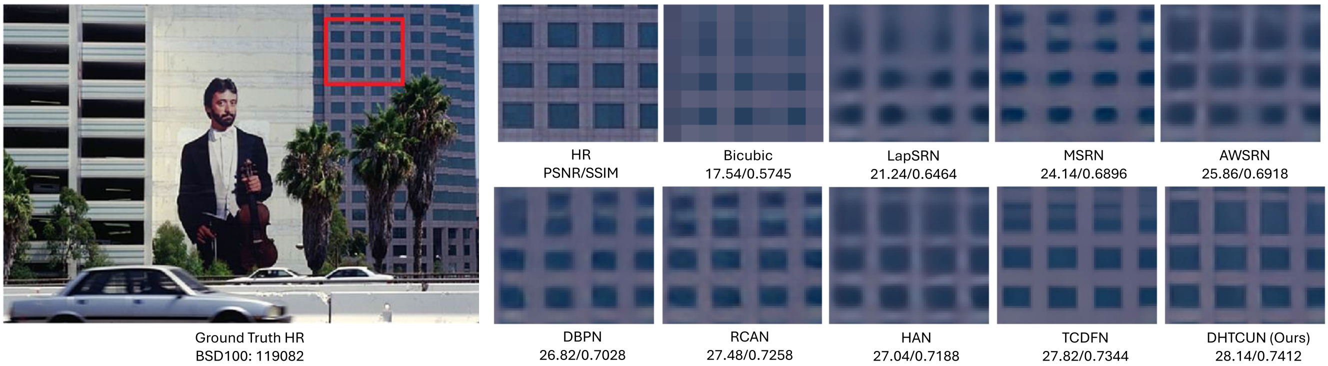

FIGURE 18. Img_119082 quality improvement on scale factor of ×8 from the BSD100 [51] dataset.

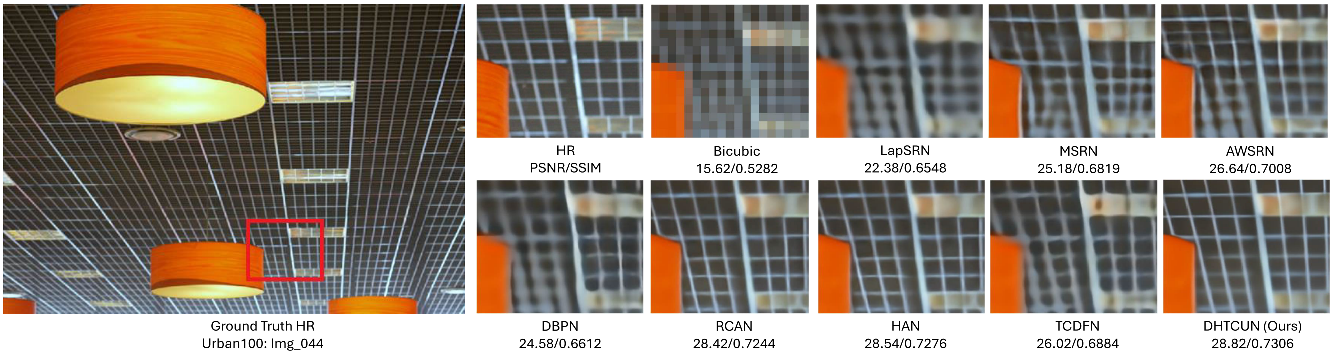

FIGURE 19. Img_044 quality improvement on the scale factor ×8 from the Urban100 [52] dataset.

