This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCSVT.2025.3614371,

Representation Sampling and Hybrid Transformer Network for Image Compressed Sensing

Heping Song, Jingyao Gong, Hongjie Jia, Xiangjun Shen, Jianping Gou, *Senior Member, IEEE*, Hongying Meng, *Senior Member, IEEE*, and Le Wang, *Senior Member, IEEE*

***Abstract*—Deep unrolling networks (DUNs) have attracted substantial attention in the field of image compressed sensing (CS) due to their superior performance and good interpretability by recasting optimization algorithms as deep networks. However, existing DUNs suffer from low sampling efficiency, and the improvement in reconstruction quality heavily relies on large model complexity. To address these issues, we propose a lightweight Representation Sampling and Hybrid Transformer Network (RHT-Net). Firstly, we propose a Representation-CS (RCS) model to extract high-level features to achieve efficient sampling. This sampling strategy leads to highly dense, semantically rich and extremely compact features without observing the original pixels, which also reduces the cross-domain loss during iteration. Secondly, we design a Tri-Scale Sparse Denoising (TSSD) module in the deep unrolling stages to extend sparse proximal projections, leveraging multi-scale auxiliary variables to enhance multi-feature flow and memory effects. Thirdly, we develop a hybrid Transformer module that includes a Global Cross Attention (GCA) block and a Window Local Attention (WLA) block, using the measurements to cross-estimate the reconstruction error, thereby generating finer spatial details and improving local recovery. Experiments demonstrate that RHT-Net enhanced version outperforms the current state-of-the-art methods by up to 1.17dB in PSNR. The lightweight RHT-Net achieves a 0.43dB gain while reducing model parameters by up to 22 times. The code will be released publicly at https://github.com/songhp/RHTNet.**

***Index Terms*—Compressed Sensing, Deep Unrolling Network, Representation Sampling**

I. Introduction

COMPRESSED Sensing (CS) [1], [2] is a signal processing theory that overcomes the limitations of the traditional Shannon-Nyquist sampling theorem [3]. It can recover sparse or compressible signals from far fewer samples than traditionally required, significantly reducing the cost and data storage demands. Thus, CS has found wide applications in various fields, including snapshot compressed imaging [4], [5], [6], [7], medical imaging [8], [9], [10], [11], hyperspectral compressed imaging [15], [16], [17], [18], and laser scanning imaging [19]. The CS sampling process can be expressed as $\mathbf{y}$=$\mathbf{Ax}$, where $\mathbf{x} \in \mathbb{R}^{n\times 1}$ represents the original signal, $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the sampling matrix, and $\mathbf{y} \in \mathbb{R}^{m\times 1}$ denotes the measurements, with a sampling rate of r=m/n. Typically, $m \ll n$, thus the core objective of CS [20], [21] is to recover the signal $\mathbf{x}$ in a highly underdetermined system.

Traditional CS methods [22], [23] typically formulate it as solving the following optimization problem:

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{Ax}-\mathbf{y}\|_2^2 + \lambda \mathrm{R}(\mathbf{x}), \tag{1}$$

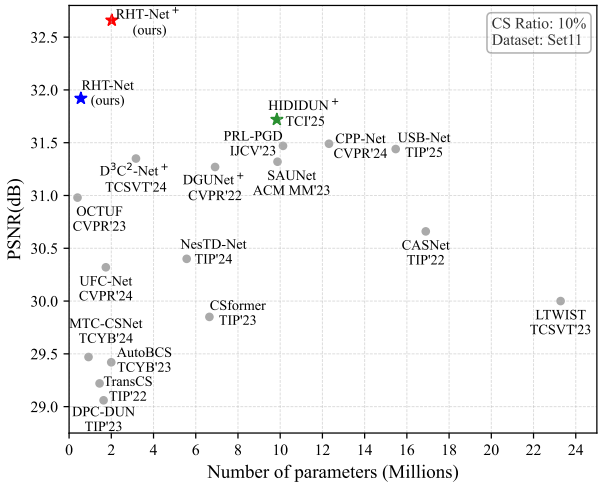


Fig. 1. PSNR performance comparison between our proposed RHT-Net and recent state-of-the-art methods demonstrates that RHT-Net achieves superior performance while maintaining a lightweight architecture.

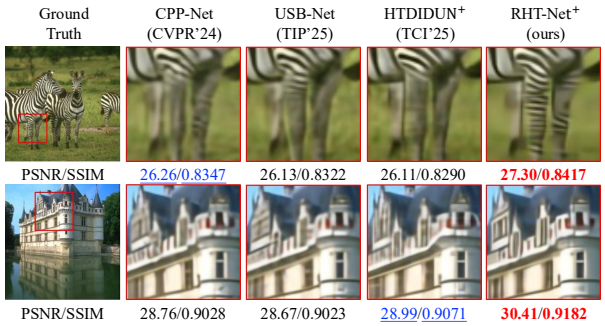


Fig. 2. Visual evaluation at a CS ratio of 0.10 demonstrates that RHT-Net outperforms CPP-Net [12], USB-Net [13], and HTDIDUN$^+$ [14] in preserving fine image details. While the compared methods exhibit noticeable detail loss in critical regions—such as the zebra's body and window structures—RHT-Net maintains a high degree of visual fidelity.

This work was supported partly by the National Natural Science Foundation of China (Nos. 62472201 and 62172193)

Heping Song, Jingyao Gong, Hongjie Jia and Xiangjun Shen are with the School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China (Email: songhp@ujs.edu.cn; gongjy@stmail.ujs.edu.cn; jiahj@ujs.edu.cn; xjshen@ujs.edu.cn)

Jianping Gou is with the College of Computer and Information Science, Southwest University, Chongqing 400715, China (Email: cherish.gjp@gmail.com)

Hongying Meng is with the Department of Electronic and Electrical Engineering, Brunel University London, UB8 3PH, Uxbridge, London, U.K. (Email: hongying.meng@brunel.ac.uk)

Le Wang is with the Institute of Artificial Intelligence and Robotic, Xi'an Jiaotong University, Xi'an 710049, China (Email: lewang@xjtu.edu.cn)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCSVT.2025.3614371,

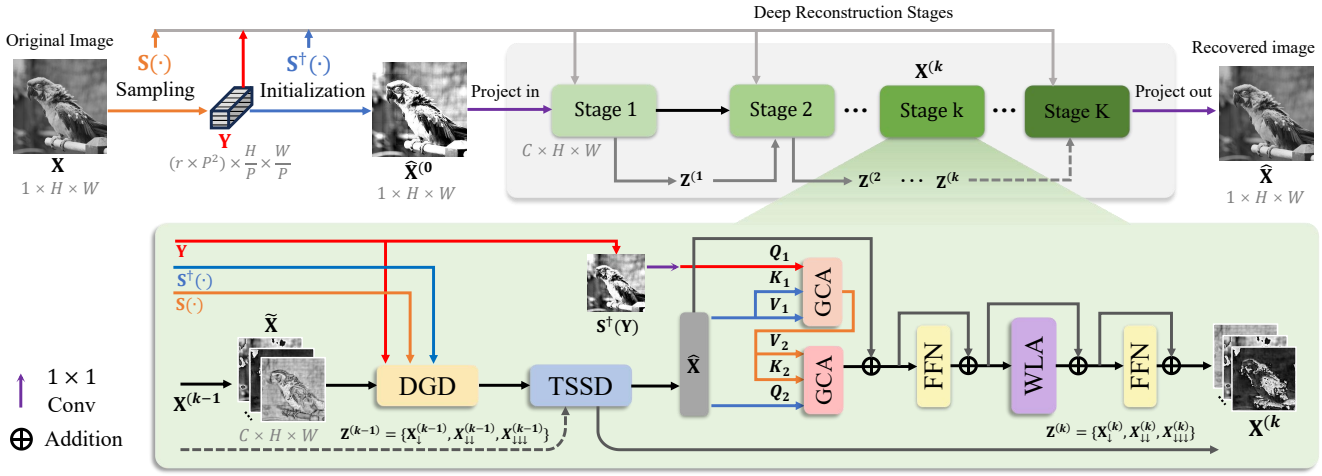


Fig. 3. Overall architecture of RHT-Net. The original image $\mathbf{X}$ is compressed by the sampling module $\mathbf{S}(\cdot)$ to obtain a low-dimensional feature representation $\mathbf{Y}$. This representation is then reshaped into an initial estimation $\hat{\mathbf{X}}^{(0)}$ via the pseudo-inverse transformation sub-network $\mathbf{S}^{\dagger}(\cdot)$. Subsequently, deep reconstruction is performed through a cascade of K iterative stages. Each stage takes $\mathbf{Y}$, $\mathbf{S}(\cdot)$, $\mathbf{S}^{\dagger}(\cdot)$, and auxiliary multi-scale variables $\mathbf{Z}^{(k)}$ as inputs, and consists of sub-modules such as DGD, TSSD, GCA, WLA, and FFN.

where $\frac{1}{2}\|\mathbf{Ax}-\mathbf{y}\|_2^2$ represents the data fidelity term, which evaluates the fit to the measurements, and $\lambda \mathrm{R}(\mathbf{x})$ is the regularization term with parameter λ. CS recovery solvers not only enforce consistency between the reconstruction and the measurements but also generally promote sparsity in the solution. Therefore, the prior term often involves sparsity-inducing operators related to predefined transform bases (e.g., ℓ_1 regularization for ISTA [24], [25]). These methods perform well in terms of convergence and mathematical analysis, but they suffer from high computational complexity and poor adaptability [26], [27].

Recently, deep learning-based compressed sensing (DCS) algorithms have demonstrated remarkable capabilities. A class of methods [28], [29], [30], [31], [32] advocate for directly learning the latent inverse mapping using an end-to-end learning approach. However, these black-box models address the complex inverse mapping problem through extensive trial and error, resulting in extremely low efficiency. In contrast, Deep Unrolling Networks (DUNs) propose an equivalent reformulation of optimization algorithms, where these approaches [21], [12], [33], [34], [35], [36], [37] replace handcrafted optimization solvers with learnable networks, allowing DCS methods to achieve a balance between recovery performance and mathematical interpretability. However, current DUNs still face three key challenges: 1). Previous direct sampling methods based on low-level, low-information-density pixels struggle with efficiency, while deep reconstruction based on high-level feature domains exhibits significant feature alignment loss. 2). Feature representation fails to adequately integrate multi-scale information, neglecting the neural network's ability to learn sparse prior terms. 3). Relying heavily on a large number of parameters to improve recovery quality severely hinders application on low-performance devices, particularly in common scenarios such as LiDAR [19] and MRI [9].

To address the above challenges, this paper proposes a Representation Sampling and Hybrid Transformer Network, termed RHT-Net. We propose an innovative representation-domain-based compressed sensing (RCS) model that differs from traditional DCS approaches, which directly sample pixel-level signals. Instead, our model first extracts semantically rich, high-information-density compact representations, and then performs a more efficient CS process within this high-level representation domain—without relying on the original pixels. ~~RCS~~**The RCS model** reduces alignment loss during cross-domain correction for deep reconstruction, supporting recovery at a fine semantic level. In the deep reconstruction phase, RHT-Net enhances signal fidelity and sparse prior constraint performance through cascaded Deep Gradient Descent (DGD) updates, combined with a Tri-Scale Sparse Denoising (TSSD) module. ~~TSSD~~**The TSSD module** learns multi-scale latent priors, expanding traditional proximal operators via a deep U-shaped network and further introduces scale auxiliary variables to enhance multi-scale feature flow and memory effects. Moreover, RHT-Net designs a hybrid Transformer with minimal computational overhead, which includes Global Sparse Cross-Attention (GCA) block and Window Local Attention (WLA) block, using cross-attention between measurements and the estimated reconstruction to inject additional and finer spatial information and local recovery. As shown in Fig. 1 and Fig. 2, RHT-Net significantly outperforms state-of-the-art methods in terms of reconstruction performance and human perceptual quality, while also maintaining a notably smaller parameter size. **The overall architecture of our proposed RHT-Net is illustrated in Fig. 3, which demonstrates the integration of representation sampling, deep unrolling, and hybrid Transformer components.**

The primary four-fold contributions can be summarized as follows:

- We propose a representation-domain CS model that facilitates efficient sampling of deep semantic features, significantly reducing the cross-domain loss between measurement and reconstruction.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCSVT.2025.3614371,

- We design a ~~Tri-Scale Sparse Denoising (TSSD)~~**TSSD** network, leveraging deep networks to extend proximal operators and learn stronger sparse prior constraints.
- We develop an efficient hybrid Transformer module that integrates ~~Global Sparse Cross-Attention (GCA) block and Window Local Attention (WLA) block~~**GCA and WLA blocks** to infuse additional support for deep reconstruction stage.
- Experiments demonstrate that RHT-Net consistently outperforms current state-of-the-art methods in terms of visual quality, robustness, and model efficiency.

