A Cluster Tree Network for Image Super-Resolution

Qi Zhang, Weiqiang Xin, Shuai Wu, Qi Zhu, Qiya Song, and Shichao Zhang, *Senior Member, IEEE*

***Abstract*—Image super-resolution (SR) enhances the clarity and detail of images produced by consumer electronics, particularly in high-end devices, offering a superior visual experience. In this paper, a SR technique is proposed to enable effective and reliable processing in consumer electronics. Deep networks extract hierarchical structural feature through multiple layers for improving image SR results, however, their robustness may be limited in complex scenes. To address this, a cluster tree network for image SR (CTSRNet) was developed. The proposed tree architecture leverages hierarchical information by branching feature extraction into pathways that process varying, levels of abstraction and complexity, unlike monolithic convolutional neural networks (CNNs), where deep features may suppress shallower ones. Adaptability to diverse scenes is enhanced through conditional parameterized convolutions (Condconv), which learn salient features. By clustering image regions based on their properties, the networks learns inter-region to recover finer details. Experimental results demonstrate that CTSRNet surpasses popular SR methods. With reduced complexity and resource demand, the proposed approach is well-suited for edge computing in consumer electronics. The code is available at https://github.com/xwq325/CTSRNet.**

***Index Terms*—Tree network, adaptive convolution, cluster, image super-resolution.**

Received 2 March 2025; revised 3 May 2025 and 7 June 2025; accepted 16 July 2025. Date of publication 23 July 2025; date of current version 8 December 2025. This work was supported in part by the Natural Science Foundation of Shandong Province of China under Grant ZR2023QG074; in part by the State Key Laboratory for Novel Software Technology, Nanjing University under Grant KFKT2024B08; in part by the National Natural Science Foundation of China under Grant 62302378; in part by the Leading Talents in Gusu Innovation and Entrepreneurship under Grant ZXL2023170; and in part by the Basic Research Programs of Taicang 2024 under Grant TC2024JC32. *(Corresponding authors: Shuai Wu; Qiya Song.)*

Qi Zhang is with the School of Economics and Management, Harbin Institute of Technology at Weihai, Weihai 264209, China, and also with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210016, China (e-mail: hit_zq910057@163.com).

Weiqiang Xin is with the School of Software, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Yangtze River Delta Research Institute, Northwestern Polytechnical University, Taicang 215400, China (e-mail: xinweiqiang325@gmail.com).

Shuai Wu is with the School of Computer Science and Technology, Xidian University, Xi'an 710071, China (e-mail: shuaiwu9@gmail.com).

Qi Zhu is with the College of Artificial Intelligence, Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: zhuqi@nuaa.edu.cn).

Qiya Song is with the College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China (e-mail: sqyunb@hnu.edu.cn).

Shichao Zhang is with Guangxi Key Laboratory of Multisource Information Mining, Security College of Computer Science and Engineering, Guangxi Normal University, Guilin 541004, China (e-mail: zhangsc@mailbox.gxnu.edu.cn).

Digital Object Identifier 10.1109/TCE.2025.3591767

I. Introduction

THE RISE of the Internet of Things and the widespread adoption of smart devices have driven a shift toward edge computing in consumer electronics. This shift introduces challenges, such as limited computational resources, stringent power constraints, and the need for scalable, device-adaptive models. In this context, efficient and robust image super-resolution (SR) techniques are essential for improving visual quality, which is critical to user experience [1]. Specifically, single image SR (SISR) improves image clarity and detail, providing significant benefits for high-end devices. A SR approach is proposed in this study to meet the specific requirements of consumer electronics. When integrated with cloud service, the method enables efficient deployment across a broad range of mobile devices, including digital cameras, smartphones, and compact consumer cameras imaging systems.

SISR aims to reconstruct high-quality images by learning degradation models from low-resolution inputs. Traditional machine learning approaches, such as interpolation methods [2], and regression-based models [3], have been employed to infer high-resolution (HR) representations. While these methods are capable of image restoration, this adaptability is constrained by manually defined parameters and optimization strategies tailored to specific models.

Deep neural networks address the image restoration tasks, such as image denoising [4], image super-resolution [5] and image watermark removal [6], [7] to extract complex structural relationships through forward and backwards propagation. Dong et al. [5] first applied CNNs to SISR with SR convolutional neural network (SRCNN), a three-layer network that directly mapped low-resolution inputs to HR outputs. Although effective, the network's shallow depth constrained its performance. Subsequent studies established a strong link between network depth and SR accuracy, with deeper architectures yielding improved results [8]. Kim et al. [9] introduced VDSR, a very deep network that enhanced training stability in SISR. Kim et al. [10] further improved SISR performance by adopting a residual architecture with recursive supervision and skip connections to ease training difficulty. To extend this approach, Tai et al. [11] proposed deep recursive residual networks (DRRN), which integrated global and local residual learning to enhance image reconstruction quality.

Although these deep learning methods outperform traditional approaches in SISR tasks, they often involve significant computational overhand [12]. To mitigate this problem, an alternative strategy was introduced: processing low-resolution inputs directly through CNNs and deferring the upsampling operation later stages in the network [13]. For instance, Lai et al. [14] proposed LapSRN, a multi-scale

pyramid-structured network that progressively reconstructs HR images. Deformable and attentive network (DANet) [15] enhanced visual reconstruction by preserving low-frequency texture information through the integration of deformable convolutions and attention mechanisms. Similarly, Tian et al. [16] designed an asymmetric architecture that employs asymmetric convolutions and multi-level feature fusion to effectively enhance low-frequency components SISR.

With the advancement of consumer electronics, increasing attention has been directed towards the performance of SR techniques on mobile devices. Rai and Rajput [17] introduced a localized regularized facial SR method, based on artificial neural networks (ANNLR-FSR), that enhances facial image reconstruction under high-noise conditions by accurately estimating noise levels in low-resolution inputs. Wei et al. [18] proposed a transducer-based multilevel alignment network (TAVSR), designed to recover high and low-frequency information in compressed video. To address issues of bias and efficiency, Liao et al. [1] developed the minimax concave penalty-based SR (MCPSR).

Although the aforementioned deep learning methods effectively capture both high and low-frequency information for SISR, their performance often degrades in complex scenes due to limited adaptability. To address this limitation, a cluster tree network is proposed in this study for image SR. The approach utilized a tree-based architecture to fully exploit hierarchical information by modeling node relationships. Unlike conventional monolithic CNNs, the proposed structure processes features at varying levels of abstraction through separate branches, preventing deeper features from suppressing shallower ones. To enhance the robustness of the proposed image super-resolution model, conditional parameterized convolutions (CondConv) are introduced to dynamically extract salient features based on regional differenced within the image, thereby improving the model's adaptability. Additionally, a clustering mechanism is incorporated to learn the similarity among different image regions, enabling more accurate reconstruction of fine details.

The primary contributions of this study are outlined as follows:

(1) A novel tree-based architecture is introduced to effectively utilize hierarchical information for SISR based on node relationships.

(2) A conditionally parameterized convolution is employed to adaptively capture scene-specific features, thereby enhancing model robustness.

(3) A hybrid approach combining a clustering technique with CNN is utilized to capture regional similarities and extract fine-grained details for SISR.

The rest of this paper is structured as follows. Section II reviews the related literature; Section III details the proposed cluster tree network for SISR; Section IV presents the experimental evaluation; and Section V concludes the study.

II. Related Work

Deep learning-based SR methods have demonstrated strong performance and potential due to their expressive capabilities. However, the effectiveness of CNNs often degrades with increased depth, primarily due to issues, such as gradient vanishing or explosion. Moreover, deeper networks typically face difficulties in capturing informative interactions between shallow and deeper layers for SISR [19]. To address this limitation, enhancing both high- and low-frequency features within the network structure is an effective strategy to improve feature interaction across layers.