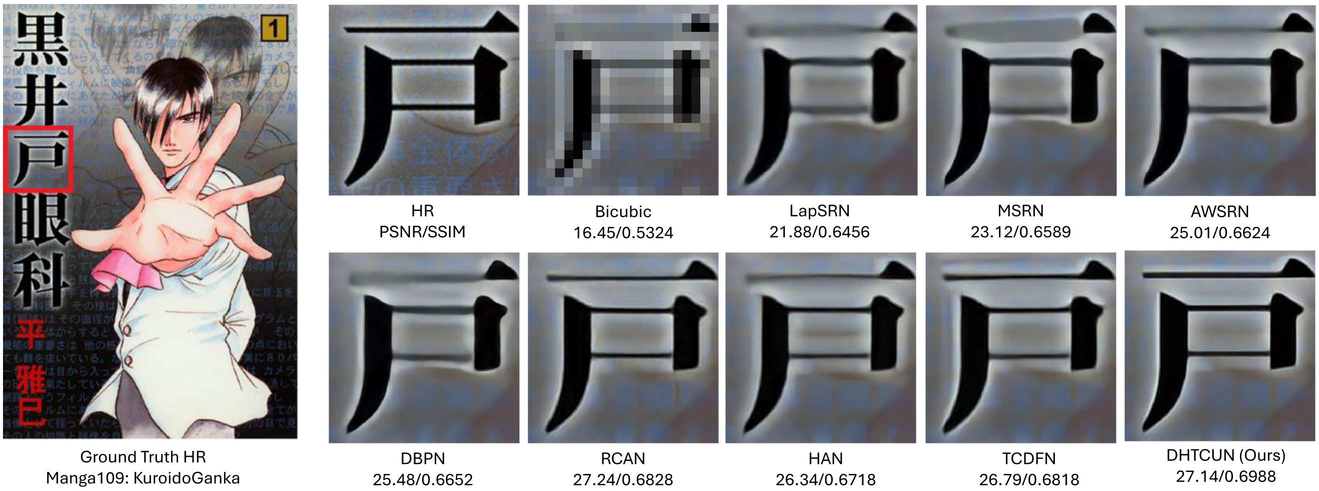

FIGURE 20. KuroidoGanka image quality improvement on the scale factor of ×8 from the Manga109 [53] dataset.

BM3D [55], Weighted Nuclear Norm Minimization with Application to Image Denoising (WNNM) [56], Denoising Convolutional Neural Network (DnCNN) [57], and Fast and Flexible Solution for CNN-Based Image Denoising (FFDNet) [58]. The Gaussian noise keeping noise level (σ) is used in Table 4 to compare performance in terms of PSNR for the values of $\sigma = 5$, $\sigma = 10$, and $\sigma = 15$. Our suggested DHTCUN model performs better at noise level $\sigma = 5$, as seen in Table 4, proving that it reduces noise amplification during the reconstruction. Thus, our model also shows better performance for noisy images.

TABLE 5. Assessment of image noise degradation performance on Set 5 [49] for scale factor ×2. Bold text with the color Red indicates the best quantitative value. The blue color and underline indicate the second-best quantitative value.

Methods / Noise Level	Factor	BM3D [55]	WNNM [56]	DnCNN [57]	FFDNet [58]	DHTCUN (Our)
$\sigma = 5$	×2	33.82	33.88	33.93	33.96	**33.98**
$\sigma = 10$	×2	32.58	32.64	32.70	32.75	32.88
$\sigma = 15$	×2	31.82	31.84	31.88	31.94	31.98

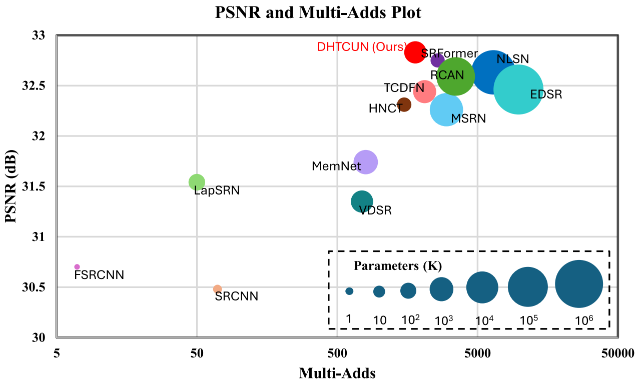

FIGURE 21. Metric assessment of PSNR (dB) versus Multi-adds where the circle size represents the number of parameters on scale factor ×4.