The remainder of the paper is organized as follows: Section II provides a brief overview of related work. Section III details the formulation of the proposed RHT-Net. Section IV presents a comparative evaluation of our method against state-of-the-art approaches. Finally, Section V concludes the paper with a summary of key findings and insights.

II. Related Works

A. Deep Black-box Networks

This category treats image reconstruction as an end-to-end learning task, learning the inverse mapping from measurements to original signals by training on large datasets. This approach avoids the explicit construction of mathematical models. The optimization goal is:

$$\min_{\Theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\mathbf{X}_i, \mathcal{M}(\mathbf{Y}_i)), \tag{2}$$

where N represents the number of samples $\mathbf{X}_i$ in the dataset, $\mathcal{Y}_i$ is the corresponding measurements, $\mathcal{M}$ is the network model, Θ denotes the learnable parameters, and $\mathcal{L}$ is the loss function. Representative methods, such as CSNet [47], MAC-Net [29], MR-CCSNet [30], NL-CSNet [48], and AutoBCS [31], employ convolutional neural networks (CNN) to address CS problems. Another class of methods, such as DPA-Net [49] and CSformer [42], leverages the power of transformer architectures to achieve improved performance, albeit at a higher computational cost. These CS models are often considered black-box models, lacking interpretability and requiring extensive and inefficient trial-and-error for parameter tuning.

B. Deep Unrolling Networks

DUNs recast traditional CS iterative algorithms as deep networks, thereby enhancing both performance and interpretability. These networks typically address bi-level optimization problems based on Proximal Gradient Descent (PGD) and its variants. As a result, several prominent DUN-based methods have been proposed, including ISTA-Net [33], OPINE-Net [26], TransCS [38], DGUNet [39], AMS-Net [50], LTwIST [43], PRL-PGD [36], D^{3}C^2-Net$^+$ [21], MTC-CSNet [44], NesTD-Net [46], UFC-Net [45], HTDIDUN$^+$ [14], USB-Net [13], and HUNet [51]. Recent research has increasingly focused on boosting the efficiency and modeling power of DUNs. OCTUF [34] incorporates

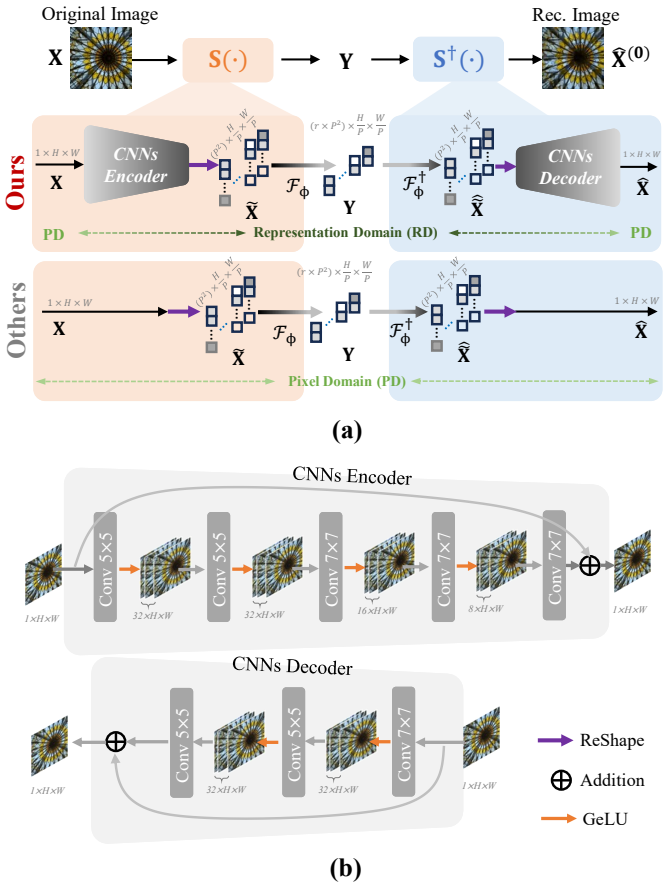


Fig. 4. ~~The proposed representation domain compressed sampling (RCS) model.~~**Detailed architecture of the proposed Representation-domain Compressed Sampling (RCS) model.** (a) Overview of RCS, which performs sampling and recovery in a learned representation domain, ~~unlike other methods that operate in the pixel domain~~**fundamentally differing from conventional methods that operate directly on the pixel domain**. We use a block size of $P = 32$ for non-overlapping block sampling, consistent with most methods [38], [39], [40], [41], [34], [36], [42], [43], [32], [21], [44], [45], [46], [12], [14]. (b) ~~Architecture of the CNNs Encoder and Decoder, which implements non-linear transformations to effectively project images into a learned representation space.~~**Detailed structure of the CS Encoder and Decoder. Both modules employ a series of convolutional layers with large kernels (5×5 and 7×7), GeLU activations, and skip connections to effectively project images into a compact, semantically rich representation space and back.**

cross-feature attention to accelerate reconstruction, while CPP-Net [12] leverages the Chambolle-Pock proximal point algorithm to improve modeling capabilities. Advancing the field further, several recent works propose novel architectures. HTDIDUN$^+$ [14] introduces a decomposition-inspired network designed for high-throughput reconstruction. USB-Net [13] unfolds the Split Bregman method, integrating multi-phase features to enhance information flow. Similarly, HUNet [51] employs a homotopy unfolding strategy, offering a new perspective for solving inverse problems. However, despite these advancements, a fundamental challenge remains: existing models often operate directly on sparse, low-information pixels. This approach limits computational efficiency and hinders the preservation of high-level semantic features. Consequently, achieving substantial improvements

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCSVT.2025.3614371,

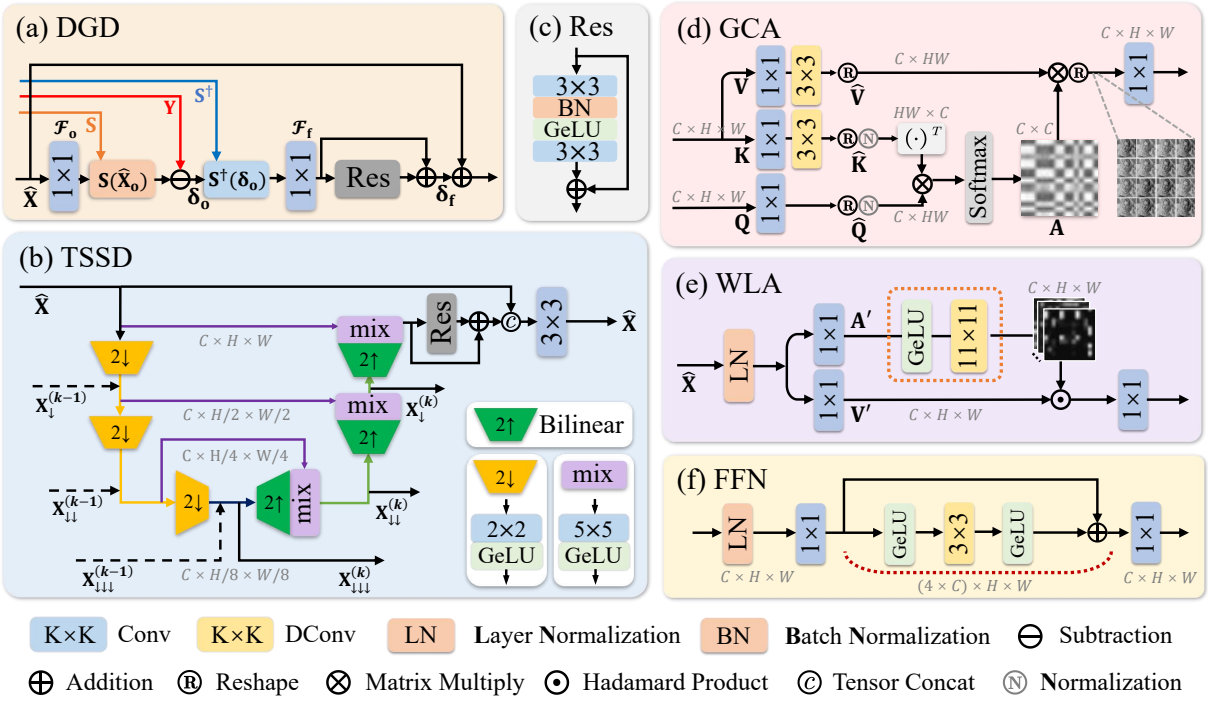


Fig. 5. Illustration of the detailed design of our RHT-Net. (a) Deep Gradient Descent (DGD) module; (b) Tri-Scale Sparse Denoising (TSSD) module; (c) Residual (Res) module; (d) Global Cross Attention (GCA) block for projecting sampled and reconstructed features, (e) Window-like Attention (WLA) block for simulating local attention with large convolution kernels, and (f) the Feed-Forward Network (FFN) module for integrating Transformer features.

in CS quality typically requires a disproportionate increase in computational resources, revealing a critical trade-off that remains central to ongoing research efforts.

III. Methodology

A. Overall Architecture

In RHT-Net, the original image $\mathbf{X}$ is initially compressed into low-dimensional measurements $\mathbf{Y}$ by the representation sampling network $\mathbf{S}(\cdot)$, and the initial reconstruction $\hat{\mathbf{X}}^{(0)}$ is obtained through the pseudo-inverse transformation $\mathbf{S}^{\dagger}(\cdot)$, i.e., $\hat{\mathbf{X}}^{(0)} = \mathbf{S}^{\dagger}(\mathbf{S}(\mathbf{X}))$. The framework then performs K cascaded stages on $\hat{\mathbf{X}}^{(0)}$ to progressively refine the reconstruction $\hat{\mathbf{X}}$. Each reconstruction stage incorporates sub-modules such as Deep Gradient Descent (DGD) and Tri-Scale Sparse Denoising (TSSD), with inputs from $\mathbf{Y}$, $\mathbf{S}(\cdot)$, and $\mathbf{S}^{\dagger}(\cdot)$. In the hybrid Transformer module, Global Cross-Attention (GCA) block and Window Local Attention (WLA) block are employed to further estimate the reconstruction error, leading to enhanced spatial information and localized recovery. The Feed Forward Network (FFN) sub-module is employed to integrate features. Additionally, TSSD improves the reconstruction reference and memory effect of Tri-scale features by passing auxiliary variables $\mathbf{Z}^{(k)} = \{\mathbf{X}_{\downarrow}^{(k)}, \mathbf{X}_{\downarrow\downarrow}^{(k)}, \mathbf{X}_{\downarrow\downarrow\downarrow}^{(k)}\}$. Fig. 3 illustrates the detailed RHT-Net framework.

B. Representation-based Sampling and Initialization

Deep representations of signals typically offer higher compression efficiency than low-level linear structures [52], [53]. Inspired by this, we develop a representation domain compressed sampling (RCS) model, as illustrated in Fig. 4(a). ~~Unlike conventional DCS methods that rely on direct sampling, the core principle of our approach is to learn a powerful nonlinear mapping to a data-adaptive, semantically rich latent space, where the signal can be compressed far more efficiently.~~ **The fundamental distinction of our approach lies in shifting the sampling process from the pixel domain to a learned, deep representation space. Instead of sampling raw pixels, RCS first employs a nonlinear encoder to transform the image into a compact and semantically rich latent representation. The sampling is then performed in this high-level domain, capturing essential information more efficiently and resulting in more meaningful measurements.**

~~The representation-domain compressed sensing (RCS) module is architecturally compact yet essential to the high performance of our method, embodying a design philosophy that balances efficiency with simplicity. Its CS Encoder, detailed in Fig. 4(b), utilizes a cascade of convolutional layers with large kernels (5×5 and 7×7) and GeLU activation function [54] to distill the essential semantic content of an image into a compact and dense representation, $\tilde{\mathbf{X}}$. The large receptive fields of the kernels enable the model to capture long-range spatial dependencies, while skip connections ensure the preservation and aggregation of both high-level semantic information and low-level textural details. This process effectively projects the image from the pixel domain onto a learned non-linear manifold where its information is more compactly organized.~~ **The RCS module is architecturally compact yet highly effective. As detailed**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCSVT.2025.3614371,

in Fig. 4(b), the CS Encoder consists of three convolutional layers. The first layer uses a 7×7 kernel, followed by two layers with 5×5 kernels, all with a stride of 1 and same padding to maintain feature map resolution. Each convolutional layer is followed by a GeLU activation function [54]. This design, featuring large receptive fields, distills the essential semantic content of an image into a compact representation, $\tilde{\mathbf{X}}$. Skip connections are integrated to ensure that both high-level semantic information and low-level textural details are preserved and aggregated. This process effectively projects the image onto a learned nonlinear manifold where its information is more compactly organized for efficient sampling.