The enhancement of high-frequency structural features is typically carried out in two stages [20]. In the first stage, interpolation operations are used to upscale the low-resolution image, generating preliminary high-frequency information. In the second stage, these features are integrated through a designed CNN to extract richer structural representations. Building on this concept, Dong et al. [5] were the first to introduce an end-to-end CNN for the SR task. Subsequently, Wang et al. [21] proposed an asymmetric pyramid structure employing varying depths for feature extraction and reconstruction across different scales to enable efficient multi-scale progressive SR. To simplify training, Mao et al. [22] utilized a symmetric encoder-decoder architecture to enhance image recovery abilities. Iterative up- and down-sampling has also proven effective in progressively refining SR features. Haris et al. [23] introduced a deep back-projection network that incorporates alternating up- and down-sampling layers, enabling an error feedback mechanism to iteratively estimate reconstruction errors and guide the network toward improved SR performance.

The enhancement of low-frequency structural features is typically achieved by feeding low-resolution images directly into a CNN to extract low-frequency information, followed by upsampling within a deep network to reconstruct high-quality outputs [24]. For example, multi-path residual networks [25] improve representational capacity by employing local feature fusion, enabling the extraction of richer low-frequency features for SISR. This not only strengthens the robustness of the extracted information but also reduces computational overhead. To further minimize model complexity, the cascaded residual network (CARN) [8] incorporates multi-path information flow and a recursive structure across layers, achieving competitive SR performance with fewer parameters and reduced computation. Li et al. [26] enhanced feature representation learning through an iterative up-down sampling feedback module, supported by densely connected skip layers. Building on this, Zhang et al. [27] improved SISR reconstruction by integrating local residual learning with sparsely connected local layers. All-in-one image restoration models have demonstrated the ability to address diverse degradations by using specific prompts to guide restoration. However, universal architecture often suffer from unpredictability and a lack of control. To mitigate these issues, Conde et al. [28] proposed a novel method that leverages human-written instructions to guide the image restoration process more effectively.

Although CNN-based SR models and their various training strategies have achieved notable improvements in the quality and accuracy of HR image reconstruction, challenges remain when addressing SR tasks in complex scenes. To address this, a tree-based framework is introduced in this work to guide the

CNN, thereby enhancing the robustness of the resulting SR model.

III. Proposed Method

A. Network Architecture

Image SR methods commonly often employ deep CNNs to capture complex image features, however, this typically results in increased model parameters and computational overhead [29]. Additionally, limitations in feature extraction and fusion may hinder the network's ability to effectively capture diverse image details. To address these challenges, a cluster tree network for image SR (CTSRNet) is proposed (Fig. 1). In CTSRNet, the root node first maps the low-resolution image to a high-dimensional feature space, generating a shallow feature map. This map is then processed through four parallel branches (B1-B4), forming the core of the tree-structured architecture. These branches operate in parallel with varying depths and complexities (base, middle, right subtree layers), thereby enabling the capture of a richer and more diverse hierarchy of features than conventional sequential architectures. Each branch is configured with a distinct network structure and depth. The base layer refines the initial shallow features to preserve essential image details. The middle subtree layer increases the network depth to capture more complex and diverse feature representations. The right subtree layer then integrates and enhances these features, improving the model's expressive capacity. Finally, the output from the fourth branch is passed through an upsampling module to generate the HR image. The feature extraction and fusion strategies for the four branches of CTSRNet are defined by Eqs. (1)-(5):

$$O_{HR} = CTSRNet(I_{LR}) = L_{UP}\left(F_{B_4}\right) \tag{1}$$

$$F_{B_4} = L_{RL,B_4}\left(L_{BL,B_4+ML,B_4}(F_0) \times F_{B_3} + F_0\right) \tag{2}$$

$$F_{B_3} = L_{RL,B_3}\left(L_{BL,B_3+ML,B_3}(F_0) \times F_{B_2} + F_0\right) \tag{3}$$

$$F_{B_2} = L_{RL,B_2}\left(L_{BL,B_2}(F_0) \times F_{B_1} + F_0\right) \tag{4}$$

$$F_{B_1} = L_{BL,B_1}(F_0) \tag{5}$$

where O_{HR} represents the HR output image, I_{LR} denotes the low-resolution input, $CTSRNet()$ refers to the cluster tree network for image SR, F_0 signifies the shallow feature map produced by the initial convolutional layer, L_{UP} is the upsampling module, F_{B_i} denotes the output of the i-th branch, L_{RL,B_i} represents the right subtree layer of the i-th branch($i = 2, 3, 4$), L_{BL,B_i+ML,B_i} indicates the base layer and the middle subtree layer of the i-th branch($i = 3, 4$), and L_{BL,B_i} is the base layer of the i-th branch($i = 1, 2$). The branches corresponding to Eqs. (2)-(5) are trained in parallel. Additionally, during exection, Eqs. (2)-(4) utilize the output of the subsequent branch in a channel-concatenated manner to enable progressive feature enrichment.

B. Loss Function

The L_1 norm is commonly employed as a loss function in SR and other low-level image processing tasks due to its robustness to outliers and its effectiveness in preserving image details [30]. Accordingly, it is adopted in this wore to construct the loss function for training the CTSRNet model. The training set is $\{I^i_{LR}, I^i_{HR}\}_{i=1}^N$, where I^i_{LR} and I^i_{HR} denote the i-th low-resolution and high-resolution training image segments, respectively, and N indicates the total number of image segments. The final loss function of CTSRNet is expressed as Eq. (6):

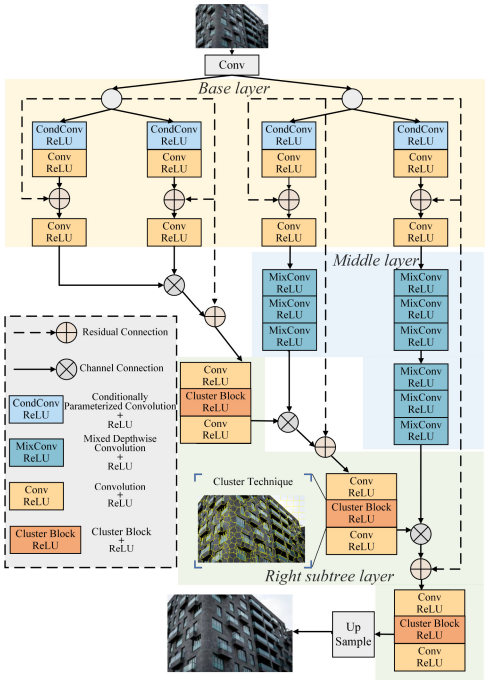

Fig. 1. Network architecture of CTSRNet.

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left\| CTSRNet\left(I^i_{LR}\right) - I^i_{HR}\right\|_1 \tag{6}$$

where L denote the L_1 loss function, and θ represents the parameter set of the CTSRNet model. The loss function is optimized using the Adam optimizer.