5) CONVERGENCE ANALYSIS

Training convergence refers to the process by which a neural network's learning algorithm iteratively adjusts the model parameters to minimize the loss function, thereby improving performance on a given task. Achieving training convergence is a critical aspect of developing effective deep-learning models.

In this subsection, we discuss performance evaluation as our model is being trained. Figure 23 displays the average PSNR (dB) for each epoch, illustrating how our model outperforms existing SR models, namely SRFormer [17], HNCT [20], and TCDFN [19], regarding training convergence. To ensure a fair comparison, we have used the same GPU for the training, and the hyperparameters remain unchanged. This analysis is computed for 200 training epochs with a ×4 enlargement factor on the DIV2K [48] dataset.

V. DISCUSSION

The proposed hybrid model that combines Convolutional Neural Networks (CNNs) and Transformers in a U-shaped architecture with skip connections has demonstrated notable improvements in single-image super-resolution (SISR) tasks. This novel approach leverages the complementary strengths of CNNs and Transformers to address several inherent challenges in SISR, such as computational expense, noise amplification, and the generation of artifacts like jagged patterns or blurring. By integrating the capabilities of both architectures, the model efficiently captures long-range dependencies and global context while enhancing the quality of noisy images without compromising computational efficiency.

Introducing the Parallel Hybrid Transformer CNN Block (PHTCB) within the proposed framework is a significant innovation. This block synergizes the strengths of CNNs and Transformers, enabling the model to capture intricate image details and reduce artifacts. The Transformer component excels at modeling long-range dependencies and global context, which is essential for reconstructing high-fidelity images from low-resolution counterparts. Concurrently, the CNN component enhances image quality, particularly in noisy scenarios, by mitigating noise amplification during the super-resolution process.

Furthermore, implementing of the Triple Enhanced Spatial Attention (TESA) block contributes to the model's performance by selectively focusing on relevant image regions while suppressing irrelevant or noisy areas. This selective attention mechanism ensures that the model prioritizes crucial features, thereby improving the super-resolved images' overall quality and perceptual realism. The TESA block's ability to enhance the model's attention to pertinent details significantly reduces common SISR artifacts.

The proposed U-shaped backbone with skip connections plays a crucial role in terms of computational efficiency. This design allows the model to extract features at different levels of abstraction without substantially increasing the computational burden. The skip connections facilitate the flow of information across layers, ensuring that the model maintains a balance between capturing detailed features and preserving computational efficiency. This design choice is particularly beneficial for processing large and complex images, where computational resources are often a limiting factor.

The study's comparative analysis and experimental results support the efficacy of the suggested strategy. The model consistently outperforms the most advanced techniques regarding perceptual quality and quantitative metrics. The suggested approach performs exceptionally well with various input image types, including extremely noisy and low-resolution images. This resilience demonstrates the model's adaptability and usefulness in real-world situations where computational effectiveness and image quality are crucial.

The proposed hybrid model represents a significant advancement in the field of SISR. The model addresses several critical challenges by effectively combining the strengths of CNNs and Transformers within a U-shaped architecture, achieving superior image quality with reduced computational demands. The innovations introduced in the PHTCB and TESA blocks further enhance the model's