The sampling process then occurs within this learned representation space. The feature map $\tilde{\mathbf{X}}$ is partitioned into non-overlapping blocks, and a simple, learnable linear projection $\mathcal{F}_\Phi$ is applied to generate the measurements $\mathbf{Y}$. The kernel Φ has a size of $P \times P$ with a stride of P, and the sampling rate is denoted by $r \in (0, 1]$. In our configuration, P is set to 32. This sampling subnetwork, $\mathbf{S}(\cdot)$, is formulated as:

$$\mathbf{Y} = \mathbf{S}(\mathbf{X}) = \mathcal{F}_\Phi(\text{Encoder}(\mathbf{X})). \tag{3}$$

Initial recovery is performed symmetrically. A deconvolution operation $\mathcal{F}_\Phi^\dagger$, which shares parameters with the sampling kernel Φ, provides an initial estimate of the latent representation, $\hat{\tilde{\mathbf{X}}} = \mathcal{F}_\Phi^\dagger(\mathbf{Y})$. This representation is then mapped back to the image domain by the CS Decoder, which mirrors the encoder's structure. The initial recovery subnetwork $\mathbf{S}^\dagger(\cdot)$, designed as a learnable pseudo-inverse of $\mathbf{S}(\cdot)$, is thus formulated as:

$$\hat{\mathbf{X}} = \mathbf{S}^\dagger(\mathbf{Y}) = \text{Decoder}(\mathcal{F}_\Phi^\dagger(\mathbf{Y})). \tag{4}$$

Crucially, the entire sampling and initial recovery pipeline is trained end-to-end with the subsequent deep reconstruction network. This joint optimization ensures that the sampler learns to produce measurements that are highly informative for the reconstructor.

A key advantage of the ~~RCS~~**proposed** framework is its inherent domain coherence. In many deep CS models, iterative reconstruction algorithms (e.g., PGD-based) require a fidelity term to enforce consistency with the measurements. This often requires switching between the measurement domain and the image/feature domain, a process prone to introducing misalignments and information loss. In contrast, RCS performs sampling on a deep representation that resides in the same domain as the features manipulated in the subsequent reconstruction stages. Consequently, when the fidelity term is applied (as detailed in Eq. (5)), the comparison and correction steps occur between features within the same domain. This mitigates alignment errors and leads to a more stable and effective optimization process. This straightforward yet principled design is a key contributor to the superior performance of RHT-Net.

C. Deep Reconstruction Stages

In each stage of the deep reconstruction, we apply the Deep Gradient Descent (DGD) module to iteratively optimize the reconstructed features. As shown in Fig. 5(a), the DGD module unfolds the computation of the fidelity update into a neural network:

$$\mathcal{M}_{\text{DGD}}(\hat{\mathbf{X}}) = \hat{\mathbf{X}} + \mathcal{F}_{res}(\mathcal{F}_f(\mathbf{S}^\dagger(\mathbf{S}(\mathcal{F}_o(\hat{\mathbf{X}})) - \mathbf{Y}))), \tag{5}$$

where the nonlinear network $\mathcal{F}_{res}(\cdot)$ strengthens the capacity to capture residual information from the observations, as detailed in Fig. 5(c). The $\mathcal{F}_o(\cdot)$ and $\mathcal{F}_f(\cdot)$ serve as projection mappings between the observation domain and the deep feature domain, both implemented through simple biased 1×1 convolutions. The DGD module aligns the deep reconstruction features with the observation-domain residual information at each stage, ensuring consistency between the optimized feature $\hat{\mathbf{X}}$ and the observation $\mathbf{Y}$.

We designed a tri-scale sparse denoising (TSSD) subnetwork to replace the proximal projection, as depicted in Fig. 5(b). Traditional proximal projection ensures feature fidelity through the intermediate variable $\mathbf{R}^{(k)}$ and incorporates a regularization term $\psi(\mathbf{X})$ and a step size λ to enforce sparsity in the features:

$$\hat{\mathbf{X}}^{(k)} = \min_{\mathbf{X}} \frac{1}{2} \parallel \mathbf{X} - \mathbf{R}^{(k)} \parallel_2^2 + \lambda\psi(\mathbf{X}). \tag{6}$$

In contrast, the TSSD module learns more flexible image priors from training data for constraint, thereby avoiding fixed-threshold proximal operations. TSSD performs three downsampling operations on the features using stride-2 2×2 convolutions with GeLU activation function [54], maintaining the channel number constant, while successively reducing the feature map sizes to $\hat{\mathbf{X}}_\downarrow \in \mathbb{R}^{\mathrm{C} \times \frac{\mathrm{H}}{2} \times \frac{\mathrm{W}}{2}}$, $\hat{\mathbf{X}}_{\downarrow\downarrow} \in \mathbb{R}^{\mathrm{C} \times \frac{\mathrm{H}}{4} \times \frac{\mathrm{W}}{4}}$, and $\hat{\mathbf{X}}_{\downarrow\downarrow\downarrow} \in \mathbb{R}^{\mathrm{C} \times \frac{\mathrm{H}}{8} \times \frac{\mathrm{W}}{8}}$, , thereby generating multi-scale feature representations. Subsequently, these low-resolution features are upsampled by a factor of $\times 2$ using linear interpolation and concatenated with their corresponding resolution features through residual connections. This is followed by 5×5 convolutions to aggregate feature information over a larger receptive field. In addition, we introduce auxiliary features $\mathbf{Z}^{(k)} = \{\mathbf{X}_\downarrow^{(k)}, \mathbf{X}_{\downarrow\downarrow}^{(k)}, \mathbf{X}_{\downarrow\downarrow\downarrow}^{(k)}\}$ for cross-level parallel feature fusion, thereby optimizing the feature representation. Finally, the residual block merges the optimized features with the original $\mathbf{X}^{(k)}$ along the channel dimension and outputs the result through a 3×3 convolution. The TSSD network efficiently and lightly processes features at three resolutions in parallel, further enhancing denoising effects across multiple iterations and reinforcing the memory effect of tri-scale reconstruction information.

To overcome the limitations of the convolutional window while maintaining a coherent mathematical interpretability, we have designed a Transformer module for projected dual-domain feature modeling and enhanced feature dependencies that performs a refined complementary fusion of gradient descent terms. Specifically, this module includes a Global Sparse Cross-Attention (GCA) module, a Window-based Local Attention (WLA) module, and a Feed Forward Network (FFN) module. The specific inputs to the GCA module, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, are respectively projected into new components $\hat{\mathbf{Q}}$, $\hat{\mathbf{K}}$, and $\hat{\mathbf{V}}$, with additional details as shown in Fig. 5(d) and Fig. 5(e). This leads to the generation of a transposed attention map

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCSVT.2025.3614371,

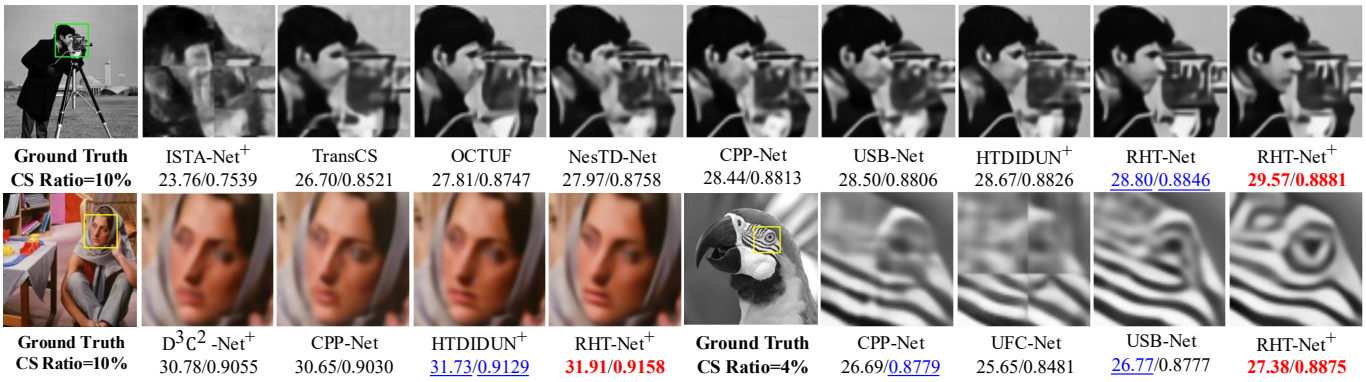


Fig. 6. Comparison of image reconstructions with SOTA methods at low sampling rates (10% and 4%) reveals the superior performance of RHT-Net. RHT-Net consistently outperforms in terms of facial fidelity, texture preservation, artifact reduction, and noise suppression.

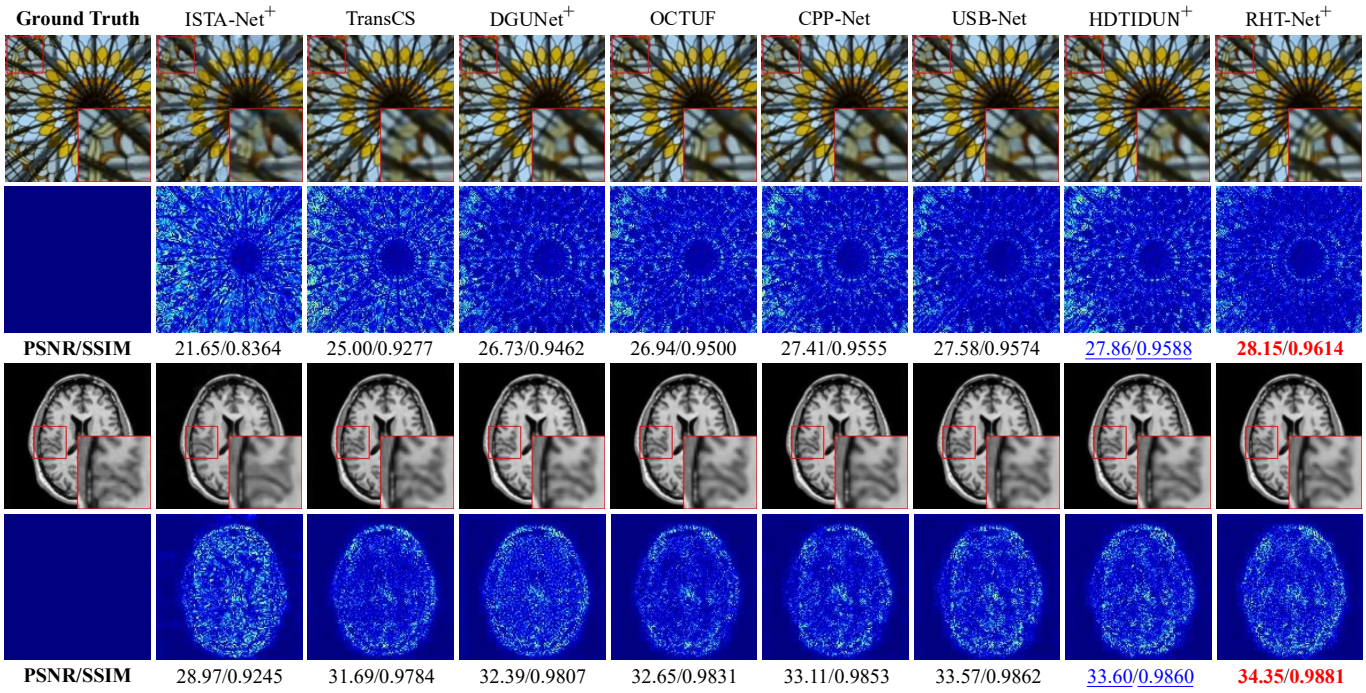


Fig. 7. Comparisons of the recovery of two images from Urban100 [55] (top) and brain [9] (bottom) at the sampling rate 30%, along with the residual to the ground truth images.