C. Base Layer Block

The extraction of detailed features is critical for image SR tasks. In CTSRNet, While the initial 3×3 convolution layer generates shallow features containing basic structural information, these features are often coarse and insufficient for representing fine details. Therefore, the base layer is introduced to refine these preliminary features and enhance the network's ability to capture subtle and complex patterns. The base layer comprises one CondConv [31] layer followed by two standard convolutional layers, each accompanied by a ReLU activation function. A residual connection is also induced to facilitate feature preservation and gradient flow. Condconv enhances representational capacity by dynamically generating a customized convolution kernel for each input. This is achieved by maintaining a set of K learnable expert kernels and employing a lightweight routing function, typically using global average pooling and fully connected layers, to predict K input-specific weights. These weights are then used to form a weighted combination of the expert kernels, producing an adaptive kernel tailored to each sample. This

mechanism improves model expressiveness while keeping computational cost low. The structure and precessing flow of the base layer are expressed as Eqs. (7) and (8):

$$F_{BL,Bi} = L_{BL}(F_0) \tag{7}$$

$$L_{BL} = L_{CR}(L_{CR}(L_{CondConv+ReLU}(x))) + x \tag{8}$$

where F_0 represents the shallow feature map from the input, L_{BL} indicates the composite function of the base layer, $F_{BL,Bi}$ signifies the base layer output of the i-th branch($i = 1, 2, 3, 4$), $L_{CondConv+ReLU}$ and L_{CR} refer to the corresponding Condconv layer and the standard convolutional layer followed by the ReLU activation function, respectively, and x is the input to the base layer. Eq (7) describes the application of the base layer before the point of significant structural divergence within each of the four conceptual branches. The parameters across the four branches are not shared, enabling parallel processing pathways with distinct depths and complexities.

D. Middle Subtree Layer Block

To enhance feature representation in deeper branches, middle subtree layers are integrated into CTSRNet. For branches with indices 1 and 2, which primarily focus on lower-level or simpler features, additional refinement through middle layers is deemed unnecessary. In contrast, branches with indices greater than 2 are responsible for processing more complex, high-level features. To support this, middle subtree layers are introduced in these branches to enrich feature representation and facilitate the capture of multiscale information in higher dimensions. Each middle subtree layer comprises three mixed depthwise convolutions (MixConv) blocks. Mixconv combines depthwise separable convolution and multi-scale convolution. Depthwise convolution first extracts spatial features independently for each input channel, followed by pointwise convolution to fuse inter-channel information, thereby approximating the effectiveness of standard convolution with significantly reduced computational cost. Additionally, convolution kernels of varying sizes are applied in parallel to capture both local and global features smaller kernels focus on fine details, while larger kernels emphasize broader context. Through the fusion of multi-scale features, MixConv enables efficient and comprehensive feature extraction. This process is formulated in Eqs. (9)-(11):

$$F_{ML,B3} = L_{ML}\left(F_{BL,B3}\right) \tag{9}$$

$$F_{ML,B4} = L_{ML}\left(L_{ML}(F_{BL,B4})\right) \tag{10}$$

$$L_{ML} = L_{Mix+ReLU}(L_{Mix+ReLU}(L_{Mix+ReLU}(x))) \tag{11}$$

where L_{ML} represents the composite function of the middle subtree layer, $F_{ML,Bi}$ indicates the output of the middle subtree layer in the i-th branch, and $L_{Mix+ReLU}$ refers to the MixConv applied within the layer followed by the ReLU activation function.

The total number of output channels is distributed across the parallel branches using different kernel sizes, with the allocation designed to ensure that each branch contributes approximately the same number of parameters. As a result, branches with larger kernels are assigned fewer output channels, while those with smaller kernels receive more. This strategy balances representational capacity and computational efficiency by constraining the dimensionality of broader contextual features and enhancing the representation of fine details. For images with diverse structures, such as those containing both intricate textures and large homogeneous regions, this design enables the network to adaptively exploit the most relevant scale information. Smaller kernels are more effective at capturing local details, whereas larger kernels contribute to modeling broader spatial context. During training, the network learns to optimally leverage the concatenated multi-scale features across different regions of an image. Although this approach efficiently facilitates multi-scale feature extraction within a block, its effectiveness may depend on the suitability of the selected kernel sizes to the dominant structural characteristics of the input image.

E. Right Subtree Layer Block

To effectively fuse and optimize multi-level information, the right subtree layer is incorporated into all branches with indexes greater than 1. This later comprises an initial standard convolution layer, followed by N Cluster Blocks, and a final standard convolution layer. Initially, the input feature maps are processed through a standard convolution and ReLU activation to perform spatial refinement and normalization. Recognizing that conventional SR models typically operate on small, disjointed image patches, often resulting in fragmented detail reconstruction, a superpixel token interaction block [32] is introduced. This block enhances the capture of structural information by segmenting feature maps into superpixels, performing local clustering, and applying a superpixel-based cross-attention mechanism. This approach improves the alignment between pixel-level data and structural features, enabling more accurate recovery of complex image details, particularly under high magnification factors. The right subtree layer further strengthens feature representation by utilizing semantically meaningful superpixel regions rather than raw patches, enhancing the extraction of both local and global features, particularly for complex structures often degraded at higher zoom factors ($\times 3$, $\times 4$). Residual connections are used to incorporate shallow features, while channel concatenation is employed to integrate broader contextual information from other branches. The processed features are finally refined through a standard convolution layer, improving both detail preservation and global coherence in the output feature map. The corresponding operations are detailed in Eqs. (12)-(15):

$$F_{RL,B2} = L_{RL}\left(F_{BL,B1} \times F_{BL,B2} + F_0\right) \tag{12}$$

$$F_{RL,B3} = L_{RL}\left(F_{RL,B2} \times F_{ML,B3} + F_0\right) \tag{13}$$

$$F_{RL,B4} = L_{RL}\left(F_{RL,B3} \times F_{ML,B4} + F_0\right) \tag{14}$$

$$L_{RL} = L_{CR}(L_{Cluster}(L_{CR}(x))) \tag{15}$$

where L_{RL} represents the composite function at the right subtree layer, $F_{RL,Bi}$ indicates the output of the right subtree layer in the i-th branch, L_{CR} refer to the standard convolutional layer followed by the ReLU activation function and $L_{Cluster}$ refers to the superpixel token interaction block.

F. Cluster Technique

Traditional SR models predominantly focus on local feature extraction, often failing to capture comprehensive pixel similarities in complex scenes, which can result in the loss of structural information. To address these limitations, CTSRNet incorporates a clustering block [32] to enhance the fusion of local and global features through superpixel-based aggregation and interaction. This mechanism is integrated within the right subtree layer block, where clustering serves as a structural framework rather than relying on conventional patch-based operations. The clustering process segments the input feature map into multiple superpixels $S = \{S_1, S_2, \ldots, S_k\}$, enabling the grouping of similar pixels. This approach facilitates not only the recognition of local similarities but also mitigates artifacts commonly introduced by traditional patch segmentation. Within each superpixel, an intra-superpixel attention mechanism (ISPA) is applied to reinforce interactions among similar pixels and enhance fine-grained details. Concurrently, the superpixel cross-attention (SPCA) module enables long-range feature communication across superpixels, allowing structurally relevant information from distant regions to be effectively integrated. Finally, the aggregated features are refined through a standard convolutional layer, ensuring both the preservation of local detail and the consistency of global structure. The process is mathematically represented by Eqs. (16)-(18):

$$F_{ISPA} = Attention(F_{S_i}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{16}$$

$$F_{SPCA} = \sum_{i=1}^{k} Attention(F_{S_i}, F_S) \text{ for } S_i \in S \tag{17}$$

$$F_{cluster} = Conv(F_{ISPA} + F_{SPCA}) \tag{18}$$

IV. Experimental Analysis and Results

A. Datasets

The DIV2K dataset [33] is utilized for the development and training of the proposed SR framework. It comprised 800 training images, 100 validation images, and 100 test images. Specifically, the first 800 images are employed for training, while images 801-810 are used for validation.