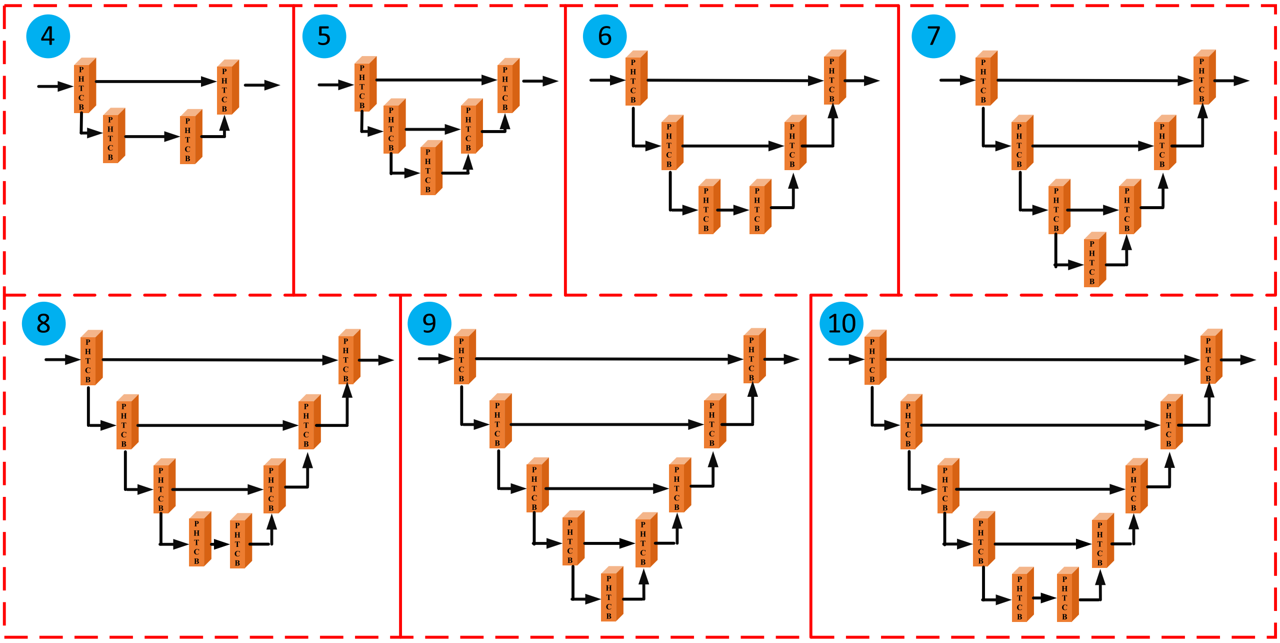

FIGURE 22. Combination of different number of PHTCBs in network as given in Table 3.

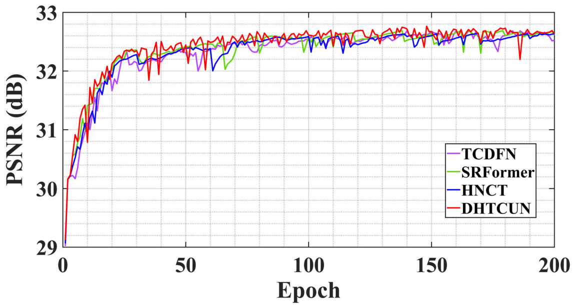

FIGURE 23. Convergence analysis of our model with the compared to another approaches.

capabilities, making it a promising solution for high-fidelity image super-resolution tasks.

VI. CONCLUSION AND FUTURE WORK

In this work, we introduced a novel approach for single-image super-resolution (SISR) that synergizes the strengths of Convolutional Neural Networks (CNNs) and Transformers within a U-shaped architecture with skip connections. This hybrid model effectively addresses key challenges in SISR, including computational expense, noise amplification, and artifact generation. By leveraging the Parallel Hybrid Transformer CNN Block (PHTCB) and the Triple Enhanced Spatial Attention (TESA) block, the model captures long-range dependencies and global context while enhancing the quality of noisy images. The U-shaped backbone with skip connections ensures efficient feature extraction at multiple levels of abstraction without increasing computational complexity. Our extensive experiments on benchmark datasets demonstrate that the proposed model consistently outperforms state-of-the-art quantitative metrics and perceptual quality methods. The robust performance across various image types, including those with extreme noise and low resolution, highlights our approach's practical applicability and versatility. This work represents a significant advancement in the field of SISR, providing a promising solution for high-fidelity image reconstruction in various real-world applications. Subsequent research endeavors may delve into more refinements and expansions of this methodology, conceivably integrating supplementary attention mechanisms or sophisticated training tactics to bolster efficacy and present our model for introducing real-time and video super-resolution applications in intricate scenarios.