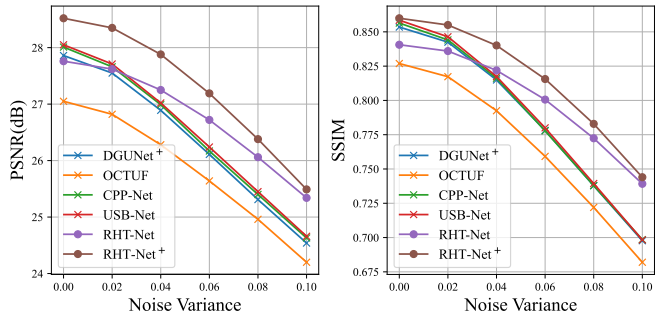


Fig. 8. Comparison of robustness to Gaussian noise.

for cross-CS inter-stage features $\mathbf{A} = \text{Softmax}(\hat{\mathbf{Q}}\hat{\mathbf{K}}^\top)$. The computation for the GCA module $\mathcal{M}_{\text{GCA}}(\cdot)$ is defined as

$$\mathcal{M}_{\text{GCA}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Conv}_{1\times1}(\mathcal{R}(\mathbf{A}\hat{\mathbf{V}})), \tag{7}$$

where $\mathcal{R}(\cdot)$ represents the reshape operation. In the specific iterations of the network, we use the pseudo-inverse of the measurement features $\mathbf{S}^\dagger(\mathbf{Y})$ as the query component $\mathbf{Q}_1$ for cross-attention calculation. The key-value pairs $\mathbf{K}_1$ and $\mathbf{V}_1$ are derived from the current reconstruction result $\hat{\mathbf{X}}$:

$$\mathbf{Q}_1,\ (\mathbf{K}_1,\ \mathbf{V}_1) = \text{Conv}_{1\times1}(\mathbf{S}^\dagger(\mathbf{Y})),\ \hat{\mathbf{X}}, \tag{8}$$

$$\mathbf{Q}_2,\ (\mathbf{K}_2,\ \mathbf{V}_2) = \hat{\mathbf{X}},\ \mathcal{M}_{\text{GCA}}(\mathbf{Q}_1, \mathbf{K}_1, \mathbf{V}_1). \tag{9}$$

By computing the cross-domain feature interaction $\mathcal{M}_{\text{GCA}}(\mathbf{Q}_1, \mathbf{K}_1, \mathbf{V}_1)$, we interpret it as a potential error term for the PGD fidelity constraint, then feed it back as the $\mathbf{K}_2$ and $\mathbf{V}_2$ components for the second stage GCA, with the current $\hat{\mathbf{X}}$ serving as the new $\mathbf{Q}_2$ for a secondary

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCSVT.2025.3614371,

TABLE I
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART (SOTA) METHODS ON THE BSD68 AND DIV2K DATASETS. PERFORMANCE IS EVALUATED USING PSNR (DB)(↑) AND SSIM(↑). INFERENCE TIME (INF. TIME (MS)(↓)), NUMBER OF PARAMETERS (PARAMS (M)(↓)), AND COMPUTATIONAL COMPLEXITY (FLOPS (G)(↓)) ARE ALSO REPORTED (FOR THESE THREE METRICS, LOWER VALUES ARE BETTER). THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN **RED** AND BLUE, RESPECTIVELY.

dataset	methods	PSNR/SSIM Ratio=50%	Ratio=40%	Ratio=30%	Ratio=25%	Ratio=10%	Inf. Time (ms)	Params (M)	FLOPs (G)
BSD68	ISTA-Net$^+$ (CVPR'18)	34.92/0.9489	32.94/0.9234	31.00/0.8860	29.95/0.8595	25.73/0.7055	7.13	0.34	36.11
	AMP-Net (TIP'20)	37.39/0.9755	35.35/0.9608	33.30/0.9373	32.21/0.9204	28.27/0.8162	73.98	0.97	28.01
	MADUN (ACM MM'21)	36.70/0.9642	34.40/0.9427	32.40/0.9126	31.33/0.8914	27.29/0.7685	27.38	3.02	207.77
	COAST (TIP'21)	35.68/0.9559	33.74/0.9335	31.81/0.9011	-/-	26.90/0.7530	22.93	1.12	79.84
	DGUNet$^+$ (CVPR'22)	37.73/0.9761	-/-	-/-	32.77/0.9258	28.97/0.8324	24.28	7.04	98.06
	TransCS (TIP'22)	37.43/0.9743	35.32/0.9588	32.96/0.9323	32.18/0.9178	28.15/0.8123	298.38	1.79	383.23
	DPC-DUN (TIP'23)	36.36/0.9621	34.30/0.9415	32.29/0.9109	31.23/0.8900	27.25/0.7676	29.12	1.63	90.17
	OCTUF (CVPR'23)	37.76/0.9760	35.68/0.9616	33.73/0.9399	32.70/0.9243	28.85/0.8267	38.00	0.62	21.82
	PRL (IJCV'23)	37.83/0.9759	35.76/0.9618	33.78/0.9405	-/-	29.06/0.8334	86.48	10.37	61.02
	CSformer (TIP'23)	37.70/0.9788	-/-	-/-	32.30/0.9259	28.60/0.8296	1424.23	6.67	379.82
	LTWIST (TCSVT'23)	-/-	-/-	-/-	32.21/0.9169	28.39/0.8151	171.83	23.28	139.83
	TCS-Net (TCI'23)	-/-	-/-	-/-	31.46/0.9146	27.94/0.8100	7.10	0.58	7.04
	NesTD-Net (TIP'24)	-/-	35.94/0.9629	33.76/0.9407	32.82/0.9261	28.75/0.8298	19.28	5.85	370.57
	MTC-CSNet (TCYB'24)	-/-	-/-	-/-	32.02/0.9210	28.38/0.8206	13.13	0.98	20.61
	UFC-Net (CVPR'24)	-/-	-/-	-/-	32.27/0.9153	28.46/0.8149	1458.20	1.77	108.79
	CPP-Net (CVPR'24)	37.67/0.9751	35.64/0.9608	33.69/0.9395	32.69/0.9245	28.95/0.8308	101.25	12.48	153.66
	D^{3}C^2-Net$^+$ (TCSVT'24)	37.88/0.9764	-/-	33.80/0.9411	-/-	28.99/0.8326	68.45	3.17	142.86
	HTDIDUN$^+$ (TCI'25)	38.04/0.9762	35.86/0.9620	33.86/0.9407	32.83/0.9252	29.07/0.8325	44.18	10.06	607.82
	USB-Net (TIP'25)	37.71/0.9753	35.69/0.9612	33.74/0.9403	32.71/0.9249	28.95/0.8312	81.83	15.64	96.09
	RHT-Net (ours)	38.62/0.9784	36.55/0.9658	34.33/0.9445	33.34/0.9315	29.46/0.8371	36.47	0.73	134.34
	RHT-Net$^+$ (ours)	**39.00/0.9789**	**36.76/0.9662**	**34.69/0.9472**	**33.65/0.9334**	**29.74/0.8417**	94.30	2.20	271.02
DIV2K	ISTA-Net$^+$ (CVPR'18)	38.26/0.9634	36.17/0.9450	33.98/0.9164	32.72/0.8942	27.60/0.7501	5.88	0.34	36.11
	AMP-Net (TIP'20)	40.66/0.9812	38.63/0.9710	36.37/0.9539	35.14/0.9411	30.36/0.8536	73.38	0.97	28.01
	MADUN (ACM MM'21)	40.28/0.9743	38.34/0.9624	36.21/0.9426	34.96/0.9272	29.98/0.8267	26.20	3.02	207.77
	COAST (TIP'21)	39.22/0.9691	37.27/0.9543	35.15/0.9310	-/-	29.31/0.8056	22.23	1.12	79.84
	DGUNet$^+$ (CVPR'22)	42.24/0.9843	-/-	-/-	36.58/0.9508	31.79/0.8780	21.55	7.04	98.06
	TransCS (TIP'22)	41.06/0.9808	38.91/0.9702	35.99/0.9491	35.33/0.9405	30.26/0.8523	292.30	1.79	383.23
	DPC-DUN (TIP'23)	40.28/0.9745	38.23/0.9615	36.05/0.9412	34.80/0.9258	29.93/0.8258	28.22	1.63	85.49
	OCTUF (CVPR'23)	41.99/0.9835	39.79/0.9742	37.54/0.9595	36.35/0.9487	31.49/0.8706	34.36	0.62	21.82
	PRL (IJCV'23)	42.11/0.9838	39.97/0.9750	37.76/0.9610	-/-	31.91/0.8800	64.48	10.37	61.02
	CSformer (TIP'23)	41.17/0.9833	-/-	-/-	35.31/0.9452	30.79/0.8666	1418.90	6.67	379.82
	LTWIST (TCSVT'23)	-/-	-/-	-/-	35.38/0.9404	30.76/0.8576	167.10	23.28	139.83
	TCS-Net (TCI'23)	-/-	-/-	-/-	34.33/0.9365	30.11/0.8497	6.90	0.58	7.04
	NesTD-Net (TIP'24)	-/-	39.85/0.9745	37.37/0.9590	36.44/0.9496	31.21/0.8708	18.42	5.85	370.57
	MTC-CSNet (TCYB'24)	-/-	-/-	-/-	34.73/0.9383	30.37/0.8541	11.83	0.98	20.61
	UFC-Net (CVPR'24)	-/-	-/-	-/-	35.78/0.9421	31.02/0.8612	1381.37	1.77	108.79
	CPP-Net (CVPR'24)	42.01/0.9834	39.85/0.9743	37.65/0.9601	36.41/0.9496	31.70/0.8752	99.52	12.48	153.66
	D^{3}C^2-Net$^+$ (TCSVT'24)	42.30/0.9842	-/-	37.89/0.9617	-/-	31.89/0.8794	56.20	3.17	142.86
	HTDIDUN$^+$ (TCI'25)	42.48/0.9842	40.34/0.9757	38.11/0.9622	36.87/0.9518	32.09/0.8809	43.56	10.06	607.82
	USB-Net (TIP'25)	42.04/0.9836	39.94/0.9748	37.72/0.9609	36.46/0.9500	31.70/0.8755	78.85	15.64	96.09
	RHT-Net (ours)	42.49/0.9845	40.43/0.9763	38.07/0.9620	36.82/0.9525	32.14/0.8791	35.05	0.73	134.34
	RHT-Net$^+$ (ours)	**43.08/0.9853**	**40.74/0.9770**	**38.52/0.9646**	**37.31/0.9549**	**32.59/0.8854**	94.78	2.20	271.02

cross-attention calculation. At this point, $\mathcal{M}_{\text{GCA}}(\mathbf{Q}_2, \mathbf{K}_2, \mathbf{V}_2)$ serves as the refined supplementary term for the first stage of PGD after global information modeling. This approach enables enhanced information interaction in the projection step, guiding further updates of $\hat{\mathbf{X}}$:

$$\hat{\mathbf{X}}_{\text{GCA}} = \hat{\mathbf{X}} \oplus \mathcal{M}_{\text{GCA}}(\mathbf{Q}_2, \mathbf{K}_2, \mathbf{V}_2), \tag{10}$$

$$\hat{\mathbf{X}} = \hat{\mathbf{X}}_{\text{GCA}} \oplus \mathcal{M}_{\text{FFN}}(\hat{\mathbf{X}}_{\text{GCA}}). \tag{11}$$

Depth-wise convolutions (DConv) effectively counterbalance the high-frequency bias introduced by self-attention operations [56], [57]. Therefore, our customized FFN module, $\mathcal{M}_{\text{FFN}}(\cdot)$, includes Layer Normalization layers, linear layers, and an effective $\text{DConv}_{3\times3}$, as shown in Fig. 5(f). To promote sparsity in the Transformer model, we employ GeLU activation function in the hidden layers, thus partially replicating the effect of proximal projections, similar to a Gaussian denoiser.