To ensure an objective evaluation of SR performance, four widely recognized benchmark datasets [34] are employed: Set 5, containing 5 natural images [35]; Set14, with 14 natural images [36]; BSD100 (B100), comprising 100 natural images [37]; and Urban100 (U100), including 100 urban images [38]. These datasets support scale factors of $\times 2$, $\times 3$, and $\times 4$, with all images stored in PNG format.

B. Experimental Settings

For the model configuration, shallow feature extraction is initiated using a standard convolution layer with 3×3 kernel. In each branch, the base layer consists of one CondConv followed by two standard 3×3 convolutions, each succeeded by a ReLU activation. The right subtree layer in each branch includes a standard 3×3 convolution, a superpixel token interaction block with $N = 1$, and an additional 3×3 convolution. The middle subtree layer comprises three stacked MixConv, utilizing an exponentially decreasing structure with a combination of 3×3, 5×5, and 7×7 kernels. Each convolutional layer is followed by a ReLU activation. For upsampling, the efficient sub-pixel convolution network [39] is employed to convert coarse-resolution features into HR RGB outputs.

For the hyperparameter configuration, a batch size of 16 is adopted, and, the image block size is set to 96. Data augmentation is performed through random horizontal and vertical flips, as well as 90-degree rotations. Additionally, the image mean subtraction technique introduced in EDSR [40] is applied both before model input and after output. Training is conducted over 350 rounds, with each round comprising 1,000 iterations. The Adam optimizer is employed with an initial learning rate of 1×10^{-4}, which is reduced by half every 200 rounds. SR performance is evaluated using peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). Ablation studies are also conducted over 350 rounds using the same iteration schedule.

For the experimental environment, the CTSRNet code is executed on a system running Ubuntu 20.04, equipped with 128 GB of RAM and an NVIDIA RTX 3090 GPU. CUDA version 11.6 is utilized for training the SR model.

TABLE I
Ablation Study Results of CTSRNet on B100 (PSNR/SSIM)

Scale	CTSRNet (BaseLine)	CTSRNet (+CondConv)	CTSRNet (+MixConv)	CTSRNet (+Cluster Block)
×2	32.03/0.8978	32.05/0.8981	32.08/0.8982	32.15/0.8992
×3	28.98/0.8029	29.01/0.8032	29.03/0.8031	29.09/0.8051
×4	27.48/0.7334	27.50/0.7336	27.52/0.7341	27.57/0.7362

C. Ablation Study

The proposed clustered tree-like network, CTSRNet, is designed to enhance the robustness of image SR, particularly in complex scenes. Efficient reconstruction is achieved through hierarchical feature extraction and fusion enabled by its branched architecture. The core of the tree structure lies in the distinct configurations of the base layer, middle subtree layer, and right subtree layer within each branch, which facilitate the effective capture and integration of fine-grained image details. While traditional CNNs often improve SR performance by increasing network depth, such approaches may fail to represent the full image and lack complementary information. While increasing network width can extract complementary information, it significantly increases model complexity and struggles to exploit the differentiated representations across structures. In contrast, the tree-like design of CTSRNet enhances interaction among all nodes, promoting the integration of width, depth, and heterogeneous fractures, and thereby ensuring comprehensive and efficient feature representation.

This subsection presents an ablation study performed on the B100 dataset to evaluate the effectiveness of the key components (Table I). Additional ablation results on the U100 dataset are provided in Table II, where each module demonstrates the anticipated performance improvement; thus, further elaboration is omitted. It is worth noting that baseline model outperforms most classical image SR methods on both B100

TABLE II
ABLATION STUDY RESULTS OF CTSRNET ON U100 (PSNR/SSIM)

Scale	CTSRNet (BaseLine)	CTSRNet (+CondConv)	CTSRNet (+MixConv)	CTSRNet (+Cluster Block)
×2	31.63/0.9236	31.67/0.9240	31.78/0.9248	32.11/0.9282
×3	27.82/0.8454	27.85/0.8457	27.92/0.8469	28.20/0.8533
×4	25.81/0.7771	25.85/0.7774	25.91/0.7791	26.15/0.7875

and U100, highlighting the effectiveness of the proposed tree-based structure.

Base layer block: In image SR tasks, varying feature distributions across input images can challenge a model's generalization ability. Traditional approaches typically rely on fixed convolutional kernels for feature extraction, which constrains the model's ability to handle complex and diverse image samples, limiting adaptability and performance. To address this issue, CondConv is employed in the base layer to replace standard convolution. This technique enables the dynamic generation of convolution kernels based the specific characteristics of the input image, thereby overcoming the limitations of fixed kernels. As a result, the model is better equipped to capture fine image details and textures, enhancing its adaptability and overall performance.

The effectiveness of incorporating CondConv in the base layer is validated through a comparison between by "CTSRNet (baseline)" and "CTSRNet (+CondConv)" (Table I), where red and blue markings denotes the best and second-best results for SISR, respectively. "CTSRNet (baseline)" refers to the version without CondConv, while "CTSRNet (+CondConv)" integrates this module into the base layer. For a scaling factor of ×2, slight improvements in both PSNR and SSIM are observed with "CTSRNet (+CondConv)", indicating that the module effectively refines shallow features at the root node (Table I). The benefits become more pronounced at a scaling factor of ×3, confirming that the module can dynamically adapt convolution operations based on input characteristics, thereby enhancing feature extraction. At a scaling factor of ×4, improvements in performance metrics are observed despite the increased reconstruction difficulty. This confirms that CondConv adaptively refines shallow features at the root node across all scales, offering a robust foundation for subsequent processing. Its ability to dynamically adjust convolution operations based on input characteristics consistently enhances the network's feature extraction capacity, even under more demanding conditions.

Middle subtree layer block: While the base layer enhances shallow features, it is insufficient on its own to capture the full range of structure and fine details required for complex image SR tasks. Therefore, the middle subtree layer plays a critical role in the network design. To strengthen multi-scale feature representation, MixConv is introduced in this layer. Small kernels are employed to capture fine textures and edges, whereas large kernels extract broader contextual information. This parallel application of multi-scale kernels allows simultaneous attention to both local details and global context within a single layer, effectively mitigating the information loss typically associated with sing-scale convolution.

The effectiveness of the MixConv in the middle subtree layer is evaluated by comparing "CTSRNet (+CondConv)" and "CTSRNet (+MixConv)" in Table I. "CTSRNet (+CondConv)" excludes MixConv, while "CTSRNet (+MixConv)" incorporates it into the middle subtree layer. At a scaling factor of ×2, improvements in PSNR and SSIM are observed with MixConv, attributed to its integration of convolution operations with varying receptive fields. As the scaling factor increases, such as ×3 and ×4, leading to greater detail loss and reconstruction complexity, MixConv continues to provide measurable benefits, particularly in SSIM, by effectively capturing and integrating spatial information. This highlights the strength of MixConv's multi-scale kernel design, which enables simultaneous extraction of fine-grained local features and broader contextual information. Such capability allows the network to adapt to structural variations across scales, enhancing its ability to process complex image context. Consequently, richer and more accurate multilevel feature representations are generated, supporting more effective feature fusion at the right subtree level.

Right subtree layer block: The right subtree layer is a critical component of the proposed CTSRNet, designed to enhance SR reconstruction by deeply integrating multi-scale features. Its primary objective is ensure the effective fusion of global structural patterns and local textures. To achieve this, a superpixel token interaction block is introduced. This module segments the image into superpixel regions with similar features, enabling a tokenization process that reduces computational overhead while preserving essential structural information. These superpixel tokens then undergo interaction for information fusion, strengthening the network's capacity to model both global context and local detail. By facilitating long-range dependencies and complex pixel relationships, this approach addresses the limitations of traditional convolution layers in capturing broad structural information.