REFERENCES

[1] H. Greenspan, "Super-resolution in medical imaging," *Comput. J.*, vol. 52, no. 1, pp. 43–63, Jan. 2009.

[2] S. Shakya, S. Kumar, and M. Goswami, "Deep learning algorithm for satellite imaging based cyclone detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 827–839, 2020.

[3] L. Zhang, H. Zhang, H. Shen, and P. Li, "A super-resolution reconstruction algorithm for surveillance images," *Signal Process.*, vol. 90, no. 3, pp. 848–859, Mar. 2010.

[4] K. Malczewski and R. Stasiski, "Super-resolution for multimedia, image, and video processing applications," in *Recent Advances in Multimedia Signal Processing and Communications*. Berlin, Germany: Springer, 2009, pp. 171–208.

[5] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[6] D. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 391–407.

[7] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[8] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5835–5843.

[9] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4539–4547.

[10] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.

[11] Y. Zhang, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[12] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3516–3525.

[13] J. Talreja, S. Aramvith, and T. Onoye, "DANS: Deep attention network for single image super-resolution," *IEEE Access*, vol. 11, pp. 84379–84397, 2023.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is 11worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–11.

[15] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 1833–1844.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[17] Y. Zhou, Z. Li, C.-L. Guo, L. Liu, M.-M. Cheng, and Q. Hou, "SRFormerV2: Taking a closer look at permuted self-attention for image super-resolution," 2023, *arXiv:2303.09735*.

[18] X. Zhang, H. Zeng, S. Guo, and L. Zhang, "Efficient long-range attention network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2022, pp. 649–667.

[19] Z. Zhou, G. Li, and G. Wang, "A hybrid of transformer and CNN for efficient single image super-resolution via multi-level distillation," *Displays*, vol. 76, Jan. 2023, Art. no. 102352.

[20] J. Fang, H. Lin, X. Chen, and K. Zeng, "A hybrid network of CNN and transformer for lightweight image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1102–1111.

[21] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12294–12305.

[22] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 370–378.

[23] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9446–9454.

[24] C. Zhang, L. Zhang, and J. Ye, "Generalization bounds for domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1–12.

[25] W. Zhang, Z. Tan, Q. Lv, J. Li, B. Zhu, and Y. Liu, "An efficient hybrid CNN-transformer approach for remote sensing super-resolution," *Remote Sens.*, vol. 16, no. 5, p. 880, Mar. 2024.

[26] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3862–3871.

[27] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.

[28] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2790–2798.

[29] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 517–532.

[30] C. Wang, Z. Li, and J. Shi, "Lightweight image super-resolution with adaptive weighted learning network," 2019, *arXiv:1904.02358*.

[31] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.

[32] J. Li, F. Fang, J. Li, K. Mei, and G. Zhang, "MDCN: Multi-scale dense cross network for image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2547–2561, Jul. 2021.

[33] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[34] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photorealistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.

[35] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. ECCV Workshop*, Sep. 2018, pp. 63–79.

[36] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4491–4500.

[37] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11057–11066.

[38] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5689–5698.

[39] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 191–207.

[40] W. Ruangsang, S. Aramvith, and T. Onoye, "Multi-FusNet of cross channel network for image super-resolution," *IEEE Access*, vol. 11, pp. 56287–56299, 2023.

[41] W. Muhammad, S. Aramvith, and T. Onoye, "SENext: Squeeze-and-ExcitationNext for single image super-resolution," *IEEE Access*, vol. 11, pp. 45989–46003, 2023.