To enhance the efficiency of local spatial feature modeling, we design the Window Local Attention (WLA) module $\mathcal{M}_{\text{WLA}}(\cdot)$. The WLA module leverages depthwise convolution with a sliding window to generate dynamic local attention maps $\mathbf{A}'$, and aggregates local features through channel-wise convolution, thereby avoiding the quadratic complexity introduced by traditional self-attention computation. Specifically:

$$\mathbf{A}', \ \mathbf{V}' = \text{DConv}_{\text{k}\times\text{k}}(\mathcal{F}_\sigma(\mathbf{W}_{\text{A}'}\hat{\mathbf{X}})), \ \mathbf{W}_{\text{V}'}\hat{\mathbf{X}}, \tag{12}$$

$$\mathcal{M}_{\text{WLA}}(\hat{\mathbf{X}}) = \text{Conv}_{1\times1}(\mathbf{A}' \odot \mathbf{V}'), \tag{13}$$

where $\mathbf{W}'_{\text{A}}$ and $\mathbf{W}'_{\text{V}}$ both represent 1×1 convolutional linear projections, and $\mathcal{F}_\sigma$ is the GeLU activation function. This configuration produces smoother attention distributions and enhances the local information representation capability of the input $\hat{\mathbf{X}}$, with the large window convolution kernel size, k, set to 11. The WLA output is then combined with the original features via residual connection and further processed by a customized Feed-Forward Network (FFN):

$$\hat{\mathbf{X}}_{\text{WLA}} = \hat{\mathbf{X}} + \mathcal{M}_{\text{WLA}}(\hat{\mathbf{X}}), \tag{14}$$

$$\hat{\mathbf{X}} = \hat{\mathbf{X}}_{\text{WLA}} + \mathcal{M}_{\text{FFN}}(\hat{\mathbf{X}}_{\text{WLA}}). \tag{15}$$

This ingenious Transformer architecture, combining global and local modeling strategies, effectively enhances the interaction and fusion of the dual-term features in Eq. (5) and Eq. (6),

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCSVT.2025.3614371,

not only deeply encoding feature details but also maintaining low computational overhead.

D. Loss Function

Similar to approaches in studies such as [40], [34], [12], the RHT-Net model defines the error between image pairs as $\Delta_i = \mathbf{X}_i - \hat{\mathbf{X}}_i$ and employs the mean squared error (MSE) loss function for end-to-end optimization:

$$\mathcal{L}(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \|\Delta_i\|_2^2, \tag{16}$$

where N represents the number of training sample pairs, and Θ denotes all trainable parameters.

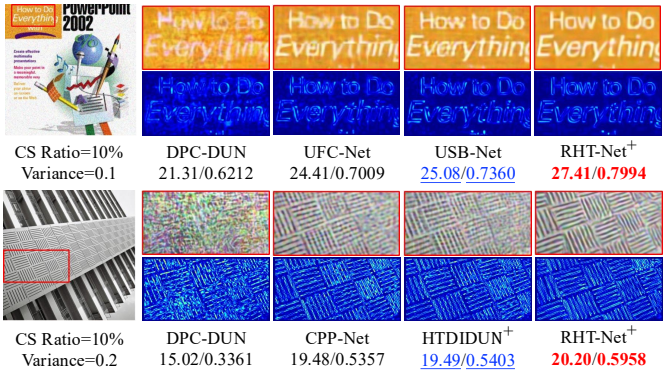


Fig. 9. Comparison of noise robustness with SOTA methods highlights the superior performance of RHT-Net. The arrows emphasize the significant advantages in detail preservation, with RHT-Net exhibiting the most accurate error map and achieving the highest denoising fidelity.

IV. Experiments

A. Implementation Details

Consistent with previous DCS methods [13], [34], [47], this study employs 400 images from the BSD500 dataset [58] to construct the training set, generating 80,000 training subimages with a resolution of 96×96 pixels through random flipping and cropping strategies. All color images are converted to the YCbCr color space, with only the Y channel extracted as training and testing input. For parameter optimization, we adopt the Adam optimizer [59], with momentum and weight decay coefficients set to 0.9 and 0.99999 respectively, batch size of 16, and a cosine annealing learning rate schedule [60], [28] implemented across all 150 training epochs, gradually reducing the learning rate from an initial value of 5×10^{-5} to 5×10^{-6}. Model performance evaluation is based on four general image CS reconstruction benchmark datasets: Set11 [61], BSDS [58], DIV2K [62], and Urban100 [55], using peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) as objective metrics to measure reconstruction quality. Additionally, this study comprehensively evaluates model computational efficiency from three dimensions: floating point operations (FLOPs), parameter size, and inference latency. For model configuration, we first constructed a lightweight model RHT-Net with 10 stages and 16 feature channel dimensions, followed by developing an enhanced version RHT-Net$^+$ by extending the number of stages to 14 and increasing feature channel dimensions to 32, aiming to achieve superior CS performance while maintaining reasonable computational complexity.

TABLE II
Quantitative comparison of noise robustness on the Urban100 dataset with a CS sampling rate of 10%.

Urban100		DGUNet$^+$	OCTUF	CPP-Net	USB-Net	RHT-Net	RHT-Net$^+$
PSNR	σ=0	27.86	27.05	28.01	28.05	27.76	**28.52**
	σ=0.02	27.55	26.82	27.66	27.71	27.62	**28.35**
	σ=0.04	26.89	26.28	26.99	27.02	27.25	**27.88**
	σ=0.06	26.11	25.64	26.17	26.24	26.72	**27.19**
	σ=0.08	25.31	24.96	25.4	25.45	26.06	**26.38**
	σ=0.1	24.54	24.2	24.63	24.66	25.34	**25.49**
	decay	-3.32	-2.85	-3.38	-3.39	**-2.42**	-3.03
	decay-rate	11.92%	10.54%	12.07%	12.09%	**8.72%**	10.62%
SSIM	σ=0	0.8535	0.8269	0.8564	0.8585	0.8406	**0.8599**
	σ=0.02	0.8425	0.8173	0.8441	0.8464	0.836	**0.855**
	σ=0.04	0.8149	0.7925	0.8162	0.8174	0.8219	**0.8401**
	σ=0.06	0.7781	0.7592	0.7778	0.78	0.8006	**0.8156**
	σ=0.08	0.7378	0.722	0.7379	0.7394	0.7724	**0.7829**
	σ=0.1	0.6977	0.682	0.6985	0.6983	0.7392	**0.744**
	decay	-0.1558	-0.1449	-0.1579	-0.1602	**-0.1014**	-0.1159
	decay-rate	18.25%	17.52%	18.44%	18.66%	**12.06%**	13.48%

B. Comparison with State-of-the-Art Methods

We conducted a comprehensive evaluation of our RHT-Net against a suite of state-of-the-art CS methods across five sampling rates: $r \in \{50\%, 40\%, 30\%, 25\%, 10\%\}$. The compared methods include ISTA-Net$^+$ [33], AMP-Net [63], MADUN [64], COAST [27], TransCS [38], DGUNet$^+$ [39], CASNet [40], DPC-DUN [41], OCTUF [34], PRL-PGD [36], CSformer [42], LTwIST [43], TCS-Net [32], D^{3}C^2-Net$^+$ [21], MTC-CSNet [44], UFC-Net [45], NesTD-Net [46], CPP-Net [12], HTDIDUN$^+$ [14], and USB-Net [13]. For a fair comparison, all competing models were evaluated using their publicly available source code and pre-trained weights, adhering to the default settings provided by the authors. **Following the standard practice in modern deep CS research [12], [14], [13], all compared methods adopt their respective optimized sampling strategies and pre-trained models as provided by the original authors. This approach ensures that each method operates under its intended design conditions with customized sampling matrices, which is the established evaluation protocol in the CS community [33], [39], [34]. Such methodology allows for fair assessment of the inherent reconstruction capabilities of different algorithms under their optimal configurations.**

As presented in Table I, both RHT-Net and its enhanced version, RHT-Net$^+$, consistently outperform all other methods across all sampling rates in terms of PSNR and SSIM. The performance advantage is particularly pronounced at very low sampling rates. For instance, on the DIV2K dataset at a 10% sampling rate, RHT-Net$^+$ surpasses the next-best method, HTDIDUN$^+$ [14], by a significant margin of 0.50 dB in PSNR, while the lightweight RHT-Net also shows competitive results. Notably, when compared to USB-Net [13], RHT-Net achieves a 0.44 dB gain while using only $1/20$ of the parameters and boasting a $2.25\times$ speedup in inference time. Moreover, the more efficient Transformer also brings competitive FLOPs.

Fig. 6 displays visual comparisons of reconstructed images at low CS ratios, demonstrating that RHT-Net produces visually superior reconstructions with sharper textures and finer

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCSVT.2025.3614371,

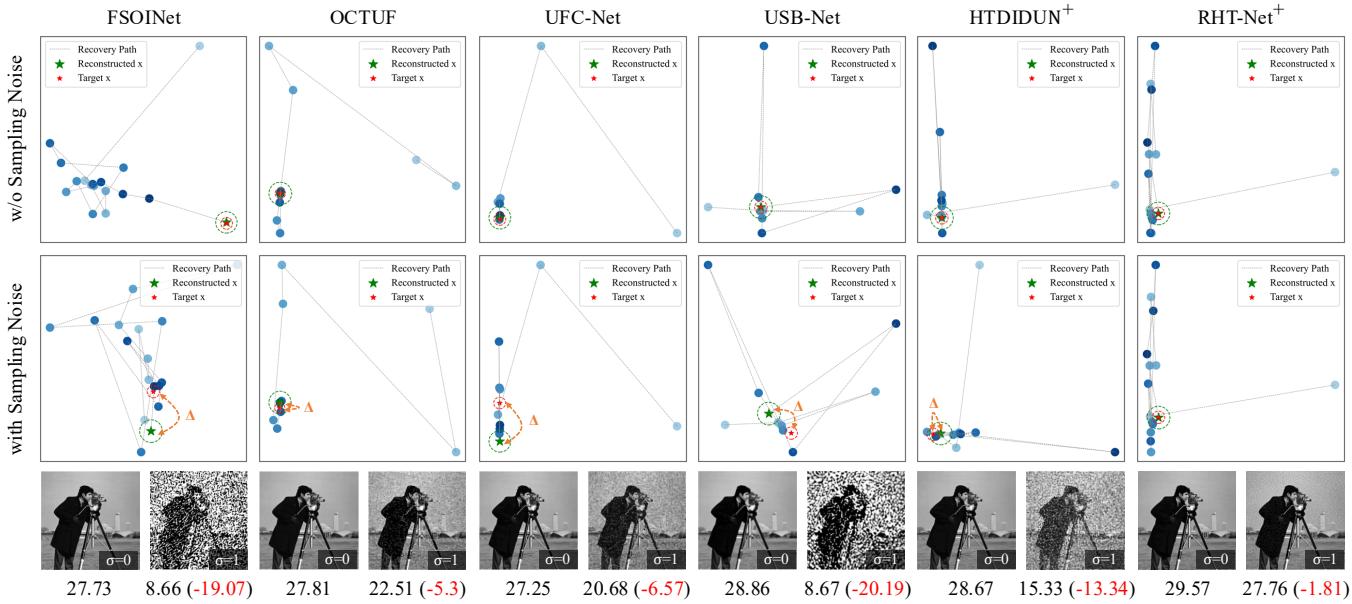


Fig. 10. Visualization of the recovery trajectories using PCA at a sampling rate of 10%. The sampled measurements are given by $\mathbf{y} = \mathbf{A}\mathbf{x} + \epsilon$, where ϵ denotes the noise term. Under noise standard deviations $\sigma = 0, 1$, RHT-Net demonstrates superior stability during the recovery process under strong interference compared to existing state-of-the-art unfolding networks.

details, alongside a marked reduction in noise and artifacts. Further analysis reveals that our method excels on images with complex structures and rich textures. Our RHT-Net$^+$ achieves up to a 0.9 dB improvement over the next-best method (HTDIDUN$^+$ [14]), demonstrating enhanced fidelity in edge and detail preservation. This superior performance stems from the synergistic effects of our representation-domain sampling and adaptive proximal projection mechanism, which effectively mitigate the ill-posed nature of the CS problem, especially at low sampling rates.

We also conducted visualization comparisons of various mainstream methods, including residual images between reconstructed images and original images, to comprehensively verify the effectiveness of the proposed method. Specifically, we evaluated the model's zero-shot generalization capability on the CS-MRI (Compressed Sensing Magnetic Resonance Imaging) dataset (i.e., without fine-tuning on MRI-specific sampling matrices and discrete Fourier transform-based sampling methods). As shown in Fig. 7, it demonstrates reconstruction results at the sampling ratio of 30% on the Urban100 [55] and brain MR dataset [9]. Experimental results show that the proposed method outperforms existing methods in both edge and detail reconstruction, while other methods often exhibit detail distortion or artifacts. These results further verify the effectiveness of our method in image compressed sensing tasks and establish a solid foundation for its application in real-world CS-MRI, video compressed sensing, and other inverse problems.