The effectiveness of the right subtree layer in modeling structural information is validated by comparing "CTSRNet (+MixConv)" and "CTSRNet (+Cluster Block)" in Table I. "CTSRNet (+MixConv)" excludes the superpixel token interaction module, while "CTSRNet (+Cluster Block)" incorporates it within the right subtree layer. Across both low and high scaling factors, "CTSRNet (+Cluster Block)" consistently outperforms "CTSRNet (+MixConv)" in terms of PSNR and SSIM. These results confirm that the superpixel token interaction block significantly enhances feature interconnectivity by promoting information exchange between global structures and local pixel regions. This mechanism not only strengthens the network's comprehension of the overall image structure, but also improves the coherence and expressiveness of the reconstructed images. By effectively capturing relevant pixel dependencies, its leads to visually superior SR outputs.

Cluster technique: The integration of clustering techniques in image SR reconstruction is designed to effectively leverage both local structural details and global contextual information. Traditional methods, which rely on fixed convolutional kernels, often lack adaptability to varying image inputs. In contrast, the clustering technique enhances feature extraction by segmenting the input feature map into multiple superpixels, grouping similar pixels, and capturing meaningful local relationship. Within each superpixel, the ISPA is employed to dynamically highlight important internal features, thereby

TABLE III
AVERAGE PSNR (DB) AND SSIM FOR DIFFERENT METHODS WITH THREE SCALING FACTORS ON THE SET5 DATASET

Datasets	Methods	×2	×3	×4
Set5	Bicubic [41]	33.66/0.9299	30.39/0.8682	28.42/0.8104
	SRCNN [5]	36.66/0.9542	32.75/0.9090	30.48/0.8628
	VDSR [9]	37.53/0.9587	33.66/0.9213	31.35/0.8838
	DRRN [11]	37.74/0.9591	34.03/0.9244	31.68/0.8888
	FSRCNN [12]	37.00/0.9558	33.16/0.9140	30.71/0.8657
	CARN-M [42]	37.53/0.9583	33.99/0.9236	31.92/0.8903
	IDN [43]	37.83/0.9600	34.11/0.9253	31.82/0.8903
	A+ [44]	36.54/0.9544	32.58/0.9088	30.28/0.8603
	JOR [45]	36.58/0.9543	32.55/0.9067	30.19/0.8563
	RFL [46]	36.54/0.9537	32.43/0.9057	30.14/0.8548
	SelfEx [38]	36.49/0.9537	32.58/0.9093	30.31/0.8619
	CSCN [47]	36.93/0.9552	33.10/0.9144	30.86/0.8732
	RED [22]	37.56/0.9595	33.70/0.9222	31.33/0.8847
	DnCNN [48]	37.58/0.9590	33.75/0.9222	31.40/0.8845
	TNRD [49]	36.86/0.9556	33.18/0.9152	30.85/0.8732
	FDSR [50]	37.40/0.9513	33.68/0.9096	31.28/0.8658
	RCN [51]	37.17/0.9583	33.45/0.9175	31.11/0.8736
	DRCN [10]	37.63/0.9588	33.82/0.9226	31.53/0.8854
	CNF [52]	37.66/0.9590	33.74/0.9226	31.55/0.8856
	LapSRN [14]	37.52/0.9590	-	31.54/0.8850
	WaveResNet [53]	37.57/0.9586	33.86/0.9228	31.52/0.8864
	CPCA [54]	34.99/0.9469	31.09/0.8975	28.67/0.8434
	NDRCN [55]	37.73/0.9596	33.90/0.9235	31.50/0.8859
	MemNet [56]	37.78/0.9597	34.09/0.9248	31.75/0.8893
	LESRCNN [57]	37.65/0.9586	33.93/0.9231	31.88/0.8903
	ScSR [58]	35.78/0.9485	31.34/0.8869	29.07/0.8263
	DSRCNN [59]	37.73/0.9588	34.17/0.9247	31.89/0.8909
	DAN [60]	37.34/0.9526	34.04/0.9199	31.89/0.8864
	DCLS [61]	37.63/0.9554	34.21/0.9218	32.12/0.8890
	HDSRNet [62]	37.95/0.9604	34.31/0.9263	32.15/0.8927
	CTSRNet(Ours)	37.88/0.9602	34.35/0.9268	32.22/0.8950

TABLE IV
AVERAGE PSNR (DB) AND SSIM FOR DIFFERENT METHODS WITH THREE SCALING FACTORS ON THE SET14 DATASET

Datasets	Methods	×2	×3	×4
Set14	Bicubic [41]	30.24/0.8688	27.55/0.7742	26.00/0.7027
	SRCNN [5]	32.42/0.9063	29.28/0.8209	27.49/0.7503
	VDSR [9]	33.03/0.9124	29.77/0.8314	28.01/0.7674
	DRRN [11]	33.23/0.9136	29.96/0.8349	28.21/0.7720
	FSRCNN [12]	32.63/0.9088	29.43/0.8242	27.59/0.7535
	CARN-M [42]	33.26/0.9141	30.08/0.8367	28.42/0.7762
	IDN [43]	33.30/0.9148	29.99/0.8354	28.25/0.7730
	A+ [44]	32.28/0.9056	29.13/0.8188	27.32/0.7491
	JOR [45]	32.38/0.9063	29.19/0.8204	27.27/0.7479
	RFL [46]	32.26/0.9040	29.05/0.8164	27.24/0.7451
	SelfEx [38]	32.22/0.9034	29.16/0.8196	27.40/0.7518
	CSCN [47]	32.56/0.9074	29.41/0.8238	27.64/0.7578
	RED [22]	32.81/0.9135	29.50/0.8334	27.72/0.7698
	DnCNN [48]	33.03/0.9128	29.81/0.8321	28.04/0.7672
	TNRD [49]	32.51/0.9069	29.43/0.8232	27.66/0.7563
	FDSR [50]	33.00/0.9042	29.61/0.8179	27.86/0.7500
	RCN [51]	32.77/0.9109	29.63/0.8269	27.79/0.7594
	DRCN [10]	33.04/0.9118	29.76/0.8311	28.02/0.7670
	CNF [52]	33.38/0.9136	29.90/0.8322	28.15/0.7680
	LapSRN [14]	33.08/0.9130	-	28.19/0.7720
	WaveResNet [53]	33.09/0.9129	29.88/0.8331	28.11/0.7699
	CPCA [54]	31.04/0.8951	27.89/0.8038	26.10/0.7296
	NDRCN [55]	33.20/0.9141	29.88/0.8333	28.10/0.7697
	MemNet [56]	33.28/0.9142	30.00/0.8350	28.26/0.7723
	LESRCNN [57]	33.32/0.9148	30.12/0.8380	28.44/0.7772
	ScSR [58]	31.64/0.8940	28.19/0.7977	26.40/0.7218
	DSRCNN [59]	33.43/0.9157	30.24/0.8402	28.46/0.7796
	DAN [60]	33.08/0.9041	30.09/0.8287	28.42/0.7687
	HDSRNet [62]	33.52/0.9170	30.27/0.8412	28.56/0.7799
	CTSRNet(Ours)	33.60/0.9184	30.34/0.8416	28.61/0.7812

optimizing information flow and enhancing local feature interaction. Additionally, the SPCA mechanism facilitates the exchange of information between distant superpixels, reinforcing global feature correlations.