[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3-19.

[43] X. Li, M. Xie, Y. Zhang, G. Ding, and W. Tong, "Dual attention convolutional network for action recognition," *IET Image Process.*, vol. 14, no. 6, pp. 1059–1065, May 2020.

[44] W. Ullah, T. Hussain, F. U. M. Ullah, M. Y. Lee, and S. W. Baik, "TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106173.

[45] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 6, pp. 1153–1160, Dec. 1981.

[46] D. Yang, Z. Li, Y. Xia, and Z. Chen, "Remote sensing image super-resolution: Challenges and approaches," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Jul. 2015, pp. 196–200.

[47] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2492–2501.

[48] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1122–1131.

[49] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L.-A. Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 135.1–135.10.

[50] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surf.*, Oslo, Norway, Jul. 2012, pp. 711–730.

[51] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Jul. 2001, pp. 416–423.

[52] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 5197–5206.

[53] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, ''Sketch-based Manga retrieval using manga109 dataset,'' *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21811–21838, Oct. 2017.

[54] W. Ai, X. Tu, S. Cheng, and M. Xie, ''Single image super-resolution via residual neuron attention networks,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 1586–1590.

[55] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, ''Color image denoising via sparse 3D collaborative filtering with grouping constraint in luminance-chrominance space,'' in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2007, p. 313.

[56] S. Gu, L. Zhang, W. Zuo, and X. Feng, ''Weighted nuclear norm minimization with application to image denoising,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2862–2869.

[57] J. Chen and F. Li, ''Denoising convolutional neural network with mask for salt and pepper noise,'' *IET Image Process.*, vol. 13, no. 13, pp. 2604–2613, Nov. 2019.

[58] K. Zhang, W. Zuo, and L. Zhang, ''FFDNet: Toward a fast and flexible solution for CNN-based image denoising,'' *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4608–4622, Sep. 2018.

JAGRATI TALREJA (Graduate Student Member, IEEE) received the B.Tech. degree in electronics and communication engineering networks from the Pranveer Singh Institute of Technology, Kanpur, Uttar Pradesh, India, in 2019. She is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, Chulalongkorn University, Bangkok, Thailand. Her research interests include electrical engineering, neural networks, and machine learning, specifically in deep learning image super-resolution.

SUPAVADEE ARAMVITH (Senior Member, IEEE) received the B.S. degree (Hons.) in computer science from Mahidol University, in 1993, and the M.S. and Ph.D. degrees in electrical engineering from the University of Washington, Seattle, WA, USA, in 1996 and 2001, respectively. In June 2001, she joined Chulalongkorn University, where she is currently an Associate Professor with the Department of Electrical Engineering, specializing in video technology. She has successfully advised 32 bachelor's, 27 master's, and ten Ph.D. graduates. She published more than 130 articles in international conference proceedings and journals with four international book chapters. She has rich project management experience as a project leader and a former Technical Committee Chair to the Thailand Government bodies in Telecommunications and ICT. She is very active in the international arena with leadership positions in the international network, such as the JICA Project for AUN/SEEDNet, and professional organizations, such as the IEEE, IEICE, APSIPA, and ITU.

TAKAO ONOYE (Senior Member, IEEE) received the B.E. and M.E. degrees in electronic engineering and the Dr.Eng. degree in information systems engineering from Osaka University, Osaka, Japan, in 1991, 1993, and 1997, respectively. He was an Associate Professor with the Department of Communications and Computer Engineering, Kyoto University, Kyoto, Japan. Since 2003, he has been a Professor with the Department of Information Systems Engineering, Osaka University. He has published more than 200 research papers in VLSI design and multimedia signal processing in reputed journals and proceedings of international conferences. His research interests include media-centric low-power architecture and its SoC implementation. He has served as a member of the CAS Society Board of Governors, since 2008. He is also a member of IEICE, IPSJ, and ITE-J.