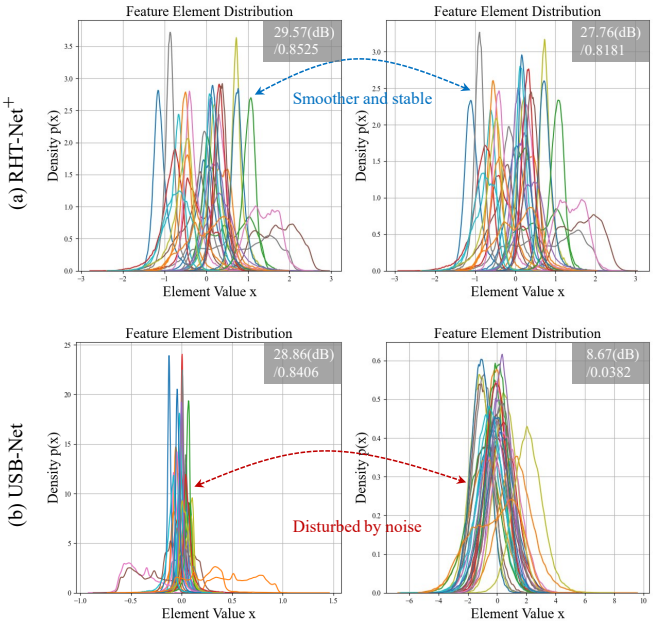


Fig. 11. Visualization of the feature distributions for RHT-Net$^+$ (a) and USB-Net [13] (b) at the penultimate stage, under both noise-free and noisy conditions. Under identical measurements noise interference, RHT-Net exhibits a consistent, stable, and smooth feature distribution. This robustness is attributed to its advanced feature domain sampling and the U-shaped module's effectiveness in preserving multi-scale structures and feature shapes.

C. Noise Robustness

In practical application scenarios, compressive imaging models often face challenges from unknown noise interference. To evaluate the robustness of RHT-Net in image reconstruction under noisy environments, we injected Gaussian noise with standard deviations σ into the original images of the Urban100 dataset, and performed sampling and reconstruction under a 10% sampling rate using RHT-Net and its main competitive methods. Table II and Fig. 8 show the PSNR performance changes of each method under different levels

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCSVT.2025.3614371,

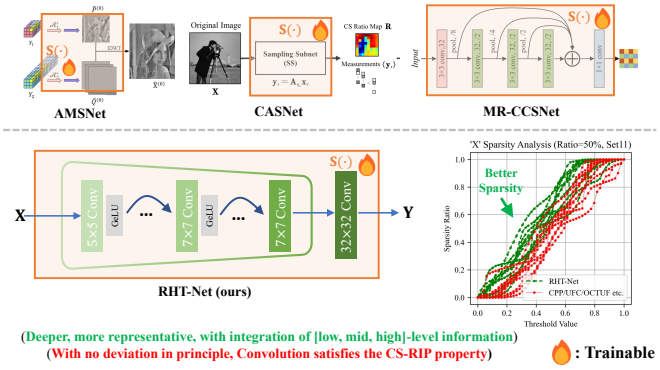


Fig. 12. Key distinctions between our representation domain sampling and other trainable sampling schemes. While some existing methods attempt to move beyond simple linear convolutions or projection matrices in pursuit of improved representations, they remain limited by their reliance on low-level features. The proposed sampling scheme achieves sparser representation and compact measurements than prior methods.

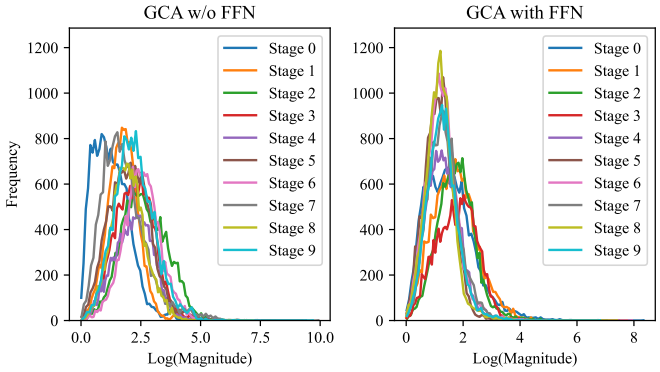


Fig. 13. The visualization of the frequency distribution of intermediate feature maps, after applying the Fourier shift transformation across 10 iterative stages of RHT-Net, demonstrates that the FFN significantly amplifies the low-pass filtering effect induced by the self-attention mechanism.

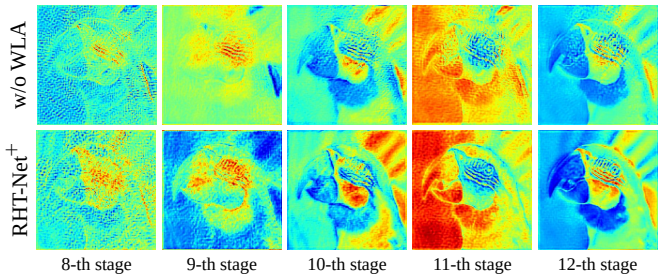


Fig. 14. Visual analysis of the heatmap changes in iterative stages of RHT-Net for simulating local attention. It reveals that the WLA blocks improve the fidelity of local dependency focus.

of Gaussian noise. The results indicate that RHT-Net exhibits significant robustness advantages under noise interference, maintaining the highest reconstruction quality under all test conditions. This advantage is primarily attributed to the adaptive proximal projection TSSD subnetwork in the model, which can efficiently process multi-scale noise and achieve cross-scale information reconstruction. The visualizations presented in Fig. 9 provide additional evidence supporting the aforementioned observations.

In addition to inherent noise in the original signal, noise interference introduced during the measurement process is equally prevalent. To deeply understand the reconstruction characteristics of different unfolded architectures, we conducted visualization analysis of image and feature processing processes of various networks from the perspective of noise robustness. We selected barbara.tif from the Set11 dataset as a test sample, and demonstrated the recovery trajectories of six advanced unfolding networks in Fig. 10: including FSOINet [28], OCTUF [34], UFC-Net [45], USB-Net [13], HTDIDUN$^+$ [14], and our proposed RHT-Net, all methods using a unified 10% sampling rate. We employed Principal Component Analysis (PCA) for dimensionality reduction and visualization, chosen for its simplicity, stability, and independence from hyperparameters or stochastic factors. We projected high-dimensional features in the RHT-Net network through feature projection from the feature domain to the image domain uniformly, and converted them into $1 \times (H \times W)$ format for PCA calculation, to maintain dimensional consistency with other methods. The visualization results in Fig. 10 indicate that, under noise-free conditions, all methods are capable of achieving relatively accurate reconstruction. However, in the presence of noise, all models experience performance degradation to varying extents. This effect is particularly pronounced in shallow pixel-domain sampling architectures, where the observation noise ϵ in the sampled measurements $\mathbf{y} = \mathbf{A}\mathbf{x} + \epsilon$ induces a cumulative amplification effect, leading to a significant "butterfly effect" on low-level features. Specifically, the reconstruction results of USB-Net [13] and FSOINet [28] exhibit severe block, honeycomb, and ripple artifacts, rendering the reconstructed image difficult to recognize. Correspondingly, the PSNR drops significantly to 20.19dB and 19.07dB, respectively. Similar artifacts were observed in other networks, including HTDIDUN$^+$ [14], UFC-Net [45], and OCTUF [34], with progressively smaller PSNR drops in that order. Notably, our RHT-Net achieved the best performance in suppressing recovery trajectory deviations. Under identical measurement noise interference, its measurement representation $\mathbf{y}$—which encodes rich hierarchical feature attributes, including high-level semantics, contours, and critical edge details—enabled a minimal PSNR drop of only 1.81dB, demonstrating the strongest noise robustness among all compared methods.

It should be noted that USB-Net [13] and FSOINet [28] employ restricted soft-thresholding or simple denoising mechanisms to maintain feature sparsity, resulting in significant performance degradation in noisy environments. Experimental results verify the effectiveness of our proposed compact and efficient multi-scale denoising structure and long-range

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCSVT.2025.3614371,

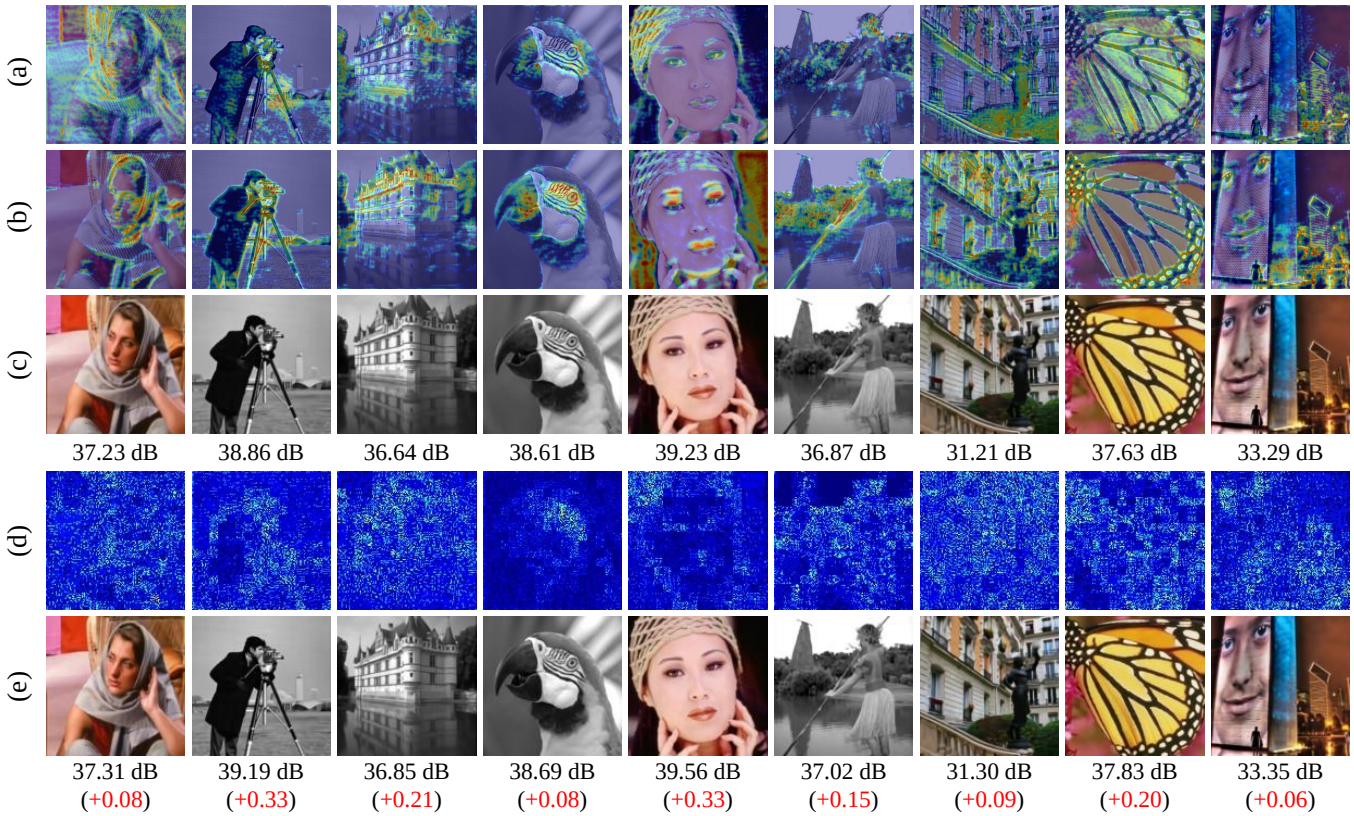


Fig. 15. Visual effects of WLA module. (a) and (b) are attention maps generated by the WLA module at the 4th and 8th stages of RHT-Net, respectively; (c) shows the reconstruction result without WLA; (e) represents the final reconstruction result of RHT-Net; (d) is the residual map between (c) and (e). This indicates that WLA can tailor-make localized attention to improve reconstruction quality, leading to enhanced detail in finer-grained areas and edge information.