The effectiveness of the clustering technique is demonstrated by comparing "CTSRNet (+BASELINE)" and "CTSRNet (+Cluster Block)" in Table I. At a scaling factor of ×2, improvements in the PSNR and SSIM of "CTSRNet (+Cluster Block)" confirm enhanced detail recovery. With a scaling factor of ×3, performance gains become more obvious, as the PSNR increases from 28.98 to 29.09, indicating improved adaptability to diverse image features. Even at is ×4, where reconstruction becomes more challenging, a PSNR of 27.57 is achieved, validating the robustness and efficiency of the clustering technique across varying scales.

Overall, the ablation study confirms that each proposed component incrementally enhances the network's performance. Among them, the cluster block yields the most substantial improvements across all scaling factors, proving especially effective in high-magnification scenarios (e.g., ×4) by strengthening structure-aware feature fusion and relational modeling.

D. Experimental Results

Both quantitative and qualitative analyses were conducted to evaluate the SR performance of CTSRNet. Quantitative evaluation was performed by comparing the PSNR and SSIM of CTSRNet against those of classical SR methods on the Set5, Set14, B100, and Urban100 datasets. Qualitative analysis assessed the visual quality of the reconstructed images to further validate the model's effectiveness.

CTSRNet demonstrates strong performance on the Set5 dataset across scaling factors ×2, ×3, and ×4, consistently achieving the best or second best results in terms of PSNR and SSIM (Table III). The red and blue indicators denote the top and second -best scores, respectively. For instance, at a ×2 scale, CTSRNet attains a PSNR of 37.88 dB, surpassing classical methods, such as VDSR [9], FSRCNN [12], and LapSRN [14]. At ×3 and ×4 scales, it achieves PSNR values of 34.35 dB and 32.22 dB, outperforming even recent methods, including heterogeneous dynamic convolutional network (HDSRNet) [62]. In terms of SSIM, CTSRNet also exhibits notable improvements over other methods, further confirming its effectiveness in image SR tasks.

CTSRNet outperforms several classical image SR methods on the Set14 dataset across all scaling factors (Table IV). At ×2, it achieves a PSNR of 33.60 dB and a SSIM of 0.9184, surpassing approaches such as SRCNN [5] and VDSR [9]. For ×3, CTSRNet attains 30.34 dB (PSNR) and 0.8416 (SSIM), demonstrating superior reconstruction quality over methods including Bicubic [41], DRRN [11], and FSRCNN [12]. At ×4, it achieves 28.61 dB and 0.7812, outperforming the CARN-M [42], IDN [43], and DRCN [10]. Notably, CTSRNet consistently exceeds the performance of the state-of-the-art HDSRNet [62] across all scales. These results confirm the robustness and effectiveness of CTSRNet, even on datasets with diverse image contents.

CTSRNet demonstrates strong robustness and generalization on the B100 dataset, which is characterized by complex natural scenes and fine-grained details (Table V). At a scaling factor of ×2, CTSRNet achieves a PSNR of 32.15 dB and an SSIM of 0.8992, surpassing most comparative methods, including a 0.79 dB gain over SRCNN [5]. At ×3, the model attains

TABLE V
AVERAGE PSNR (DB) AND SSIM FOR DIFFERENT METHODS WITH THREE SCALING FACTORS ON THE B100 DATASET

Datasets	Methods	×2	×3	×4
B100	Bicubic [41]	29.56/0.8431	27.21/0.7385	25.96/0.6675
	SRCNN [5]	31.36/0.8879	28.41/0.7863	26.90/0.7101
	VDSR [9]	31.90/0.8960	28.82/0.7976	27.29/0.7251
	DRRN [11]	32.05/0.8973	28.95/0.8004	27.38/0.7284
	FSRCNN [12]	31.53/0.8920	28.53/0.7910	26.98/0.7150
	CARN-M [42]	31.92/0.8960	28.91/0.8000	27.44/0.7304
	IDN [43]	32.08/0.8985	28.95/0.8013	27.41/0.7297
	A+ [44]	31.21/0.8863	28.29/0.7835	26.82/0.7087
	JOR [45]	31.22/0.8867	28.27/0.7837	26.79/0.7083
	RFL [46]	31.16/0.8840	28.22/0.7806	26.75/0.7054
	SelfEx [38]	31.18/0.8855	28.29/0.7840	26.84/0.7106
	CSCN [47]	31.40/0.8884	28.50/0.7885	27.03/0.7161
	RED [22]	31.96/0.8972	28.88/0.7993	27.35/0.7276
	DnCNN [48]	31.90/0.8961	28.85/0.7981	27.29/0.7253
	TNRD [49]	31.40/0.8878	28.50/0.7881	27.00/0.7140
	FDSR [50]	31.87/0.8847	28.82/0.7797	27.31/0.7031
	DRCN [10]	31.85/0.8942	28.80/0.7963	27.23/0.7233
	CNF [52]	31.91/0.8962	28.82/0.7980	27.32/0.7253
	LapSRN [14]	31.80/0.8950	-	27.32/0.7280
	NDRCN [55]	32.00/0.8975	28.86/0.7991	27.30/0.7263
	MemNet [56]	32.08/0.8978	28.96/0.8001	27.40/0.7281
	LESRCNN [57]	31.95/0.8964	28.91/0.8005	27.45/0.7313
	ScSR [58]	30.77/0.8744	27.72/0.7647	26.61/0.6983
	DSRCNN [59]	32.05/0.8978	29.01/0.8029	27.50/0.7341
	DAN [60]	31.76/0.8858	28.94/0.7919	27.51/0.7248
	HDSRNet [62]	32.14/0.8994	29.06/0.8048	27.55/0.7357
	CTSRNet(Ours)	32.15/0.8992	29.09/0.8051	27.57/0.7362

TABLE VI
AVERAGE PSNR (DB) AND SSIM FOR DIFFERENT METHODS WITH THREE SCALING FACTORS ON THE URBAN100 DATASET

Datasets	Methods	×2	×3	×4
Urban100	Bicubic [41]	26.88/0.8403	24.46/0.7349	23.14/0.6577
	SRCNN [5]	29.50/0.8946	26.24/0.7989	24.52/0.7221
	VDSR [9]	30.76/0.9140	27.14/0.8279	25.18/0.7524
	DRRN [11]	31.23/0.9188	27.53/0.8378	25.44/0.7638
	FSRCNN [12]	29.88/0.9020	26.43/0.8080	24.62/0.7280
	CARN-M [42]	31.23/0.9193	27.55/0.8385	25.62/0.7694
	IDN [43]	31.27/0.9196	27.42/0.8359	25.41/0.7632
	A+ [44]	29.20/0.8936	26.03/0.7973	24.32/0.7183
	JOR [45]	29.25/0.8951	25.97/0.7972	24.29/0.7181
	RFL [46]	29.11/0.8904	25.86/0.7900	24.19/0.7096
	SelfEx [38]	29.54/0.8967	26.44/0.8088	24.79/0.7374
	RED [22]	30.91/0.9159	27.31/0.8303	25.35/0.7587
	DnCNN [48]	30.74/0.9139	27.15/0.8276	25.20/0.7521
	TNRD [49]	29.70/0.8994	26.42/0.8076	24.61/0.7291
	FDSR [50]	30.91/0.9088	27.23/0.8190	25.27/0.7417
	DRCN [10]	30.75/0.9133	27.15/0.8276	25.14/0.7510
	LapSRN [14]	30.41/0.9100	-	25.21/0.7560
	WaveResNet [53]	30.96/0.9169	27.28/0.8334	25.36/0.7614
	NDRCN [55]	31.06/0.9175	27.23/0.8312	25.16/0.7546
	MemNet [56]	31.31/0.9195	27.56/0.8376	25.50/0.7630
	LESRCNN [57]	31.45/0.9206	27.70/0.8415	25.77/0.7732
	ScSR [58]	28.26/0.8828	-	24.02/0.7024
	DSRCNN [59]	31.83/0.9252	27.99/0.8483	25.94/0.7815
	DAN [60]	30.60/0.9060	27.65/0.8352	25.86/0.7721
	HDSRNet [62]	32.00/0.9267	28.02/0.8493	26.01/0.7827
	HAFRN [63]	32.20/0.9289	28.16/0.8528	26.02/0.7849
	CTSRNet(Ours)	32.11/0.9282	28.20/0.8533	26.15/0.7875

a balanced performance with 29.09 dB (PSNR) and 0.8051 (SSIM). Even at the challenging ×4 scale, CTSRNet reaches an SSIM of 0.7362, slightly outperforming HDSRNet [62], highlighting its ability to preserve structural information under higher scaling multiples. These results confirm that CTSRNet effectively restores image details and textures across various scaling factors, demonstrating both high accuracy and strong adaptability in real-world SR tasks.