TABLE III
COMPREHENSIVE MODULE ABLATION STUDY: PROGRESSIVE ADDITION AND COMPONENT REMOVAL (BSD68, $r = 10\%$).

Case	Configuration	PSNR (dB)	SSIM	Params (M)	Performance Change
Part A: Progressive Module Addition					
A1	Baseline Network	27.34	0.7400	0.51	-
A2	+RCS Sampling	27.97	0.7850	0.62	+0.63
A3	+RCS+TSSD	28.24	0.8200	1.56	+0.27
A4	+RCS+TSSD+GCA	28.35	0.8310	1.79	+0.11
A5	+RCS+TSSD+GCA+WLA	28.43	0.8420	1.92	+0.08
A6	**RHT-Net Complete (+FFN)**	**28.51**	**0.8597**	**2.02**	**+0.08**
Part B: Component Removal Verification					
B1	**RHT-Net Complete**	**28.51**	**0.8597**	**2.02**	**-**
B2	w/o RCS (Random Gaussian Sampling)	27.85	0.8150	1.91	-0.66
B3	w/o TSSD (Dual-scale U-Net)	28.21	0.8350	0.62	-0.30
B4	w/o GCA (Remove Global Cross Attention)	28.43	0.8420	1.79	-0.08
B5	w/o WLA (Remove Window Local Attention)	28.35	0.8310	1.92	-0.16
B6	w/o FFN (Remove Feed Forward Network)	28.42	0.8410	1.92	-0.09
Total Improvement (A1→A6)		**+1.17 dB**	**+0.1197**	**+1.51 M**	**-**

connections in enhancing noise robustness, which is crucial for achieving stable 14-stage reconstruction. Additionally, we extracted recovery features from the penultimate stage of RHT-Net$^+$ and USB-Net [13] to analyze their data distribution characteristics. For noise standard deviation $\epsilon \in \{0, 1\}$, Fig. 11 shows two feature distributions curves. The results indicate that under noise interference, the feature-domain recovery signal distribution of USB-Net [13] becomes sharp and unstable, leading to significant degradation in image content and structural consistency, with PSNR and SSIM dropping by 20.19 dB and 0.8024, respectively. In contrast, our RHT-Net$^+$ (decreased by only 1.81dB/0.0344) maintains spatial structure smoothly and stably by decomposing the image into multiple channels and maintaining a more balanced distribution. RHT-Net$^+$ presents a broader and sparser mean distribution in 32-dimensional channel space, which helps enhance edge and texture details. These results further confirm the necessity of feature domain sampling and recovery, supporting high-throughput information transmission and retaining sufficient representation freedom to enhance network robustness.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCSVT.2025.3614371,

TABLE IV
EFFECT OF THE FEATURE CHANNELS AND STAGES OF OUR MODEL UNDER THE SET11 DATASET.

Cases	Channels	Stages	PSNR(dB)	SSIM	Parameters(M)	
					Sampling	Reconstruction
1	8	8	36.59	0.9646	0.058	0.391
2	16	8	37.57	0.9685	0.058	0.588
3	16	10	37.74	0.9694	0.058	0.657
4	32	10	38.27	0.9713	0.058	1.606
5	32	14	38.53	0.9722	0.058	2.123
6	48	16	38.67	0.9731	0.058	4.865
7	64	16	39.01	0.9740	0.058	8.316

D. Ablation Studies

To further explore the effect of input/output channel numbers in each stage on performance, we conducted experiments with $C \in \{8, 16, 32, 48, 64\}$ under different reconstruction stages. As shown in Table IV, when $C \leq 32$, increasing the number of channels significantly improves reconstruction performance, indicating that the number of channels is crucial for the model's feature expression capability. Increasing the number of deep reconstruction stages also helps improve performance, but when the number of stages exceeds 10, the gain tends to saturate, indicating that there is still room for optimization in parameter compression. Considering both performance and complexity, we finally selected Case 3 and Case 5 as the recommended configurations for RHT-Net and RHT-Net$^+$.

To systematically evaluate the effectiveness of each proposed component, we conduct comprehensive ablation studies using both progressive addition and component removal strategies. Following the methodology of competing methods such as OCTUF [34], D^{3}C^2-Net$^+$ [21], CPP-Net [12], and USB-Net [13], we adopt the sampling and proximal optimization modules from FSOINet [28] as the baseline for fair comparison.

The comprehensive ablation study results, presented in Table III, quantify the individual contribution of each component through two complementary approaches: (1) Progressive Addition (Part A): starting from the baseline network and incrementally adding each module to demonstrate cumulative improvements; and (2) Component Removal Verification (Part B): removing individual components from the complete model to validate their necessity. This dual-approach methodology ensures robust validation of each component's contribution.

The progressive addition analysis reveals that the RCS module provides the largest single-module benefit (+0.63 dB, A1→A2), demonstrating the fundamental importance of representation-domain sampling. The TSSD module contributes an additional +0.27 dB improvement (A2→A3), while the hybrid Transformer components (GCA, WLA, FFN) collectively contribute +0.27 dB (A3→A6). The component removal verification confirms these findings, with RCS removal causing the largest performance drop (-0.66 dB, B1→B2), TSSD removal resulting in -0.30 dB loss (B1→B3), and individual Transformer sub-components showing losses ranging from -0.08 dB to -0.16 dB. As shown in Fig. 12, this highlights a key distinction from other trainable sampling schemes. While some prior methods attempt to go beyond simple linear projections, they are often still limited by operating on low-level features. In contrast, our RCS module samples from a deep, semantically rich representation space, resulting in more compact and meaningful measurements. This shift is crucial for mitigating the domain gap and is a primary factor contributing to the observed performance improvements.

To further validate the parameter efficiency of our proposed components, we conduct a detailed sub-module efficiency analysis as shown in Table V. This analysis demonstrates that each component introduces only minimal overhead while achieving substantial performance gains.

The efficiency analysis reveals that the RCS module achieves exceptional parameter efficiency (5.73 dB/M), with mixed kernels ($5 \times 5, 7 \times 7$) providing optimal performance-parameter trade-offs. The TSSD module demonstrates consistent efficiency (0.27 dB/M) through its tri-scale design, while the hybrid Transformer maintains competitive efficiency (0.59 dB/M) despite its complexity. Notably, as shown in Fig. 13, the FFN module effectively compensates for the attention frequency distribution of GCA, and Fig. 14 demonstrates that the WLA module enhances fine-grained attention, validating the balanced design of our lightweight architecture. To further clarify the role of components within the TSSD and WLA modules, we also conducted more fine-grained component-level ablation experiments, with results summarized in Table VI and Table VII. Extensive experimental results demonstrate that RHT-Net's architecture achieves optimal performance across multiple evaluation metrics: (1) deeper and wider convolutional kernels significantly enhance attention map construction, (2) three-scale proximal operators surpass dual-scale alternatives and (3) unified multi-scale information circulation improves global feature interaction. These systematic improvements collectively contribute to the model's superior performance.

To further validate the effectiveness of the local spatial attention module WLA, we conducted ablation experiments with visual comparisons. Fig. 15 presents the results of five ablation experiments, where (a) and (b) correspond to the visualization of the WLA attention maps in the 4th and 8th stages respectively. The results indicate that WLA adaptively adjusts the local attention mechanism across different reconstruction stages, effectively enhancing reconstruction quality and further highlighting the superiority of the proposed WLA module. Furthermore, (c) and (e) compare the reconstruction results with and without WLA. The quantitative PSNR results demonstrate that WLA offers a significant advantage in detail recovery, while maintaining extremely low computational overhead. The residual image in (d) further highlights that the reconstruction results using WLA excel in fine-grained areas and edge preservation, leading to a substantial overall performance improvement.

V. CONCLUSION

In this paper, we propose a representation-domain compressed sensing model based on deep unrolling

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCSVT.2025.3614371,

TABLE V
SUB-MODULE PARAMETER EFFICIENCY ANALYSIS (BSD68, $r = 10\%$).

Module	Configuration	Key Changes	PSNR (dB)	Params (M)	Efficiency (+dB/M)
+RCS	Baseline (Random Sampling)	-	27.34	0.51	-
	Small Kernels	3×3 only	27.68	0.58	4.86
	Mixed Kernels	$5 \times 5, 7 \times 7$	27.97	0.62	5.73
	w/o Activation	Remove GeLU	27.66	0.62	2.91
	w/o Skip Connections	Remove residual	27.72	0.62	3.45
	RCS Complete	**Optimal config**	**27.97**	**0.62**	**5.73**
+TSSD	Single-scale Basic	1 scale, no flow	28.05	0.85	-
	Dual-scale Standard	2 scales + skip	28.12	1.15	0.23
	Tri-scale Basic	3 scales + skip	28.18	1.35	0.26
	Tri-scale+Feature Flow	+ cross-scale flow	28.22	1.45	0.28
	TSSD Complete	**Full tri-scale**	**28.24**	**1.56**	**0.27**
+HT	No Transformer	Baseline	28.24	1.56	-
	GCA Only	Global cross attention	28.35	1.79	0.48
	WLA Only	Window local attention	28.32	1.69	0.62
	FFN Only	Feed-forward network	28.30	1.66	0.60
	GCA+WLA	Dual attention	28.43	1.92	0.53
	GCA+FFN	Global + FFN	28.41	1.89	0.52
	WLA+FFN	Local + FFN	28.38	1.82	0.54
	Complete Transformer	**GCA+WLA+FFN**	**28.51**	**2.02**	**0.59**

TABLE VI
ABLATION STUDY OF THE TSSD MODULE.

Case	mix(5×5)	mix(3×3)	UNet-2	UNet-3	feature flow	PSNR
7	–	–	–	–	–	37.12
8	–	√	√	–	–	37.21
9	√	–	√	–	–	37.28
10	√	–	–	√	–	37.34
11	–	√	–	√	√	37.33
TSSD	√	–	–	√	√	37.37

TABLE VII
ABLATION STUDY OF THE WLA MODULE.

Case	DConv(11×11)	DConv(5×5)	LN	GeLU	ReLU	PSNR
12	–	–	–	–	–	37.22
13	–	√	√	–	√	37.29
14	√	–	–	–	√	37.34
15	√	–	√	–	–	37.32
16	√	–	–	√	–	37.38
WLA	√	–	√	√	–	37.37

framework, dubbed RHT-Net. RHT-Net significantly reduces cross-domain loss by performing compact sampling in high-level representations. The model combines a tri-scale downsampling denoising module and a hybrid Transformer architecture, simulating sparse operators through a lightweight network and estimating finer recovery information via cross-attention. Experimental results show that RHT-Net significantly outperforms existing methods in terms of reconstruction quality, model efficiency, noise robustness, and inference speed, providing a novel solution for image inverse problems.

REFERENCES

[1] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006. 1

[2] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006. 1

[3] A. Liutkus, D. Martina, S. Popoff, G. Chardon, O. Katz, G. Lerosey, S. Gigan, L. Daudet, and I. Carron, "Imaging with nature: Compressive imaging using a multiply scattering medium," *Scientific Reports*, vol. 4, no. 1, p. 5552, 2014. 1

[4] Y. Liu, X. Yuan, J. Suo, D. J. Brady, and Q. Dai, "Rank minimization for snapshot compressive imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 2990–3006, 2018. 1

[5] Z. Cheng, B. Chen, R. Lu, Z. Wang, H. Zhang, Z. Meng, and X. Yuan, "Recurrent neural networks for snapshot compressive imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2264–2281, 2022. 1

[6] Y. Sun, X. Chen, M. S. Kankanhalli, Q. Liu, and J. Li, "Video snapshot compressive imaging using residual ensemble network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5931–5943, 2022. 1

[7] Z. Meng, X. Yuan, and S. Jalali, "Deep unfolding for snapshot compressive imaging," *International Journal of Computer Vision*, vol. 131, no. 11, pp. 2933–2958, 2023. 1

[8] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressive sensing mri," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 72–82, 2008. 1

[9] Y. Yang, J. Sun, H. Li, and Z. Xu, "ADMM-CSNet: A deep learning approach for image compressive sensing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 521–538, 2018. 1, 2, 6, 9

[10] D. Liang, J. Cheng, Z. Ke, and L. Ying, "Deep magnetic resonance image reconstruction: Inverse problems meet neural networks," *IEEE Signal Processing Magazine*, vol. 37, no. 1, pp. 141–151, 2020. 1

[11] P. Kong, A. Li, D. Guo, L. Zhou, and C. Qin, "Joint lossless compression and encryption for medical images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 6, pp. 4786–4799, 2023. 1

[12] Z. Guo and H. Gan, "CPP-Net: Embracing multi-scale feature fusion into deep unfolding CP-PPA network for compressive sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 25 086–25 095. 1, 2, 3, 8, 12

[13] ——, "USB-Net: Unfolding split bregman method with multi-phase feature integration for compressive imaging," *IEEE Transactions on Image Processing*, vol. 34, pp. 925–938, 2025. 1, 3, 8, 9, 10, 11, 12

[14] T. Li, Q. Yan, Y. Li, and J. Yan, "High-throughput decomposition-inspired deep unfolding network for image compressed sensing," *IEEE Transactions on Computational Imaging*, vol. 11, pp. 89–100, 2025. 1, 3, 8, 9, 10

[15] J. Hahn, C. Debes, M. Leigsnering, and A. M. Zoubir, "Compressive sensing and adaptive direct sampling in hyperspectral imaging," *Digital Signal Processing*, vol. 26, pp. 113–126, 2014. 1

[16] S. Zhang, H. Huang, and Y. Fu, "Fast parallel implementation of dual-camera compressive hyperspectral imaging system," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 11, pp. 3404–3414, 2018. 1

[17] T. Xie, L. Liu, and L. Zhuang, "Plug-and-play priors for multi-shot compressive hyperspectral imaging," *IEEE Transactions on Image Processing*, vol. 32, pp. 5326–5339, 2023. 1

[18] X. Zhang, B. Chen, W. Zou, S. Liu, Y. Zhang, R. Xiong, and J. Zhang, "Progressive content-aware coded hyperspectral snapshot compressive imaging," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1

[19] S. Liu, B. Chen, W. Zou, H. Sha, X. Feng, S. Han, X. Li, X. Yao, J. Zhang, and Y. Zhang, "Compressive confocal microscopy imaging at the single-photon level with ultra-low sampling ratios," *Communications Engineering*, vol. 3, no. 1, p. 88, 2024. 1, 2

[20] J. Zhang, B. Chen, R. Xiong, and Y. Zhang, "Physics-inspired compressive sensing: Beyond deep unrolling," *IEEE Signal Processing Magazine*, vol. 40, no. 1, pp. 58–72, 2023. 1

[21] W. Li, B. Chen, S. Liu, S. Zhao, B. Du, Y. Zhang, and J. Zhang, "D^3C^2-Net: Dual-domain deep convolutional coding network for compressive sensing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 10, pp. 9341–9355, 2024. 1, 2, 3, 8, 12

[22] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009. 1

[23] J. A. Tropp and S. J. Wrightt, "Computational methods for sparse solution o linear inverse problems," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 948–958, 2010. 1

[24] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009. 2

[25] L. Zeng, P. Yu, and T. K. Pong, "Analysis and algorithms for some compressed sensing models based on l1/l2 minimization," *SIAM Journal on Optimization*, vol. 31, no. 2, pp. 1576–1603, 2021. 2

[26] J. Zhang, C. Zhao, and W. Gao, "Optimization-inspired compact deep compressive sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 4, pp. 765–774, 2020. 2, 3

[27] D. You, J. Zhang, J. Xie, B. Chen, and S. Ma, "COAST: Controllable arbitrary-sampling network for compressive sensing," *IEEE Transactions on Image Processing*, vol. 30, pp. 6066–6080, 2021. 2, 8

[28] W. Chen, C. Yang, and X. Yang, "FSOINET: Feature-space optimization-inspired network for image compressive sensing," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 2460–2464. 2, 8, 10, 12

[29] J. Chen, Y. Sun, Q. Liu, and R. Huang, "Learning memory augmented cascading network for compressive sensing of images," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 513–529. 2, 3

[30] Z. Fan, F. Lian, and J. Quan, "Global sensing and measurements reuse for image compressive sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 8954–8963. 2, 3

[31] H. Gan, Y. Gao, C. Liu, H. Chen, T. Zhang, and F. Liu, "AutoBCS: Block-based image compressive sensing with data-driven acquisition and noniterative reconstruction," *IEEE Transactions on Cybernetics*, vol. 53, no. 4, pp. 2558–2571, 2021. 2, 3

[32] H. Gan, M. Shen, Y. Hua, C. Ma, and T. Zhang, "From patch to pixel: A transformer-based hierarchical framework for compressive image sensing," *IEEE Transactions on Computational Imaging*, vol. 9, pp. 133–146, 2023. 2, 3, 8

[33] J. Zhang and B. Ghanem, "ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1828–1837. 2, 3, 8

[34] J. Song, C. Mou, S. Wang, S. Ma, and J. Zhang, "Optimization-inspired cross-attention transformer for compressive sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6174–6184. 2, 3, 8, 10, 12

[35] H. Song, J. Gong, H. Meng, and Y. Lai, "Multi-cross sampling and frequency-division reconstruction for image compressed sensing," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024, pp. 4909–4917. 2

[36] B. Chen, J. Song, J. Xie, and J. Zhang, "Deep physics-guided unrolling generalization for compressive sensing," *International Journal of Computer Vision*, vol. 131, no. 11, pp. 2864–2887, 2023. 2, 3, 8

[37] J. Z. Jiechong Song, Bin Chen, "Deep memory-augmented proximal unrolling network for compressive sensing," *International Journal of Computer Vision*, vol. 131, no. 6, pp. 1477–1496, 2023. 2

[38] M. Shen, H. Gan, C. Ning, Y. Hua, and T. Zhang, "TransCS: A transformer-based hybrid architecture for image compressed sensing," *IEEE Transactions on Image Processing*, vol. 31, pp. 6991–7005, 2022. 3, 8

[39] C. Mou, Q. Wang, and J. Zhang, "Deep generalized unfolding networks for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 399–17 410. 3, 8

[40] B. Chen and J. Zhang, "Content-aware scalable deep compressed sensing," *IEEE Transactions on Image Processing*, vol. 31, pp. 5412–5426, 2022. 3, 8

[41] J. Song, B. Chen, and J. Zhang, "Dynamic path-controllable deep unfolding network for compressive sensing," *IEEE Transactions on Image Processing*, vol. 32, pp. 2202–2214, 2023. 3, 8

[42] D. Ye, Z. Ni, H. Wang, J. Zhang, S. Wang, and S. Kwong, "CSformer: Bridging convolution and transformer for compressive sensing," *IEEE Transactions on Image Processing*, vol. 32, pp. 2827–2842, 2023. 3, 8

[43] H. Gan, X. Wang, L. He, and J. Liu, "Learned two-step iterative shrinkage thresholding algorithm for deep compressive sensing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 3943–3956, 2024. 3, 8

[44] M. Shen, H. Gan, C. Ma, C. Ning, H. Li, and F. Liu, "MTC-CSNet: Marrying transformer and convolution for image compressive sensing," *IEEE Transactions on Cybernetics*, vol. 54, no. 9, pp. 4949–4961, 2024. 3, 8

[45] X. Wang and H. Gan, "UFC-Net: Unrolling fixed-point continuous network for deep compressive sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 25 149–25 159. 3, 8, 10

[46] H. Gan, Z. Guo, and F. Liu, "NesTD-Net: Deep NESTA-inspired unfolding network with dual-path deblocking structure for image compressive sensing," *IEEE Transactions on Image Processing*, vol. 33, pp. 1923–1937, 2024. 3, 8

[47] W. Shi, F. Jiang, S. Liu, and D. Zhao, "Image compressive sensing using convolutional neural network," *IEEE Transactions on Image Processing*, vol. 29, pp. 375–388, 2019. 3, 8

[48] W. Cui, S. Liu, F. Jiang, and D. Zhao, "Image compressive sensing using non-local neural network," *IEEE Transactions on Multimedia*, vol. 25, pp. 816–830, 2021. 3

[49] Y. Sun, J. Chen, Q. Liu, B. Liu, and G. Guo, "Dual-path attention network for compressive sensing image reconstruction," *IEEE Transactions on Image Processing*, vol. 29, pp. 9482–9495, 2020. 3

[50] K. Zhang, Z. Hua, Y. Li, Y. Chen, and Y. Zhou, "AMS-Net: Adaptive multi-scale network for image compressive sensing," *IEEE Transactions on Multimedia*, vol. 25, pp. 5676–5689, 2022. 3

[51] F. Shen and H. Gan, "HUNet: Homotopy unfolding network for image compressive sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 12 799–12 808. 3

[52] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *International Conference on Learning Representations (ICLR)*, 2017. 4

[53] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006. 4

[54] D. Hendrycks and K. Gimpel, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units," in *International Conference on Learning Representations (ICLR)*, 2017. 4, 5

[55] J. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5197–5206. 6, 8, 9

[56] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2021, pp. 12 116–12 128. 7

[57] Y. Zhou, Z. Li, C.-L. Guo, S. Bai, M.-M. Cheng, and Q. Hou, "SRFormer: Permuted self-attention for single image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 12 780–12 791. 7

[58] 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 *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2001, pp. 416–423. 8

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2015. 8

[60] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016. 8

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCSVT.2025.3614371,

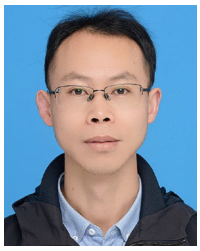
[61] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, and A. Ashok, "Reconnet: Non-iterative reconstruction of images from compressively sensed measurements," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 449–458. 8
[62] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 136–144. 8
[63] Z. Zhang, Y. Liu, J. Liu, F. Wen, and C. Zhu, "AMP-Net: Denoising-based deep unfolding for compressive image sensing," *IEEE Transactions on Image Processing*, vol. 30, pp. 1487–1500, 2020. 8
[64] J. Song, B. Chen, and J. Zhang, "Memory-augmented deep unfolding network for compressive sensing," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2021, pp. 4249–4258. 8

Heping Song received the Ph.D. degree in computer application technology from Sun Yat-sen University, Guangzhou, China. He is currently a Associate Professor with the Department of Software Engineering, School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China. His current research interests include computer vision, pattern recognition, and deep learning.

Jingyao Gong is currently working toward the master's degree with the School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China. Her research interests include compressed sensing, deep learning, and image processing.

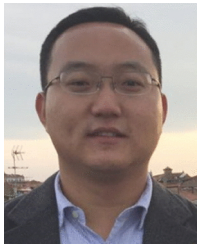
Hongjie Jia received the Ph.D. degree in computer science and technology from China University of Mining Technology - Xuzhou, China. He is currently an Associate Professor with the School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China. His research interests include clustering, deep clustering, subspace clustering, density clustering, federated learning, and contrastive learning.

Xiangjun Shen received the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China. He is currently a Professor with the School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China. He worked as a senior visiting scholar at the University of North Carolina Charlotte from 2013 to 2014. His research interests include large-scale image and video classification and recognition, multimodal social multimedia processing, distributed network architecture, and social media computing technology.



Jianping Gou (Senior Member, IEEE) received the PhD degree in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2012. He was a post-doctoral research fellow with The University of Sydney. He is currently a professor with the College of Computer and Information Science, College of Software, Southwest University, Chongqing, China. His current research interests include pattern classification and machine learning. So far, he has published more than 140 papers on international journals or conferences, such as IJCV and CVPR.

Hongying Meng (Senior Member, IEEE) received the Ph.D. degree in communication and electronic systems from Xi'an Jiaotong University, Xi'an China. He is currently a Professor with the Department of Electronic and Electrical Computer Engineering, Brunel University London, U.K. He has authored over 200 academic papers with more than 8000 citations. His research interests include signal processing, computer vision, affective computing, artificial intelligence, neuromorphic computing, and Internet of Things. He is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology and IEEE Transactions on Cognitive and Developmental Systems.



Le Wang (Senior Member, IEEE) received the B.S. and Ph.D. degrees in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2008 and 2014, respectively. From 2013 to 2014, he was a Visiting Ph.D. Student with the Stevens Institute of Technology, Hoboken, NJ, USA. From 2016 to 2017, he was a Visiting Scholar with Northwestern University, Evanston, IL, USA. He is currently a Professor with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include computer vision, pattern recognition, and machine learning. He is the author of more than 80 peer-reviewed publications in prestigious international journals and conferences. He is an Associate Editor of Pattern Recognition, Machine Vision and Applications, and Physical Research Laboratory. He is the Area Chair of WACV'2024&2025, ICPR'2022&2024, and CVPR'2022.