The Urban100 dataset includes a wide range of urban scenes featuring buildings, roads, and other man-made structures, characterized by complex textures and abundant high-frequency details. This diversity makes it an effective

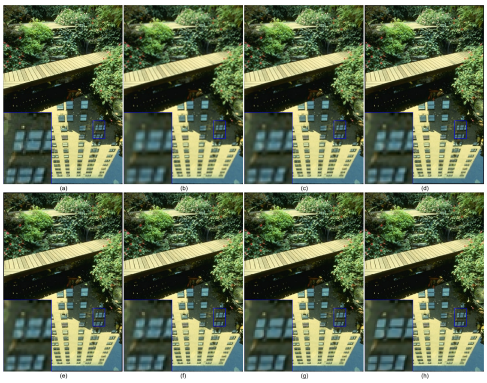

Fig. 2. Super-resolution visualization results of different methods with scaling factor of ×2 on B100-148026. (a) A HR image, (b)Bicubic, (c) ScSR, (d) A+, (e) SRCNN, (f) CARN-M, (g) HDSRNet and (h) CTSRNet (Ours).

benchmark for evaluating the capability of SR models in handling structurally complex and visually rich content. CTSRNet demonstrates strong performance across all scaling factors (Table VI). At ×2 magnification, the model maintains stable reconstruction quality, effectively preserving structural details. With scaling factors of ×3 and ×4, CTSRNet consistently outperforms most competing models, including HAFRN [63] and DRRN [11], highlighting its superior ability to restore fine-grained textures and reduce blurring in high-magnification settings. These results validate the model's robustness in urban environments and its practical value in real-world applications requiring high-fidelity visual reconstruction.

As demonstrated in quantitative analysis, CTSRNet achieved notable improvements in PSNR and SSIM. However, such metrics alone are insufficient to fully reflect the perceptual quality of super-resolved images, especially regarding detail and texture fidelity. To complement the quantitative results, a qualitative analysis was conducted using the B100, Set14, and Urban100 datasets to visually compare the reconstruction performance of CTSRNet against six established SR methods. These include three traditional approaches, i.e., bicubic interpolation (Bicubic) [41], sparse coding-based SR (ScSR) [58] and anchored neighborhood regression (A+) [44], and three deep learning-based methods, such as SRCNN [5], CARN-M [42], and HDSRNet [62]. In the visualization results, subfigures (a)–(h) represent, respectively: the ground-truth HR images, outputs of the six comparative methods, and the result produced by CTSRNet.

A visualization comparison using an image from the B100 dataset demonstrates the effectiveness of CTSRNet (Fig. 2). The results indicate that CTSRNet aligns more closely with human visual perception. In the magnified region, classical interpolation methods fail to reconstruct the window frame structure, while deep learning-based methods exhibit varying levels of blurring and distortion. In contrast, CTSRNet more accurately restores the true window frame structure, preserving clearer and more faithful image details.

Figure 3 presents a comparative visualization using a zebra image from the Set14 dataset. The results demonstrate that CTSRNet significantly outperforms other methods in recovering fine details. In the magnified region, traditional interpolation produces highly blurred results, with the zebra's

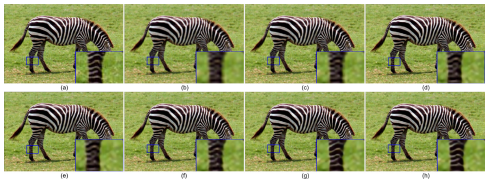

Fig. 3. Super-resolution visualization results of different methods with scaling factor of $\times 3$ on Set14-zebra. (a) A HR image, (b)Bicubic, (c) ScSR, (d) A+, (e) SRCNN, (f) CARN-M, (g) HDSRNet and (h) CTSRNet (Ours).

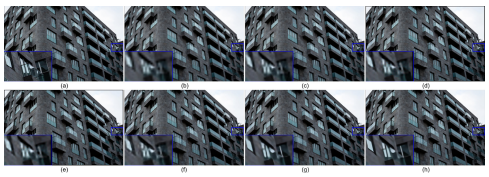

Fig. 4. Super-resolution visualization results of different methods with scaling factor of $\times 4$ on Urban100-001. (a) A HR image, (b)Bicubic, (c) ScSR, (d) A+, (e) SRCNN, (f) CARN-M, (g) HDSRNet and (h) CTSRNet (Ours).

texture appearing indistinct. Classical deep learning methods also fail to accurately reconstruct the fifth zebra, causing it to blend with the fourth. In contrast, CTSRNet effectively preserves the separation and restores the original texture details, resulting in a more realistic and structurally accurate reconstruction.

Figure 4 illustrates a visual comparison using an image from the Urban100 test set. The analysis shows that the proposed CTSRNet exhibits superior SR performance. In the magnified region, the classical interpolation method fails to reconstruct window frames clearly, and deep learning-based methods introduce noticeable distortions that do not reflect the true scene. In contrast, CTSRNet more accurately restores the building's geometry, effectively preserving structural integrity and avoiding deformation.

Crucially, the qualitative analysis reveals that CTSRNet consistently produces visually superior results compared to other methods, demonstrating exceptional capability in recovering fine details and preserving structural integrity across diverse scene types. Unlike traditional interpolation methods that often result in blurring and loss of detail, or some deep learning approaches that may introduce artifacts or distortion, CTSRNet's reconstructions are characterized by enhanced clarity, sharper textures, and more accurate geometries. This suggests that the proposed tree-based architecture effectively capture and leverage hierarchical features and regional similarities, aligning well with human visual perception.

V. Conclusion

This paper presents a cluster tree network designed for image SR. The model leverages a hierarchical structure with parallel branches to enable diverse feature extraction. Conditionally parameterized convolution is introduced to adaptively modulate convolution kernels based on input characteristics. Furthermore, clustering techniques are employed to capture regional similarities, facilitating detailed texture recovery. Experimental results confirm that the proposed method effectively reconstructs fine details, outperforming existing methods in SR tasks.

References

[1] X. Liao, X. Wei, and M. Zhou, "Minimax concave penalty regression for Superresolution image reconstruction," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 2999–3007, Feb. 2024.

[2] A. Sanchez-Beato and G. Pajares, "Noniterative interpolation-based super-resolution minimizing aliasing in the reconstructed image," *IEEE Trans. Image Process.*, vol. 17, pp. 1817–1826, 2008.

[3] G. Riegler, S. Schulter, M. Rüther, and H. Bischof, "Conditioned regression models for non-blind single image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 522–530.

[4] C. Tian, M. Zheng, C.-W. Lin, Z. Li, and D. Zhang, "Heterogeneous window transformer for image denoising," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 54, no. 11, pp. 6621–6632, Nov. 2024.

[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] C. Tian, M. Zheng, T. Jiao, W. Zuo, Y. Zhang, and C.-W. Lin, "A self-supervised CNN for image watermark removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7566–7576, Aug. 2024.

[7] C. Tian, M. Zheng, B. Li, Y. Zhang, S. Zhang, and D. Zhang, "Perceptive self-supervised learning network for noisy image watermark removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7069–7079, Aug. 2024.

[8] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 252–268.

[9] 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)*, 2016, pp. 1646–1654.

[10] 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)*, 2016, pp. 1637–1645.

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

[12] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 391–407.

[13] C. Yang and G. Lu, "Deeply recursive low-and high-frequency fusing networks for single image super-resolution," *Sensors*, vol. 20, no. 24, p. 7268, 2020.

[14] 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)*, 2017, pp. 5835–5843.

[15] Y. Huang et al., "Learning deformable and attentive network for image restoration," *Knowl.-Based Syst.*, vol. 231, Nov. 2021, Art. no. 107384.

[16] C. Tian, Y. Xu, W. Zuo, C.-W. Lin, and D. Zhang, "Asymmetric CNN for image superresolution," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 6, pp. 3718–3730, Jun. 2022.

[17] D. Rai and S. S. Rajput, "ANNLR-FSR: Artificial neural network-based locality Regularization for face super-resolution," *IEEE Trans. Consum. Electron.*, vol. 70, no. 3, pp. 5053–5064, Aug. 2024.

[18] L. Wei, M. Ye, L. Ji, Y. Gan, S. Li, and X. Li, "Multi-level alignments for compressed video super-resolution," *IEEE Trans. Consum. Electron.*, vol. 70, no. 3, pp. 5101–5114, Aug. 2024.

[19] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electron. Markets*, vol. 31, no. 3, pp. 685–695, 2021.

[20] C. Tian, Y. Zhang, W. Zuo, C.-W. Lin, D. Zhang, and Y. Yuan, "A heterogeneous group CNN for image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 5, pp. 6507–6519, May 2024.

[21] Y. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, and C. Schroers, "A fully progressive approach to single-image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2018, pp. 977–97709.

[22] X. J. Mao, C. Shen, and Y. B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. 30th Conf. Neural Inf. Process. Syst.*, 2016, pp. 1–9.

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

[24] C. Tian, Y. Xu, W. Zuo, B. Zhang, L. Fei, and C.-W. Lin, "Coarse-to-fine CNN for image super-resolution," *IEEE Trans. Multimedia*, vol. 23, pp. 1489–1502, 2020.

[25] Q. Wang, Q. Gao, L. Wu, G. Sun, and L. Jiao, "Adversarial multi-path residual network for image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 6648–6658, 2021.

[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)*, 2019, pp. 3867–3876.

[27] 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. (CVPR)*, 2018, pp. 2472–2481.

[28] M. V. Conde, G. Geigle, and R. Timofte, "Instructir: High-quality image restoration following human instructions," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 1–21.

[29] C. Tian, M. Song, X. Fan, X. Zheng, B. Zhang, and D. Zhang, "A tree-guided CNN for image super-resolution," *IEEE Trans. Consum. Electron.*, early access, May 22, 2025, doi: 10.1109/TCE.2025.3572732.

[30] L. Abrahamyan, A. M. Truong, W. Philips, and N. Deligiannis, "Gradient variance loss for structure-enhanced image super-resolution," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 3219–3223.

[31] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "CondConv: Conditionally parameterized convolutions for efficient inference," in *Proc. 33rd Conf. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1307–1318.

[32] A. Zhang, W. Ren, Y. Liu, and X. Cao, "Lightweight image super-resolution with superpixel token interaction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 12682–12691.

[33] 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)*, 2017, pp. 1122–1131.

[34] C. Tian, X. Zhang, Q. Zhang, M. Yang, and Z. Ju, "Image super-resolution via dynamic network," *CAAI Trans. Intell. Technol.*, vol. 9, no. 4, pp. 837–849, 2024.

[35] M. Bevilacqua, A. Roumy, C. Guillemot, and A. Morel, "Low-complexity single image super-resolution based on nonnegative neighbor embedding," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–10.

[36] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surfaces*, 2010, pp. 711–730.

[37] 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)*, 2001, pp. 416–423.

[38] 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)*, 2015, pp. 5197–5206.

[39] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1874–1883.

[40] 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)*, 2017, pp. 1132–1140.

[41] R. E. Carlson and F. N. Fritsch, "Monotone piecewise bicubic interpolation," *SIAM J. Numer. Anal.*, vol. 22, no. 2, pp. 386–400, 1985.

[42] Y. Li, E. Agustsson, S. Gu, R. Timofte, and L. Van, "CARN: Convolutional anchored regression network for fast and accurate single image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 166–181.

[43] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 723–731.

[44] R. Timofte, V. Desmet, and L. Vangool, "A+: Adjusted anchored Neighborhood regression for fast super-resolution," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 111–126.

[45] D. Dai, R. Timofte, and L. Van Gool, "Jointly optimized regressors for image super-resolution," *Comput. Graph. Forum*, vol. 34, no. 2, pp. 95–104, 2015.

[46] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3791–3799.

[47] 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)*, 2015, pp. 370–378.

[48] K. Zhang, X. Gao, D. Tao, and X. Li, "Single image super-resolution with non-local means and steering kernel regression," *IEEE Trans. Image Process.*, vol. 21, pp. 4544–4556, 2012.

[49] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1256–1272, Jun. 2017.

[50] Z. Lu et al., "Fast single image super-resolution via dilated residual networks," *IEEE Access*, vol. 7, pp. 109729–109738, 2019.

[51] Y. Shi, K. Wang, C. Chen, L. Xu, and L. Lin, "Structure-preserving image super-resolution via Contextualized Multitask learning," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2804–2815, Dec. 2017.

[52] H. Ren, M. El-Khamy, and J. Lee, "Image super resolution based on fusing multiple convolution neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017, pp. 1050–1057.

[53] W. Bae, J. Yoo, and J. C. Ye, "Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017, pp. 1141–1149.

[54] J. Xu, M. Li, J. Fan, X. Zhao, and Z. Chang, "Self-learning super-resolution using convolutional principal component analysis and random matching," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1108–1121, May 2019.

[55] F. Cao and B. Chen, "New architecture of deep recursive convolution networks for super-resolution," *Knowl.-Based Syst.*, vol. 178, pp. 98–110, Aug. 2019.

[56] 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)*, 2017, pp. 4549–4557.

[57] C. Tian et al., "Lightweight image super-resolution with enhanced CNN," *Knowl.-Based Syst.*, vol. 205, Oct. 2020, Art. no. 106235.

[58] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[59] J. Song, J. Xiao, C. Tian, Y. Hu, L. You, and S. Zhang, "A dual CNN for image super-resolution," *Electronics*, vol. 11, no. 5, p. 757, 2022.

[60] Y. Huang, S. Li, L. Wang, T. Tan et al., "Unfolding the alternating optimization for blind super resolution," in *Proc. 34th Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 5632–5643.

[61] Z. Luo, H. Huang, L. Yu, Y. Li, H. Fan, and S. Liu, "Deep constrained least squares for blind image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 17621–17631.

[62] C. Tian, X. Zhang, J. Ren, W. Zuo, Y. Zhang, and C.-W. Lin, "A heterogeneous dynamic convolutional neural network for image super-resolution," 2024, *arXiv:2402.15704*.

[63] K. Wang, X. Yang, and G. Jeon, "Hybrid attention feature refinement network for lightweight image super-resolution in Metaverse Immersive display," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 3232–3244, Feb. 2024.